

**Methods and Informatics to Analyze Intact Protein Sequence and Structure by Ion
Mobility-Mass Spectrometry**

by

Daniel A. Polasky

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemistry)
in the University of Michigan
2019

Doctoral Committee:

Professor Brandon T. Ruotolo, Chair
Professor Philip C. Andrews
Professor Kristina L. Håkansson
Professor Alexey I. Nesvizhskii

Daniel A. Polasky

dpolasky@umich.edu

ORCID iD: [0000-0002-0515-1735](https://orcid.org/0000-0002-0515-1735)

© Daniel A. Polasky 2019

Acknowledgements

I would like to thank my advisor, Prof. Brandon Ruotolo. Your tireless support and positivity have made these past five years a truly enjoyable experience. I would also like to thank my current and former committee members, Professors Kristina Hakansson, Philip Andrews, Alexey Nesvizhskii, and Georgios Skiniotis. Your support and insight has helped shape this thesis and I thank you for all your input.

I've had the privilege of working with an incredible group of colleagues in the Ruotolo lab and throughout the Michigan Chemistry department. I'd like to thank the all the group members who came before me for building such a positive and cooperative culture in the group, and particularly Dr. Joe Eschweiler and Dr. Jessica Rabuck-Gibbons for teaching me everything when I joined the lab. Sugyan Dixit, I can't imagine a better person to work alongside for a PhD. You've been the best sounding board for ideas, both scientific and personal, and a fantastic friend. This thesis would not be what it is, nor these past five years the same without you. I'd also like to thank all the current members of the group. There's never a boring day in the lab and that's a testament to your curiosity and friendliness – it's been a real pleasure to watch the lab grow and change over these past five years.

In addition to coworkers, I've been fortunate to work with many fantastic collaborators at the University of Michigan and beyond. Special thanks to Phil Andrews, Lolita Piersimoni, Susan Hagen, Hollis Showalter, Dmitry Avtonomov, Alexey Nesvizhskii, Sarah Haynes, Brent Martin, Samantha Schon, Sue Hammoud, Ruwan Kurulugama, and John Fjeldsted for giving me the opportunity to work on a fascinating range of projects and for all your support and input to

the work presented here. Without your generous contributions of expertise, time, and samples, this thesis would not be what it is today.

I'd like to thank my friends and family for their support and encouragement throughout this long journey. Thank you to my parents for making all of this possible, and for being the most supportive and loving parents anyone could hope to have. Finally, I'd like to thank my wife Melissa for all your love and support, both personal and scientific, over these past seven years. I'm incredibly fortunate to have such a wonderful partner.

Table of Contents

Acknowledgements	ii
List of Tables	viii
List of Figures	x
List of Appendices	xiv
Abstract	xv
Chapter 1 Introduction	1
1.1 Methods for Analysis of Protein Sequence and Modifications	1
1.2 Methods for Determination of Protein Structure and Interactions	6
1.3 Mass Spectrometry-based Analysis of Protein Structure	8
1.4 Ion mobility-mass spectrometry (IM-MS) for Protein Structure Analysis	11
1.4.1 Protein Ion Generation	13
1.4.2 Tandem MS	14
1.4.3 IM Separation	15
1.5 Collision-induced Unfolding (CIU)	16
1.5.1 Generation and Analysis of CIU Data	18
1.5.2 Probing Protein Structure and Stability using CIU	20
1.5.3 Applications of CIU	23
1.6 Summary	26
1.7 References	29
Chapter 2 Fixed-Charge Trimethyl Pyrilium Modification for Enabling Enhanced Top-Down Mass Spectrometry Sequencing of Intact Protein Complexes	39
2.1 Abstract	39
2.2 Introduction	40
2.3 Experimental Methods	42
2.3.1 Chemical Modification	42
2.3.2 Ion Mobility-Mass Spectrometry	43
2.3.3 FT-ICR mass spectrometry	44
2.3.4 Data Analysis	44
2.4 Results	45

2.4.1 Fixed-charge modification enhances sequence coverage in a model protein complex	47
2.4.2 Modification extends proteomic sequencing to large protein complexes	51
2.4.3 High resolution MS analysis of large complexes	54
2.5 Conclusions	56
2.6 Acknowledgements	57
2.7 References	57
Chapter 3 Chemical Derivatization Illuminates the Links Between Controlled Sequence Informative Fragmentation Chemistry and Gas-phase Protein Ion Structure	62
3.1 Introduction	62
3.2 Experimental Methods	66
3.2.1 Chemical modification of primary amines	66
3.2.2 Chemical modification of carboxyl groups	66
3.2.3 Successive modification of primary amines and carboxyl groups	67
3.2.4 Ion mobility-mass spectrometry (IM-MS)	68
3.2.5 High resolution mass spectrometry	68
3.2.6 Data analysis	69
3.2.7 Molecular dynamics simulations	70
3.3 Results and Discussion	71
3.4 Conclusions	82
3.5 References	84
Chapter 4 IMTBX and Grppr: Software for Top-Down Proteomics Utilizing Ion Mobility-Mass Spectrometry	88
4.1 Abstract	88
4.2 Introduction	88
4.3 Methods	91
4.3.1 Experimental Section	91
4.3.2 Data Processing	92
4.3.3 Raw Feature Detection (IMTBX)	93
4.3.4 Isotopic Clustering (Grppr)	96
4.4 Results and Discussion	97
4.5 Conclusion	102
4.6 References	103
Chapter 5 CIUSuite 2: Next-Generation Software for the Analysis of Gas-phase Protein Unfolding Data	106
5.1 Abstract	106

5.2 Introduction	107
5.3 Methods	111
5.3.1 Experimental Section	111
5.3.2 Raw Data Extraction	112
5.3.3 CIUSuite 2 Overview	112
5.3.4 Stability Shift (“CIU50”) Analysis	113
5.3.5 Classification	114
5.3.6 Gaussian Fitting and Automated De-noising	115
5.4 Results	116
5.4.1 CIU Stability Shift Analysis	116
5.4.2 Classification of CIU data	118
5.4.3 Classifying Noisy, Low Intensity CIU Data	120
5.4.4 Gaussian Fitting and Automated Denoising	122
5.5 Conclusions	124
5.6 Acknowledgements	125
5.7 References	125
Chapter 6 A Modified Drift Tube Ion Mobility-Mass Spectrometer for Charge Multiplexed Collision Induced Unfolding	131
6.1 Abstract	131
6.2 Introduction	132
6.3 Experimental Methods	134
6.3.1 Sample Preparation	134
6.3.2 Protein Charge Stripping Analysis	135
6.3.3 A Modified Agilent 6560 for Collision Induced Unfolding Experiments	136
6.3.4 Mass Spectrometry Data Analysis	137
6.4 Results and Discussion	139
6.5 Conclusions	148
6.6 References	149
Chapter 7 An Algorithm for Building Multi-State Classifiers Based on Collision Induced Unfolding Data	154
7.1 Introduction	154
7.2 Methods	156
7.2.1 Sample Preparation	156
7.2.2 CIU Acquisition	157
7.2.3 Classification	157

7.3 Results and Discussion	159
7.4 Conclusions	165
7.5 References	165
Chapter 8 Conclusions and Future Directions	168
8.1 Conclusions	168
8.2 Future Directions	171
8.2.1 Charge manipulation for optimized fragmentation of intact proteins	171
8.2.2 Development of top-down sequence annotation software for PTM localization	173
8.2.3 Advanced CIU software: width analysis and direct incorporation of mass information	174
8.2.4 High-throughput CIU with microfluidics for rapid sample introduction and automated acquisition	176
8.3 References	177
Appendices	180

List of Tables

Table II-1 Mass assignments for EDC-modified 6+ SERF (Figure 3-3B).....	190
Table II-2 Mass assignments for TMP + EDC-modified SERF (Figure 3-3G).....	191
Table II-3 . Experimental CCS of unmodified SERF ions..	196
Table II-4 Experimental CCS of modified SERF ions..	197
Table II-5 Theoretical CCS of models from each cluster in the unmodified version of SERF ..	200
Table II-6 Theoretical CCS of models from each cluster in the modified version of SERF.....	200
Table II-7 Table of average and standard deviation of fraction of secondary structure elements in each cluster for unmodified SERF.....	200
Table II-8 Table of average and standard deviation of fraction of secondary structure elements in each cluster for unmodified SERF.....	201
Table III-1 User-definable parameters used for IMTBX processing.....	210
Table III-2 User-definable parameters used for Grppr processing	210
Table III-3 Estimated processing time for top-down data by IMTBX workflows.	211
Table IV-1 Test data probability values for classification of IgG subclasses.....	215
Table IV-2 Benchmarks for Gaussian fitting and Classification analysis time.....	216
Table V-1 CCS measurements extracted from IgG subclass CIU.....	222
Table V-2 Calculated charge stripping for select charge states in 6560 mass spectra	222
Table V-3 RMSD values for Figure 6-3 G	224
Table V-4 RMSD values for Figure 6-3 H	225

Table V-5 RMSD values for Figure 6-4 G	226
Table V-6 RMSD values for Figure 6-4 H	226
Table V-7 RMSD values for Figure 6-5 C.....	227
Table V-8 RMSD values for Figure 6-5 D	228
Table VI-1 Synthesized compounds tested and results of testing.....	230

List of Figures

Figure 1-1 . Sources of proteoform diversity.....	2
Figure 1-2 Top-down vs bottom-up proteomics.....	4
Figure 1-3 Representation of methods to characterize protein sequence and structure.....	5
Figure 1-4 Selected structural mass spectrometry approaches.....	10
Figure 1-5 Schematic diagram of the Synapt G2 instrument.....	12
Figure 1-6 A depiction of positive mode nanoESI.....	13
Figure 1-7 Schematic depiction of ion mobility separation.....	16
Figure 1-8 Diagrams of CIU of proteins and common methods of analysis.....	19
Figure 1-9 Applications of CIU.....	22
Figure 2-1 Charge-fixing chemical modification scheme with TMP.....	46
Figure 2-2 Enhanced sequencing of the Avidin tetramer following TMP modification.....	48
Figure 2-3 Mapping locations of peptide bond cleavage on Avidin.....	50
Figure 2-4 Enhanced sequencing of large protein complexes ADH and Ovalbumin.....	52
Figure 2-5 FT-ICR MS sequencing of large protein complexes with chemical modification.....	54
Figure 3-1 Charge fixing chemical modification drives charge remote fragmentation in intact proteins.....	72
Figure 3-2 TMP modification of SERF protein reveals fragmentation at fixed charge sites.....	75
Figure 3-3 Effect of blocking acidic residues with amides on protein fragmentation.....	77
Figure 3-4 Modeled SERF structures with and without chemical modification.....	80

Figure 4-1 Filtering IM-MS data with IMTBX.	94
Figure 4-2 Graphical interface for viewing raw IM-MS data.....	95
Figure 4-3 Schematic Overview of the Data Processing by IMTBX for Top-down IM-MS	96
Figure 4-4 Isotopic clusters that cannot be resolved without IM separation	98
Figure 4-5 Trends in fragment ion populations in IM-MS	99
Figure 4-6 IMTBX and Grppr processing of 8 protein standards by top-down IM-MS	101
Figure 5-1 CIUSuite 2 overview.....	113
Figure 5-2 CIU50 analysis mAb glycoforms using CIUSuite 2.....	116
Figure 5-3 Classification of different IgG standards	119
Figure 5-4 CIU50 analysis and classification TSPO-lipid complexes.....	121
Figure 5-5 Automated de-noising of membrane protein CIU data using Gaussian fitting	123
Figure 6-1 A diagram of the modified Agilent 6560 IM-MS instrument.....	140
Figure 6-2 Characterizing charge stripping in monoclonal antibodies.....	141
Figure 6-3 CIU fingerprints of IgG subtypes.....	144
Figure 6-4 CIU fingerprints of IgG light chain variants.....	146
Figure 6-5 Snapshot CIU fingerprints of IgG 1 and IgG 4.....	147
Figure 7-1 Multiple charge state classification of IgGs.....	160
Figure 7-2 Multiple charge state classification of Src kinase.....	162
Figure 7-3 Stress multi-state classification distinguishes Avastin and Avegra.....	163
Figure I-1 TMP modification of Substance P peptide (RPKPQQFFGLM-amide)	180
Figure I-2 A) Mass spectrum of a monomer of unmodified avidin.....	181
Figure I-3 Number of TMP modifications to avidin monomer from model analysis described in Figure I-2	182

Figure I-4 Ion mobility arrival time distributions of replicate analyses of Avidin.....	183
Figure I-5 Ion mobility arrival time distributions of replicate analyses of ADH	184
Figure I-6 Ion mobility arrival time distributions of replicate analyses of Ovalbumin	185
Figure I-7 Collision induced unfolding (CIU) profiles of Avidin with and without TMP	186
Figure II-1 Linear trend fit (alternative fit to Figure 3-1G).....	187
Figure II-2 CID-50 values for EDC-modified SERF vs unmodified SERF	188
Figure II-3 Fragmentation propensity maps by amino acid from all charge states of chemically modified SERF.....	189
Figure II-4 Fixed charge modification reveals extensive charge solvation by intact proteins ...	192
Figure II-5 Numbered cluster structures for unmodified SERF.	198
Figure II-6 Numbered cluster structures for modified SERF.	198
Figure II-7 Plot of TM CCS values determined via IMPACT against IMoS	199
Figure II-8 Plots of fraction of secondary structure vs. residue number for unmodified SERF. 201	
Figure II-9 Plots of fraction of secondary structure vs. residue number for modified SERF.....	202
Figure II-10 Contact map for unmodified SERF lysines.....	203
Figure II-11 Contact map for modified SERF lysines.....	204
Figure II-12 Z-score contact map for unmodified SERF lysines.....	205
Figure II-13 Z-score contact map for modified SERF lysines.....	206
Figure III-1 IMTBX Algorithm 1	209
Figure IV-1 Depiction of querying each possible combination of input data between classes ..	212
Figure IV-2 Cross-validation workflow.....	213
Figure IV-3 Feature selection and cross-validation for IgG1, IgG2, IgG3, and IgG4 classification	214

Figure V-1 Illustration of charge stripping determination..... 220

Figure V-2 CIU fingerprints of IgGs on the Synapt G2 system. 221

Figure VII-1 Naming scheme used to describe protamine isoforms 231

Figure VII-2 Intact masses of protamine proteoforms observed from pooled samples..... 231

List of Appendices

Appendix I: Supporting Information for Chapter 2	180
Appendix II: Supporting Information for Chapter 3	187
Appendix III: Supporting Information for Chapter 4	207
Appendix IV: Supporting Information for Chapter 5	212
Appendix V: Supporting Information for Chapter 6	220
Appendix VI: Compounds Synthesized for Charged Labeling of Proteins	229
Appendix VII: Protamine Analysis Procedure	240
Appendix VIII: References for all appendices.....	243

Abstract

Methods for rapid assessment of intact protein sequence and structure are increasingly important to understanding biology and treating disease. Mass spectrometry and ion mobility have emerged as effective tools for assessing protein sequence and structure, but face significant technical challenges in evaluating intact proteins and converting data from gas-phase ions to solution-relevant information.

Protein complexes preserved intact for analysis by mass spectrometry typically generate low quality fragmentation, in part due to their relatively low charge. In Chapter 2, we develop a chemical modification method that affixes stable, intrinsically-charged reagents to proteins and demonstrate improvements in sequence coverage for protein complexes. In Chapter 3, we use the reagents developed in Chapter 2 to alter the competition between fragmentation pathways of intact proteins and demonstrate the immense capacity of protein ions to accommodate excess charge through charge solvation. We show that fragmentation can be directed to the site of fixed charge, a novel observation for intact, multiply-charged protein ions, and that mobile proton-mediated fragmentation can be restored by capping carboxylic acids, blocking the charge-remote pathway and resulting in improved sequence coverage.

The conclusions reached in Chapters 2 and 3 depended on the development of software tools to analyze the complex data generated during fragmentation of intact proteins on an IM-MS platform. Ion mobility provides an additional dimension of separation when coupled to MS, enabling resolution of overlapping fragment ion signals that could not be resolved by MS alone.

To utilize this additional information, we adapted tools being developed for bottom-up IM-MS proteomics, IMTBX and Grppr, to analyze intact protein fragmentation by IM-MS.

IM-MS also has the potential to provide structural insight into intact proteins and complexes, including through the use of collision-induced unfolding (CIU). Similar to analysis of protein fragmentation by IM-MS, CIU experiments generate complex datasets requiring informatics to recover useful information about protein structure. In Chapter 5, we develop a software suite to provide automated annotation of gas-phase stability shifts and statistical classification of CIU data to support the development of high-throughput screening methods using CIU. We also utilize Gaussian fitting to distinguish and remove chemical noise from protein signal in membrane protein CIU to enable stability shift and classification analyses in these challenging datasets.

In Chapter 6, we develop a framework to include all observed charge states in CIU analyses rather than a single charge state. Adducts associated with intact proteins can be lost as charged species, reducing the charge of the protein and contaminating nearby charge states through “charge stripping.” We characterize charge stripping of intact monoclonal antibodies and develop an algorithm to predict and remove the contaminating signals. The software tools for CIU data developed in Chapter 5 annotate only a single charge state of CIU data at a time. In Chapter 7, we expand the classification method of Chapter 5 to incorporate data from all charge states, or any other perturbation that alters the CIU pathway, into a single “multi-state” classifier. We are able to generate robust multi-state classifiers to distinguish kinase inhibitor binding modes and a highly similar innovator/biosimilar biotherapeutic pair, which had each proven challenging to differentiate using a single charge state.

In total, the work presented in this thesis describes improved methods for analysis of protein complexes by IM-MS, including derivatization for improved sequence analysis and informatics for both sequencing and structural analyses.

Chapter 1 Introduction

The cellular machinery responsible for the biological processes of life is composed primarily of noncovalently associated complexes of proteins.^{1,2} Identifying the proteins involved in these complexes and annotating their function has been a major research goal for fields spanning wide ranging topics in cellular biology and human disease over the past several decades.³ The sequencing of the human genome coupled with emerging technologies in mass spectrometry (MS) has enabled the rapid identification of proteins present in a sample,⁴ but fully characterizing their modifications, interactions, and three-dimensional structure remains a fundamental challenge. Developing technologies capable of rapidly generating and integrating these various levels of protein information is of great importance to transforming molecular medicine, in which the specific molecular mechanisms of disease are discovered and treated, to precision medicine, which adapts this paradigm to incorporate the specific genetic and proteomic status of an individual.

1.1 Methods for Analysis of Protein Sequence and Modifications

Proteins are composed of long chains of amino acids, with the sequence of amino acids encoded by DNA. If every protein-coding DNA sequence resulted in the synthesis of exactly one protein form, genetic information alone would be sufficient to determine much of biology. However, in the years following the sequencing of the human and other genomes, it has become clear this is not the case. A range of processes result in the formation of many different forms of proteins,⁵ collectively referred to as “proteoforms.”⁶ Mutations and polymorphisms in DNA, alternative

RNA splicing,⁷⁻⁹ and errors in the translation process¹⁰ result in the generation of many additional amino acid sequences from a single canonical gene. Furthermore, the completed proteins are modified throughout their lifecycles with post-translational modifications (PTMs), which can result in thousands of potential proteoforms from a single amino acid sequence (Figure 1-1, adapted with permission from ⁵). PTMs provide crucial regulatory and feedback functions within the cell, making their characterization crucial to understanding protein function. Perhaps best example of this process currently known is the example of the ‘histone code,’¹¹ in which post-translational modification of histone protein tails provides epigenetic regulation by remodeling chromatin to enable or disable access to DNA, altering the gene expression and ultimately the phenotype of an organism.

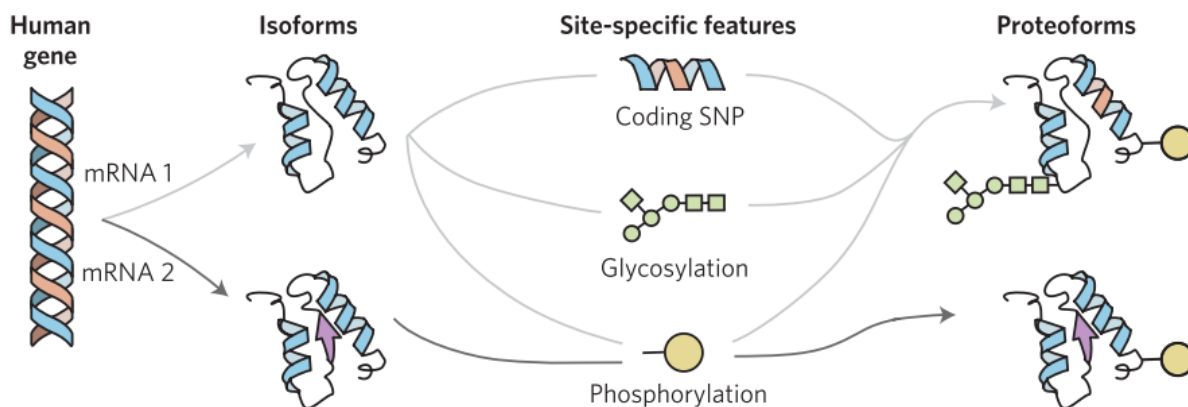


Figure 1-1 . Sources of proteoform diversity. A single gene (far left) can result in many different proteoforms due to a range of processes. Alternative splicing of RNA can result in different isoforms, which can be further modified with PTMs (including, for example, glycosylation and phosphorylation). Single nucleotide polymorphisms (SNPs) can also alter the protein sequence at specific sites.

Direct assessment of both protein sequences and modifications is thus critical to understanding cellular machinery and its dysregulation in disease states. Following the advent of electrospray ionization (ESI)¹² and improvements in mass spectrometer accuracy and instrument control in the 1990s, MS emerged as a powerful method to analyze protein sequence.¹³ While direct analysis of intact proteins to obtain sequence information is possible with MS, the many

challenges associated with separating intact proteins by liquid chromatography (LC) and their comprehensive analysis by MS led to non-optimal protein identification workflows early in the development of proteomics. Instead, an approach now termed “bottom-up” or shotgun proteomics emerged, in which proteins are first digested by proteolytic enzymes into peptides. Peptides can be readily separated by LC and analyzed by available MS instruments, which, when combined with bioinformatics tools that are able to match observed peptides against a database of protein sequences known from genome sequencing,¹⁴ resulted in a powerful method to identify proteins.⁴ Ultimately, complex mixtures of proteins, up to complete cell lysates containing over 10,000 proteins, can be digested, separated, analyzed, and identified, giving rise to “proteomic” analyses (*i.e.* of the entire proteome).¹⁵⁻¹⁸ Bottom-up proteomics has since been applied to a wide range of biological and clinical applications.¹⁹

Despite the technological success and widespread adoption of bottom-up proteomics, key constraints have limited its clinical relevance.²⁰ Because proteins are digested into peptides prior to analysis, quantifying PTM states can be challenging, and it cannot be determined if PTMs detected on different peptides from the same protein were present together (for example, one protein with two modifications) or separately (two copies of the protein with one modification each), e.g. proteoforms.⁶ For example, histones utilize a combination of many PTMs as a means of influencing protein expression epigenetically, meaning that the translation of PTM information to biological relevance requires complete annotation of all PTMs and linking this information back to the relevant proteoforms. An alternative approach, termed “top-down” proteomics, has emerged in recent years in an attempt to address these limitations. Top-down proteomics analyzes intact proteins without digestion into peptides, in principle enabling

complete annotation of all PTMs present (Figure 1-2, adapted with permission from ²¹),²² but faces significant technical challenges in separation and MS-based analysis of intact proteins.

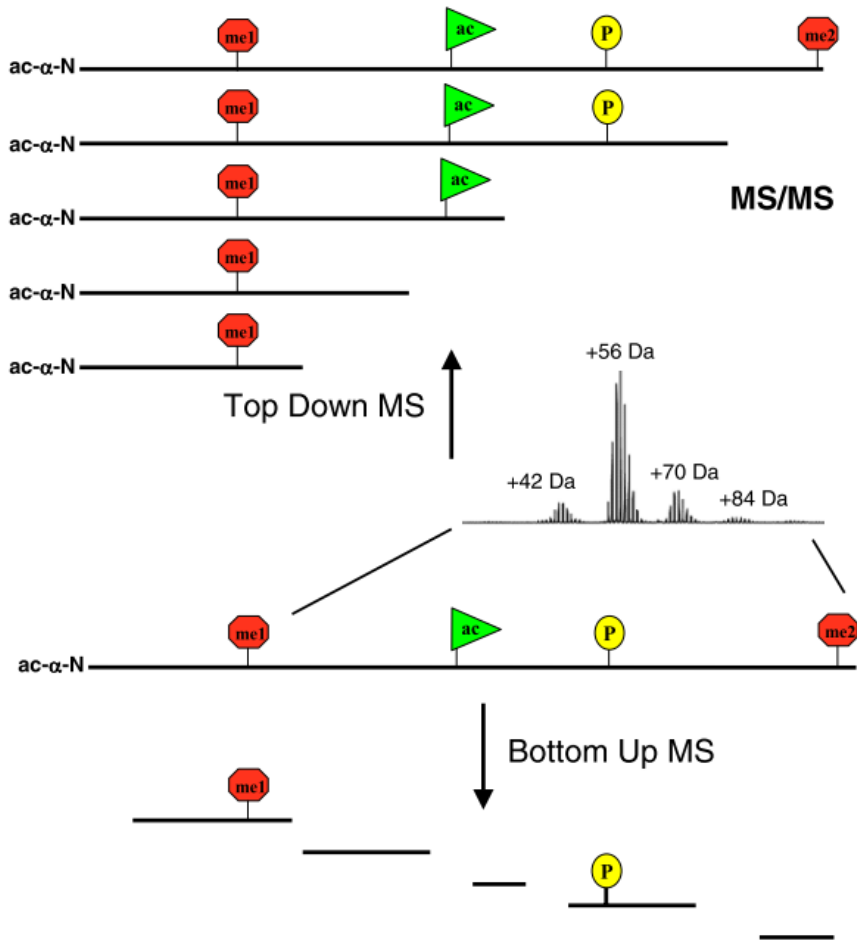


Figure 1-2 Top-down vs bottom-up proteomics. An example protein (represented by the solid black line) containing four PTMs is shown in the middle. Because the protein is digested into peptides prior to MS analysis in bottom-up proteomics (bottom), whether all four PTMs existed together on one protein or separately on different copies of the protein cannot be determined. The figure also highlights that missing peptides can preclude identification of some PTMs, though this is a challenge for top-down proteomics as well. Because top-down proteomics (top five lines) considers the entire protein sequence without digestion, a complete picture of PTMs can be determined, in this case showing that all four PTMs existed together on the same protein molecule.

Improved separations utilizing MS-compatible electrophoresis methods for both denaturing^{23,24} and native²⁵ conditions have been employed to perform top-down proteomics on samples approaching the complexity of bottom-up analyses.²⁶⁻²⁸ These analyses have necessitated the use of alternative ion activation methods, including electron capture/transfer dissociation (ECD^{29,30},

ETD^{31,32}) and ultraviolet photodissociation (UVPD³³⁻³⁵), improved mass spectrometer design, and informatics to process the resulting data.³⁶⁻³⁹ Despite these advances, significant challenges remain in achieving complete fragmentation of proteins larger than 20-30 kDa.⁴⁰ Incomplete fragmentation means PTMs may not be confidently localized to a single site, or even missed

entirely if sections of the protein are not observed in any fragment ions. While focused studies on individual purified proteins have been capable of identifying proteins much larger than this limit,^{41,42} achieving complete fragmentation of larger proteins to localize PTMs in large-scale studies has remained elusive. This deficiency is amplified by the fact that most proteins form multi-protein assemblies in order to accomplish their biological function.^{1,2} Detecting and analyzing these complexes requires the characterization of not just individual intact proteins, but the preservation and analysis of the non-covalent complexes they form in the cell (Figure 1-3, adapted with permission from ⁴³). Both bottom-up and typical top-down proteomics canonically require denaturation and/or enzymatic digestion of proteins, often precluding any analysis of the structure and dynamics of these complex assemblies.

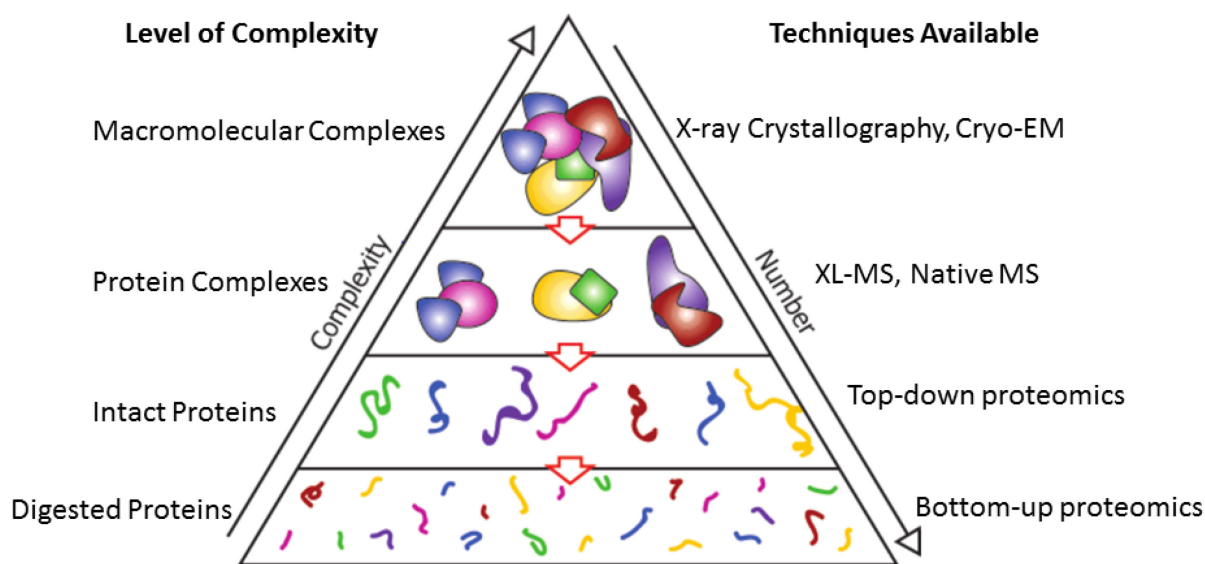


Figure 1-3 Representation of methods to characterize protein sequence and structure. Biological activity is typically accomplished by macromolecular complexes (top level) made up of many individual proteins with unique sequences, modifications, and three-dimensional structures. Directly characterizing these complexes is sometimes possible with atomic resolution structural characterization methods, but remains challenging and low-throughput. MS-based approaches have been developed to characterize protein sequence (bottom two levels) and interactions (second level from top) to attempt to build towards the goal of identifying and understanding the biology of proteins and complexes.

Direct characterization of intact protein assemblies using native MS⁴⁴ offers a promising alternative for the characterization of functional multiprotein machines.⁴⁵⁻⁴⁷ Furthermore, native

MS has the potential to provide protein sequence and structural information in the context of the same experiment, dramatically increasing the throughput of modern protein structure/function studies. However, sequencing technology coupled to native MS experiments lags far behind complementary bottom-up and top-down approaches targeting small, monomeric proteins. Incomplete fragmentation of proteins limits the ability of native MS to identify unknown proteins within complexes and prevents detailed analysis of the proteoforms incorporated within such assemblies. While collision induced dissociation (CID) is a widely available and effective technology for peptide sequencing, CID information extracted from large proteins and protein complexes analyzed under native conditions is often limited.⁴⁸ In many cases, sequence coverage is concentrated into a few labile regions, e.g. flexible or terminal loop areas.⁴¹ Achieving full sequence coverage for such large protein systems is one of the key challenges facing top-down proteomics, as well as the establishment of native MS workflows for wide-ranging structural proteomics.

1.2 Methods for Determination of Protein Structure and Interactions

The fundamental relationship between protein structure and function makes their study critical in ongoing efforts to understand fundamental elements of biochemistry and human disease.⁴⁹ While efforts to predict protein three-dimensional structure from amino acid sequence⁵⁰ have seen some success in small, well-ordered protein systems,⁵¹ systematic prediction of structure from sequence remains elusive.^{52,53} As a result, experimental assessment of protein structure is essential, in addition to genomic and proteomic assessment, to generate a complete understanding of biological machinery, including protein sequence, modifications, interactions, and three-dimensional structure.

Several techniques are capable of generating three-dimensional structures of proteins and protein complexes with atomic or nearly atomic resolution. X-ray diffraction⁵⁴ of crystallized proteins has been the primary method for generating protein structures since the 1950s,⁵⁵ and nearly 90% of known structures of biomolecules today have been determined by x-ray crystallography.⁵⁶ Macromolecular assemblies as complex as the complete 80S ribosome of *S. cerevisiae* have been characterized at atomic resolution,⁵⁷ enabling discoveries in cellular biology, drug discovery, and more. However, x-ray crystallography requires generation of high quality protein crystals, a process that requires large quantities of high-concentration, purified protein and time. The need to generate static protein crystals also precludes direct evaluation of protein dynamics, as well as assessment of proteins without stable, ordered structures. Nuclear magnetic resonance spectroscopy (NMR⁵⁸) is also capable of generating atomic resolution structures, and can evaluate structural dynamics.⁵⁹ However, NMR is generally limited to proteins smaller than ~50 kDa,⁶⁰ and suffers from similar requirements for large amounts of concentrated, purified protein as x-ray crystallography. Finally, recent advances in cryo-electron microscopy (cryo-EM) have enabled generation of atomic resolution structures, particularly for large, symmetric molecular assemblies.^{61,62} Cryo-EM has the potential to provide structures with lower sample requirements than x-ray crystallography or NMR, but still requires significant sample preparation, restricting the possible throughput to near x-ray or NMR methods for now.

Atomic resolution structures remain the preferred method to determine the molecular mechanisms of protein machines, but the number of known structures (~150,000 in 2019)⁵⁶ lags far behind the number of computationally predicted protein sequences (currently approaching 150,000,000),⁶³ which does not account for the unknown number of putative complexes assembled from these proteins to accomplish biological tasks.⁶⁴ Dynamic interactions between

proteins and changes to structure from sequence permutations and PTMs are further complications that remain largely beyond the current capabilities of structural biology. Development of high speed, lower resolution methods to assess protein structure are thus in great demand. A number of biophysical techniques, such as circular dichroism (CD)^{65,66} and small angle x-ray scattering (SAXS)⁶⁷ are used to determine protein secondary and low-resolution tertiary structure, providing relatively rapid measurements with lower sample requirements than atomic resolution techniques. However, the total information content provided by these techniques remains relatively limited.

In order to understand protein structure, its role in defining function, and any changes that may occur in disease states, it is essential to explore the parameters that link such elements of biophysics together.⁶⁸ One such element is protein stability, often reported as a free energy of protein unfolding and represents one of the most widely utilized descriptors of protein structure.⁶⁹ Calorimetry experiments, such as differential scanning calorimetry (DSC) and isothermal titration calorimetry (ITC) can measure global stability of proteins and thermodynamics of binding interactions, respectively, but require substantial amounts of purified protein and provide averaged measurements across ensembles of protein states.⁷⁰ Given the significance of protein stability in the framework of understanding protein structure and function, new experimental techniques that can extract such values with improved figures of merit are needed.

1.3 Mass Spectrometry-based Analysis of Protein Structure

Mass spectrometry (MS) has recently experienced a proliferation of structural biology related research, focusing primarily on heterogeneous proteins, protein complexes, and protein-ligand

complexes due to its ability to access such mixtures with sensitivity, speed, and low limits of detection.⁷¹ While MS cannot measure structure directly, a variety of methods have been developed to encode structural information into a mass measurement to leverage the power of MS, and particularly of bottom-up proteomics, to structure determination. Chemical crosslinking mass spectrometry (XL-MS) uses bifunctional reagents to link pairs of residues within a protein together (Figure 1-4), followed by enzymatic digestion and bottom-up proteomics methods to identify the crosslinked peptides that result.⁷² The length of the crosslinking reagent determines a distance constraint, so any crosslinked residues are expected to be within that distance of each other in the 3D protein structure.^{73–75} Recent reports have utilized this basic workflow to restrain integrative modeling of protein structures,^{76,77} assess protein dynamics,⁷⁸ and crosslink whole cell lysates to identify interactions between proteins.⁷⁹ Hydrogen-deuterium exchange (HDX),^{80,81} oxidative footprinting (*e.g.* FPOP),^{82,83} and covalent labeling⁸⁴ all probe the solvent accessibility of reactive sites on a protein by modifying reactive residues exposed to solvent (Figure 1-4), digesting the protein, and using bottom-up proteomics methods to identify the modified residues. The degree of modification can be used to map the exposed surface of the protein and assess structural dynamics, typically in conjunction with molecular modeling approaches by restraining generated structural models using the experimental data. The methods differ in the reactive groups they target: HDX probes amide hydrogens along the peptide backbone, which can be more or less protected by the secondary and tertiary structure surrounding them. All amino acids except proline contain an exchangeable backbone hydrogen, meaning that HDX can generate nearly individual amino acid-resolved information with appropriate data processing.^{85,86} The variable degree of protection from various secondary and tertiary structure elements also enables substantial structural insight, making HDX an

increasingly popular method for structural assessment.⁸⁷ FPOP uses hydroxyl radicals to oxidize 14 of the 20 naturally occurring amino acids, in principle yielding nearly the spatial resolution of HDX, but without the challenges associated with back-exchange of deuterium for hydrogen prior to measurement or H-D scrambling in the gas phase.

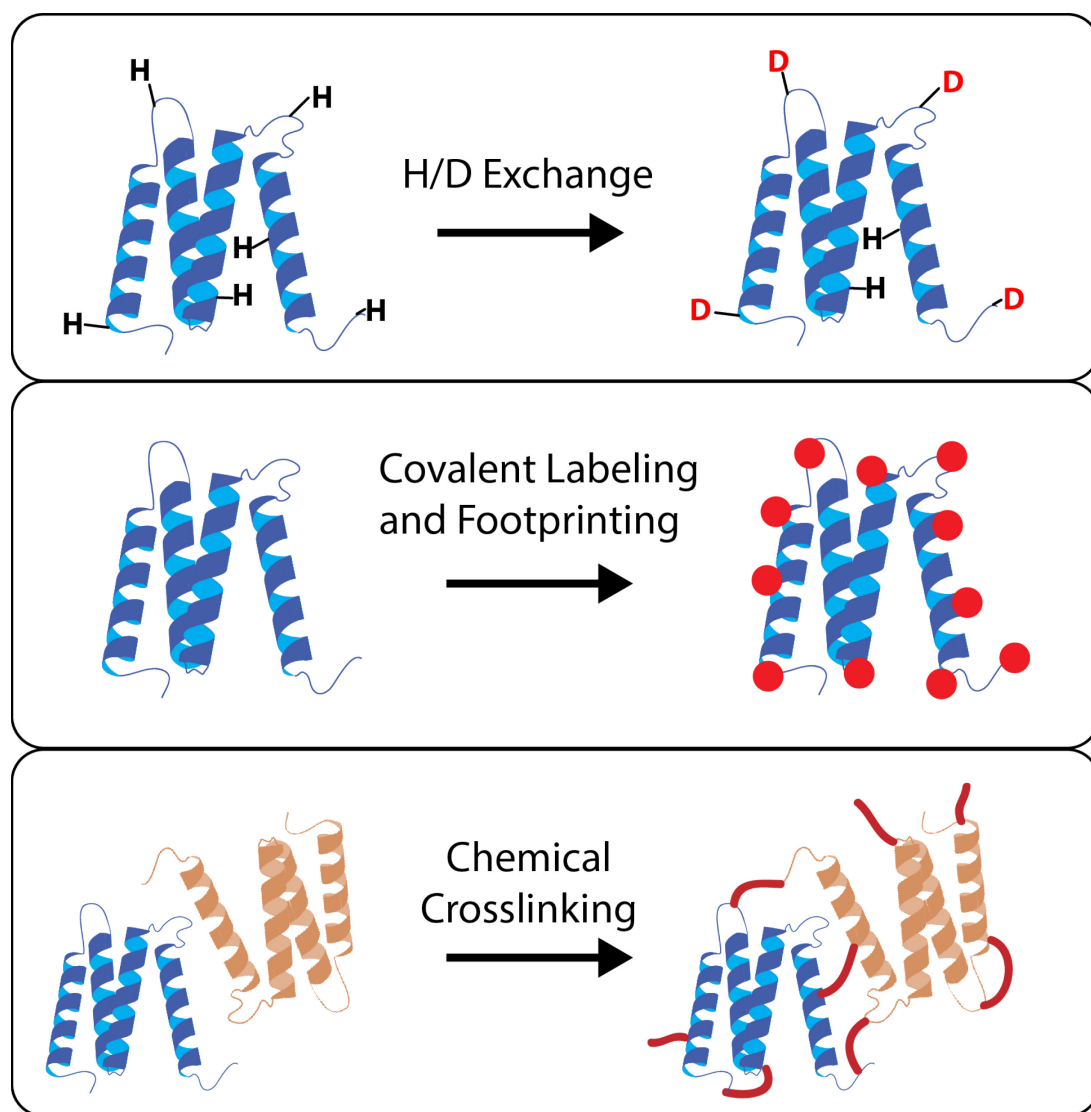


Figure 1-4 Selected structural mass spectrometry approaches. H/D exchange (top) exchanges exposed hydrogen (H) atoms for deuterium (D). Covalent labeling and oxidative footprinting approaches (FPOP, for example) label solvent accessible regions of the protein using various reagents. Chemical crosslinking uses bifunctional reagents to link to reactive sites (most commonly lysine residues) together, enabling determination of protein-protein interactions and distance determination to constrain structural models.

There are many covalent labeling approaches for assessing protein structure that follow the same general principle: modification of solvent-exposed amino acid residues of proteins

under physiological conditions followed by bottom-up proteomics to identify the modified sites.⁸⁴ Of particular interest to the work related in this thesis are methods to modify Lys, Asp, and Glu side chains, as these were the targets of the chemical modifications described in Chapters 2 and 3. Lysine remains perhaps the most often targeted amino acid for chemical modification due to the reactivity of its primary amine side chain and the high frequency with which it occurs in protein sequences and particularly in surface exposed regions.⁸⁴ A range of reagents have been used to target Lys, including organic acid anhydrides and *N*-hydroxysuccinimide derivatives, several of which are compatible with labeling under physiological conditions to examine native protein structures.⁸⁴ The acidic side chains of Asp and Glu can be targeted using carbodiimide chemistry,⁸⁸ however, this requires lowering pH to around 5 to proceed efficiently. Alternative reagents based on dihydrazides have been proposed to enable labeling to proceed under physiological conditions.⁸⁹

1.4 Ion mobility-mass spectrometry (IM-MS) for Protein Structure Analysis

Ion mobility (IM), which separates ions based on their size to charge ratio and reports ion size in terms of an orientationally-averaged collision cross section (CCS), has also been widely deployed in combination with MS as a platform for structural biology.⁹⁰ While orientational averaging and the resolution of modern IM spectrometers limits the structural information to

much lower resolution than XRD or NMR, the combination of speed, sensitivity, capability to handle complex mixtures has resulted in a growing field of IM-MS structural biology.

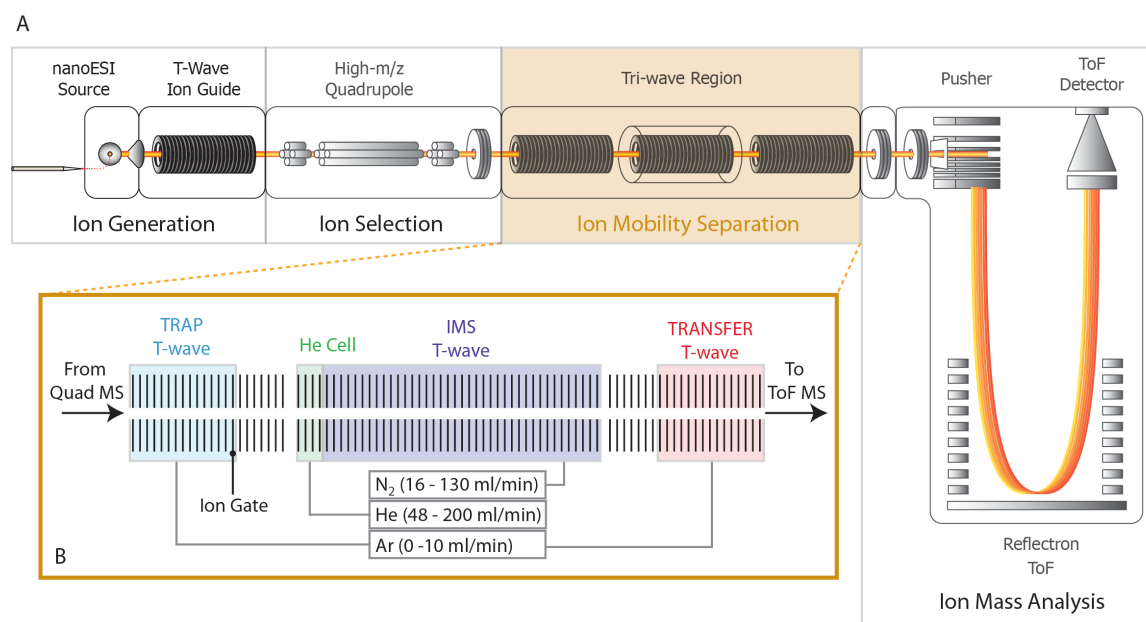


Figure 1-5 A: Schematic diagram of the Synapt G2 quadrupole ion mobility-time of flight instrument, indicating the four main regions of the instrument. Ions are generated using a nanoESI source and transferred to the quadrupole after several differential pumping stages. The quadrupole has been modified to enable selection of ions up to m/z 32000 and can operate in selection or full transmission modes. The “tri-wave region” includes two collision cells before and after the traveling wave IMS, enabling mobility separation and collisional activation. Finally, the ToF mass analyzer determines the m/z ratio of ions with resolving power of approx. 10,000 in standard operating modes employed in this work. B: Detailed view of gas flows in the IMS region. The “trap” and “transfer” collision cells are pressurized with argon to pressures in the high 10^{-3} to low 10^{-2} mbar range. The helium (He) cell reduces ion activation in transit from the low pressure trap cell to the high pressure (approx. 3.4 mbar in this work) IMS cell, which is pressurized with nitrogen.

The instrument used in the majority of work described in this thesis is a Synapt G2 HDMS (Waters, Milford, MA) shown in Figure 1-5, with the exception of some data in Chapter 6 that was generated on an Agilent 6560 (that system will be described in detail in Chapter 6). The Synapt instrument consists of a nanoelectrospray ionization (nESI) source, a quadrupole (Q) mass analyzer, a traveling wave ion mobility separator (TWIMS), and a time of flight (ToF) mass analyzer. The following sections provide specific details as to the operation of modern IM-MS instrument platforms.

1.4.1 Protein Ion Generation

To analyze a sample with IM-MS, the analyte(s) of interest must be ionized and transferred to the gas phase. To evaluate solution phase structure, proteins must be transferred from solution and ionized without (substantially) disrupting their structures, a challenging task. In ESI,¹² a high voltage is applied between a very small orifice containing the solution to be ionized and the inlet of the mass spectrometer, forming a Taylor cone and jet of highly charged droplets (Figure 1-6).⁹¹ The droplets evaporate (often assisted by heated gas flows in the source region) until the Coulombic repulsion of the charges in the droplet exceeds the surface tension holding the droplet together (a point termed the “Rayleigh limit”) and the droplet fissions into smaller droplets.⁹² After many cycles of evaporation and fission, the final droplets approach the size of individual

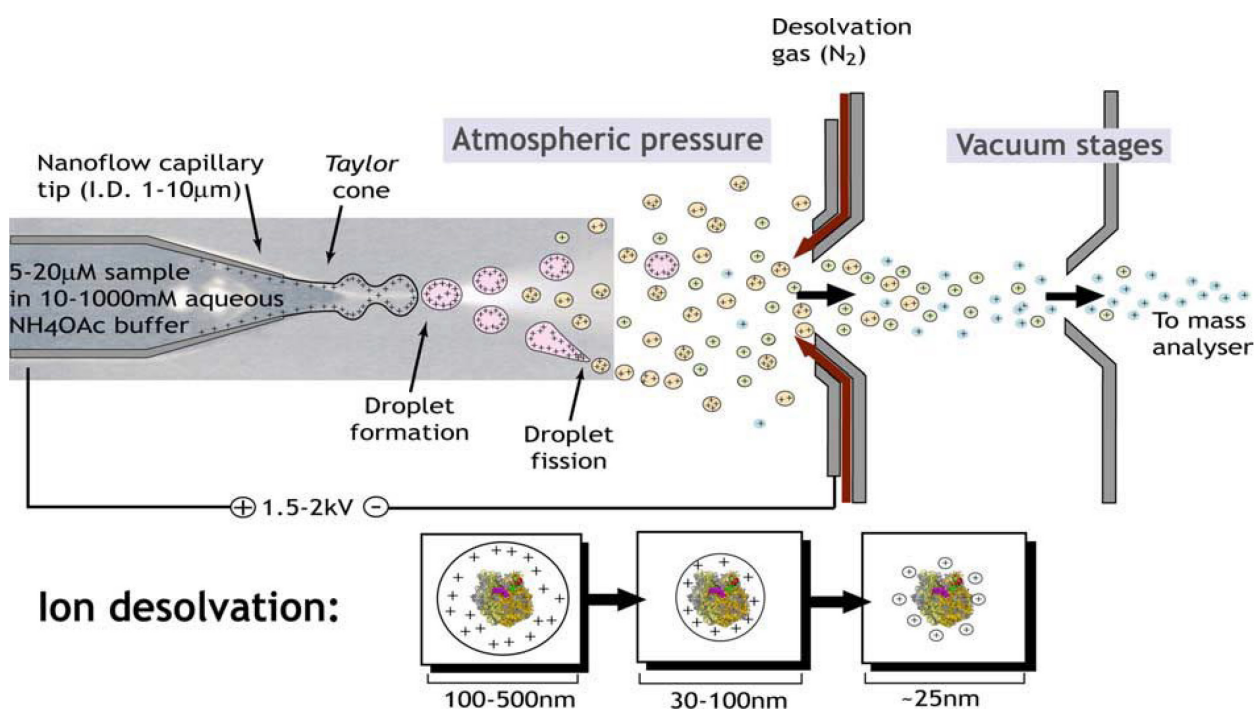


Figure 1-6 A depiction of positive mode nanoESI. The high voltage applied to the capillary generates a Taylor cone and droplets, which evaporate and fission to eventually leave charged analyte (*e.g.* protein) ions that are directed into the mass spectrometer.

protein molecules, which are left behind in the gas phase as the final solvent molecules evaporate.⁹²⁻⁹⁴

The number of evaporation and fission cycles depends on the initial size of the droplet, with fewer cycles generally associated with improved transmission of ions into the instrument.⁹⁵

NanoESI (nESI) utilizes significantly smaller orifices and lower flow rates to generate smaller starting droplets, resulting in improved ionization efficiency and increased tolerance for salts and other ion-suppressive species.^{96–98} The extended evaporation and fission process ensures that the majority of the energy imparted during the ionization and transfer process is not deposited with the protein, enabling sufficiently gentle transfer to preserve “native-like” structure, including noncovalently associated protein complexes and protein-ligand interactions, into the gas phase.⁹⁹ The exact degree to which the protein remains in its solution-phase conformation is a matter of considerable debate, but an extensive body of literature indicates that large scale structural changes can be minimized with appropriate instrument tuning.^{99–105}

1.4.2 Tandem MS

Measuring the mass of a molecule is often insufficient to identify it, particularly for large biopolymers like proteins, which are composed of up to hundreds of amino acids arranged in a particular sequence. To identify molecules, tandem MS involves selecting a “precursor” ion with an initial (nondestructive) stage of MS, most often using a quadrupole or ion trap mass analyzer, then activating that precursor to break it apart into fragments (“product” ions) that are then measured by a second stage of MS. The combination of the precursor mass measured in the first stage of MS and all of the observed product masses can, in many cases, be sufficient to determine the identity of the molecule. For protein ions, many types of activation, including IRMPD,¹⁰⁶ ECD,^{29,30} ETD,³¹ UVPD,^{33–35} and CID¹⁰⁷ can be utilized to generate fragmentation along the peptide backbone, allowing reconstruction of the amino acid sequence by reading the

addition of amino acids from smaller to larger fragment ions. In CID, ions are accelerated into (or through) a region of inert gas. Collisions with the gas convert the ions' kinetic energy into internal energy, which is rapidly redistributed throughout the molecule.¹⁰⁸ This “slow heating” process continues until the internal energy of the ion becomes sufficient to break the weakest chemical bond, resulting in dissociation. For protonated peptides and proteins, the peptide bond is typically among the weakest molecular bonds (in the gas phase), resulting in dissociation in between amino acid residues, yielding sequence-informative product ions. On the Synapt G2, tandem MS can be performed using the high mass quadrupole to select precursor ions, followed by CID either before or after IM, in the trap and transfer collision cells, respectively (Figure 1-5). For the top-down experiments described, CID is always performed in the trap collision cell prior to IM to enable IM separation of the resulting fragment ions.

1.4.3 IM Separation

IM separates ions through the opposed forces of an electric field and momentum transfer from collisions with an inert, neutral gas. Ions with larger cross sections experience more collisions, resulting in greater opposition to the electric field and, for example, a longer time to transit a drift tube under the influence of a pushing electric field (Figure 1-7).^{109,110}

There are a variety of IMS platforms involving different configurations of electric field and gas. The simplest conceptually is a drift tube instrument, in which a linear electric field gradient is maintained to push ions through a “drift tube” filled with stationary gas. The Agilent 6560 instrument discussed in Chapter 6 utilizes a drift tube IMS. The IMS device in the Synapt

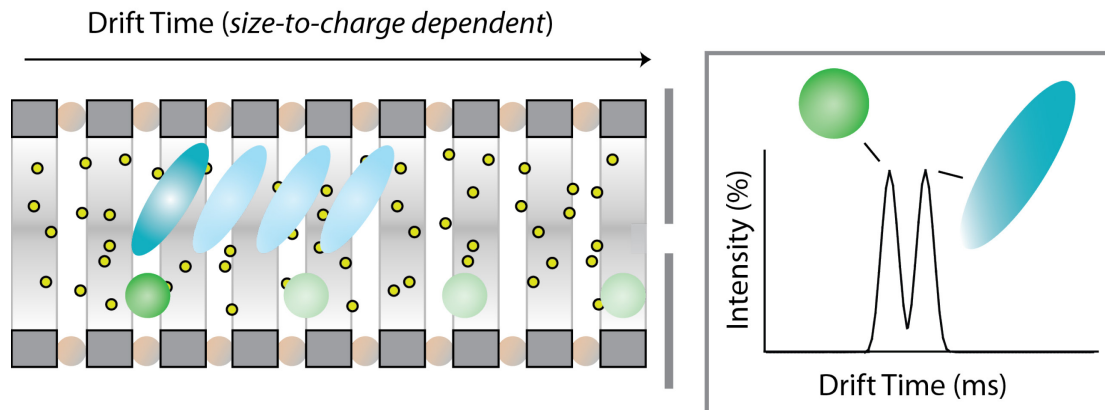


Figure 1-7 Schematic diagram of basic ion mobility. The ions, represented by the large blue oval and smaller green circle, are accelerated towards the right of the device by an electric field. Collisions with the gas in the cell, represented by the very small yellow circles, slows the progress of the ions. The larger blue oval experiences more collisions than the green circle, and thus moves more slowly through the device, allowing the ions to be separated by their time to transit the entire device.

instrument is a traveling wave IMS (TWIMS),¹¹¹ in which ions are propelled through a tube filled with stationary gas by a series of low voltage “waves” that move through the device from beginning to end. Ions are briefly carried by the waves until collisions with the gas cause them to “roll over” the wave, so ions with larger cross sections experience more rollover events and thus take a longer time to transit the device.^{112,113} TWIMS devices typically require lower applied electric fields and shorter device lengths to achieve high resolution separations when compared to equivalent drift tube IM analyzers, but currently require calibration to convert from measured arrival time to collision cross section due to the complexity of ion motion resulting from the non-linear electric field used to propel the ions.¹¹⁴

1.5 Collision-induced Unfolding (CIU)

The first commercial IM-MS system became available in the mid-2000s, sparking a period of rapid growth in the use of IM for protein structural analysis and other applications. Prior to this recent period of expansion, IM-MS was used to primarily assign the conformations of peptides¹¹⁵ and small proteins¹¹⁶ in the gas phase. However, as the size and complexity of biomolecules

increases, IM-derived CCS values alone often yield insufficient information to define the structures of proteins in detail.¹¹⁶ Collisional activation has long been used to probe the structure and stability of protein ions in the gas phase.^{43,117} Collision induced unfolding (CIU) represents an extension of this earlier work, and is best viewed as a gas-phase analog of differential scanning calorimetry experiments often carried out in solution. In a typical CIU experiment, isolated biomolecular ions are activated through energetic collisions with a background gas (e.g. Argon) in order to increase their internal energy and cause them to change conformation (unfold) in the gas-phase, without providing sufficient energy to cause the significant dissociation of covalent bonds.¹¹⁸ The progress of this CIU process is followed through IM-MS, with the former stage providing direct measurement of protein unfolding through changes in ion CCS and the latter analyzing the composition of the isolated biomolecules and enabling any collision induced dissociation (CID) products to be excluded from the analysis. Early examples of CIU include the observation of cytochrome c¹¹⁶ and apomyoglobin unfolding in the gas phase.¹¹⁹ Modern implementations of the technology have been extended well beyond these examples, to include detailed analyses of the CIU mechanism and applications to a range of therapeutically-relevant targets.

The potential of CIU as an analytical fingerprinting technique to study the structures and stabilities of proteins, protein complexes, and protein-ligand complexes is now emerging. The collisional activation of protein assemblies often yields a multitude of partially folded intermediates stable on the millisecond time scale that can provide a range of diagnostic information related to the structures of the isolated protein complexes.^{120,121} In addition, CIU has been used to assay the stabilities of proteins and protein-ligand complexes in the gas phase. Although the stability measurements offered by CIU data for biomolecular ions are relative, and

allow for comparisons of protein states rather than determination of absolute thermodynamic properties, they also provide valuable insight into the structure and native binding interactions of proteins and their complexes.^{117,122–125}

1.5.1 Generation and Analysis of CIU Data

Typical CIU experiments are performed by sequentially increasing an accelerating potential difference that serves to activate ions prior to ion mobility separation. As such, IM arrival time distributions (ATDs) are acquired at each stepped potential (Figure 1-8 A), creating a multi-dimensional dataset. The changes in measured ATD correspond to structural transitions of the protein ion in the gas phase which, while not directly assessing solution phase structures, can be used to generate unique fingerprints (Figure 1-8 B) that can reflect such native state structure information. Several methods to generate these fingerprints have been described,^{126–129} offering quantitative metrics for rapidly distinguishing subtle structural changes in proteins and protein complexes.

To generate a CIU fingerprint, the arrival time distribution of the m/z corresponding to the analyte ion must be extracted from the raw data at each collision voltage applied to create a matrix for analysis. Manual generation of this matrix can be time-consuming, and recent CIU experiments have relied upon automated extraction tools capable of creating such file structures rapidly.^{126,129} Once generated, replicates can be used to assign statistical confidence to observed deviations between fingerprints representing different protein forms, e.g. between ligand-bound and unbound states. These quantitative comparisons can be leveraged to classify binding events and structural changes into biologically relevant categories, such as differentiating functional

from nonspecific lipids bound to membrane proteins,¹²⁶ or determining the binding site of a ligand in systems with multiple known binding pockets.¹³⁰ As these workflows become routine and advance towards automated and high-throughput analyses, continued development of automated extraction and processing tools will be essential to realizing the full potential of CIU experiments.

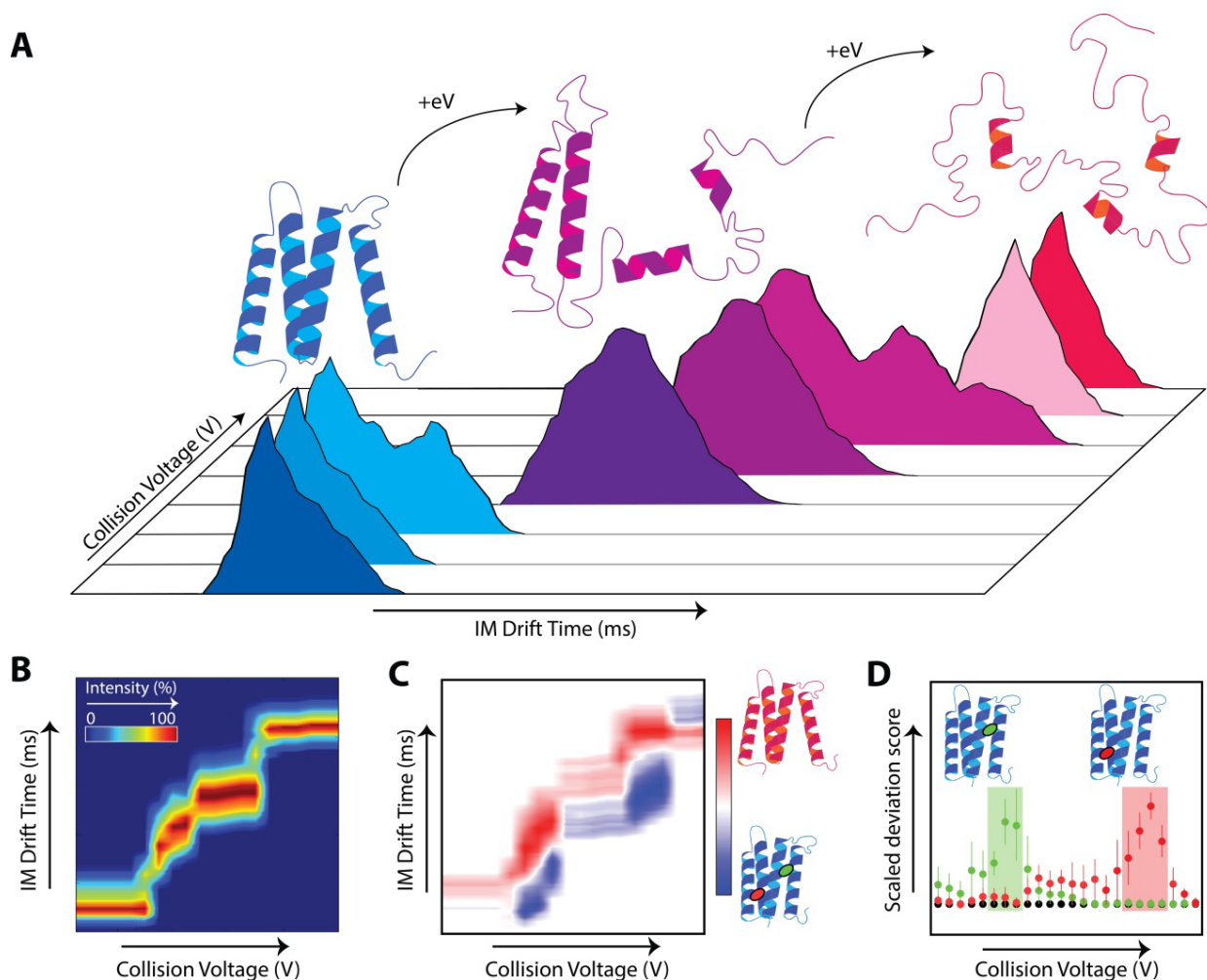


Figure 1-8 A: Diagrams and cartoons depicting the CIU of proteins and common methods of analysis. As collision energy (eV) is increased, an isolated protein ion unfolds in the gas phase. B: CIU fingerprint with collision voltage on the x-axis, arrival time on y-axis, and intensity shown using a color scale. C: CIU comparison plot analysis depicting an unbound (apo) and a doubly bound protein-ligand complex (red and green oval) with collision voltage on x-axis, arrival time on y-axis, and color scheme representing the differential intensities of the apo (red) and ligand bound (blue) states. D: A scaled deviation score analysis depicting a comparison of two different ligand bound states with CIU data acquired for the apo protein. A score is computed that statistically assesses fingerprint similarity at each voltage, enabling a narrow window of collision voltage to be defined that maximizes dissimilarity between analytes, as shown by green shaded area for the green (●) ligand and red shaded area for the red (●) ligand.

1.5.2 Probing Protein Structure and Stability using CIU

Collisional activation followed by IM-MS has been used to probe the conformations of proteins in the gas phase for nearly two decades.^{118,119} For example, early CIU experiments probed the activation energy barriers associated the gas-phase folding and unfolding of apomyoglobin following charge manipulation, revealing clear evidence of both Coulombic and structural components for the barriers detected between the gas-phase conformers.¹¹⁹ Tandem IM technology^{131,132} combined with collisional activation has been used to examine similar activation energy barriers in greater detail, revealing connectivity maps between the multitude of intermediate states populated during the CIU of small proteins.¹³³ Overall, these early CIU experiments were aimed primarily at uncovering the biophysical rules governing gas-phase protein ions, and succeeded in significantly advancing our understanding of protein stability and structure in a solvent-free environment.

Following on from this earlier work, CIU has been implemented to study the structure and dissociation behavior of protein complexes.¹³⁴ For example, early work¹³⁵ proposed an unfolding-based mechanism for protein complex CID, in which a single subunit unfolds and is ejected bearing a large portion of the total charge of the assembly, largely through collecting indirect evidence of protein CIU. The introduction of CIU enabled the direct observation of collisionally-activated protein assemblies, confirming that they populate partially folded intermediates that are stable on the millisecond timescale.^{43,102,120} Other structural rearrangements of protein complexes have been shown in the gas phase via collisional activation and IM-MS. For instance, many reports have shown evidence of compaction upon the activation of ring-like protein complexes that contain significant internal cavities.^{102,136} Moreover, computational chemistry has been used to probe protein complex CIU, reproducing many of the

general features of experimental data.¹³⁷⁻¹³⁹ Recent computational approaches in this area incorporate charge hopping within coarse-grained models and mobile protons within all-atom MD simulations.¹⁴⁰ Despite these recent advances, however, a complete model capable of predicting the unfolding transitions of heated protein complexes in CIU remains elusive.

Extending from these mechanistically-framed studies, CIU has been used to quantify shifts in protein complex stability upon binding large populations of both anions and cations. Early CIU work in this area indicated that buffer components of low volatility bound to intact protein complexes can act to stabilize protein complex in the gas phase.¹⁴¹ These initial results were expanded upon by screening a wider range of solution additives for their ability to stabilize gas-phase protein complexes.^{120,142-144} For example, CIU and CID studies incorporating a broad range anions and cations bound to significant number of protein complexes revealed that stabilizing anions act primarily through evaporative cooling, whereas stabilizing cations act to bind tightly to protein complexes and limit charge mobility.^{143,144} More direct efforts to stabilize protein complexes have been prosecuted through chemical cross-linking, where the CIU of intact protein complexes modified with charge-bearing chemical agents revealed significant increases in gas-phase stabilities.¹⁴⁵ Overall, this work demonstrates the clear utility of CIU data in building next-generation IM-MS technologies aimed at measuring labile protein complexes and structures.

More recent experiments have aimed provide a detailed mechanism for CIU in the context of monomeric proteins. For instance, a survey of proteins ranging from 8 to 78 kDa, and

containing between one and four domains, produced evidence of a strong correlation between native domain structure and the number of CIU transitions observed for low charge state protein

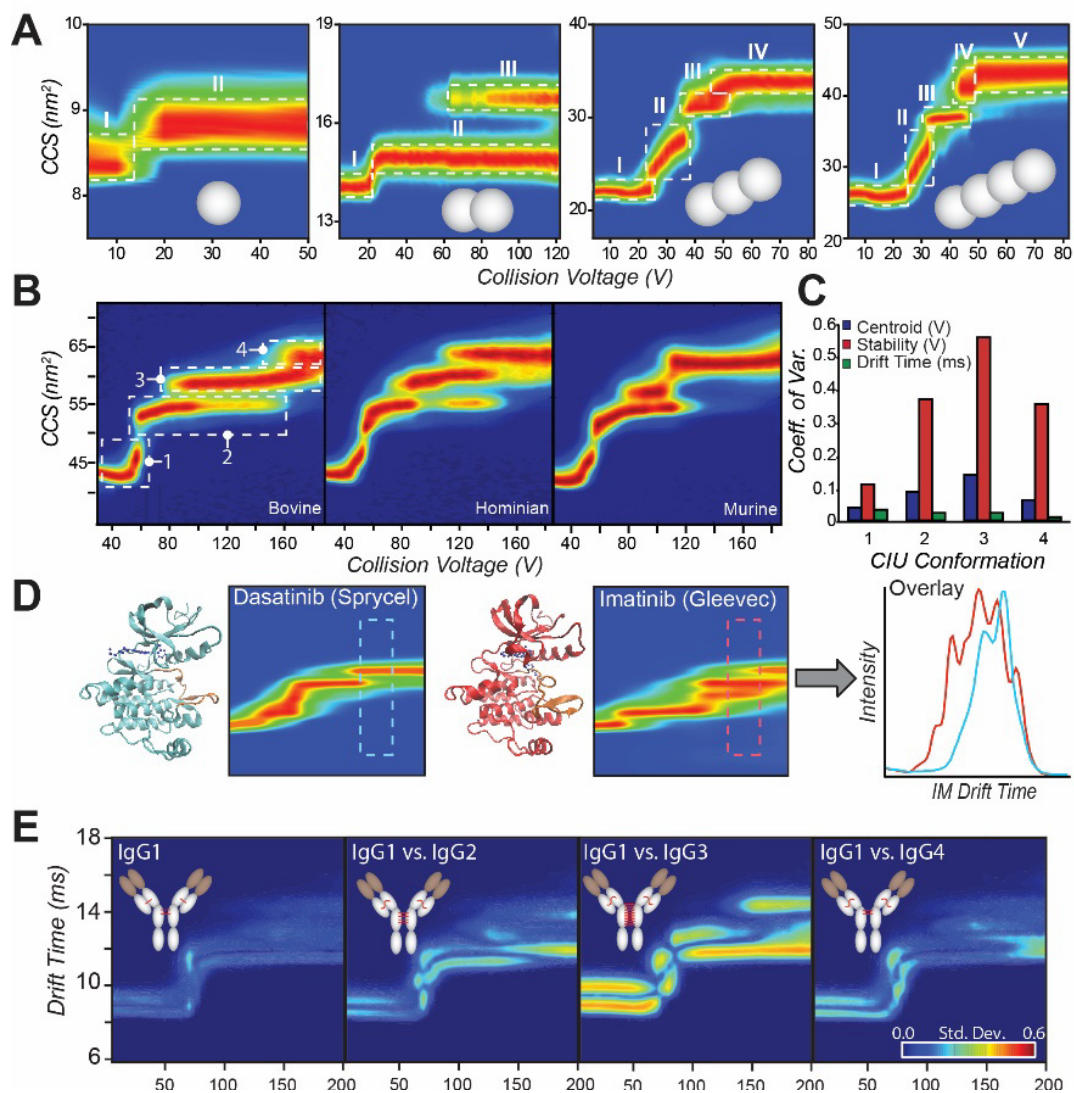


Figure 1-9 A: A series of covalently linked poly-ubiquitin proteins (1-4 ubiquitins, gray spheres) is probed by CIU. Single domain ubiquitin results in a single CIU transition, from an initial native-like state (I) to a more extended state (II) upon collisional activation. Each additional domain added results in an additional CIU transition, indicating that the transitions are representative of the domain structure of the protein in solution. B: Bovine, human, and murine serum albumin proteins CIU fingerprints are compared. Despite high sequence homology and globally similar three-dimensional structures, CIU readily distinguishes each variant, demonstrating sensitivity towards subtle alterations in protein isoforms. C: Coefficient of variation (CV) across the bovine, human, and murine albumins represented in (B) for centroid voltage (blue), stability or horizontal length (red), and center drift time (green) for each feature. High CVs indicate significant differences between fingerprints. D: Comparison of type I (Dasatinib, left) and type II (Imatinib, right) inhibitors bound to Abl kinase. CIU distinguishes the binding location of inhibitors to the kinase, enabling a screening assay based on the region of maximal difference in the CIU fingerprint (far right). E: IgG subtypes 1-4 (left to right) are quantitatively distinguished by CIU. Each subtype exhibits different patterns of disulfide bonding in a broadly conserved overall structure, resulting in different CIU fingerprints.

ions (Figure 1-9 A).¹⁴⁶ A follow-on study in this area used both domain-specific ligand binding

and noncovalent constructs to build the first detailed CIU mechanism for a multi-domain protein.¹²¹ This same report highlighted both the similarities and differences in the CIU of iso-CCS homologous protein variants, demonstrating both a strong correlation between quantified CIU similarity and sequence identity as well as identifying the stability of CIU features as the main element of variation in unfolding data acquired across sequence variants (Figure 1-9 B). Similarly, the structural differences of ubiquitin ions having similar ground state CCS values produced from cation-to-anion proton transfer reaction (CAPTR) experiments targeting a broad range of precursor ion charge states were detected by CIU.¹⁴⁷ These studies, taken together, begin to paint a detailed picture of the CIU mechanism as well as point toward future applications in protein engineering, where the stability of individual protein domains within larger constructs can be measured without need of labelling or surface attachment.

1.5.3 Applications of CIU

Characterizing the binding of ligands to proteins and protein complexes is a rapidly growing application area for CIU measurements, as the information content of such experiments can be used to rapidly provide binding affinities, inform on the nature of ligand attachment, and elucidate the location of binding. Binding locations can be differentiated by CIU, as the binding of a ligand to different sites in a protein results in differential alterations to its unfolding pathway. By comparing against CIU fingerprints acquired for ligands with known binding sites, specific binding locations for uncharacterized ligands can be determined. Building on early work in this area,¹¹⁷ a number of reports now utilize CIU to probe allosteric and conformationally-selective binding modes in the context of both inhibitor screening and analysis as well as probing membrane protein-lipid interactions.^{124,148} Because discrete ligand bound states can be resolved and analyzed separately in CIU, stability shifts detected and compared between binding states

can be used as evidence to support a cooperative stabilization mechanism. For example, such a mechanism was detected in the Concanavalin A tetramer upon polysaccharide binding.¹²⁴ In another report, CIU indicated that a compact conformation of a ligand-bound protein was highly stabilized, suggesting a possible allosteric binding mode in the context of the protein system in question, which was confirmed by hydrogen-deuterium exchange.¹⁴⁸

CIU has been widely utilized to assess binding of inhibitors and drugs to enzymes.^{130,149,150} For example, the kinase domain of BCR-Abl, a target implicated in chronic myeloid leukemia, was screened against a small library of kinase inhibitors using CIU.¹³⁰ Inhibitors having known selectivities for the active or inactive states of the kinase produced significantly different CIU fingerprints (Figure 1-9 D), enabling the development of a classification system based on narrow regions of the acquired fingerprints where the two classes produced maximally different unfolding patterns. Another kinase, protein kinase A (PKA), was probed by CIU,¹⁴⁹ similarly revealing significant differences in gas-phase unfolding upon binding different kinase inhibitor classes. CIU was also used to probe binding of HIV drugs to the membrane protein ZMPSTE24,¹⁵⁰ demonstrating that shifts in gas-phase protein stability can be directly correlated to solution phase K_d values.

CIU measurements are uniquely suited to analyze the role of lipids and other stabilizing molecules bound to heterogeneous membrane proteins. An early example of such work utilized CIU to classify a range of lipids interacting with membrane protein channels, where gas-phase unfolding data provided predictive information allowing the authors to identify functional lipids that bore structural and functional consequences when attached to the proteins studied.¹⁵¹ This type of CIU assessment is now part of an automated workflow,¹²⁶ enabling the rapid quantitative analysis of membrane protein stabilization through lipid and ligand binding. Most CIU studies of

membrane protein lipid binding are carried out over the entire ensemble of binding stoichiometries detected, leading to uncertainty surrounding the role of individual lipid bound states in contributing to overall protein stability. However, recent work utilizing a heated electrospray ion source coupled to IM-MS demonstrates that most lipids dissociate from model membrane proteins in CIU experiments as neutral species, confirming the validity of such ensemble analysis and pointing the way toward improved IM-MS instrumentation tailored for the CIU analysis of membrane protein ligand binding.¹⁵²

As biotherapeutics have emerged as a multibillion dollar industry, their analytical characterization has received proportional interest. Characterization of monoclonal antibodies is highly challenging given their size and the dynamic nature of their post-translational modifications, the state of which directly influences their function and efficacy. Given their critical importance and complex nature, as well as the need for high-throughput analysis and quality control metrics, CIU is ideally poised to be a part of future biotherapeutic analysis workflows. Recently, CIU methods have been applied to the characterization of the NIST monoclonal antibody standard,¹⁵³ comparative analyses of immunoglobulins,¹⁵⁴ active innovator and biosimilar therapeutics,^{155,156} and antibody-drug conjugates (ADCs),¹⁵⁷ indicating the rapid expansion of CIU applications in this area.

CIU has been used to rapidly distinguish subtle differences in large antibodies, such as between IgG subclasses with different disulfide bonding patterns.¹⁵⁴ The differences in disulfide bridging, despite identical sequences and other post-translational modifications, resulted in nearly identical mass and arrival time information, but could be quantitatively differentiated using CIU (Figure 1-9 E). Glycosylated and deglycosylated IgGs were also distinguished by CIU, indicating that CIU has broad applicability to rapidly distinguish subtle changes to large

proteins that are otherwise highly challenging to characterize. More recent work compares an innovator biotherapeutic, Remicade, with Remsima, the first FDA-approved antibody-based biosimilar.¹⁵⁶ CIU was used as part of a multi-attribute monitoring (MAM) workflow, and provided a rapid assessment of therapeutic similarity, with the differences detected amongst biotherapeutic lots of Remsima linked to variations in antibody glycoforms using bottom-up proteomics. Other recent studies have extended CIU into the analysis of antibody-drug conjugates (ADCs), an area of intense pharmaceutical interest. Recent work has demonstrated drug conjugation serves to stabilize monoclonal antibodies in a manner readily detectable by CIU.¹⁵⁷ As such, the rapid analysis of biotherapeutics is clearly a growth area for CIU methods, where the technique promises to provide key solutions to the growing challenges surrounding the quality control and similarity assessment of intact protein therapeutics.

1.6 Summary

This dissertation presents a series of methods aimed at improving efforts to determine the sequence and structure of intact proteins and protein complexes. The work presented is divided into two main approaches, the first focused on improving sequencing analyses for top-down proteomics and the second on improving structural analyses, specifically native IM-MS and CIU, with the combined goal of merging sequence and structural analysis of intact proteins using IM-MS. Chapters 2 and 3 focus on chemical modification of proteins for improved fragmentation and sequencing, particularly in the context of native proteins and protein complexes, which are challenging targets for current sequencing methods. Chapter 2 introduces the use of trimethyl pyrylium to affix stable intrinsic charges to proteins, and demonstrates that the resulting products display enhanced and orthogonal fragmentation in comparison to the unmodified protein. This

work has been previously published as **Polasky, D. A.; Lermyte, F.; Nshanian, M.; Sobott, F.; Andrews, P. C.; Loo, J. A.; Ruotolo, B. T. Fixed-Charge Trimethyl Pyrylium Modification for Enabling Enhanced Top-Down Mass Spectrometry Sequencing of Intact Protein Complexes. *Anal. Chem.* 2018, 90 (4), 2756–2764.** Chapter 3 continues with chemical modification, elucidating the molecular mechanism for altered fragmentation following derivatization with fixed charges and generating a method to predict the fragmentation pathways of intact proteins based on their charge state and primary sequence. A novel fragmentation pathway is observed for an extensively modified small protein, and future directions for improving sequencing of proteins via chemical derivatization of acidic residues are discussed. Finally, substantial alterations to gas phase protein structure are observed with charged modifiers, ultimately defining the physical limitations of current sequencing techniques.

The work presented in Chapters 2 and 3 depended on the development of data analysis tools and methods for processing top-down ion mobility-mass spectrometry data. Chapter 4 describes the design and implementation of software tools for peak detection and clustering (“de-isotoping”) specifically for IM-MS data. Peak analysis using both the IM and MS dimensions of the data greatly improves the effective resolution and peak capacity of IM-MS, particularly for top-down proteomics data, in which complex isotopic distributions and large numbers of fragment ions complicate data analysis. This work was been previously published as **Avtonomov, D. M.; Polasky, D. A.; Ruotolo, B. T.; Nesvizhskii, A. I. IMTBX and Grppr: Software for Top-Down Proteomics Utilizing Ion Mobility-Mass Spectrometry. *Anal. Chem.* 2018, 90 (3), 2369–2375.**

Chapters 5, 6, and 7 focus on the development of methods and software tools to process IM-MS data from native mass spectrometry and collision-induced unfolding (CIU) experiments.

Chapter 5 describes the development of CIUSuite 2, a software suite to process CIU data, particularly from challenging datasets. Using the analytical tools in the software, we demonstrate a 60-fold improvement in data acquisition speed over previously published work, and enable stability analysis of noisy data from membrane proteins that had proven refractory to previous analysis tools. A supervised classification workflow was also developed to enable screening of CIU data into defined classes with improved statistical analysis over existing methods. This work has been previously published as **Polasky, D. A.; Dixit, S. M.; Fantin, S. M.; Ruotolo, B. T. CIUSuite 2: Next-Generation Software for the Analysis of Gas-Phase Protein Unfolding Data. *Anal. Chem.* 2019, 91 (4), 3147–3155.** Chapter 6 describes a joint effort to implement CIU analysis on a new instrument, the Agilent 6560, including instrument modifications, characterization of adduct loss from precursor ions, and the development of data processing software to support data acquired on the new instrument and correction of adduct losses. The use of simultaneous CIU analysis of data from multiple charge states (“charge-multiplexed CIU”) was characterized and used to demonstrate improvements in information content and throughput for CIU experiments. This work is currently undergoing peer review as **Vallejo, D. D.; Polasky, D. A.; Kurulugama, R. T.; Eschweiler, J. D.; Fjeldsted, J. C.; Ruotolo, B. T. A Modified Drift Tube Ion Mobility-Mass Spectrometer for Charge Multiplexed Collision Induced Unfolding. *Anal. Chem.* 2019.** Finally, Chapter 7 describes extending the classification methods developed for high-throughput screening in Chapter 5 to include data from all charge states of a protein. This combines work from Chapter 5 with the new instrumentation from Chapter 6, in which data from all charge states are always collected due to the design of the instrument. Analyzing all charge states improves classification accuracy and can be used to reduce acquisition time required for analysis.

Part of the content in this introduction was previously published as a review article

(Dixit, S. M.; Polasky, D. A.; Ruotolo, B. T. Collision Induced Unfolding of Isolated Proteins in the Gas Phase: Past, Present, and Future. *Curr. Opin. Chem. Biol.* 2018, 42, 93–100.)

1.7 References

- (1) Robinson, C. V.; Sali, A.; Baumeister, W. The Molecular Sociology of the Cell. *Nature* **2007**, 450 (7172), 973–982.
- (2) Alberts, B. The Cell as a Collection of Protein Machines: Preparing the next Generation of Molecular Biologists. *Cell* **1998**, 92 (3), 291–294.
- (3) Gavin, A. C.; Aloy, P.; Grandi, P.; Krause, R.; Boesche, M.; Marzioch, M.; Rau, C.; Jensen, L. J.; Bastuck, S.; Dümpelfeld, B.; et al. Proteome Survey Reveals Modularity of the Yeast Cell Machinery. *Nature* **2006**, 440 (7084), 631–636.
- (4) Yates, J. R. The Revolution and Evolution of Shotgun Proteomics for Large-Scale Proteome Analysis. *J. Am. Chem. Soc.* **2013**, 135 (5), 1629–1640.
- (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.; et al. How Many Human Proteoforms Are There? *Nat. Chem. Biol.* **2018**, 14 (3), 206–214.
- (6) Smith, L. M.; Kelleher, N. L. Proteoform: A Single Term Describing Protein Complexity. *Nat. Methods* **2013**, 10 (3), 186–187.
- (7) Xia, Y.; Duran-Frigola, M.; Trigg, S. A.; Boone, C.; Begg, B. E.; Iakoucheva, L. M.; Tan, G.; Vidal, M.; Sun, S.; Charlotiaux, B.; et al. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **2016**, 164 (4), 805–817.
- (8) Wang, E. T.; Sandberg, R.; Luo, S.; Khrebtkova, I.; Zhang, L.; Mayr, C.; Kingsmore, S. F.; Schroth, G. P.; Burge, C. B. Alternative Isoform Regulation in Human Tissue Transcriptomes. *Nature* **2008**, 456 (7221), 470–476.
- (9) Li, Y. I.; van de Geijn, B.; Raj, A.; Knowles, D. A.; Petti, A. A.; Golan, D.; Gilad, Y.; Pritchard, J. K. RNA Splicing Is a Primary Link between Genetic Variation and Disease. *Science* (80-.). **2016**, 352 (6285), 600–604.
- (10) Loftfield, R. B.; Vanderjagt, D. The Frequency of Errors in Protein Biosynthesis. *Biochem. J.* **1972**, 128 (5), 1353–1356.
- (11) Jenuwein, T. Translating the Histone Code. *Science* (80-.). **2001**, 293 (5532), 1074–1080.
- (12) Fenn, J.; Mann, M.; Meng, C.; Wong, S.; Whitehouse, C. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science* (80-.). **1989**, 246 (4926), 64–71.
- (13) Yates, J. R. Mass Spectrometry and the Age of the Proteome. *J. Mass Spectrom.* **1998**, 33 (1), 1–19.
- (14) Eng, J. K.; McCormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **1994**, 5 (11), 976–989.

- (15) Liu, H.; Sadygov, R. G.; Yates, J. R. A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. *Anal. Chem.* **2004**, *76* (14), 4193–4201.
- (16) Nesvizhskii, A. I.; Aebersold, R. Interpretation of Shotgun Proteomic Data. *Mol. Cell. Proteomics* **2005**, *4* (10), 1419–1440.
- (17) Aebersold, R.; Mann, M. Mass Spectrometry-Based Proteomics. *Nature* **2003**, *422* (6928), 198–207.
- (18) Beck, M.; Claassen, M.; Aebersold, R. Comprehensive Proteomics. *Curr. Opin. Biotechnol.* **2011**, *22* (1), 3–8.
- (19) Heffner, K. M.; Hizal, D. B.; Kumar, A.; Shiloach, J.; Zhu, J.; Bowen, M. A.; Betenbaugh, M. J. Exploiting the Proteomics Revolution in Biotechnology: From Disease and Antibody Targets to Optimizing Bioprocess Development. *Curr. Opin. Biotechnol.* **2014**, *30*, 80–86.
- (20) Savaryn, J. P.; Catherman, A. D.; Thomas, P. M.; Abecassis, M. M.; Kelleher, N. L. The Emergence of Top-down Proteomics in Clinical Research. *Genome Med.* **2013**, *5* (6), 53.
- (21) Garcia, B. A. What Does the Future Hold for Top Down Mass Spectrometry? *J. Am. Soc. Mass Spectrom.* **2010**, *21* (2), 193–202.
- (22) Siuti, N.; Kelleher, N. L. Decoding Protein Modifications Using Top-down Mass Spectrometry. *Nat. Methods* **2007**, *4* (10), 817–821.
- (23) Tran, J. C.; Doucette, A. A. Gel-Eluted Liquid Fraction Entrapment Electrophoresis: An Electrophoretic Method for Broad Molecular Weight Range Proteome Separation. *Anal. Chem.* **2008**, *80* (5), 1568–1573.
- (24) Lee, J. E.; Kellie, J. F.; Tran, J. C.; Tipton, J. D.; Catherman, A. D.; Thomas, H. M.; Ahlf, D. R.; Durbin, K. R.; Vellaichamy, A.; Ntai, I.; et al. A Robust Two-Dimensional Separation for Top-Down Tandem Mass Spectrometry of the Low-Mass Proteome. *J. Am. Soc. Mass Spectrom.* **2009**, *20* (12), 2183–2191.
- (25) Skinner, O. S.; Do Vale, L. H. F.; Catherman, A. D.; Havugimana, P. C.; Sousa, M. V. De; Compton, P. D.; Kelleher, N. L. Native GELFrEE: A New Separation Technique for Biomolecular Assemblies. *Anal. Chem.* **2015**, *87* (5), 3032–3038.
- (26) 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.; et al. Mapping Intact Protein Isoforms in Discovery Mode Using Top-down Proteomics. *Nature* **2011**, *480* (7376), 254–258.
- (27) Cleland, T. P.; Dehart, C. J.; Fellers, R. T.; Vannispén, A. J.; Greer, J. B.; LeDuc, R. D.; Parker, W. R.; Thomas, P. M.; Kelleher, N. L.; Brodbelt, J. S. High-Throughput Analysis of Intact Human Proteins Using UVPD and HCD on an Orbitrap Mass Spectrometer. *J. Proteome Res.* **2017**, *16* (5), 2072–2079.
- (28) Cannon, J. R.; Cammarata, M. B.; Robotham, S. A.; Cotham, V. C.; Shaw, J. B.; Fellers, R. T.; Early, B. P.; Thomas, P. M.; Kelleher, N. L.; Brodbelt, J. S. Ultraviolet Photodissociation for Characterization of Whole Proteins on a Chromatographic Time Scale. *Anal. Chem.* **2014**, *86* (4), 2185–2192.
- (29) 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.
- (30) McLafferty, F. W.; Horn, D. M.; Breuker, K.; Ge, Y.; Lewis, M. A.; Cerda, B.; Zubarev, R. A.; Carpenter, B. K. Electron Capture Dissociation of Gaseous Multiply Charged Ions by Fourier-Transform Ion Cyclotron Resonance. *J. Am. Soc. Mass Spectrom.* **2001**, *12* (3),

- 245–249.
- (31) Syka, J. E. P.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and Protein Sequence Analysis by Electron Transfer Dissociation Mass Spectrometry. *Proc. Natl. Acad. Sci.* **2004**, *101* (26), 9528–9533.
 - (32) Riley, N. M.; Westphall, M. S.; Coon, J. J. Activated Ion-Electron Transfer Dissociation Enables Comprehensive Top-Down Protein Fragmentation. *J. Proteome Res.* **2017**, *16* (7), 2653–2659.
 - (33) Bowers, W. D.; Delbert, S. S.; Hunter, R. L.; McIver, R. T. Fragmentation of Oligopeptide Ions Using Ultraviolet Laser Radiation and Fourier Transform Mass Spectrometry. *J. Am. Chem. Soc.* **1984**, *106* (23), 7288–7289.
 - (34) Reilly, J. P. Ultraviolet Photofragmentation of Biomolecular Ions. *Mass Spectrom. Rev.* **2009**, *28* (3), 425–447.
 - (35) Brodbelt, J. S. Photodissociation Mass Spectrometry: New Tools for Characterization of Biological Molecules. *Chem. Soc. Rev.* **2014**, *43* (8), 2757–2783.
 - (36) Leduc, R. D.; Fellers, R. T.; Early, B. P.; Greer, J. B.; Thomas, P. M.; Kelleher, N. L. The C-Score: A Bayesian Framework to Sharply Improve Proteoform Scoring in High-Throughput Top down Proteomics. *J. Proteome Res.* **2014**, *13* (7), 3231–3240.
 - (37) D. LeDuc, R.; L. Kelleher, N. Using ProSight PTM and Related Tools for Targeted Protein Identification and Characterization with High Mass Accuracy Tandem MS Data. In *Current Protocols in Bioinformatics*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2007; Vol. 19, p 13.6.1-13.6.28.
 - (38) 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.; et al. Informed-Proteomics: Open-Source Software Package for Top-down Proteomics. *Nat. Methods* **2017**, *14* (9), 909–914.
 - (39) Skinner, O. S.; Havugimana, P. C.; Haverland, N. A.; Fornelli, L.; Early, B. P.; Greer, J. B.; Fellers, R. T.; Durbin, K. R.; Do Vale, L. H. F.; Melani, R. D.; et al. An Informatic Framework for Decoding Protein Complexes by Top-down Mass Spectrometry. *Nat. Methods* **2016**, *13* (3), 237–240.
 - (40) 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.
 - (41) Han, X.; Jin, M.; Breuker, K.; McLafferty, F. W. Extending Top-down Mass Spectrometry to Proteins with Masses Great than 200 Kilodaltons. *Science (80-.)*. **2006**, *314* (5796), 109–112.
 - (42) Karabacak, N. M.; Li, L.; Tiwari, A.; Hayward, L. J.; Hong, P.; Easterling, M. L.; Agar, J. N. Sensitive and Specific Identification of Wild Type and Variant Proteins from 8 to 669 KDa Using Top-down Mass Spectrometry. *Mol. Cell. Proteomics* **2009**, *8* (4), 846–856.
 - (43) Benesch, J. L. P.; Aquilina, J. A.; Ruotolo, B. T.; Sobott, F.; Robinson, C. V. Tandem Mass Spectrometry Reveals the Quaternary Organization of Macromolecular Assemblies. *Chem. Biol.* **2006**, *13* (6), 597–605.
 - (44) Leney, A. C.; Heck, A. J. R. Native Mass Spectrometry: What Is in the Name? *J. Am. Soc. Mass Spectrom.* **2017**, *28* (1), 5–13.
 - (45) Pukala, T. L.; Ruotolo, B. T.; Zhou, M.; Politis, A.; Stefanescu, R.; Leary, J. A.; Robinson, C. V. Subunit Architecture of Multiprotein Assemblies Determined Using Restraints from Gas-Phase Measurements. *Structure* **2009**, *17* (9), 1235–1243.
 - (46) Sharon, M.; Taverner, T.; Ambroggio, X. I.; Deshaies, R. J.; Robinson, C. V. Structural

- Organization of the 19S Proteasome Lid: Insights from MS of Intact Complexes. *PLoS Biol.* **2006**, *4* (8), 1314–1323.
- (47) Chait, B. T.; Cadene, M.; Olinares, P. D.; Rout, M. P.; Shi, Y. Revealing Higher Order Protein Structure Using Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (6), 952–965.
- (48) Yin, S.; Loo, J. A. Top-down Mass Spectrometry of Supercharged Native Protein-Ligand Complexes. *Int. J. Mass Spectrom.* **2011**, *300* (2–3), 118–122.
- (49) Osadchy, M.; Kolodny, R. Maps of Protein Structure Space Reveal a Fundamental Relationship between Protein Structure and Function. *Proc. Natl. Acad. Sci.* **2011**, *108* (30), 12301–12306.
- (50) Baker, D.; Sali, A. Protein Structure Prediction and Structural Genomics. *Science*. 2001, pp 93–96.
- (51) Ambrish Roy; Alper Kucukural; & Yang Zhang. I-TASSER: A Unified Platform for Automated Protein Structure and Function Prediction. *Nat. Protoc.* **2010**, *5* (4), 725–738.
- (52) Lee, J.; Freddolino, P. L.; Zhang, Y. Ab Initio Protein Structure Prediction BT - From Protein Structure to Function with Bioinformatics; J. Rigden, D., Ed.; Springer Netherlands: Dordrecht, 2017; pp 3–35.
- (53) Kinch, L. N.; Li, W.; Monastyrskyy, B.; Kryshtafovych, A.; Grishin, N. V. Evaluation of Free Modeling Targets in CASP11 and ROLL. *Proteins Struct. Funct. Bioinforma.* *84*, 51–66.
- (54) Hull, A. W. A New Method of Chemical Analysis. *J. Am. Chem. Soc.* **1919**, *41* (8), 1168–1175.
- (55) KENDREW, J. C.; PARRISH, R. G.; MARRACK, J. R.; ORLANS, E. S.; J C Kendrew, B. D.; G Parrish, D. R.; R Marrack, P. J.; S ORLANSt, D. E. The Species Specificity of Myoglobin. *Nature* **1954**, *174* (4438), 946–949.
- (56) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.
- (57) Ben-Shem, A.; Garreau de Loubresse, N.; Melnikov, S.; Jenner, L.; Yusupova, G.; Yusupov, M. The Structure of the Eukaryotic Ribosome at 3.0 Å Resolution. *Science* (80-). **2011**, *334* (6062), 1524–1529.
- (58) Rabi, I. I.; Zacharias, J. R.; Millman, S.; Kusch, P. A New Method of Measuring Nuclear Magnetic Moment. *Physical Review*. 1938, p 318.
- (59) Palmer III, A. G. NMR Probes of Molecular Dynamics: Overview and Comparison with Other Techniques. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30* (1), 129–155.
- (60) Fiaux, J.; Bertelsen, E. B.; Horwich, A. L.; Wüthrich, K. NMR Analysis of a 900K GroEL-GroES Complex. *Nature*. Nature Publishing Group July 2002, pp 207–211.
- (61) Bai, X.; McMullan, G.; Scheres, S. H. . How Cryo-EM Is Revolutionizing Structural Biology. *Trends Biochem. Sci.* **2015**, *40* (1), 49–57.
- (62) Cheng, Y. Single-Particle Cryo-EM at Crystallographic Resolution. *Cell* **2015**, *161* (3), 450–457.
- (63) Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; et al. The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL in 2003. *Nucleic Acids Res.* **2003**, *31* (1), 365–370.
- (64) Sali, A.; Glaeser, R.; Earnest, T.; Baumeister, W. From Words to Literature in Structural

- Proteomics. *Nature* **2003**, 422 (6928), 216–225.
- (65) Brahms, S.; Brahms, J. Determination of Protein Secondary Structure in Solution by Vacuum Ultraviolet Circular Dichroism. *J. Mol. Biol.* **1980**, 138 (2), 149–178.
- (66) Kelly, S. M.; Jess, T. J.; Price, N. C. How to Study Proteins by Circular Dichroism. *Biochim. Biophys. Acta - Proteins Proteomics* **2005**, 1751 (2), 119–139.
- (67) Förster, F.; Webb, B.; Krukenberg, K. A.; Tsuruta, H.; Agard, D. A.; Sali, A. Integration of Small-Angle X-Ray Scattering Data into Structural Modeling of Proteins and Their Assemblies. *J. Mol. Biol.* **2008**, 382 (4), 1089–1106.
- (68) Redfern, O. C.; Dessailly, B.; Orengo, C. A. Exploring the Structure and Function Paradigm. *Curr. Opin. Struct. Biol.* **2008**, 18 (3), 394–402.
- (69) Nick Pace, C.; Martin Scholtz, J.; Grimsley, G. R. Forces Stabilizing Proteins. *FEBS Lett.* **2014**, 588 (14), 2177–2184.
- (70) Jelesarov, I.; Bosshard, H. R. Isothermal Titration Calorimetry and Differential Scanning Calorimetry as Complementary Tools to Investigate the Energetics of Biomolecular Recognition. *J. Mol. Recognit.* **1999**, 12 (1), 3–18.
- (71) Konermann, L.; Vahidi, S.; Sowole, M. A. Mass Spectrometry Methods for Studying Structure and Dynamics of Biological Macromolecules. *Anal. Chem.* **2014**, 86 (1), 213–232.
- (72) Sinz, A. Chemical Cross-Linking and Mass Spectrometry for Mapping Three-Dimensional Structures of Proteins and Protein Complexes. *J. Mass Spectrom.* **2003**, 38 (12), 1225–1237.
- (73) Leitner, A.; Walzthoeni, T.; Kahraman, A.; Herzog, F.; Rinner, O.; Beck, M.; Aebersold, R. Probing Native Protein Structures by Chemical Cross-Linking, Mass Spectrometry, and Bioinformatics. *Mol. Cell. Proteomics* **2010**, 9 (8), 1634–1649.
- (74) Singh, P.; Panchaud, A.; Goodlett, D. R. Chemical Cross-Linking and Mass Spectrometry as a Low-Resolution Protein Structure Determination Technique. *Anal. Chem.* **2010**, 82 (7), 2636–2642.
- (75) Rappsilber, J. The Beginning of a Beautiful Friendship: Cross-Linking/Mass Spectrometry and Modelling of Proteins and Multi-Protein Complexes. *J. Struct. Biol.* **2011**, 173 (3), 530–540.
- (76) Ward, A. B.; Sali, A.; Wilson, I. A. Integrative Structural Biology. *Science*. 2013, pp 913–915.
- (77) Zeng-Elmore, X.; Gao, X. Z.; Pellarin, R.; Schneidman-Duhovny, D.; Zhang, X. J.; Kozacka, K. A.; Tang, Y.; Sali, A.; Chalkley, R. J.; Cote, R. H.; et al. Molecular Architecture of Photoreceptor Phosphodiesterase Elucidated by Chemical Cross-Linking and Integrative Modeling. *J. Mol. Biol.* **2014**, 426 (22), 3713–3728.
- (78) Schmidt, C.; Robinson, C. V. A Comparative Cross-Linking Strategy to Probe Conformational Changes in Protein Complexes. *Nat. Protoc.* **2014**, 9 (9), 2224–2236.
- (79) Liu, F.; Rijkers, D. T. S.; Post, H.; Heck, A. J. R. Proteome-Wide Profiling of Protein Assemblies by Cross-Linking Mass Spectrometry. *Nat. Methods* **2015**, 12 (12), 1179–1184.
- (80) Engen, J. R.; Wales, T. E. Analytical Aspects of Hydrogen Exchange Mass Spectrometry. *Annu. Rev. Anal. Chem.* **2015**, 8 (1), 127–148.
- (81) Katta, V.; Chait, B. T. *Hydrogen/Deuterium Exchange Electrospray Ionization Mass Spectrometry: A Method for Probing Protein Conformational Changes in Solution*; 1993; Vol. 115.

- (82) Wang, L.; Chance, M. R. Protein Footprinting Comes of Age: Mass Spectrometry for Biophysical Structure Assessment. *Mol. Cell. Proteomics* **2017**, *16* (5), 706–716.
- (83) Gau, B. C.; Sharp, J. S.; Rempel, D. L.; Gross, M. L. Fast Photochemical Oxidation of Protein Footprints Faster than Protein Unfolding. *Anal. Chem.* **2009**, *81* (16), 6563–6571.
- (84) Mendoza, V. L.; Vachet, R. W. Probing Protein Structure by Amino Acid-Specific Covalent Labeling and Mass Spectrometry. *Mass Spectrom. Rev.* **2009**, *28* (5), 785–815.
- (85) Landgraf, R. R.; Chalmers, M. J.; Griffin, P. R. Automated Hydrogen/Deuterium Exchange Electron Transfer Dissociation High Resolution Mass Spectrometry Measured at Single-Amide Resolution. *J. Am. Soc. Mass Spectrom.* **2012**, *23* (2), 301–309.
- (86) Saltzberg, D. J.; Broughton, H. B.; Pellarin, R.; Chalmers, M. J.; Espada, A.; Dodge, J. A.; Pascal, B. D.; Griffin, P. R.; Humblet, C.; Sali, A. A Residue-Resolved Bayesian Approach to Quantitative Interpretation of Hydrogen–Deuterium Exchange from Mass Spectrometry: Application to Characterizing Protein–Ligand Interactions. *J. Phys. Chem. B* **2017**, *121* (15), 3493–3501.
- (87) Deng, B.; Lento, C.; Wilson, D. J. Hydrogen Deuterium Exchange Mass Spectrometry in Biopharmaceutical Discovery and Development – A Review. *Anal. Chim. Acta* **2016**, *940*, 8–20.
- (88) Zhang, H.; Wen, J.; Huang, R. Y. C.; Blankenship, R. E.; Gross, M. L. Mass Spectrometry-Based Carboxyl Footprinting of Proteins: Method Evaluation. *Int. J. Mass Spectrom.* **2012**, *312*, 78–86.
- (89) Leitner, A.; Joachimiak, L. A.; Unverdorben, P.; Walzthoeni, T.; Frydman, J.; Förster, F.; Aebersold, R. Chemical Cross-Linking/Mass Spectrometry Targeting Acidic Residues in Proteins and Protein Complexes. *Proc. Natl. Acad. Sci.* **2014**, *111* (26), 9455–9460.
- (90) Lanucara, F.; Holman, S. W.; Gray, C. J.; Eyers, C. E. The Power of Ion Mobility-Mass Spectrometry for Structural Characterization and the Study of Conformational Dynamics. *Nat. Chem.* **2014**, *6* (4), 281–294.
- (91) Wilm, M. S.; Mann, M. Electrospray and Taylor-Cone Theory, Dole’s Beam of Macromolecules at Last? *Int. J. Mass Spectrom. Ion Process.* **1994**, *136* (2–3), 167–180.
- (92) Hogan, C. J.; Carroll, J. A.; Rohrs, H. W.; Biswas, P.; Gross, M. L. Combined Charged Residue-Field Emission Model of Macromolecular Electrospray Ionization. *Anal. Chem.* **2009**, *81* (1), 369–377.
- (93) Konermann, L.; Ahadi, E.; Rodriguez, A. D.; Vahidi, S. Unraveling the Mechanism of Electrospray Ionization. *Anal. Chem.* **2013**, *85* (1), 2–9.
- (94) Breuker, K.; McLafferty, F. W. Stepwise Evolution of Protein Native Structure with Electrospray into the Gas Phase, 10-12 to 102 S. *Proc. Natl. Acad. Sci.* **2008**, *105* (47), 18145–18152.
- (95) Schmidt, A.; Karas, M.; Dülcks, T. Effect of Different Solution Flow Rates on Analyte Ion Signals in Nano-ESI MS, or: When Does ESI Turn into Nano-ESI? *J. Am. Soc. Mass Spectrom.* **2003**, *14* (5), 492–500.
- (96) Susa, A. C.; Xia, Z.; Williams, E. R. Small Emitter Tips for Native Mass Spectrometry of Proteins and Protein Complexes from Nonvolatile Buffers That Mimic the Intracellular Environment. *Anal. Chem.* **2017**, *89* (5), 3116–3122.
- (97) Juraschek, R.; Dulcks, T.; Karas, M. *Nanoelectrospray--More Than Just a Minimized-Flow Electrospray Ionization Source*; 1999; Vol. 10.
- (98) El-Faramawy, A.; Siu, K. W. M.; Thomson, B. A. Efficiency of Nano-Electrospray Ionization. *J. Am. Soc. Mass Spectrom.* **2005**, *16* (10), 1702–1707.

- (99) Ruotolo, B. T.; Robinson, C. V. Aspects of Native Proteins Are Retained in Vacuum. *Curr. Opin. Chem. Biol.* **2006**, *10* (5), 402–408.
- (100) Hall, Z.; Robinson, C. V. Do Charge State Signatures Guarantee Protein Conformations? *J. Am. Soc. Mass Spectrom.* **2012**, *23* (7), 1161–1168.
- (101) Suckau, D.; Shi, Y.; Beu, S. C.; Senko, M. W.; Quinn, J. P.; Wampler, F. M.; McLafferty, F. W. Coexisting Stable Conformations of Gaseous Protein Ions. *Proc. Natl. Acad. Sci.* **1993**, *90* (3), 790–793.
- (102) Ruotolo, B. T.; Giles, K.; Campuzano, I.; Sandercock, A. M.; Bateman, R. H.; Robinson, C. V. Biochemistry: Evidence for Macromolecular Protein Rings in the Absence of Bulk Water. *Science (80-.)*. **2005**, *310* (5754), 1658–1661.
- (103) Seo, J.; Hoffmann, W.; Warnke, S.; Bowers, M. T.; Pagel, K.; von Helden, G. Retention of Native Protein Structures in the Absence of Solvent: A Coupled Ion Mobility and Spectroscopic Study. *Angew. Chemie - Int. Ed.* **2016**, *55* (45), 14173–14176.
- (104) Daly, S.; Knight, G.; Halim, M. A.; Kulesza, A.; Choi, C. M.; Chirot, F.; MacAleese, L.; Antoine, R.; Dugourd, P. Action-FRET of a Gaseous Protein. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (1), 38–49.
- (105) Barr, J. D.; Shi, L.; Russell, D. H.; Clemmer, D. E.; Holliday, A. E. Following a Folding Transition with Capillary Electrophoresis and Ion Mobility Spectrometry. *Anal. Chem.* **2016**, *88* (22), 10933–10939.
- (106) 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), 2809–2815.
- (107) Mitchell Wells, J.; McLuckey, S. A. Collision-Induced Dissociation (CID) of Peptides and Proteins. In *Methods in Enzymology*; Academic Press, 2005; Vol. 402, pp 148–185.
- (108) Stannard, P. R.; Gelbart, W. M. Intramolecular Vibrational Energy Redistribution. *J. Phys. Chem.* **1981**, *85* (24), 3592–3599.
- (109) Collins, D. C.; Lee, M. L. Developments in Ion Mobility Spectrometry-Mass Spectrometry. *Analytical and Bioanalytical Chemistry*. 2002, pp 66–73.
- (110) Clemmer, D. E.; Jarrold, M. F. Ion Mobility Measurements and Their Applications to Clusters and Biomolecules. *J. Mass Spectrom.* **1997**, *32* (6), 577–592.
- (111) Giles, K.; Pringle, S. D.; Worthington, K. R.; Little, D.; Wildgoose, J. L.; Bateman, R. H. Applications of a Travelling Wave-Based Radio-Frequency-Only Stacked Ring Ion Guide. *Rapid Commun. Mass Spectrom.* **2004**, *18* (20), 2401–2414.
- (112) Shvartsburg, A. A.; Smith, R. D. Fundamentals of Traveling Wave Ion Mobility Spectrometry. *Anal. Chem.* **2008**, *80* (24), 9689–9699.
- (113) Richardson, K.; Langridge, D.; Giles, K. Fundamentals of Travelling Wave Ion Mobility Revisited: I. Smoothly Moving Waves. *Int. J. Mass Spectrom.* **2018**, *428*, 71–80.
- (114) Shvartsburg, A. A.; Smith, R. D. Fundamentals of Traveling Wave Ion Mobility Spectrometry. *Anal. Chem.* **2008**, *80* (24), 9689–9699.
- (115) Wyttenbach, T.; Von Helden, G.; Bowers, M. T. Gas-Phase Conformation of Biological Molecules: Bradykinin. *J. Am. Chem. Soc.* **1996**, *118* (35), 8355–8364.
- (116) Clemmer, D. E.; Hudgins, R. R.; Jarrold, M. F. Naked Protein Conformations: Cytochrome c in the Gas Phase. *J. Am. Chem. Soc.* **1995**, *117* (40), 10141–10142.
- (117) Hyung, S. J.; Robinson, C. V.; Ruotolo, B. T. Gas-Phase Unfolding and Disassembly Reveals Stability Differences in Ligand-Bound Multiprotein Complexes. *Chem. Biol.* **2009**, *16* (4), 382–390.

- (118) Shelimov, K. B.; Clemmer, D. E.; Hudgins, R. R.; Jarrold, M. F. Protein Structure in Vacuo: Gas-Phase Conformations of BPTI and Cytochrome C. *J. Am. Chem. Soc.* **1997**, *119* (9), 2240–2248.
- (119) Shelimov, K. B.; Jarrold, M. F. Conformations, Unfolding, and Refolding of Apomyoglobin in Vacuum: An Activation Barrier for Gas-Phase Protein Folding. *J. Am. Chem. Soc.* **1997**, *119* (13), 2987–2994.
- (120) Ruotolo, B. T.; Hyung, S. J.; Robinson, P. M.; Giles, K.; Bateman, R. H.; Robinson, C. V. Ion Mobility-Mass Spectrometry Reveals Long-Lived, Unfolded Intermediates in the Dissociation of Protein Complexes. *Angew. Chemie - Int. Ed.* **2007**, *46* (42), 8001–8004.
- (121) Eschweiler, J. D.; Martini, R. M.; Ruotolo, B. T. Chemical Probes and Engineered Constructs Reveal a Detailed Unfolding Mechanism for a Solvent-Free Multidomain Protein. *J. Am. Chem. Soc.* **2017**, *139* (1), 534–540.
- (122) Hopper, J. T. S.; Oldham, N. J. Collision Induced Unfolding of Protein Ions in the Gas Phase Studied by Ion Mobility-Mass Spectrometry: The Effect of Ligand Binding on Conformational Stability. *J. Am. Soc. Mass Spectrom.* **2009**, *20* (10), 1851–1858.
- (123) Reading, E.; Liko, I.; Allison, T. M.; Benesch, J. L. P.; Laganowsky, A.; Robinson, C. V. The Role of the Detergent Micelle in Preserving the Structure of Membrane Proteins in the Gas Phase. *Angew. Chemie - Int. Ed.* **2015**, *54* (15), 4577–4581.
- (124) Niu, S.; Ruotolo, B. T. Collisional Unfolding of Multiprotein Complexes Reveals Cooperative Stabilization upon Ligand Binding. *Protein Sci.* **2015**, *24* (8), 1272–1281.
- (125) Niu, S.; Rabuck, J. N.; Ruotolo, B. T. Ion Mobility-Mass Spectrometry of Intact Protein-Ligand Complexes for Pharmaceutical Drug Discovery and Development. *Curr. Opin. Chem. Biol.* **2013**, *17* (5), 809–817.
- (126) Allison, T. M.; Reading, E.; Liko, I.; Baldwin, A. J.; Laganowsky, A.; Robinson, C. V. Quantifying the Stabilizing Effects of Protein-Ligand Interactions in the Gas Phase. *Nat. Commun.* **2015**, *6*, 8551.
- (127) Eschweiler, J. D.; Rabuck-Gibbons, J. N.; Tian, Y.; Ruotolo, B. T. CIUSuite: A Quantitative Analysis Package for Collision Induced Unfolding Measurements of Gas-Phase Protein Ions. *Anal. Chem.* **2015**, *87* (22), 11516–11522.
- (128) Sivalingam, G. N.; Yan, J.; Sahota, H.; Thalassinou, K. Amphitrite: A Program for Processing Travelling Wave Ion Mobility Mass Spectrometry Data. *Int. J. Mass Spectrom.* **2013**, *345–347*, 54–62.
- (129) Haynes, S. E.; Polasky, D. A.; Dixit, S. M.; Majmudar, J. D.; Neeson, K.; Ruotolo, B. T.; Martin, B. R. Variable-Velocity Traveling-Wave Ion Mobility Separation Enhancing Peak Capacity for Data-Independent Acquisition Proteomics. *Anal. Chem.* **2017**, *89* (11), 5669–5672.
- (130) Rabuck, J. N.; Hyung, S. J.; Ko, K. S.; Fox, C. C.; Soellner, M. B.; Ruotolo, B. T. Activation State-Selective Kinase Inhibitor Assay Based on Ion Mobility-Mass Spectrometry. *Anal. Chem.* **2013**, *85* (15), 6995–7002.
- (131) Koeniger, S. L.; Merenbloom, S. I.; Valentine, S. J.; Jarrold, M. F.; Udseth, H. R.; Smith, R. D.; Clemmer, D. E. An IMS-IMS Analogue of MS-MS. *Anal. Chem.* **2006**, *78* (12), 4161–4174.
- (132) Merenbloom, S. I.; Koeniger, S. L.; Valentine, S. J.; Plasencia, M. D.; Clemmer, D. E. IMS-IMS and IMS-IMS-IMS/MS for Separating Peptide and Protein Fragment Ions. *Anal. Chem.* **2006**, *78* (8), 2802–2809.
- (133) Pierson, N. A.; Clemmer, D. E. An IMS-IMS Threshold Method for Semi-Quantitative

- Determination of Activation Barriers: Interconversion of Proline Cis \rightleftharpoons Trans Forms in Triply Protonated Bradykinin. *Int. J. Mass Spectrom.* **2015**, 377 (1), 646–654.
- (134) Mehmood, S.; Allison, T. M.; Robinson, C. V. Mass Spectrometry of Protein Complexes: From Origins to Applications. *Annu. Rev. Phys. Chem.* **2015**, 66 (1), 453–474.
- (135) Felitsyn, N.; Kitova, E. N.; Klassen, J. S. Thermal Decomposition of a Gaseous Multiprotein Complex Studied by Blackbody Infrared Radiative Dissociation. Investigating the Origin of the Asymmetric Dissociation Behavior. *Anal. Chem.* **2001**, 73 (19), 4647–4661.
- (136) Hall, Z.; Politis, A.; Bush, M. F.; Smith, L. J.; Robinson, C. V. Charge-State Dependent Compaction and Dissociation of Protein Complexes: Insights from Ion Mobility and Molecular Dynamics. *J. Am. Chem. Soc.* **2012**, 134 (7), 3429–3438.
- (137) Fegan, S. K.; Thachuk, M. A Charge Moving Algorithm for Molecular Dynamics Simulations of Gas-Phase Proteins. *J. Chem. Theory Comput.* **2013**, 9 (6), 2531–2539.
- (138) Fegan, S. K.; Thachuk, M. Controlling Dissociation Channels of Gas-Phase Protein Complexes Using Charge Manipulation. *J. Am. Soc. Mass Spectrom.* **2014**, 25 (5), 722–728.
- (139) Thachuk, M.; Fegan, S. K.; Raheem, N. Description and Control of Dissociation Channels in Gas-Phase Protein Complexes. *J. Chem. Phys.* **2016**, 145 (6), 65101.
- (140) Popa, V.; Trecroce, D. A.; McAllister, R. G.; Konermann, L. Collision-Induced Dissociation of Electrosprayed Protein Complexes: An All-Atom Molecular Dynamics Model with Mobile Protons. *J. Phys. Chem. B* **2016**, 120 (23), 5114–5124.
- (141) Freeke, J.; Robinson, C. V.; Ruotolo, B. T. Residual Counter Ions Can Stabilise a Large Protein Complex in the Gas Phase. *Int. J. Mass Spectrom.* **2010**, 298 (1–3), 91–98.
- (142) Wagner, N. D.; Kim, D.; Russell, D. H. Increasing Ubiquitin Ion Resistance to Unfolding in the Gas Phase Using Chloride Adduction: Preserving More “Native-Like” Conformations Despite Collisional Activation. *Anal. Chem.* **2016**, 88 (11), 5934–5940.
- (143) Han, L.; Hyung, S. J.; Mayers, J. J. S.; Ruotolo, B. T. Bound Anions Differentially Stabilize Multiprotein Complexes in the Absence of Bulk Solvent. *J. Am. Chem. Soc.* **2011**, 133 (29), 11358–11367.
- (144) Han, L.; Hyung, S. J.; Ruotolo, B. T. Bound Cations Significantly Stabilize the Structure of Multiprotein Complexes in the Gas Phase. *Angew. Chemie - Int. Ed.* **2012**, 51 (23), 5692–5695.
- (145) Samulak, B. M.; Niu, S.; Andrews, P. C.; Ruotolo, B. T. Ion Mobility-Mass Spectrometry Analysis of Cross-Linked Intact Multiprotein Complexes: Enhanced Gas-Phase Stabilities and Altered Dissociation Pathways. *Anal. Chem.* **2016**, 88 (10), 5290–5298.
- (146) Zhong, Y.; Han, L.; Ruotolo, B. T. Collisional and Coulombic Unfolding of Gas-Phase Proteins: High Correlation to Their Domain Structures in Solution. *Angew. Chemie - Int. Ed.* **2014**, 53 (35), 9209–9212.
- (147) Laszlo, K. J.; Munger, E. B.; Bush, M. F. Folding of Protein Ions in the Gas Phase after Cation-to-Anion Proton-Transfer Reactions. *J. Am. Chem. Soc.* **2016**, 138 (30), 9581–9588.
- (148) Beveridge, R.; Migas, L. G.; Payne, K. A. P.; Scrutton, N. S.; Leys, D.; Barran, P. E. Mass Spectrometry Locates Local and Allosteric Conformational Changes That Occur on Cofactor Binding. *Nat. Commun.* **2016**, 7, 12163.
- (149) Byrne, D. P.; Vonderach, M.; Ferries, S.; Brownridge, P. J.; Evers, C. E.; Evers, P. A. CAMP-Dependent Protein Kinase (PKA) Complexes Probed by Complementary

- Differential Scanning Fluorimetry and Ion Mobility-Mass Spectrometry. *Biochem. J.* **2016**, *473* (19), 3159–3175.
- (150) Mehmood, S.; Marcoux, J.; Gault, J.; Quigley, A.; Michaelis, S.; Young, S. G.; Carpenter, E. P.; Robinson, C. V. Mass Spectrometry Captures Off-Target Drug Binding and Provides Mechanistic Insights into the Human Metalloprotease ZMPSTE24. *Nat. Chem.* **2016**, *8* (12), 1152–1158.
- (151) Laganowsky, A.; Reading, E.; Allison, T. M.; Ulmschneider, M. B.; Degiacomi, M. T.; Baldwin, A. J.; Robinson, C. V. Membrane Proteins Bind Lipids Selectively to Modulate Their Structure and Function. *Nature* **2014**, *510* (7503), 172–175.
- (152) Liu, Y.; Cong, X.; Liu, W.; Laganowsky, A. Characterization of Membrane Protein–Lipid Interactions by Mass Spectrometry Ion Mobility Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (4), 579–586.
- (153) Campuzano, I. D. G.; Larriba, C.; Bagal, D.; Schnier, P. D. Ion Mobility and Mass Spectrometry Measurements of the Humanized IgGk NIST Monoclonal Antibody. *ACS Symp. Ser.* **2015**, *1202*, 75–112.
- (154) Tian, Y.; Han, L.; Buckner, A. C.; Ruotolo, B. T. Collision Induced Unfolding of Intact Antibodies: Rapid Characterization of Disulfide Bonding Patterns, Glycosylation, and Structures. *Anal. Chem.* **2015**, *87* (22), 11509–11515.
- (155) Ferguson, C. N.; Gucinski-Ruth, A. C. Evaluation of Ion Mobility-Mass Spectrometry for Comparative Analysis of Monoclonal Antibodies. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (5), 822–833.
- (156) Pisupati, K.; Tian, Y.; Okbazghi, S.; Benet, A.; Ackermann, R.; Ford, M.; Saveliev, S.; Hosfield, C. M.; Urh, M.; Carlson, E.; et al. A Multidimensional Analytical Comparison of Remicade and the Biosimilar Remsima. *Anal. Chem.* **2017**, *89* (9), 4838–4846.
- (157) Botzanowski, T.; Erb, S.; Hernandez-Alba, O.; Ehkirch, A.; Colas, O.; Wagner-Rousset, E.; Rabuka, D.; Beck, A.; Drake, P. M.; Cianfèrani, S. Insights from Native Mass Spectrometry Approaches for Top- and Middle- Level Characterization of Site-Specific Antibody-Drug Conjugates. *MAbs* **2017**, *9* (5), 801–811.

Chapter 2 Fixed-Charge Trimethyl Pyrylium Modification for Enabling Enhanced Top-Down Mass Spectrometry Sequencing of Intact Protein Complexes

Adapted with permission from: Daniel A. Polasky, Frederik Lermyte; Michael Nshanian, Frank Sobott, Philip C. Andrews, Joseph A. Loo, and Brandon T. Ruotolo. Fixed-Charge Trimethyl Pyrylium Modification for Enabling Enhanced Top-Down Mass Spectrometry Sequencing of Intact Protein Complexes. *Anal. Chem.* **2018**, *90* (4), 2756–2764.

2.1 Abstract

Mass spectrometry of intact proteins and protein complexes has the potential to provide a transformative level of information on biological systems, ranging from sequence and post-translational modification analysis to the structures of whole protein assemblies. This ambitious goal requires the efficient fragmentation of both intact proteins and the macromolecular, multi-component machines they collaborate to create through non-covalent interactions. Improving technologies in an effort to achieve such fragmentation remains perhaps the greatest challenge facing current efforts to comprehensively analyze cellular protein composition and is essential to realizing the full potential of proteomics. In this work, we describe the use of a trimethyl pyrylium (TMP) fixed-charge covalent labeling strategy aimed at enhancing fragmentation for challenging intact proteins and intact protein complexes. Combining analysis of TMP-modified and unmodified protein complexes results in a greater diversity of regions within the protein that give rise to fragments, and results in an up to 2.5-fold increase in sequence coverage when compared to unmodified protein alone, for protein complexes up to 148 kDa. TMP modification offers a simple and powerful platform to expand the capabilities of existing mass spectrometric instrumentation for the complete characterization of intact protein assemblies.

2.2 Introduction

The rapid characterization of proteins by mass spectrometry (MS) has emerged as a powerful platform for understanding the details of biology and biochemistry at an unprecedented level of detail. Conventional ‘bottom-up’ proteomics workflows utilize enzymatic digestion prior to separation and MS to sequence the peptides excised from proteins within whole cell lysates¹. This rapid characterization of exceptionally complex protein mixtures has enabled a revolution in biological analysis², ranging from the elucidation of human disease processes to supporting the determination of three-dimensional structures for biochemically-central multi-protein complexes.

Despite the advances described above, current ‘bottom-up’ proteomics approaches are typically unable to completely identify all the post-translational modifications and proteoforms that are crucially important for biological function³. In response to such deficiencies, ‘top-down’ MS technologies have been developed that are capable of sequencing intact proteins without enzymatic digestion⁴⁻⁶. Such tools are typically capable of capturing a wide-ranging snapshot of proteoform composition, but typically only able to completely characterize relatively small proteins^{7,8}. This deficiency is amplified by the fact that most proteins form multi-protein assemblies in order to accomplish their biological function^{9,10}. Detecting and analyzing these complexes requires the characterization of not just individual intact proteins, but the preservation and analysis of the non-covalent complexes they form in the cell. Both bottom-up and typical top-down proteomics canonically require denaturation and/or enzymatic digestion of proteins, often precluding any analysis of the structure and dynamics of these complex assemblies.

Direct characterization of intact protein assemblies using native MS¹¹ offers a promising alternative for the characterization of functional multiprotein machines¹²⁻¹⁴. Furthermore, native MS has the potential to provide protein sequence and structural information

in the context of the same experiment, dramatically increasing the throughput of modern protein structure/function studies. However, sequencing technology coupled to native MS experiments lags far behind complementary bottom-up and top-down approaches targeting small, monomeric proteins. Incomplete fragmentation of proteins limits the ability of native MS to identify unknown proteins within complexes and prevents detailed analysis of the proteoforms incorporated within such assemblies. While collision induced dissociation (CID) is a widely available and effective technology for peptide sequencing, CID information extracted from large proteins and protein complexes analyzed under native conditions is often limited by the low charge density observed for analytes under such conditions¹⁵. In many cases, sequence coverage is concentrated into a few labile regions, e.g. flexible or terminal loop areas⁸. Achieving full sequence coverage for such large protein systems is one of the key challenges facing top-down proteomics, as well as the establishment of native MS workflows for wide-ranging structural proteomics.

Approaches to improve sequence coverage in top-down proteomics have focused largely on the development of ion activation paradigms other than CID. Electron transfer and capture dissociation (ETD^{16,17} and ECD¹⁸⁻²¹), infrared multiphoton dissociation (IRMPD^{22,23}), electron ionization dissociation (EID²⁴), and ultraviolet photodissociation (UVPD^{7,25}) have demonstrated significant improvements in sequence coverage relative to and in combination with CID. Current state of the art protein ion fragmentation is typically achieved by employing several activation methods in concert, but often still provides only modest coverage for large proteins and protein complexes^{7,8}. Implementing multiple activation techniques often requires instrument modification beyond the capability of many laboratories and is expensive, thus many promising new techniques remain limited to a small subset of available MS platforms. CID

remains the most widely available fragmentation technique, making further development of CID techniques an attractive target for improving protein sequencing technology.

Chemical modification of protein complexes offers the potential for a new set of complementary methods to expand intact protein characterization by MS. Derivatization of single peptides with reagents that bear intrinsic positive charge has previously been shown to alter the dissociation pathways accessed in CID, improving sequencing for bottom-up proteomics experiments²⁶⁻²⁹. Several reports have used fixed charge derivatization to alter the charge states of electrosprayed protein ions^{30,31}, but expanding this concept to sequencing of large proteins and protein complexes has proven highly challenging due to the difficulty of maintaining a fixed charge at the energies required to cause backbone fragmentation in such systems³². In this report, we present the use of trimethyl pyrylium^{33,34} (TMP) to covalently tether a stable positive charge to protein lysine side chains, altering the energy of various dissociation pathways to enable improved sequencing for large protein complex ions. Fixed charge modification by TMP provides orthogonal sequence coverage to other forms of biomolecular activation in the gas phase, opening a new pathway for improved sequence coverage of challenging protein targets using a simple derivatization that relies upon a commercially available and inexpensive reagent.

2.3 Experimental Methods

2.3.1 Chemical Modification

Avidin from chicken egg white, Alcohol dehydrogenase (ADH) from *Saccharomyces cerevisiae*, Ovalbumin from chicken (all from Sigma Aldrich, St. Louis, MO), were dissolved in 100mM triethylammonium bicarbonate (TEAB, Sigma Aldrich), pH 8.5, to make solutions containing 25uM protein for chemical modification. 2,4,6-Trimethyl pyrylium (TMP) tetrafluoroborate (Alfa Aesar, Haverhill, MA) was dissolved in 100mM TEAB, pH 8.5, vortexed

for ten seconds to dissolve, and quickly added to protein solutions to 10- to 25-fold molar excess relative to the reactive Lysine residues present. Reaction solutions were briefly vortexed and allowed to react for 24 hours at room temperature. Following modification, proteins were buffer exchanged sequentially into 1M ammonium acetate, then 200mM ammonium acetate, both pH 7.4 (Sigma Aldrich), with P6 (Avidin) or P30 (ADH, Ovalbumin) microspin columns (BioRad Laboratories, Hercules, CA) according to manufacturer instructions. Buffer exchanged samples were either analyzed immediately or flash frozen with liquid nitrogen and stored at -80C prior to analysis.

2.3.2 Ion Mobility-Mass Spectrometry

A quadrupole ion mobility-time of flight mass spectrometer (Synapt G2 HDMS, Waters, Milford, MA) was used for all ion mobility experiments. 5uL of buffer-exchanged protein solution (20uM) was transferred to a gold-coated borosilicate capillary (0.78mm i.d., Harvard Apparatus, Holliston, MA) for direct infusion. Instrumental settings were optimized to preserve intact protein complexes prior to activation: capillary voltage 1.5 kV, sample cone 40 V, extraction cone 0 V. Gas flows (mL/min): source: 50, trap: 6 (Avidin) or 8 (ADH, Ovalbumin), helium cell: 200, IM separation: 90. IMS traveling wave settings were the same for all proteins: wave velocity: 150 m/s, wave height: 20 V, IMS bias: 5 V. Backing pressure was set to 5.5 mbar (Avidin), or 8.0 mbar (ADH, Ovalbumin). A single charge state of each protein complex was selected and collisionally activated in the trap cell (trap collision voltage: Avidin, 140 V; Ovalbumin, 160 V; ADH, 200 V) prior to ion mobility separation. Trap (collision cell) pressures were 3.8×10^{-2} mbar (Avidin, Ovalbumin) or 4.4×10^{-2} mbar (ADH). Time of flight pressure was 1.8×10^{-6} mbar for all analyses. Scans were combined for 30 seconds (Avidin) or 10 minutes (Avidin, ADH, Ovalbumin) to obtain sufficient signal to noise ratios.

2.3.3 FT-ICR mass spectrometry

After buffer exchange, protein solutions (diluted to 10 μ M in 100 mM aqueous ammonium acetate) were transferred to a metal-coated borosilicate capillary (Au/Pd coated, 1 μ m i.d., Thermo Fisher Scientific, West Palm Beach, FL, USA) and mounted in the nanospray ion source. Mass spectrometry experiments were performed using a 15T SolariX FTICR instrument equipped with an infinity cell (Bruker Daltonics, Bremen, Germany). The following instrument settings were used: ESI voltage: 1.2 – 1.3 kV, dry gas temperature: 180 °C, flow rate: 2.0 L/min, RF amplitude of ion funnels: 200 Vpp, Funnel 1 voltage: 200 V, Funnel 2 voltage: 6 V, Skimmer 1 voltage: 60 – 180 V (longer for larger precursor masses; not enough to induce fragmentation), Skimmer 2 voltage: 5 V, multipole 1 RF: 2 MHz, quadrupole RF: 1.4 MHz, transfer hexapole RF: 2 MHz, time-of-flight: 1 – 2 ms (higher for larger proteins). CID experiments were performed at collision cell voltages from 30-100V at a collision cell gas flow of 35%. Ions were accumulated for 500 ms in the collision cell before entering the infinity ICR cell. Source, quadrupole, and UHF pressures were 2.5e0, 3.7e-6, and 1.8e-9 mbar, respectively. At least 200 scans were combined to obtain a sufficiently high signal-to-noise ratio. IRMPD was performed using a 30 W CO₂ laser (Synrad, Mukilteo, WA, USA) interfaced to the back of the instrument. Laser power was held at 95% with an irradiation time of 1 s.

2.3.4 Data Analysis

For ion mobility-mass spectrometry data, slices of the 2D IM-MS data corresponding to peptide charge states were extracted from raw data to text format using TWIMExtract³⁵, a data querying tool developed for processing Waters IM-MS data. Extracted data was smoothed (Savitsky-Golay, 0.2 m/z window size, 3 cycles), peak-picked, and de-isotoped (max charge 5, isotope mass tolerance 0.05 m/z , isotope intensity tolerance 100%) using mMass v5.5.0³⁶⁻³⁸ to

generate peak lists. Peaks were identified using an in-house single protein search script written in java to allow for identification of fragments containing variable numbers of intrinsically charged TMP modifications given the starting protein sequence. *a*, *b*, and *y*-type ions and neutral losses of water and ammonia were considered for identification of peaks after examination of the data revealed little contribution from other fragmentation pathways. Only terminal fragments were considered, as statistically confident identification of internal fragments was not possible given available resolution and mass accuracy. Mass tolerance of 10 ppm was used as the cutoff for peak identification. At least three replicates were used for all fragmentation data presented (Avidin *n*=5, ADH *n*=3, Ovalbumin *n*=3). Fragments identified in fewer than the majority of replicates were excluded from analyses. Error bars for sequence coverage plots were presented as two times the standard deviation in the number of cleavage sites observed for each replicate. For FT-ICR data, Data Analysis 4.0 (Bruker Daltonics, Billerica, MA) was used to extract and process raw data into peak lists. Peak lists were then processed using the same in-house search program and parameters as ion mobility-mass spectrometry data.

2.4 Results

In CID experiments, a large number of collisions with inert gas molecules are required to impart sufficient energy for the production of sequence informative fragment ions. This slow heating, along with the relatively rapid rate of intramolecular redistribution of the imparted vibrational energy³⁹, results in fragmentation of the most labile bonds in the protein. Charge mobility plays a key role in the CID process, as “mobile protons”^{40–42} move along the peptide backbone, triggering *b*- and *y*-ion formation across a range of sites. Protein complexes preserved through the electrospray ionization (ESI) process (*i.e.* using native mass spectrometry conditions) typically have a low charge-to-mass ratio, and are furthermore stabilized by

secondary, tertiary, and quaternary structural elements, making typical mobile-proton based CID an efficient process for only a small subset of the peptide bonds available. Fixing intrinsically-charged moieties has been shown to alter the fragmentation of peptides from primarily *b*- and *y*-type ions formed through mobile proton type fragmentation to primarily *a*-ions formed through charge-remote mechanisms²⁶⁻²⁹, but prior to this report, had not yet been extended to the CID of intact proteins and their assemblies.

The compound 2,4,6-trimethyl pyrylium (TMP) reacts with primary amines to produce an intrinsically charged pyridinium salt at the nitrogen of the original primary amine^{33,34} (Figure

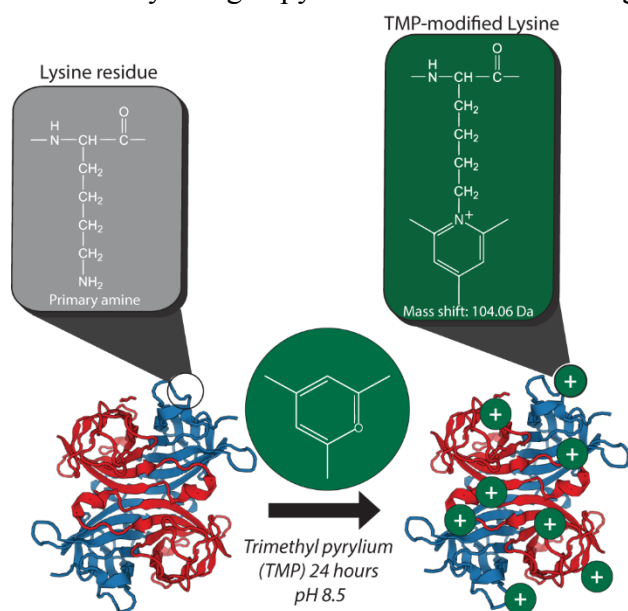


Figure 2-1 Charge-fixing chemical modification scheme with TMP. At left, the cartoon structure of an example protein complex prior to modification with the structure of a lysine residue highlighted above. At right, the effect of TMP modification on primary amines present in the protein complex (lysine residues and the N-terminus). Reaction of the pyrylium with a primary amine results in a pyridinium derivative with a fixed positive charge on the nitrogen atom of the former amine. Plus signs indicate positive charges localized to lysine residues or N-termini throughout the protein complex

2-1), resulting in the addition of $C_8H_8^+$ with a mass shift of 104.06205 Da (Figure 2-1, Figure I-1). Labeling under native conditions is typically not stoichiometric (Figure I-2, Figure I-3), resulting in a mixture of modification states. The TMP derivatization reaction proceeds under conditions that allow for native-like buffer conditions (see methods), resulting in labeling of intact protein complexes without significantly altering their structure (Figure I-4 - Figure I-7). IM profiles of the three protein complexes examined in this report before

and after labeling, demonstrate the preservation of a single, compact conformation through the labeling procedure. The drift times of the modified proteins increase slightly (5-10%), which can

be attributed to the additional mass (up to ~10 kDa) added to the complexes by the reaction, which is consistent with previous data involving the native labeling of protein complexes⁴³. Furthermore, collisional activation of TMP-modified Avidin tetramer results in an unfolding pathway that mirrors that of the unmodified tetramer (Figure I-7), indicating preservation of the existing overall structure.

Following ESI, both TMP fixed charges and mobile protons can influence the CID behavior of the complex. Unlike most commonly used, intrinsically charged modifications, such as sulfonium-based reagents^{44,45} and quaternary amines^{32,46–48}, gas-phase decomposition of TMP-modified lysine is not energetically favorable under typical CID conditions for proteins and peptides, ensuring that the charges remain fixed throughout the process. This enables TMP-based fixed charges to alter the potential energy landscape associated with protein fragmentation and thus enhance the formation of sequence-informative product ions by CID.

2.4.1 Fixed-charge modification enhances sequence coverage in a model protein complex

Avidin, a 64 kDa homo-tetramer, has been studied extensively as a CID model for noncovalent protein assemblies in native MS^{49,50}. It exhibits some of the strongest noncovalent interactions between subunits of known protein complexes, making the tetramer a challenging target for top-down sequencing. TMP-modified Avidin tetramer was compared to unmodified Avidin using top-down IM-MS to determine the benefits of chemical modification for extracting protein sequence information directly from protein complex ions (Figure 2-2). The ion mobility separation, specifically, was used to separate peptides of different charge states⁵¹, enhancing the ability of the time-of-flight mass analyzer to characterize the ion populations resulting from fragmentation of the complexes examined.

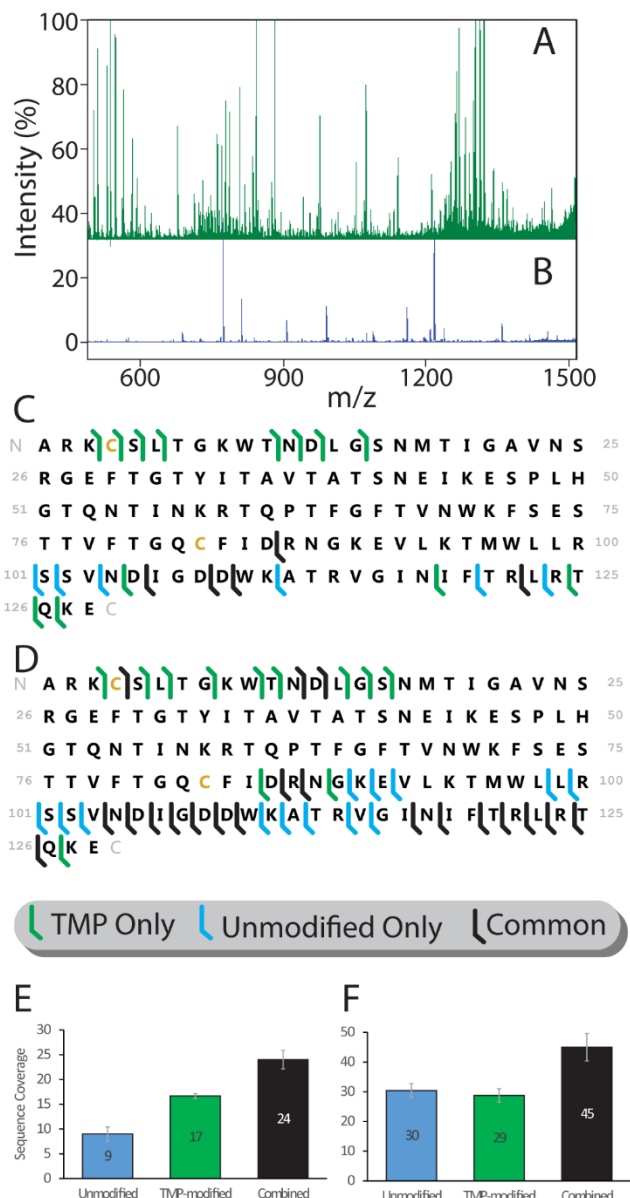


Figure 2-2 Enhanced sequencing of the Avidin tetramer following TMP modification. **(A)** Mass spectrum of fragments from CID of TMP-modified Avidin compared to **(B)**, mass spectrum of fragments from unmodified Avidin presented on the same intensity axis. Sequencing information obtained from intact Avidin tetramer for thirty-second **(C)** and ten-minute **(D)** accumulation times (N=3). Cleavage sites unique to TMP-modified Avidin are highlighted in green, those unique to unmodified Avidin in blue, and common to both states in black. **(E)** Total sequence coverage obtained from thirty-second fragment accumulation, or **(F)** ten-minute fragment accumulation, for unmodified (blue), TMP-modified (green), and both datasets combined (black).

greatest enhancements to the number of product ion populations and their orthogonality to those

Top-down analyses often require significant signal accumulation times, as extensively fragmenting a single precursor results in many low intensity fragment ions⁵². The impact of TMP modification was examined on timescales that both mimic those utilized in the context of on-line separations coupled to MS detection (30 seconds total accumulation time) and an in-depth full coverage experiment, typically associated with separation tools coupled off-line to top-down MS (10 minutes total accumulation time). While significant accumulation times offer the most information on TMP product ion populations, the majority of applied top-down proteomics is typically done on the timescale of chromatographic separations, where elution of a species typically occurs on a timescale of seconds to tens of seconds. As such, the 30 second accumulation time data shown in Figure 2-2, where we observe the

produced from unmodified protein ions, represents the most practical assessment of the ability of TMP to serve as part of current top-down sequencing workflows.

The low mass region of the mass spectra acquired for unmodified and TMP-modified Avidin tetramers over the ten-minute time frame discussed above (Figure 2-2 A-B) exhibit a more even distribution of intensity amongst many fragment peaks when compared to equivalent spectra acquired for the unmodified protein. For example, CID of unmodified Avidin typically generates approximately five main fragment peaks that exceed 20% relative intensity (Figure 2-2 B), which is typical of the CID of proteins having low, native-like charge states. The spectrum from TMP-modified Avidin, in contrast, shows intensity more evenly distributed across dozens of peaks (Figure 2-2 A). This fragment ion intensity distribution enables, in part, the substantial improvement in sequencing observed in our thirty-second accumulation runs, where TMP-modified Avidin generates nearly double the sequence coverage (17 cleavage sites vs 9) when compared to equivalent data for the unmodified protein tetramer (Figure 2-2 C, E).

Most cleavage sites observed are associated with numerous fragment ions (multiple charge states, neutral losses, and modification numbers), thus producing both a rich fragment spectra as well as lower total numbers of cleavage sites. Modified Lysine residues are not explicitly labeled in Figure 2-2 as the exact location of modification cannot be determined in all cases. Over the course of a full ten-minute accumulation, many peaks that are generated in very low abundance in the unmodified Avidin accumulate sufficient signal to be resolved and identified, reducing the difference in total coverage between modified and unmodified Avidin (Figure 2-2 D, F). The combined datasets, however, maintain a significant improvement in sequence coverage relative to any individual sequencing dataset due largely to the orthogonality of regions we observe to be covered within the Avidin sequence when using modified and

unmodified precursor ions. We note that Avidin contains a highly heterogeneous glycosylation site at residue ASN-17^{53,54}. Top-down fragmentation datasets from neither control nor TMP-modified Avidin reveal any evidence of this modification, despite extensive coverage of the n-terminus of the protein in our TMP-related data up to SER-15.

TMP modification acts to diversify the structural elements of the Avidin assembly from which sequence information can be obtained in addition to improving the total number of fragments observed. Figure 2-3 A shows the X-ray structure of Avidin⁵⁵ (PDB code: 1AVD), where highlighted residues indicate a detected fragmentation event for unmodified (Figure 2-3 B, E) and TMP-modified (Figure 2-3 C, F) tetramer. We use this representation in order to visualize the locations of fragmentation events across the protein surface, taking into account the monomer and dimer substructures of the Avidin complex, which are shown in Figure 2-3 in place of the

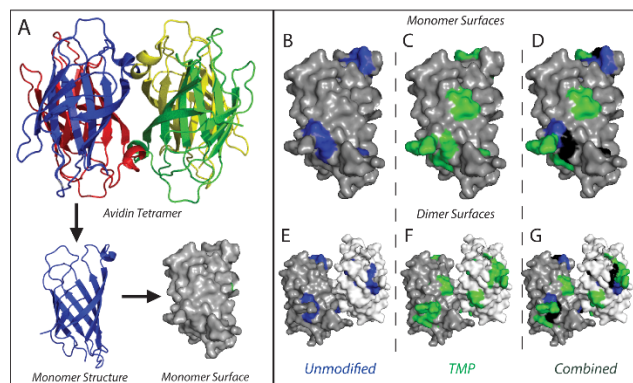


Figure 2-3 Mapping locations of peptide bond cleavage on Avidin. **(A)** Depiction of the monomer surface map used for comparison of cleavage locations, generated from Avidin crystal structure (PDB code: 1AVD). Monomer and dimer surfaces from the complex were used for ease of view. **(B-D)** Cleavage site map of fragments from thirty-second accumulation for unmodified, TMP-modified, and both combined, respectively. **(E-G)** Cleavage site maps from (B-D) presented on dimer surfaces to highlight interface locations in the complete Avidin tetrameric complex. In the combined views (D, G), blue coloring indicates cleavage sites unique to unmodified Avidin, green coloring indicates those unique to TMP-modified Avidin, and black coloring indicates sites common to both.

tetramer for simplicity. Fragmentation in the unmodified Avidin is confined to two main regions of the surface map with very little coverage of the total protein surface. In contrast, the regions of coverage in TMP-modified Avidin cover more of the protein surface and access regions left undetected in our datasets from unmodified Avidin, particularly at the n-terminus of the protein.

The orthogonality of coverage is clearly demonstrated by projecting a combined

fragmentation map of the Avidin complex (Figure 2-3 D, G), where a few small regions are

found in both modified and unmodified Avidin, but much of the observed coverage is unique to one experimental condition or the other.

2.4.2 Modification extends proteomic sequencing to large protein complexes

Many proteins of critical biological importance assemble into complexes with masses that extend to hundreds of kilodaltons and beyond^{56,57}. Such assemblies are exceptionally challenging targets for current top-down sequencing technologies, leaving important post-translational modifications located deep within these sequences inaccessible. In order to investigate the ability of TMP derivatization to bridge this technology gap, we modified two large protein complexes and carried out top-down sequencing experiments in a mode similar to those described above. Alcohol dehydrogenase (ADH) from yeast, a 148 kDa tetramer containing several phosphorylation sites, and Ovalbumin from chicken, a 170 kDa tetramer, were modified with TMP and subjected to CID for sequencing (Figure 2-4).

Sequencing data obtained for unmodified ADH results in relatively few fragment ions that originate primarily (>90%) from the n-terminal region of the protein. Such results are common for typical top-down sequencing efforts involving ADH, in which fragmentation rarely penetrates past residue 30 of the protein sequence^{17,21}, though concentrated efforts with multiple activation methods have achieved additional coverage²³. In contrast, TMP modification of ADH yielded a dramatic improvement in CID sequence coverage compared to unmodified ADH, with more than twice as many sequence-informative fragments detected for the TMP-modified protein (Figure 2-4 A, C). Furthermore, while TMP-modification led to enhanced coverage at the ADH n-terminus, on par with previous ETD¹⁷ and ECD²¹ datasets, it also unlocked significant fragmentation from the c-terminus, including 12 new c-terminal fragment sites that cover an additional 57 residues not typically accessed in unmodified ADH by CID, ECD, or ETD data.

Similar to Avidin, TMP modification resulted in orthogonal coverage information and the combined analysis of both modified and unmodified data resulted in the most total coverage, with nearly two and a half times the coverage of the unmodified ADH alone. Our TMP-modified sequencing data also covers a phosphorylation site at SER-316⁵⁸⁻⁶⁰, which was previously shown to be up-regulated in the presence of a mating pheromone⁵⁸. Phosphorylation at this site was not

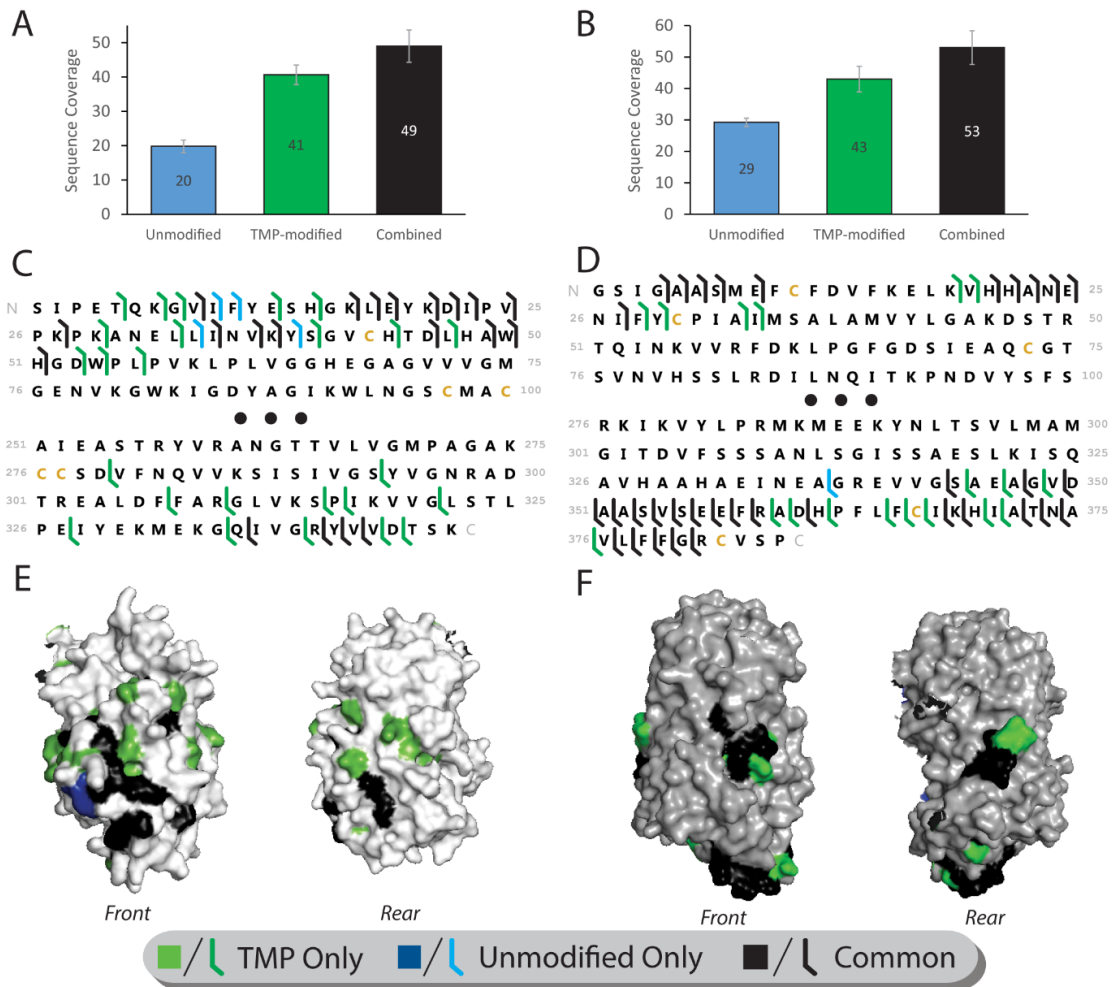


Figure 2-4 Enhanced sequencing of large protein complexes ADH and Ovalbumin. (A, D) Total sequence coverage (number of unique peptide bond cleavage sites) obtained from modified and unmodified ADH (a) and Ovalbumin (d) (N=3). (B, E) Sequence map of cleavage sites obtained from ADH (B) and Ovalbumin (E). Black dots indicate the middle 150 (b) or 200 (e) residues of the protein sequence, from which no coverage was obtained for any condition. (C, F) Cleavage location maps for ADH (C) and Ovalbumin (F). As in Figure 2-3, only a monomer is shown of the tetrameric structure to allow view of all sides. Coverage unique to unmodified protein is colored blue, TMP-modified protein is colored green, and sites common to both states are colored black

detected in our dataset, likely indicating the ADH standard used is not phosphorylated at this site, although we cannot rule out loss of the attached phosphate during CID.

Unlike ADH and Avidin, both n- and c-termini of unmodified Ovalbumin exhibited significant sequence coverage values in our experiments. The compact, low-charge monomer of Ovalbumin was analyzed, rather than the intact tetramer due to low amounts of tetramer in both unmodified and modified spectra. Despite the significant coverage already present in unmodified Ovalbumin, TMP modification still resulted in a substantial improvement in coverage (Figure 2-4 B, D), with an average over three replicates of nearly 50% more fragmentation sites than unmodified Ovalbumin. Like ADH and Avidin, combining the modified and unmodified data resulted in the best overall sequence coverage for Ovalbumin, with an 80% improvement over unmodified Ovalbumin alone, on average (n=3). Ovalbumin contains several PTM sites; however, coverage of these regions of the protein sequence was not significantly extended through TMP modification. As might be expected given the similarity of the coverage maps in Figure 2-4 D, a superposition of fragmentation data with Ovalbumin monomer structure shows an incremental, yet significant, improvement in the diversity of regions of the protein surface covered through CID fragmentation of the TMP modified protein (Figure 2-4 F).

Modification of ADH and Ovalbumin demonstrates the potential of fixed charge derivatization with TMP to expand the capabilities of intact sequencing for large proteins and protein complexes. The dramatic improvement in sequencing of the ADH tetramer, and particularly the generation of fragments from the previously intractable c-terminal region represent the potential of TMP modification for enabling sequencing of previously inaccessible regions of large protein complexes. TMP modification compares favorably to state of the art fragmentation methods such as ECD and ETD without the advanced instrumentation requirements of those techniques, and provides complementary coverage in many cases. Furthermore, modification by TMP enabled coverage of a PTM site near the c-terminus of ADH

that has not been previously accessed before by any top-down analysis to our knowledge. In the case of Ovalbumin, where essentially the same regions of the protein are fragmented in both the modified and unmodified cases, the improvement to total sequence coverage remains substantial.

2.4.3 High resolution MS analysis of large complexes

IM-MS analysis of ADH and Ovalbumin revealed significant portions of the protein sequences, despite being generally unable to isotopically resolve large (mass greater than approx. 8000 Da), highly charged ($z > 5$) fragment ions. To resolve these large fragments and further confirm the sequencing improvements offered by TMP, high resolution FT-ICR tandem MS analysis was performed on modified and unmodified ADH and Ovalbumin.

A comparison of CID fragmentation spectra acquired from the same ADH samples on both the FT-ICR and IM-MS platforms reveals clear trade-offs between the two instruments for top-down sequencing experiments. On the IM-MS platform, high energy CID and control of pressure and gas flow enabled extensive fragmentation of the large protein complex, but the

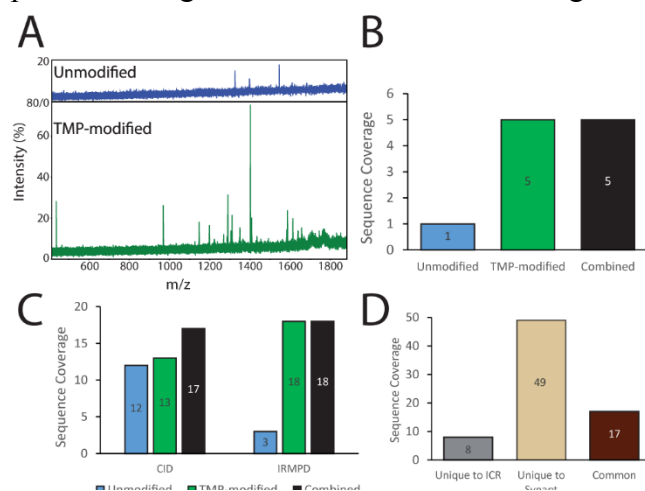


Figure 2-5 FT-ICR MS sequencing of large protein complexes with chemical modification. **(A)** Comparison of fragment mass spectra from IRMPD of intact ADH tetramer unmodified (top) and TMP-modified (bottom). **(B)** Unique cleavage sites obtained from IRMPD of ADH unmodified (blue) and TMP-modified (green). **(C)** FT-ICR sequencing of Ovalbumin using both CID (left) and IRMPD (right) for unmodified (blue), TMP-modified (green), and combined (black) analyses. **(D)** Comparison of sequence information obtained for CID of Ovalbumin on FT-ICR (gray), IM-MS (beige), and common to both instruments (brown).

mass resolution of the time of flight mass analyzer is insufficient to resolve isotopes for the largest fragments. The FT-ICR platform, in contrast, readily resolves any fragments generated, but did not bring about the degree of fragmentation observed on the IM-MS instrument used (Figure 2-5 D). Fragments identified for CID of modified and unmodified Ovalbumin precursors were nearly identical between the two platforms,

with the exception of several high mass fragments identified on the FT-ICR platform that could not be isotopically resolved on the IM-MS instrument.

Infrared multiphoton photodissociation (IRMPD) was performed in addition to CID for both ADH and Ovalbumin on the FT-ICR platform. As is the case with our CID datasets discussed above, TMP modified ADH and Ovalbumin exhibited dramatic increases in fragmentation when compared to unmodified proteins. In the case of ADH, only a single fragment could be observed at low intensity in the IRMPD spectrum of the unmodified complex, while the TMP-modified complex had nearly a dozen peaks corresponding to five sequence informative cleavage sites (Figure 2-5 A, B). Analysis of IRMPD data acquired for TMP-modified Ovalbumin revealed a similar trend, with activation of the protein producing 18 sites of coverage compared to just three for the unmodified Ovalbumin (Figure 2-5 C).

Like CID, IRMPD requires step-wise heating of the protein in order to deposit sufficient energy to elicit sequence-informative fragmentation⁶¹. TMP modification results in improved sequence coverage for both the CID and IRMPD experiments, presumably due to the fact that both methods typically cleave the weakest bonds present in the protein sequence, and that this landscape of bond energies is significantly altered in TMP-bound protein ions. Modification by TMP thus represents a strategy for enhancing sequence coverage of intact proteins and protein complexes across slow-heating fragmentation techniques available on a great many MS instrument platforms. Large protein complexes often require many activation techniques operating together to achieve sufficient sequence coverage for post-translational modification (PTM) analysis and identification purposes, and it is clear from the data reported here that TMP modification, or similar strategies based on the principles discussed here, have the potential to enable an increased role for CID/IRMPD tools within such workflows.

2.5 Conclusions

Top-down and native MS have emerged as valuable techniques for the analysis of intact proteins and protein complexes, but have been limited by incomplete sequence coverage, particularly for large proteins and complexes. Chemical modification using intrinsically-charged moieties such as TMP provides a simple and effective method to substantially enhance sequence coverage of intact proteins and protein complexes. Using this approach, we demonstrate enhancements in sequence coverage of three challenging model systems of 50-150% over analysis of the unmodified protein complexes by CID alone. We show that TMP modification provides these benefits across multiple instrument platforms that utilize multiple activation techniques, with enhancements to coverage in both CID and IRMPD datasets. TMP-enhanced CID compares favorably with standard top-down activation techniques like ECD, demonstrating both comparable coverage in ECD-accessible regions and providing coverage of previously intractable regions of the ADH tetramer.

Despite fixing many positive charges to our model proteins through TMP modification, the fragmentation we observe remains dominated by the *b*- and *y*-type ions characteristic of mobile proton-induced fragmentation, as is typical for CID of unmodified proteins. The absence of a shift to primarily *a*-type ions, contrasting to experiments with charge-derivatized peptides, is a novel and somewhat surprising result that demonstrates the persistence of mobile-proton behavior for high mass ions capable of intramolecular charge pairing. Clearly, however, the fragmentation pattern of intact proteins can be altered by fixed charges, likely through changing the relative ordering of bond strengths throughout the molecule as a result of the new locations of fixed charges imparted by TMP. We cannot rule out that some of the observed alterations to fragmentation are due to pathways associated with charge remote fragmentation events, which

was the ultimate aim of the TMP modification chemistry described here. Further data collection will be necessary to verify and quantify such channels within TMP modified proteins.

Modification with TMP is a simple, single-step procedure that can be easily incorporated into an existing experimental workflow. Improving sequencing of intact proteins and complexes without the need for extensive instrument modifications has the potential to expand the capabilities of many laboratories to analyze intact proteins and complexes within top-down workflows. A comprehensive protein complex analysis workflow, utilizing TMP modification in conjunction with one or several activation techniques, holds the potential to provide both state-of-the-art sequencing and PTM information as well as structural and stoichiometric details for protein assemblies, enabling next generation experiments in structural proteomics.

2.6 Acknowledgements

This work was supported by the National Institutes of Health (R01GM095832 to BTR and R01GM103479 and S10RR028893 to JAL) and the US Department of Energy (UCLA/DOE Institute for Genomics and Proteomics; DE-FC03-02ER63421). The authors thank M. W. Haskell for assistance with modeling of mass shift data. F.L. thanks the Research Foundation – Flanders (FWO) for funding a PhD fellowship.

2.7 References

- (1) Mayne, J.; Ning, Z.; Zhang, X.; Starr, A. E.; Chen, R.; Deeke, S.; Chiang, C. K.; Xu, B.; Wen, M.; Cheng, K.; et al. Bottom-Up Proteomics (2013-2015): Keeping up in the Era of Systems Biology. *Anal. Chem.* **2016**, *88* (1), 95–121.
- (2) Yates, J. R. The Revolution and Evolution of Shotgun Proteomics for Large-Scale Proteome Analysis. *J. Am. Chem. Soc.* **2013**, *135* (5), 1629–1640.
- (3) Beck, M.; Claassen, M.; Aebersold, R. Comprehensive Proteomics. *Curr. Opin. Biotechnol.* **2011**, *22* (1), 3–8.
- (4) Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. Top Down Proteomics: Facts and Perspectives. *Biochem. Biophys. Res. Commun.* **2014**, *445* (4), 683–693.
- (5) Savaryn, J. P.; Catherman, A. D.; Thomas, P. M.; Abecassis, M. M.; Kelleher, N. L. The Emergence of Top-down Proteomics in Clinical Research. *Genome Med.* **2013**, *5* (6), 53.

- (6) Siuti, N.; Kelleher, N. L. Decoding Protein Modifications Using Top-down Mass Spectrometry. *Nat. Methods* **2007**, *4* (10), 817–821.
- (7) Shaw, J. B.; Li, W.; 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) Han, X.; Jin, M.; Breuker, K.; McLafferty, F. W. Extending Top-down Mass Spectrometry to Proteins with Masses Great than 200 Kilodaltons. *Science (80-.)*. **2006**, *314* (5796), 109–112.
- (9) Alberts, B. The Cell as a Collection of Protein Machines: Preparing the next Generation of Molecular Biologists. *Cell* **1998**, *92* (3), 291–294.
- (10) Robinson, C. V.; Sali, A.; Baumeister, W. The Molecular Sociology of the Cell. *Nature* **2007**, *450* (7172), 973–982.
- (11) Leney, A. C.; Heck, A. J. R. Native Mass Spectrometry: What Is in the Name? *J. Am. Soc. Mass Spectrom.* **2017**, *28* (1), 5–13.
- (12) Pukala, T. L.; Ruotolo, B. T.; Zhou, M.; Politis, A.; Stefanescu, R.; Leary, J. A.; Robinson, C. V. Subunit Architecture of Multiprotein Assemblies Determined Using Restraints from Gas-Phase Measurements. *Structure* **2009**, *17* (9), 1235–1243.
- (13) Sharon, M.; Taverner, T.; Ambroggio, X. I.; Deshaies, R. J.; Robinson, C. V. Structural Organization of the 19S Proteasome Lid: Insights from MS of Intact Complexes. *PLoS Biol.* **2006**, *4* (8), 1314–1323.
- (14) Chait, B. T.; Cadene, M.; Olinares, P. D.; Rout, M. P.; Shi, Y. Revealing Higher Order Protein Structure Using Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (6), 952–965.
- (15) Yin, S.; Loo, J. A. Top-down Mass Spectrometry of Supercharged Native Protein-Ligand Complexes. *Int. J. Mass Spectrom.* **2011**, *300* (2–3), 118–122.
- (16) Syka, J. E. P.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and Protein Sequence Analysis by Electron Transfer Dissociation Mass Spectrometry. *Proc. Natl. Acad. Sci.* **2004**, *101* (26), 9528–9533.
- (17) Lermite, F.; Sobott, F. Electron Transfer Dissociation Provides Higher-Order Structural Information of Native and Partially Unfolded Protein Complexes. *Proteomics* **2015**, *15* (16), 2813–2822.
- (18) McLafferty, F. W.; Horn, D. M.; Breuker, K.; Ge, Y.; Lewis, M. A.; Cerda, B.; Zubarev, R. A.; Carpenter, B. K. Electron Capture Dissociation of Gaseous Multiply Charged Ions by Fourier-Transform Ion Cyclotron Resonance. *J. Am. Soc. Mass Spectrom.* **2001**, *12* (3), 245–249.
- (19) 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.
- (20) Xie, Y.; Zhang, J.; Yin, S.; Loo, J. A. Top-down ESI-ECD-FT-ICR Mass Spectrometry Localizes Noncovalent Protein-Ligand Binding Sites. *J. Am. Chem. Soc.* **2006**, *128* (45), 14432–14433.
- (21) Zhang, H.; Cui, W.; Wen, J.; Blankenship, R. E.; Gross, M. L. Native Electrospray and Electron-Capture Dissociation FTICR Mass Spectrometry for Top-down Studies of Protein Assemblies. *Anal. Chem.* **2011**, *83* (14), 5598–5606.
- (22) 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), 2809–2815.
- (23) Li, H.; Wongkongkathep, P.; Van Orden, S. L.; Ogorzalek Loo, R. R.; Loo, J. A. Revealing Ligand Binding Sites and Quantifying Subunit Variants of Noncovalent Protein Complexes in a Single Native Top-down Fticr Ms Experiment. *J. Am. Soc. Mass Spectrom.* **2014**, *25* (12), 2060–2068.
- (24) 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.
- (25) Bowers, W. D.; Delbert, S. S.; Hunter, R. L.; McIver, R. T. Fragmentation of Oligopeptide Ions Using Ultraviolet Laser Radiation and Fourier Transform Mass Spectrometry. *J. Am. Chem. Soc.* **1984**, *106* (23), 7288–7289.
- (26) Vath, J. E.; Biemann, K. Microderivatization of Peptides by Placing a Fixed Positive Charge at the N-Terminus to Modify High Energy Collision Fragmentation. *Int. J. Mass Spectrom. Ion Process.* **1990**, *100* (C), 287–299.
- (27) Adamczyk, M.; Gebler, J. C.; Wu, J. Charge Derivatization of Peptides to Simplify Their Sequencing with an Ion Trap Mass Spectrometer. *Rapid Commun. Mass Spectrom.* **1999**, *13* (14), 1413–1422.
- (28) Chen, W.; Lee, P. J.; Shion, H.; Ellor, N.; Gebler, J. C. Improving de Novo Sequencing of Peptides Using a Charged Tag and C-Terminal Digestion. *Anal. Chem.* **2007**, *79* (4), 1583–1590.
- (29) Roth, K. D. W.; Huang, Z. H.; Sadagopan, N.; Watson, J. T. Charge Derivatization of Peptides for Analysis by Mass Spectrometry. *Mass Spectrom. Rev.* **1998**, *17* (4), 255–274.
- (30) Frey, B. L.; Krusemark, C. J.; Ledvina, A. R.; Coon, J. J.; Belshaw, P. J.; Smith, L. M. Ion-Ion Reactions with Fixed-Charge Modified Proteins to Produce Ions in a Single, Very High Charge State. *Int. J. Mass Spectrom.* **2008**, *276* (2–3), 136–143.
- (31) Krusemark, C. J.; Frey, B. L.; Belshaw, P. J.; Smith, L. M. Modifying the Charge State Distribution of Proteins in Electrospray Ionization Mass Spectrometry by Chemical Derivatization. *J. Am. Soc. Mass Spectrom.* **2009**, *20* (9), 1617–1625.
- (32) He, Y.; Reilly, J. P. Does a Charge Tag Really Provide a Fixed Charge? *Angew. Chemie - Int. Ed.* **2008**, *47* (13), 2463–2465.
- (33) King, L. C.; Ozog, F. J. Reactions of Pyrylium and Pyridinium Salts with Amines. *J. Org. Chem.* **1955**, *20* (4), 448–454.
- (34) Li, X.; Cournoyer, J. J.; Lin, C.; O'Connor, P. B. The Effect of Fixed Charge Modifications on Electron Capture Dissociation. *J. Am. Soc. Mass Spectrom.* **2008**, *19* (10), 1514–1526.
- (35) Haynes, S. E.; Polasky, D. A.; Dixit, S. M.; Majmudar, J. D.; Neeson, K.; Ruotolo, B. T.; Martin, B. R. Variable-Velocity Traveling-Wave Ion Mobility Separation Enhancing Peak Capacity for Data-Independent Acquisition Proteomics. *Anal. Chem.* **2017**, *89* (11), 5669–5672.
- (36) Strohal, M.; Kavan, D.; Novák, P.; Volný, M.; Havlíček, V. MMass 3: A Cross-Platform Software Environment for Precise Analysis of Mass Spectrometric Data. *Anal. Chem.* **2010**, *82* (11), 4648–4651.
- (37) Strohal, M.; Hassman, M.; Košata, B.; Kudiček, M. MMass Data Miner: An Open Source Alternative for Mass Spectrometric Data Analysis. *Rapid Commun. Mass Spectrom.* **2008**, *22* (6), 905–908.

- (38) Niedermeyer, T. H. J.; Strohal, M. MMass as a Software Tool for the Annotation of Cyclic Peptide Tandem Mass Spectra. *PLoS One* **2012**, *7* (9), e44913.
- (39) Baer, T.; Mayer, P. M. Statistical Rice-Ramsperger-Kassel-Marcus Quasiequilibrium Theory Calculations in Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **1997**, *8* (2), 103–115.
- (40) Cox, K. A.; Gaskell, S. J.; Morris, M.; Whiting, A. Role of the Site of Protonation in the Low-Energy Decompositions of Gas-Phase Peptide Ions. *J. Am. Soc. Mass Spectrom.* **1996**, *7* (6), 522–531.
- (41) Palzs, B.; Suhal, S. Fragmentation Pathways of Protonated Peptides. *Mass Spectrom. Rev.* **2005**, *24* (4), 508–548.
- (42) Dongré, A. R.; Jones, J. L.; Somogyi, Á.; Wysocki, V. H. Influence of Peptide Composition, Gas-Phase Basicity, and Chemical Modification on Fragmentation Efficiency: Evidence for the Mobile Proton Model. *J. Am. Chem. Soc.* **1996**, *118* (35), 8365–8374.
- (43) Samulak, B. M.; Niu, S.; Andrews, P. C.; Ruotolo, B. T. Ion Mobility-Mass Spectrometry Analysis of Cross-Linked Intact Multiprotein Complexes: Enhanced Gas-Phase Stabilities and Altered Dissociation Pathways. *Anal. Chem.* **2016**, *88* (10), 5290–5298.
- (44) Zhou, X.; Lu, Y.; Wang, W.; Borhan, B.; Reid, G. E. “Fixed Charge” Chemical Derivatization and Data Dependant Multistage Tandem Mass Spectrometry for Mapping Protein Surface Residue Accessibility. *J. Am. Soc. Mass Spectrom.* **2010**, *21* (8), 1339–1351.
- (45) Reid, G. E.; Roberts, K. D.; Simpson, R. J.; O’Hair, R. A. J. Selective Identification and Quantitative Analysis of Methionine Containing Peptides by Charge Derivatization and Tandem Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2005**, *16* (7), 1131–1150.
- (46) Clifford-Nunn, B.; Showalter, H. D. H.; Andrews, P. C. Quaternary Diamines as Mass Spectrometry Cleavable Crosslinkers for Protein Interactions. *J. Am. Soc. Mass Spectrom.* **2012**, *23* (2), 201–212.
- (47) Ko, B. J.; Brodbelt, J. S. Enhanced Electron Transfer Dissociation of Peptides Modified at C-Terminus with Fixed Charges. *J. Am. Soc. Mass Spectrom.* **2012**, *23* (11), 1991–2000.
- (48) Krusemark, C. J.; Ferguson, J. T.; Wenger, C. D.; Kelleher, N. L.; Belshaw, P. J. Global Amine and Acid Functional Group Modification of Proteins. *Anal. Chem.* **2008**, *80* (3), 713–720.
- (49) Light-Wahl, K. J.; Schwartz, B. L.; Smith, R. D. Observation of the Noncovalent Quaternary Associations of Proteins by Electrospray Ionization Mass Spectrometry. *J. Am. Chem. Soc.* **1994**, *116* (12), 5271–5278.
- (50) Schwartz, B. L.; Light-Wahl, K. J.; Smith, R. D. Observation of Noncovalent Complexes to the Avidin Tetramer by Electrospray Ionization Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (3), 201–204.
- (51) Zinnel, N. F.; Pai, P. J.; Russell, D. H. Ion Mobility-Mass Spectrometry (IM-MS) for Top-down Proteomics: Increased Dynamic Range Affords Increased Sequence Coverage. *Anal. Chem.* **2012**, *84* (7), 3390–3397.
- (52) 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.
- (53) Lee, B. S.; Krishnanchettiar, S.; Lateef, S. S.; Lateef, N. S.; Gupta, S. Characterization of Oligosaccharide Moieties of Intact Glycoproteins by Microwave-Assisted Partial Acid

- Hydrolysis and Mass Spectrometry. *Rapid Commun. Mass Spectrom.* **2005**, *19* (18), 2629–2635.
- (54) DeLange, R. J.; Huang, T.-S. Egg White Avidin. *J. Biol. Chem.* **1971**, *246* (3), 698–709.
- (55) Pugliese, L.; Coda, A.; Malcovati, M.; Bolognesi, M. Three-Dimensional Structure of the Tetragonal Crystal Form of Egg-White Avidin in Its Functional Complex with Biotin at 2.7 Å Resolution. *J. Mol. Biol.* **1993**, *231* (3), 698–710.
- (56) Havugimana, P. C.; Hart, G. T.; Nepusz, T.; Yang, H.; Turinsky, A. L.; Li, Z.; Wang, P. I.; Boutz, D. R.; Fong, V.; Phanse, S.; et al. A Census of Human Soluble Protein Complexes. *Cell* **2012**, *150* (5), 1068–1081.
- (57) Kirkwood, K. J.; Ahmad, Y.; Larance, M.; Lamond, A. I. Characterization of Native Protein Complexes and Protein Isoform Variation Using Size-Fractionation-Based Quantitative Proteomics. *Mol. Cell. Proteomics* **2013**, *12* (12), 3851–3873.
- (58) Gruhler, A.; Olsen, J. V.; Mohammed, S.; Mortensen, P.; Færgeman, N. J.; Mann, M.; Jensen, O. N. Quantitative Phosphoproteomics Applied to the Yeast Pheromone Signaling Pathway. *Mol. Cell. Proteomics* **2005**, *4* (3), 310–327.
- (59) Li, X.; Gerber, S. A.; Rudner, A. D.; Beausoleil, S. A.; Haas, W.; Villén, J.; Elias, J. E.; Gygi, S. P. Large-Scale Phosphorylation Analysis of ??-Factor-Arrested *Saccharomyces Cerevisiae*. *J. Proteome Res.* **2007**, *6* (3), 1190–1197.
- (60) Holt, L. J.; Tuch, B. B.; Villen, J.; Johnson, A. D.; Gygi, S. P.; Morgan, D. O. Global Analysis of Cdk1 Substrate Phosphorylation Sites Provides Insights into Evolution. *Science* (80-.). **2009**, *325* (5948), 1682–1686.
- (61) McLuckey, S. A.; Goeringer, D. E. Slow Heating Methods in Tandem Mass Spectrometry. *J. Mass Spectrom.* **1997**, *32* (5), 461–474.

Chapter 3 Chemical Derivatization Illuminates the Links between Controlled Sequence Informative Fragmentation Chemistry and Gas-phase Protein Ion Structure

Daniel A. Polasky, Sugyan M. Dixit, Michael Keating, Varun V. Gadkari, Philip C. Andrews, and Brandon T. Ruotolo.

3.1 Introduction

Sequence analysis of intact proteins by tandem mass spectrometry (MS), also known as top-down proteomics, offers great promise in characterizing the proteoforms¹ involved in human disease.²⁻⁴ This growing field remains limited by the myriad challenges associated with achieving complete fragmentation of intact proteins, which is required to accurately localize all potential post-translational modification (PTM) sites within a given biopolymer. Such a high bar for sequence coverage remains difficult to achieve, rarely accomplished for proteins weighing more than 30 kDa,⁵ and especially challenging to produce when precursor ions are multi-protein assemblies, which represent a rapidly expanding group of top-down sequencing targets that promise to revolutionize our understanding of functional protein states.⁵ The tandem MS tools used to develop such sequence information rely upon ion activation technologies designed to increase the internal energy of isolated protein ions to the point of covalent bond dissociation, and despite significant developments in electron and photon-mediated activation techniques, activation through collisions with inert, neutral gas remains one of the most utilized methods for top-down proteomics due to its ease of implementation and high fragmentation efficiency. As ever larger proteins are analyzed, the prediction of protein fragmentation patterns, as is common in bottom-up proteomics methods,⁶⁻⁸ is critical to generating high-confidence results. Previous

work has indicated that the overall mobility of charges along the protein backbone plays a central role in determining its fragmentation, but gaps in fundamental knowledge of gas-phase protein structure, salt bridging, and charge solvation mean accurately predicting the charge mobility and fragmentation pathways for intact proteins remains a key challenge for the field.

Extensive study has identified the major fragmentation pathways of small peptides⁹⁻¹⁶ and enabled the use of simple fragmentation models in associated sequencing analyses, most commonly applied to tryptic peptides generated during bottom-up proteomics experiments. Peptide fragmentation is generally divided into two major mechanistic classes: “charge-directed,” in which a mobile proton¹² mediates fragmentation and “charge-remote,” in which charging protons are not involved in the fragmentation observed.⁹ In charge-directed, or mobile proton, fragmentation, charging protons can occupy many sites along the peptide backbone and fragmentation occurs at the site occupied by a proton following activation.¹² In most cases, this results in stochastic fragmentation and high sequence coverage for peptide ions. Fragmentation on the C-terminal side of proline residues can also be strongly enhanced under these conditions due to the high proton affinity of the backbone amide of proline.¹⁶ When charging protons are not mobile, *e.g.* when they are sequestered by a side chain with significant gas-phase basicity, charge-remote pathways can become lower energy than mobile proton pathways.^{17,18} The most common charge remote pathway in protonated peptides results in the “Asp-Xxx” cleavage involving the carboxylic acid group of an aspartic (or glutamic) acid residue.¹⁹ The exact mechanism for this process is debated, involving either salt bridged^{19,20} or charge solvated^{10,21,22} structures at the acidic site resulting in strongly preferential cleavage at these sites when no mobile protons are present.²³

In intact protein ions, the competition between mobile proton and charge remote fragmentation pathways as a function of proton mobility bears a general similarity to the case described above for peptides. Intact proteins can contain many basic residues and carry large numbers of charging protons, allowing for a much wider range of proton mobility conditions than in peptides. The overall propensities for fragmentation at various amino acids for a large number of proteins in both denaturing (high charge state) and native (lower charge states) top-down MS experiments indicate that precursor ions of the later class result in a greater amount of fragment ions produced through charge remote channels.²⁴ Recent studies have demonstrated that various mechanisms unique to intact protein ions can be observed at different proton mobility conditions.^{25,26} As a specific example, fragmentation of disulfide bonds, which typically preclude sequencing the regions they enclose, has been shown to be competitive with charge remote pathways under low charge mobility conditions.²⁷ The relative contributions of these processes, are highly challenging to predict *a priori* for a protein sequence, and represent a significant challenge for both experimental design and informatics in top-down proteomics.

Chemical derivatization offers a method to affect charge mobility and has a long history of use in peptides both for elucidation of fragmentation pathways and as a means to enhance fragmentation.²⁸ Early work described fixing charge to peptide termini as means to simplify the resulting mass spectra by driving fragmentation into charge remote channels,^{18,29} and provided some of the initial evidence for the role of charge mobility in what would become the mobile proton model of peptide fragmentation.¹⁷ In intact proteins, several groups have attempted to enhance fragmentation via the use of chemical derivatization. Few efforts have involved the use of fixed charges as in peptides, due to the instability of most fixed charge derivatives during precursor ion activation.³⁰ Instead, indirect modulation of charge mobility through altering the

charge states of electrosprayed protein ions, using both charged³¹ and uncharged derivatives,^{32–34} demonstrated that it can be highly challenging to alter protein charge state distributions (CSDs) via derivatization. Only by attaching fixed charges to formerly acidic or neutral side chains could charge state be increased, and it was found that appending fixed charges reduces the number of charging protons acquired during electrospray.³¹

We have previously demonstrated that protein derivatization with trimethyl pyrylium (TMP) produces sufficiently stable fixed charges to remain fixed through collision-induced dissociation (CID), and that TMP-derivatized protein fragmentation can be enhanced in some cases.³⁵ Here, we use the charge fixing capability of TMP, in conjunction with derivatization of carboxylic acids, to provide new insight into the role of charge mobility in fragmentation of native-like protein and protein complex ions. The addition of fixed charges reduces charge mobility, driving fragmentation to charge remote pathways, including Asp-Xxx cleavages and, with sufficient fixed charges, directs fragmentation to the site of the fixed charge. Alternatively, converting carboxyl groups to amides blocks charge remote fragmentation via the Asp-Xxx pathway and restores mobile proton-mediated fragmentation even under conditions of extremely low charge mobility. Together, these derivatization strategies enable unprecedented control over the fragmentation chemistry of intact proteins, allowing access to charge remote or charge directed pathways across a wide range of charge states and primary sequences. Our experiments with charge-fixed proteins and complexes also reveal that protein ions have a significant capacity to solvate excess positive charge and that canonical salt bridging interactions are not required to do so. Molecular dynamics simulations of a model protein including fixed-charge labels reveals a strong propensity to form helices that engender strong macrodipoles. TMP-modified lysine residues are shown to interact with the helix dipoles, solvating the intrinsic positive charge of the

modification. These studies reveal the extent to which intact protein ions are capable of solvating charge in the gas-phase, including through substantial rearrangements of secondary structure, providing the most complete indication to date of the extent of physical forces opposing sequencing technologies for intact proteins.

3.2 Experimental Methods

3.2.1 Chemical modification of primary amines

Proteins were purchased purified (Sigma Aldrich, St. Louis, MO) and diluted to 25 μ M in 100 mM triethylammonium bicarbonate (TEAB, Sigma-Aldrich), pH 8.5. 2,4,6-Trimethyl pyrylium (TMP) tetrafluoroborate (Alfa Aesar, Haverhill, MA) was dissolved in 100 mM TEAB, pH 8.5, vortexed for ten seconds to dissolve, and quickly added to protein solutions at 100-fold excess (proteins less than 50kDa) or 1000-fold excess (proteins weighing more than 50kDa) relative to protein concentration. For Small EDRK Rich Factor (SERF), 100 μ M protein was mixed with 25, 100, 250, or 1000-fold molar excess of TMP to generate varying degrees of modification. All reaction solutions were vortexed for ten seconds and allowed to react for 24 hours at room temperature. Reactions were quenched by addition of ammonium acetate (Sigma-Aldrich), pH 7.8, at slight molar excess (1-3) relative to TMP concentration. Quenched reactions were subsequently buffer exchanged into 100 mM ammonium acetate, pH 7.8, with P6 microspin columns (BioRad Laboratories, Hercules, CA) according to manufacturer instructions. Buffer exchanged sample were either analyzed immediately or flash frozen with liquid nitrogen and stored at -80 °C prior to analysis.

3.2.2 Chemical modification of carboxyl groups

Small EDRK Rich Factor (SERF) was dissolved in 50 mM 2-(*N*-morpholino)ethanesulfonic acid (MES, Sigma-Aldrich), pH 4.5, to make solutions containing 100 μ M protein for chemical

modification. Glycinamide-HCl (Sigma-Aldrich) was dissolved in water, vortexed for ten seconds to dissolve, and added to protein solutions to achieve 10000-fold molar excess relative to protein concentration. The reaction mixture was vortexed for one minute. 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide-HCl (EDC, Sigma-Aldrich) was dissolved in water, vortexed for ten seconds to dissolve, and added to reaction mixtures to achieve 500 molar excess relative to protein concentration. Reaction solutions were allowed to react for 2 hours at room temperature. Reactions were quenched by addition of ammonium acetate, pH 7.8, at slight molar excess (1-3) relative to EDC concentration. Quenched reactions were subsequently buffer exchanged into 100 mM ammonium acetate, pH 7.8, with P6 microspin columns, as in primary amine modification procedure.

3.2.3 Successive modification of primary amines and carboxyl groups

SERF modified with 1000-fold excess of TMP according to procedure above was buffer exchanged into 50mM MES, pH 4.5. Carboxyl group modification was the performed according to the procedure above. Following modification, reaction solutions were desalted and buffer exchanged into 100mM ammonium acetate by size exclusion chromatography using an S-75 Increase 3.2/300, 2.4 mL column (GE Healthcare, Chicago, IL) on an Akta Purifier (GE Healthcare, Chicago, IL) SEC system. 100 μ L of protein solution was injected and buffer exchanged into 100 mM ammonium acetate at a flow rate of 0.05 mL/min for 2 column volumes. 0.1 mL fractions were collected. Fractions containing modified protein were combined and concentrated using Vivaspin 500 3 kDa MWCO filters (Sartorius Stedium Lab Ltd, Stonehouse, UK) prior to analysis.

3.2.4 Ion mobility-mass spectrometry (IM-MS)

A quadrupole ion mobility time-of-flight mass spectrometer (Synapt G2 HDMS, Waters, Milford, MA) was used for all top-down mass spectrometry experiments. 5 μ L of buffer-exchanged protein solution was transferred to a gold-coated borosilicate capillary (0.78 mm i.d., Harvard Apparatus, Holliston, MA) for direct infusion. Instrumental settings were optimized to preserve intact protein complexes prior to activation: capillary voltage, 1.5 kV; sample cone, 40 V; extraction cone, 0 V. Gas flows were as follows (mL/min): source, 50; trap, 2 (proteins below 50 kDa) or 8 (proteins above 50 kDa); helium cell, 200; IM separation, 90. IMS traveling wave settings were the same for all proteins: wave velocity, 150 m/s; wave height, 20 V; IMS bias, 5 V. Backing pressure was set to 2.7 mbar (proteins below 50 kDa) or 8.0 mbar (proteins above 50 kDa). A single charge state of each protein complex was selected and collisionally activated in the trap cell at voltages optimized to dissociate 50-75% of the selected precursor prior to ion mobility separation. Time-of-flight pressure was 1.8×10^{-6} mbar for all analyses. Scans were combined for 10 minutes to improve fragment ion signal-to-noise ratios.

3.2.5 High resolution mass spectrometry

SERF protein with both primary amines and carboxylic acids modified was analyzed on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific, Waltham, MA). 5 μ L of buffer-exchanged protein solution was transferred to a gold-coated borosilicate capillary (0.78 mm i.d., Harvard Apparatus, Holliston, MA) for direct infusion using a Nanospray Flex ion source (Thermo Scientific, Waltham, MA). Capillary voltage was 1.6 kV, transfer capillary temperature was 275C. Intact mass analysis was conducted in the Orbitrap analyzer with a resolution of 240,000 at m/z 400 with an AGC target of $1e6$ and 5 microscans and data was accumulated for several minutes to ensure sufficient signal-to-noise ratio for deconvolution. Data

was analyzed using Thermo XCalibur Qual Browser and BioPharmaFinder v3.0 software. Intact mass deconvolution using BioPharmaFinder was performed in “average over selected retention time” mode with a mass range of 3000-25000. All other parameters were left at default values. Output from intact mass deconvolution was manually matched to mass of SERF and varying numbers of modifications.

3.2.6 Data analysis

Top down IM-MS data was processed to a list of monoisotopic peaks using IMTBX³⁶ via the IMTBX+Grppr GUI interface with the following parameters: filter size: 4 4 (m/z) 2 2 (drift time) 0 0 (retention time), hyper values: 1 1 1, intensity cutoff: 20, minimum signal-to-noise: 2 (peak) and 3 (most intense peak in isotopic cluster), minimum peak points: 9, noise filtering enabled with window 4 8, minimum peaks in isotopic cluster: 3, average isotopic model with cluster matching tolerance of 100 ppm. Peak lists were annotated to protein sequence using a home-built pipeline written in Python. Peak lists were searched using a multipass search strategy, in which the first pass searched for only *b*- and *y*-ions with no neutral losses and the second pass allowed *a*-ions and neutral losses. Chemical modifications were treated as variable additions to all residues capable of modification (lysine and the N-terminus for primary amine modification, aspartic and glutamic acids and the C-terminus for carboxylic acid modification). Data was calibrated during search using the median error of all hits within 100 ppm of the theoretical mass. Final match tolerance was 10 ppm. Fragment intensity plots by sequence position and amino acid were generated from matched data normalized by relative fragment ion abundance. Charge state distributions and ion mobility profiles were extracted from raw data using TWIMExtract.³⁷ Data was aggregated and comparisons performed using custom scripts written in Python.

Collision cross section (CCS) values were calibrated using Gaussian fitted peak centroids in ion mobility arrival time distributions using the Gaussian fitting module of CIUSuite 2³⁸ in signal-only mode with peak width of 1 ± 0.6 ms and a maximum of 6 peaks. CCS for modified SERF was calibrated using D, L polyalanine ions (30 – 32 alanines, 3⁺), ubiquitin 5⁺, β -lactoglobulin 8⁺ and 9⁺, and β -lactoglobulin dimer 12⁺ and 13⁺ ions at a wave height of 20 V, wave velocity of 200 m/s, and pressure of 3.4 mbar. CCS calibration for unmodified SERF was done using ubiquitin 5⁺, β -lactoglobulin 7⁺ – 9⁺, cytochrome c 6⁺ and 7⁺, insulin monomer 3⁺ and 4⁺, and denatured ubiquitin 9⁺ – 13⁺ ions at wave height of 30 V, wave velocity of 600 m/s, and pressure of 3.4 mbar. The CCS values were used to construct the calibration function. Six replicate measurements were taken in order to obtain uncertainty values in our prediction using the following equation: $\mathbf{error} = \sqrt{\sigma^2 + \mathbf{cal_rmse}^2 + \mathbf{database_error}^2}$ where σ is standard deviation of replicate measurement values, $\mathbf{cal_rmse}$ is the calibration CCS root mean square error (RMSE), and $\mathbf{database_error}$ is the uncertainty in the database values³⁹.

3.2.7 Molecular dynamics simulations

Molecular dynamics (MD) simulations were performed with CHARMM on a workstation with an Intel Xeon processor with eight CPU cores at 2.50 GHz. The CHARMM36 force field was employed with CMAP correction for improved treatment of peptide backbones to achieve accurate peptide conformations.⁴⁰ Residue topology file and parameter file for TMP-derivatized lysine was generated by combining the lysine topology file from the protein force field and TMP topology file from general force field parameters using ParamChem^{41–43} (Text II-1). Topology files for EDC capped aspartic acid and glutamic acid residues were created by combining the topology information from unmodified residues, amidated c-terminus, n-methylamide c-terminus, and glycine residue (Text II-1). Fully extended SERF structure was created both

without modification and with TMP and EDC modification. The models were energy minimized using steepest descent for 100 steps. The models were then equilibrated at 300 K for 200 ps, then subjected to a replica exchange (REX) MD simulation.⁴⁴⁻⁴⁷ Briefly, multiple copies (replicas) of the system are simulated at different temperatures independently and simultaneously. Replicas are exchanged during the simulation periodically according to a Metropolis-type algorithm. In this work, we used 20 replicas covering a temperature range of 300 K to 800 K with temperatures distributed exponentially in the specified range. SHAKE was applied to fix the lengths of all bonds with hydrogen atoms. Langevin dynamics with a friction coefficient of 5 ps⁻¹ and a time-step of 2 fs were used. REX simulations lasted 20 ns with exchanges attempted every 2 ps. The pairwise exchange ratio was greater than 30 % for each run. The coordinates were saved every 1 ps. The energy in simulations converged starting at 10 ns. Conformations sampled at the lowest temperature (300 K) during the last 10 ns were clustered to provide the final models. Details on further analysis are provided in Text II-1.

3.3 Results and Discussion

Studies of peptide and protein fragmentation accomplished through CID have determined that charge mobility plays a key role in partitioning fragmentation between charged directed and charge remote pathways.⁹⁻¹⁵ When charge mobility is low (e.g. for low charge states of intact protein ions), fragmentation is typically dominated by charge remote pathways, often by Asp-Xxx cleavages. To modulate the degree of charge mobility in intact proteins and complexes under native conditions, we utilized trimethyl pyrylium (TMP), a reagent that affixes an intrinsically charged pyridinium moiety to primary amines (Figure 3-1 A).^{29,48} TMP has been

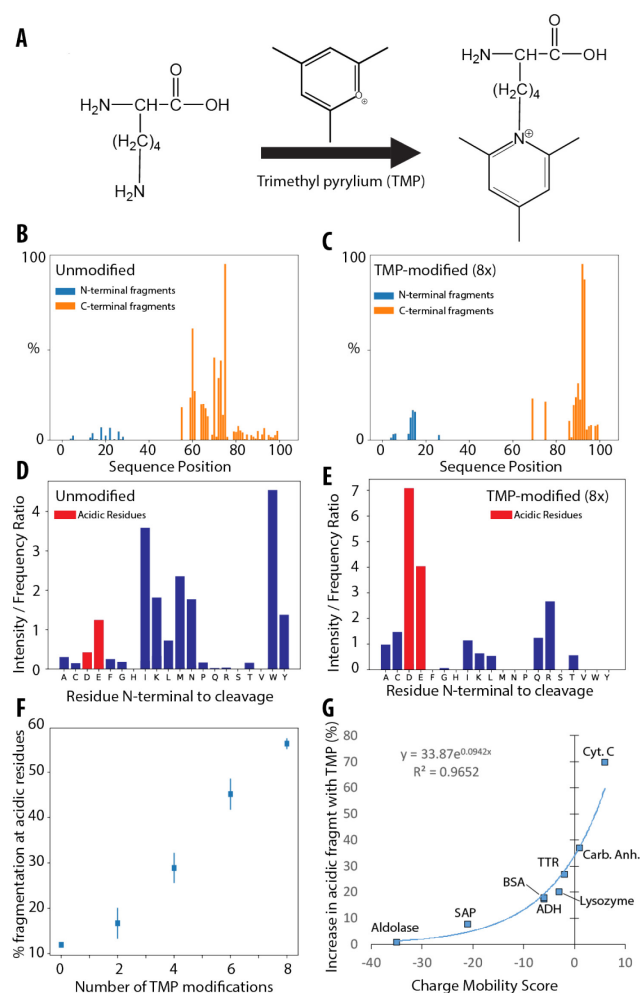


Figure 3-1 Charge fixing chemical modification drives charge remote fragmentation in intact proteins. A) Trimethyl pyrylium (TMP) derivatization of lysine residues yields an intrinsically charged trimethyl pyridinium group. B) Sequence coverage of unmodified equine cytochrome C, shown as relative intensity of all fragment ions resulting from a given sequence position from N-terminus (position 0) to C-terminus. C) Sequence coverage plot of cytochrome C with 8 TMP modifications showing both similarity and difference with unmodified protein. D) Fragmentation propensity at each amino acid, shown as a ratio of the intensity of fragmentation at each amino acid (N-terminal to the cleavage site) to the frequency of that amino acid occurring in the protein sequence for unmodified protein (D) and 8 times TMP modified protein (E). Acidic residues are highlighted in red. F) Percentage of all fragmentation occurring N-terminal to acidic residues as a function of the number of TMP modifications present. G) “TMP effect,” or the increase in fragmentation at acidic sites following TMP modification, plotted against an approximation of charge mobility, determined by subtracting the number of arginines in the protein from the observed charge (number of protons). Proteins fall along a trend with increasing effect from TMP modification occurring for proteins beginning with higher charge mobility.

shown previously to alter protein fragmentation and enable improved sequencing for some protein complex precursor ions, while remaining stable throughout the CID process.³⁵ Each fixed charge provided by a TMP modification can replace the charge provided by a potentially mobile proton, thus reducing the overall mobility of charge on the protein.

To investigate these effects in detail, we modified several proteins with varying amounts of TMP to generate a range of modification states, then selected and fragmented each state individually to assess the impact of TMP on fragmentation.

Fragmentation data from equine cytochrome C, a 12 kDa heme-containing protein, is shown in Figure 3-1 B for unmodified and Figure 3-1 C for TMP-modified protein. The location of the fragmentation observed, displayed as a relative intensity at each possible fragmentation site (amino acid position) from the N-terminus, shows only

minor differences in the TMP-modified sample compared to unmodified protein. However, when sorted by the amino acid at which fragmentation occurred (Figure 3-1 D, E) substantial differences become clear. Fragmentation is shown as the relative abundance of all fragmentation events occurring N-terminal to each amino acid type, normalized against the frequency of that amino acid in the sequence. For example, if all residues fragmented with equal likelihood, the fragmentation propensity plots in Figure 3-1 D and 1E would have a uniform magnitude of 1 for all amino acids. In the TMP-modified protein, cleavage at acidic residues (Asp and Glu) has become the dominant fragmentation pathway, representing nearly 60% of all fragment ion intensity with an enhancement factor of ~7-fold, as opposed to the 11% of fragment ion intensity in the unmodified protein observed with no significant enhancement over other amino acids. The relative intensity of fragment ions from acidic residues is very strongly correlated with the degree of TMP modification (Figure 3-1 F). Thus, as TMP is added to the protein, charge mobility is reduced, increasing the energy of mobile proton fragmentation pathways and resulting in charge remote fragmentation at acidic sites.

We extended this analysis to native proteins and complexes covering a wide range of molecular weights and expected charge mobilities. Proton mobility has been shown to be strongly influenced by arginine residues, which have a relatively high proton affinity in the gas phase and can effectively sequester a proton, preventing it from moving along the peptide backbone.^{9,12-14} Other basic residues, such as lysine and histidine, can also reduce proton mobility, but generally to a lesser extent.^{12,13} Proteins and protein complexes with 2-60 arginines were analyzed with and without TMP modification and the difference in the amount of fragmentation in charge remote (acidic residue) channels was plotted against a charge mobility score, which we defined as the observed charge of the protein minus the number of arginine

residues, resulting in a strong correlation (Figure 3-1 G, Figure II-1). Proteins and complexes with more protons than proton-sequestering arginine residues, and thus a high degree of charge mobility, e.g. cytochrome C and carbonic anhydrase, experienced dramatic changes in fragmentation following TMP modification. In contrast, aldolase tetramers and serum amyloid P-component (SAP) pentamers, which contain more arginines than protons and thus exhibit very low charge mobility, experienced almost no effect from TMP modification as charge-remote fragmentation pathways were already likely lower energy than mobile proton pathways prior to TMP modification in these systems. Fixed-charge modification by TMP thus provides a method to modulate protein fragmentation as a function of charge mobility, enabling high charge mobility proteins, e.g. highly charged proteins electrosprayed from denaturing solvents, to be fragmented under low charge mobility conditions, accessing charge remote pathways.

In all proteins probed in Figure 3-1, charge remote fragmentation occurring at acidic residues was the primary pathway in competition with mobile proton fragmentation. To probe the limits of the acidic charge remote pathway, we investigated the Small EDRK-Rich Factor (SERF) protein from *s. cerevisiae*. SERF is an intrinsically disordered protein⁴⁹ and contains many acidic and basic residues (Figure 3-2 A). Charge states from 5⁺ to 15⁺ were observed following electrospray ionization, resulting in a range of charge mobility conditions. Because of its small size (68 amino acids) and large proportion of TMP-modifiable residues (14 lysines), SERF presented a unique opportunity to study the extreme limits of low charge mobility following extensive TMP modification under native conditions. As expected for a highly basic protein, the 8⁺ charge state of unmodified SERF fragmented almost exclusively (>80% of total fragment intensity, approx. 10-fold enhancement relative to aspartic acid frequency) via charge remote fragmentation at acidic residues. Unlike the proteins surveyed in Figure 3-1, however, the

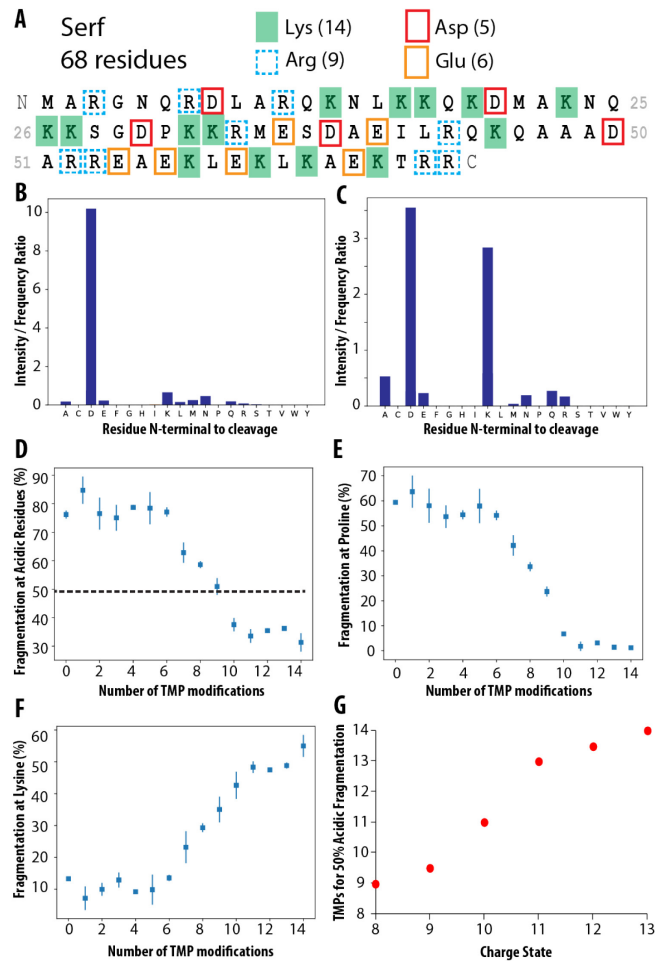


Figure 3-2 TMP modification of SERF protein reveals fragmentation at fixed charge sites. A) Sequence of yeast SERF protein with Lys, Arg, Asp, and Glu residues highlighted. B) Fragmentation propensity ratio (intensity of fragmentation at a given amino acid over frequency of that amino acid in the protein sequence) maps for 8+ SERF without (B) and with (C) TMP modification. D) Percentage of all fragment intensity occurring N-terminal to acidic residues for 8+ Serf as a function of the number of TMP modifications present. 50% acidic fragmentation line is highlighted for reference. E) Percentage of fragmentation occurring C-terminal to proline for 8+ SERF. F) Percentage of fragmentation occurring N-terminal to lysine for 8+ SERF. G) Number of TMP modifications required to pass below 50% acidic fragmentation as a function of SERF charge state.

addition of up to 14 TMPs to SERF resulted in decreased fragmentation at acidic residues (Figure 3-2 D). This corresponded with a reduction in fragmentation on the C-terminal side of proline residues, as would be expected when reducing the availability of mobile protons, indicating that the reduction in acidic fragmentation was not due to an increase in mobile proton-mediated fragmentation (Figure 2E). Further investigation revealed that fragmentation was shifting specifically to the TMP-modified lysine residues (Figure 3-2 C) containing the fixed charges. Indeed, the intensity of fragmentation at lysine residues within modified SERF tracks in an inverse manner with the fragmentation at acidic and proline residues (Figure 3-2 F), and almost no fragmentation (less than 20% of all fragment intensity in all cases) is found N-terminal to

residues other than Asp, Glu, or Lys. To our knowledge, cleavage of the peptide bond to form *b*- and *y*-ions at the site of a side chain containing a fixed charge has not previously been observed on the scale of an intact, multiply charged protein. This result indicates that directing CID to the

site(s) of introduced charges is indeed possible in intact proteins provided charge mobility can be reduced sufficiently, *e.g.* by using fixed charge reagents targeting multiple side chain chemistries in proteins with fewer basic residues than SERF.

Somewhat surprisingly, the transition from charge remote fragmentation at acidic residues to fragmentation at fixed charge sites was dependent on the SERF charge state probed. For higher charge states, more TMP modifications were required to reduce the resulting fragmentation at acidic residues to < 50%, shifting the remaining fragment ion current to charged-fixed lysines (Figure 3-2 G). Asp-Xxx fragmentation at acidic residues is thought to proceed in the absence of mobile protons, so the relatively increased amount of this pathway we observed in high-charge state TMP-modified SERF, where charge mobility would be increased, must be explained by other processes. We speculate that the charge solvated structures involved in Asp-Xxx cleavage may be more energetically favorable under higher charge conditions, lowering the barrier for this process relative to the direct fragmentation at fixed charges at high charge states.

While charge fixing modifications like TMP can reduce charge mobility and direct fragmentation to charge remote channels, the specificity of these channels presents challenges in many sequencing analyses for which the more evenly distributed fragmentation provided by mobile proton pathways is superior. We investigated the use of carbodiimide chemistry to modify acidic sites (aspartic and glutamic acids and the C-terminus) to amides⁵⁰⁻⁵² (Figure 3-3 A) as a method to block charge remote fragmentation at acidic residues. Modification of SERF by 1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide (EDC), with glycine added in excess as a nucleophile, resulted in capping of carboxylic acid groups with the glycine (Figure 3-3 B, Table II-1). Compared to unmodified SERF (Figure 3-3 C), EDC modification of SERF

resulted in a dramatic improvement in sequence coverage, with ~20 sites exhibiting more than 10% relative intensity of fragments (Figure 3-3 D), as opposed to just 2 sites in unmodified SERF. As noted above, in unmodified SERF, the large number of arginines (9) and lysines (14) results in very low charge mobility even at high charge states (*e.g.* 13⁺), and fragmentation at acidic residues dominates, with >60% of fragment intensity resulting from cleavage at aspartic acid alone, an enhancement of nearly 10-fold compared to the frequency of aspartic acid in

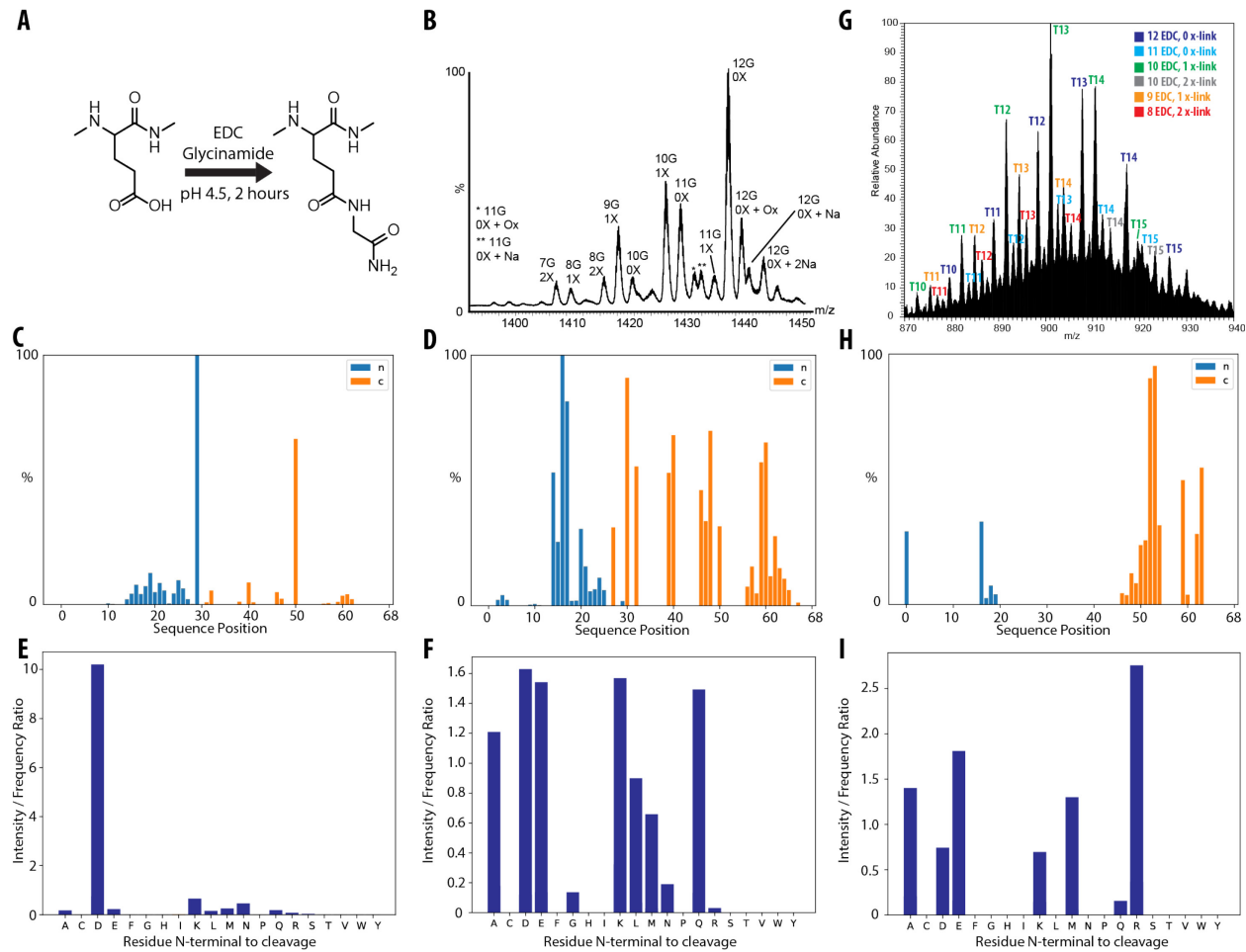


Figure 3-3 Effect of blocking acidic residues with amides on protein fragmentation. A) Modification of carboxylic acids with EDC in the presence of excess glycylglycine converts them to amides. B) Mass spectrum of EDC-modified SERF (6+) shows near complete derivatization of the 12 available carboxylic acids with minor contributions from crosslinks. C) Sequence coverage (shown by fragment relative intensity) as a function of position from the N-terminus (position 0) to C-terminus (position 68) for unmodified (C), EDC-modified (D), and TMP + EDC-modified (H) SERF. EDC-modified SERF exhibits the greatest sequence coverage. E) Fragmentation propensity ratio maps for unmodified (E) and EDC-modified (F) SERF. Inset of (F) shows the number of each amino acid present in SERF, which is similar to the distribution of fragment intensity following EDC modification. G) Mass spectrum of SERF (10+) modified with both TMP and EDC, producing a series of 10-15 TMP (“Tx^x”, where “xx” is the number of TMPs) modifications for each of 6 major EDC modification types (color coded in legend). H) Sequence coverage map of TMP + EDC modified SERF. I) Fragmentation propensity ratios of TMP + EDC-modified SERF.

SERF (Figure 3-3 E). When acidic residues are converted to amides by EDC, the charge remote Asp-Xxx fragmentation pathway is effectively blocked, and fragmentation occurs more evenly across the amino acids present in SERF (Figure 3-3 F). No single residue experiences more than a 1.6-fold enhancement in fragmentation relative to its frequency, and many residues are clustered around a ratio of 1, indicating a remarkably stochastic distribution of fragmentation (Figure 3-3 F). This distribution of fragment ions across many types of amino acids is characteristic of mobile proton fragmentation and indicates that by blocking the charge remote pathway, the protein returns to primarily mobile proton-mediated fragmentation, even under conditions of very low charge mobility. Additional energy is required to cause fragmentation following EDC modification (Figure II-2), confirming that the mobile proton pathway is less favorable than charge remote fragmentation of the unmodified SERF protein under the low charge mobility conditions analyzed here. Other charge states of EDC-modified SERF show similarly distributed fragmentation across many amino acids (Figure II-3). Additionally, TMP-modified SERF was subsequently modified using EDC to investigate the fragmentation pathways with charge remote channels blocked under conditions of low charge mobility. The combination of modifications results in a complex mass spectrum (Figure 3-3 G) that corresponds to each of the major peaks in the SERF + EDC spectrum (Figure 3-3 B, Table II-2) with a range of 9-15 TMPs. Similar to the EDC-modified SERF, SERF with both capped acidic residues (EDC) and fixed charges from TMP exhibited relatively stochastic fragmentation, without clear evidence of any residue-specific charge remote fragmentation channels (Figure 3-3 H, I). Due to the large number of modification states present, signal dilution resulted in significantly lower yields of detectable fragment ions, and the minor differences observed

between the EDC and EDC + TMP fragmentation results are largely attributed to the resulting differences in data quality.

As charge mobility is the primary factor influencing the relative favorability of charge remote vs mobile proton fragmentation pathways, predicting charge mobility (e.g. from protein sequence and observed charge, as in Figure 3-1) would provide a method to predict the expected fragmentation pathways of a protein. In Figure 3-1, it was assumed that an added fixed charge would replace a charging proton; however, observations from several TMP-modified proteins indicate this is not always the case. For SERF in particular, more intrinsically charged TMP modifications were observed than the charge of the protein in many cases. For example, up to 14 TMPs are observed on the 8+ charge state of SERF (Figure II-4 A) without adduction of anionic charge carriers (e.g. acetate or other anions). Thus, up to six charges on this 8+ SERF are being neutralized, presumably via intramolecular salt bridges or charge solvation. Salt bridges within positively charged protein ions can result from the pairing of acidic and basic residues or positive charge, resulting in a neutral net charge. Capping acidic residues with EDC prevents this common form of salt bridge from occurring, as no carboxylic acids are available following modification. Comparing the charge states of TMP-modified SERF (Figure II-4 B) to TMP+EDC-modified SERF (Figure II-4 C), minimal differences in charge states are observed, with distributions centered around 10-11⁺ and extending from 7⁺ to 15⁺ in both cases.

We performed molecular dynamics simulations of modified and unmodified SERF protein to investigate the neutralization of these charges in the absence of canonical salt bridging sites. REX simulations provided models for unmodified and TMP and EDC derivatized SERF molecules (Figure II-5, Figure II-6) with CCS values in excellent agreement with values observed from experiment. Unmodified SERF exhibits a large range of charge states and CCS

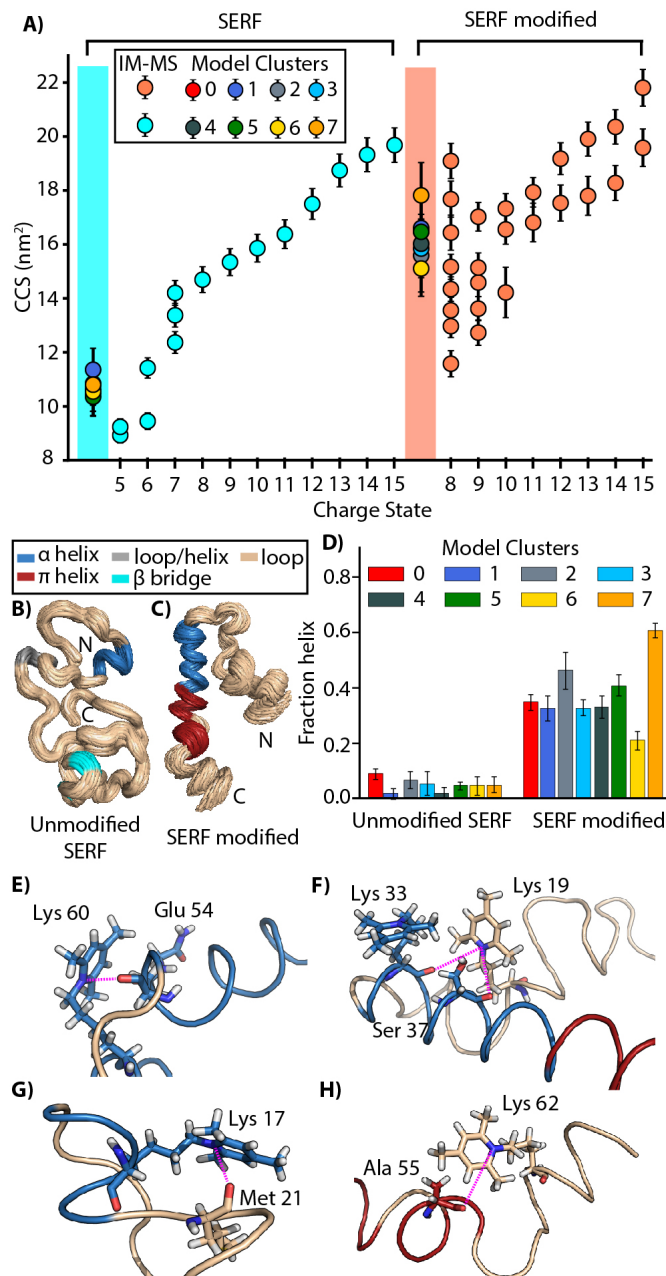


Figure 3-4 A) Plot of CCS vs. all charge states observed for unmodified and modified SERF. The shaded region in cyan and orange shows theoretical CCS calculated using IMPACT¹¹ and IMoS^{12,13} (See Supporting Info) for unmodified and modified SERF. The most dominant cluster of unmodified and modified SERF structures are shown in B) and C), respectively. The structures are color coded according to the secondary structure element. D) Plot of fraction of helix observed in all clusters for unmodified and modified SERF. E), F), G), and H) shows electrostatic interactions between lysine modified with TMP to carbonyl oxygen atoms on the peptide backbone.

values, ranging from 5⁺ to 15⁺ and 8.9 nm² to 19.6 nm², respectively (Figure 3-4 A, Table II-3). There is a general trend of increasing CCS as a function of increasing charge states for SERF species, as is common for disordered proteins. SERF cluster models from 300 K replica of REX simulations have mean CCS values ranging from 10.3 nm² to 11.3 nm² (Figure 3-4 A, Table II-4). These values are within 3 % of CCS values for SERF 5⁺, 6⁺, and the most compact 7⁺ conformers. Like unmodified SERF, SERF modified with TMP and EDC exhibits a range of charge states (8⁺ to 15⁺) and conformations, as evidenced by large distribution of CCS values, with mean value ranging from 11.5 nm² to 21.8 nm² (Figure 3-4 A, Table II-4). However, CCS values recorded for modified SERF does not increase near-linearly as a function of protein charge state as in the case of unmodified SERF.

The 8⁺ species of modified SERF presents a highly heterogeneous CCS distribution, with mean

CCS values spanning 11.5 nm² to 19.1 nm², accounting for almost 88% of CCS observed across all charge states (Table II-4). Modified SERF cluster models also exhibit a large CCS distribution with values ranging from 15.1 nm² to 17.8 nm², as shown in Figure 3-4 A (Table II-4). These values encapsulate many structural populations observed experimentally across different charge states. We further evaluated our SERF models in order to examine the structures in detail, leveraging the excellent CCS agreement achieved between theory and experiment. The most dominant cluster observed (Cluster 0) for both unmodified and modified versions of SERF are shown in Figure 3-4 B and C, respectively. Unmodified SERF adopts a mostly unstructured form in the gas phase, consisting primarily of loops, with a total helical content of < 10% (Figure 3-4 B, Table II-5, Figure II-8). However, when labeled with TMP and EDC, SERF gains substantial helical content over the unmodified version (Figure 3-4 C, 4D, Table II-6, Figure II-9). We observed regions of both α -helix and π -helix⁵³ in most of our output structures for modified SERF (Figure 3-4 C, Table II-6, Figure II-9). As shown in Figure 3-4 D, modified SERF clusters exhibit helical content in the range of 21-61%.

Our inspection of SERF model structures produced by REX simulations led us to speculate that the helices generated upon TMP/EDC modification may provide the protein an alternative means of charge solvation, thus rationalizing both the relatively invariant SERF charge state distributions and fragmentation chemistry observed in our experiments. By evaluating z-scores computed in order to track interactions between the positively charged nitrogen atom in TMP-modified lysine and carbonyl oxygen atoms in the peptide backbone within a distance of 5 Å, we observe that lysines are involved in many electrostatic interactions in both unmodified and modified SERF (Figure II-12, Figure II-13). The highly compact structures modeled for unmodified SERF show many of these electrostatic interactions with

carbonyl oxygen atoms. As these structures match the experimental CCS for very low charge states of SERF, the number of these interactions observed indicates a highly compacted structure supported by a high degree of charge solvation. In fully modified SERF, we observe electrostatic interactions specifically involved in the helical regions of the protein (Figure 3-4 E, F, G, H, and Figure II-13). For example, the fixed positive charge on derivatized lysine 60 interacts with the c-terminal end of the α -helix shown in Figure 3-4 E, stabilizing the helix macrodipole. We also observe the fixed positive charge on lysine 19 in the loop region interacting with two carbonyl oxygen atoms in the helix region (Figure 3-4 F), instances where a fixed positive charge within an α -helix region interacts with a loop region (Figure 3-4 G), and interactions in π -helix regions, for example, where the fixed positive charge on lysine 62 interacting with the carbonyl oxygen in alanine 55 (Figure 3-4 H). Our models also suggest that there are many longer range interactions (Figure II-10 - 13) within 10 Å with similar distribution to the interactions observed within a 5 Å distance threshold we used to analyze the modified SERF clusters (Figure II-13), indicating that both short and long range interactions could be involved in charge solvation. The prevalence of such electrostatic interactions in our models for both modified and unmodified SERF suggest that such forces play an important role in the gas phase structures of biomolecular ions and represent a potentially significant barrier to top-down sequencing methods for intact proteins, especially those captured in native-like conformations.

3.4 Conclusions

We present a wide-ranging data, using multiple chemical derivatization methods to control the CID fragmentation pathways of intact proteins. By affixing stable, intrinsically charged moieties with TMP, we direct fragmentation to charge remote channels by replacing mobile protons with fixed charges and show that the effects of this derivatization are predictable given the number of

arginines present in protein sequence and the charge state observed following electrospray. Following extensive labeling under conditions of very low charge mobility, we also demonstrate fragmentation of an intact protein specifically at the site of introduced charged labels, a foundation with the potential to provide a unique degree of control over protein dissociation via CID. Alternatively, by blocking acidic functional groups with EDC, we can prevent charge remote fragmentation and return to mobile proton-mediated dissociation regardless of the level of charge mobility in the protein. Combining these labeling techniques results in a paradigm in which any protein can be fragmented under charge remote or charge directed conditions regardless of the initial charge mobility of the system, provided that sufficient reactive residues are available for the derivatization in question. Further development of stable, intrinsically charged labels targeting other amino acids offers a path to ensure sufficient charge can be affixed to any protein system of interest.

Finally, experiments with intrinsically charged labels on protein ions reveal the immense capacity of these ions to accommodate excess charge. Through capping of acidic residues and molecular dynamics simulations, we find evidence for widespread solvation of charges in gas-phase protein ions, including to the point of substantially altering the secondary structure in charge-derivatized SERF. We found this result especially surprising, as proteins routinely undergo denaturation in order to accommodate excess charge, yet our data suggests that proteins possess a previous unappreciated capacity to absorb excess charge and remain compact, even when canonical salt-bridging within the structure is blocked. The observations indicate charge solvation plays a critical role in the structures of gas-phase protein ions and represents a key challenge to overcome for further development of top-down sequencing technologies.

3.5 References

- (1) Smith, L. M.; Kelleher, N. L. Proteoform: A Single Term Describing Protein Complexity. *Nat. Methods* **2013**, *10* (3), 186–187.
- (2) Kelleher, N. L.; Thomas, P. M.; Ntai, I.; Compton, P. D.; Leduc, R. D. Deep and Quantitative Top-down Proteomics in Clinical and Translational Research. *Expert Rev. Proteomics* **2014**, *11* (6), 649–651.
- (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.; et al. How Many Human Proteoforms Are There? *Nat. Chem. Biol.* **2018**, *14* (3), 206–214.
- (4) Savaryn, J. P.; Catherman, A. D.; Thomas, P. M.; Abecassis, M. M.; Kelleher, N. L. The Emergence of Top-down Proteomics in Clinical Research. *Genome Med.* **2013**, *5* (6), 53.
- (5) 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.
- (6) Eng, J. K.; McCormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976–989.
- (7) Nesvizhskii, A. I.; Aebersold, R. Interpretation of Shotgun Proteomic Data. *Mol. Cell. Proteomics* **2005**, *4* (10), 1419–1440.
- (8) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *Electrophoresis* **1999**, *20* (18), 3551–3567.
- (9) Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Brechi, L. A. Mobile and Localized Protons: A Framework for Understanding Peptide Dissociation. *J. Mass Spectrom.* **2000**, *35* (12), 1399–1406.
- (10) Tsaprailis, G.; Nair, H.; Somogyi, Á.; Wysocki, V. H.; Zhong, W.; Futrell, J. H.; Summerfield, S. G.; Gaskell, S. J. Influence of Secondary Structure on the Fragmentation of Protonated Peptides. *J. Am. Chem. Soc.* **1999**, *121* (22), 5142–5154.
- (11) Jones, J. L.; Dongre, A. R.; Somogyi, A.; Wysocki, V. H. Sequence Dependence of Peptide Fragmentation Efficiency Curves Determined by Electrospray Ionization/Surface-Induced Dissociation Mass Spectrometry. *J. Am. Chem. Soc.* **1994**, *116* (18), 8368–8369.
- (12) Dongré, A. R.; Jones, J. L.; Somogyi, Á.; Wysocki, V. H. Influence of Peptide Composition, Gas-Phase Basicity, and Chemical Modification on Fragmentation Efficiency: Evidence for the Mobile Proton Model. *J. Am. Chem. Soc.* **1996**, *118* (35), 8365–8374.
- (13) Harrison, A. G.; Yalcin, T. Proton Mobility in Protonated Amino Acids and Peptides. *Int. J. Mass Spectrom. Ion Process.* **1997**, *165–166*, 339–347.
- (14) Cox, K. A.; Gaskell, S. J.; Morris, M.; Whiting, A. Role of the Site of Protonation in the Low-Energy Decompositions of Gas-Phase Peptide Ions. *J. Am. Soc. Mass Spectrom.* **1996**, *7* (6), 522–531.
- (15) Tang, X. J.; Thibault, P.; Boyd, R. K. Fragmentation Reactions of Multiply-Protonated Peptides and Implications for Sequencing by Tandem Mass Spectrometry with Low-Energy Collision-Induced Dissociation. *Anal. Chem.* **1993**, *65* (20), 2824–2834.
- (16) Palzs, B.; Suhal, S. Fragmentation Pathways of Protonated Peptides. *Mass Spectrom. Rev.* **2005**, *24* (4), 508–548.

- (17) Johnson, R. S.; Martin, S. A.; Biemann, K. Collision-Induced Fragmentation of (M + H)⁺-ions of Peptides. Side Chain Specific Sequence Ions. *Int. J. Mass Spectrom. Ion Process.* **1988**, *86* (C), 137–154.
- (18) Jensen, N. J.; Tomer, K. B.; Gross, M. L. Gas-Phase Ion Decomposition Occurring Remote to a Charge Site. *J. Am. Chem. Soc.* **1985**, *107* (7), 1863–1868.
- (19) Yu, W.; Vath, J. E.; Huberty, M. C.; Martin, S. A. Identification of the Facile Gas-Phase Cleavage of the Asp-Pro and Asp-Xxx Peptide Bonds in Matrix-Assisted Laser Desorption Time-of-Flight Mass Spectrometry. *Anal. Chem.* **1993**, *65* (21), 3015–3023.
- (20) Price, W. D.; Schnier, P. D.; Jockusch, R. A.; Strittmatter, E. F.; Williams, E. R. Unimolecular Reaction Kinetics in the High-Pressure Limit without Collisions. *J. Am. Chem. Soc.* **1996**, *118* (43), 10640–10644.
- (21) Tsaprailis, G.; Somogyi, Á.; Nikolaev, E. N.; Wysocki, V. H. Refining the Model for Selective Cleavage at Acidic Residues in Arginine-Containing Protonated Peptides. *Int. J. Mass Spectrom.* **2000**, *195–196*, 467–479.
- (22) Paizs, B.; Suhai, S.; Hargittai, B.; Hruby, V. J.; Somogyi, Á. Ab Initio and MS/MS Studies on Protonated Peptides Containing Basic and Acidic Amino Acid Residues - I. Solvated Proton vs. Salt-Bridged Structures and the Cleavage of the Terminal Amide Bond of Protonated RD-NH₂. *Int. J. Mass Spectrom.* **2002**, *219* (1), 203–232.
- (23) Bythell, B. J.; Suhai, S.; Somogyi, Á.; Paizs, B. Proton-Driven Amide Bond-Cleavage Pathways of Gas-Phase Peptide Ions Lacking Mobile Protons. *J. Am. Chem. Soc.* **2009**, *131* (39), 14057–14065.
- (24) Haverland, N. A.; Skinner, O. S.; Fellers, R. T.; Tariq, A. A.; Early, B. P.; LeDuc, R. D.; Fornelli, L.; Compton, P. D.; Kelleher, N. L. Defining Gas-Phase Fragmentation Propensities of Intact Proteins During Native Top-Down Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (6), 1203–1215.
- (25) Wang, H.; Leeming, M. G.; Ho, J.; Donald, W. A. Origin and Prediction of Highly Specific Bond Cleavage Sites in the Thermal Activation of Intact Protein Ions. *Chem. – Eur. J.* **2018**, chem.201804668.
- (26) 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.
- (27) Chen, J.; Shiyonov, P.; Zhang, L.; Schlager, J. J.; Green-Church, K. B. Top-down Characterization of a Native Highly Intralinked Protein: Concurrent Cleavages of Disulfide and Protein Backbone Bonds. *Anal. Chem.* **2010**, *82* (14), 6079–6089.
- (28) Roth, K. D. W.; Huang, Z. H.; Sadagopan, N.; Watson, J. T. Charge Derivatization of Peptides for Analysis by Mass Spectrometry. *Mass Spectrom. Rev.* **1998**, *17* (4), 255–274.
- (29) Johnson, R. S.; Martin, S. A.; Biemann, K. Collision-Induced Fragmentation of (M + H)⁺ Ions of Peptides. Side Chain Specific Sequence Ions. *Int. J. Mass Spectrom. Ion Process.* **1988**, *86*, 137–154.
- (30) He, Y.; Reilly, J. P. Does a Charge Tag Really Provide a Fixed Charge? *Angew. Chemie - Int. Ed.* **2008**, *47* (13), 2463–2465.
- (31) Krusemark, C. J.; Frey, B. L.; Belshaw, P. J.; Smith, L. M. Modifying the Charge State Distribution of Proteins in Electrospray Ionization Mass Spectrometry by Chemical Derivatization. *J. Am. Soc. Mass Spectrom.* **2009**, *20* (9), 1617–1625.
- (32) Frey, B. L.; Krusemark, C. J.; Ledvina, A. R.; Coon, J. J.; Belshaw, P. J.; Smith, L. M. Ion-Ion Reactions with Fixed-Charge Modified Proteins to Produce Ions in a Single, Very

- High Charge State. *Int. J. Mass Spectrom.* **2008**, *276* (2–3), 136–143.
- (33) Krusemark, C. J.; Ferguson, J. T.; Wenger, C. D.; Kelleher, N. L.; Belshaw, P. J. Global Amine and Acid Functional Group Modification of Proteins. *Anal. Chem.* **2008**, *80* (3), 713–720.
- (34) Greer, S. M.; Holden, D. D.; Fellers, R.; Kelleher, N. L.; Brodbelt, J. S. Modulation of Protein Fragmentation Through Carbamylation of Primary Amines. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (8), 1587–1599.
- (35) Polasky, D. A.; Lermyte, F.; Nshanian, M.; Sobott, F.; Andrews, P. C.; Loo, J. A.; Ruotolo, B. T. Fixed-Charge Trimethyl Pyrylium Modification for Enabling Enhanced Top-Down Mass Spectrometry Sequencing of Intact Protein Complexes. *Anal. Chem.* **2018**, *90* (4), 2756–2764.
- (36) Avtonomov, D. M.; Polasky, D. A.; Ruotolo, B. T.; Nesvizhskii, A. I. IMTBX and Grppr: Software for Top-Down Proteomics Utilizing Ion Mobility-Mass Spectrometry. *Anal. Chem.* **2018**, *90* (3), 2369–2375.
- (37) Haynes, S. E.; Polasky, D. A.; Dixit, S. M.; Majmudar, J. D.; Neeson, K.; Ruotolo, B. T.; Martin, B. R. Variable-Velocity Traveling-Wave Ion Mobility Separation Enhancing Peak Capacity for Data-Independent Acquisition Proteomics. *Anal. Chem.* **2017**, *89* (11), 5669–5672.
- (38) Polasky, D. A.; Dixit, S. M.; Fantin, S. M.; Ruotolo, B. T. CIUSuite 2: Next-Generation Software for the Analysis of Gas-Phase Protein Unfolding Data. *Anal. Chem.* **2019**, *91* (4), 3147–3155.
- (39) Ruotolo, B. T.; Benesch, J. L. P.; Sandercock, A. M.; Hyung, S.-J.; Robinson, C. V. Ion Mobility–mass Spectrometry Analysis of Large Protein Complexes. *Nat. Protoc.* **2008**, *3* (7), 1139–1152.
- (40) Buck, M.; Bouguet-Bonnet, S.; Pastor, R. W.; MacKerell Jr., A. D. Importance of the CMAP Correction to the CHARMM22 Protein Force Field: Dynamics of Hen Lysozyme. *Biophys J* **2006**, *90* (4), L36-8.
- (41) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; et al. CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2009**, *31* (4), NA-NA.
- (42) Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **2012**, *52* (12), 3144–3154.
- (43) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* **2012**, *52* (12), 3155–3168.
- (44) Im, W.; Brooks Iii, C. L. *Interfacial Folding and Membrane Insertion of Designed Peptides Studied by Molecular Dynamics Simulations*; 2005; Vol. 10.
- (45) Michael Feig, †; Alexander D. MacKerell, J. . and; Charles L. Brooks, I. Force Field Influence on the Observation of π -Helical Protein Structures in Molecular Dynamics Simulations. **2003**.
- (46) Chen, J.; Brooks, C. L. Can Molecular Dynamics Simulations Provide High-Resolution Refinement of Protein Structure? *Proteins Struct. Funct. Bioinforma.* **2007**, *67* (4), 922–930.
- (47) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein

- Folding. *Chem. Phys. Lett.* **1999**, *314* (1–2), 141–151.
- (48) Li, X.; Cournoyer, J. J.; Lin, C.; O'Connor, P. B. The Effect of Fixed Charge Modifications on Electron Capture Dissociation. *J. Am. Soc. Mass Spectrom.* **2008**, *19* (10), 1514–1526.
- (49) van Ham, T. J.; Holmberg, M. A.; van der Goot, A. T.; Teuling, E.; Garcia-Arencibia, M.; Kim, H. eui; Du, D.; Thijssen, K. L.; Wiersma, M.; Burggraaff, R.; et al. Identification of MOAG-4/SERF as a Regulator of Age-Related Proteotoxicity. *Cell* **2010**, *142* (4), 601–612.
- (50) Zhang, H.; Wen, J.; Huang, R. Y. C.; Blankenship, R. E.; Gross, M. L. Mass Spectrometry-Based Carboxyl Footprinting of Proteins: Method Evaluation. *Int. J. Mass Spectrom.* **2012**, *312*, 78–86.
- (51) Mendoza, V. L.; Vachet, R. W. Probing Protein Structure by Amino Acid-Specific Covalent Labeling and Mass Spectrometry. *Mass Spectrom. Rev.* **2009**, *28* (5), 785–815.
- (52) Hoare, D. G.; Koshland, D. E. A Method for the Quantitative Modification and Estimation of Carboxylic Acid Groups in Proteins. *J. Biol. Chem.* **1967**, *242* (10), 2447–2453.
- (53) Rajagopalan Sudha; Motoya Kohtani; Gary A. Breaux, and; Jarrold*, M. F. π -Helix Preference in Unsolvated Peptides. **2004**.

Chapter 4 IMTBX and Grppr: Software for Top-Down Proteomics Utilizing Ion Mobility-Mass Spectrometry

Adapted with permission from: Dmitry M. Avtonomov*, Daniel A. Polasky*, Brandon T. Ruotolo, and Alexey I. Nesvizhskii. IMTBX and Grppr: Software for Top-Down Proteomics Utilizing Ion Mobility-Mass Spectrometry. *Anal. Chem.* **2018**, *90* (3), 2369–2375.
(*equal contribution)

4.1 Abstract

Top-down proteomics has emerged as a transformative method for the analysis of protein sequence and post-translational modifications (PTMs). Top-down experiments have historically been performed primarily on ultrahigh resolution mass spectrometers due to the complexity of spectra resulting from fragmentation of intact proteins, but recent advances in coupling ion mobility separations to faster, lower resolution mass analyzers now offer a viable alternative. However, software capable of interpreting the highly complex two-dimensional spectra that result from coupling ion mobility separation to top-down experiments is currently lacking. In this manuscript we present a software suite consisting of two programs, IMTBX (“IM Toolbox”) and Grppr (“Grouper”), that enable fully automated processing of such data. We demonstrate the capabilities of this software suite by examining a series of intact proteins on a Waters Synapt G2 ion-mobility equipped mass spectrometer and compare the results to the manual and semiautomated data analysis procedures we have used previously.

4.2 Introduction

Rapid assessment of protein sequence by mass spectrometry (MS) has enabled a revolution in the analysis of biological and biochemical samples,¹ powering the rapidly

expanding field of proteomics. Typical proteomic workflows involve the enzymatic digestion of proteins into peptides, which can be separated by chromatography to enable rapid analysis by MS.^{2,3} Despite the advances made with these techniques, the requirement for enzymatic digestion in these “bottom-up” approaches often complicates identification and complete analysis of the post-translational modifications (PTMs) and proteoforms that are critical to biological function of proteins.^{4,5} In response, “top-down” MS techniques have been developed to sequence intact proteins without enzymatic digestion.⁶⁻⁸

Top-down MS holds great promise but remains less common than bottom-up approaches due to a number of technical challenges associated with achieving efficient separation and fragmentation of intact proteins. One of the major challenges associated with top-down techniques is the low signal-to-noise ratio (SNR) of sequence-informative fragment ions.⁹ Fragmentation of intact proteins typically results in a population of ions distributed among charge states ranging from 1+ to 5+ or higher, with masses extending from 100 to over 10000 Da. For larger proteins in particular, the need to generate hundreds of fragments from a single precursor in order to achieve sufficient sequence coverage often results in poor SNR for fragment ions, as well as many instances of overlapping isotopic distributions. To resolve these overlapping distributions, most top-down MS workflows utilize Fourier transform (FT) instruments (both ion cyclotron resonance (ICR) and Orbitrap) that possess extremely high MS resolving power. While capable of producing exceptionally narrow line widths for MS measurements, such equipment also requires significant acquisition times per spectrum to achieve high performance results, creating a situation that gives rise to low instrument duty cycles and difficulties in coupling to chromatographic separations. Slow acquisitions compound the challenge of low SNR for fragment ions, as fewer spectra can be acquired to average and

improve spectral quality. Significant efforts have been made to mitigate these challenges,¹⁰⁻¹² but they remain a barrier to widespread adoption of top-down approaches.

Ion mobility-mass spectrometry (IM-MS) offers a promising alternate path to alleviate several of the challenges detailed above by providing additional separation of the product ions created during top-down sequencing experiments. IM separates ions based on their migration time through a chamber pressurized with inert neutrals under the influence of a relatively weak electric field. Ions of different orientationally averaged sizes (collision cross sections, CCSs and charges take different amounts of time to travel through the IM device, offering significant improvements to the peak capacity of a given mass analyzer for a complex fragment ion population.¹³ The improvements offered by IM have seen adoption for bottom-up proteomics,¹⁴⁻¹⁶ and initial proof-of-concept demonstrations for top-down protein sequencing have been reported,^{17,18} but the routine use of IM-MS for top-down proteomics remains rare. This trend runs counter to a dramatic increase in the use of IM-MS in many other MS-based fields, such as native and structural MS,¹⁹ metabolomics,²⁰ and drug discovery,²¹ which have appeared since the introduction of commercial IM-MS instruments.

Perhaps the most significant reason for the lack of adoption of IM-MS for top-down proteomics is the general lack of software for processing IM-MS data containing the complex fragment ion populations inherent in top-down analyses. It has been well established for MS data that top-down data requires specialized analysis tools that differ from those used for bottom-up data,^{22,23} and the same need for specialized tools is clearly required for top-down IM-MS data. While many software tools have been developed for automated handling of top-down MS data without IM, they are generally incapable of handling IM-MS data, due both to differences in the spectral characteristics of the raw data and a general inability to read raw IM-MS data from any

format in which it is natively generated. Indeed, all reports of IM-MS top-down to date, to our knowledge, rely largely or entirely on manual interpretation of spectra. The complexity of fragment ion populations generated in top-down experiments makes this a massively time-consuming process, as hundreds to thousands of peaks must be identified in each data set, severely limiting the scope and speed of viable experiments. Here, we describe a software suite capable of processing top-down IM-MS data in a rapid and automated fashion to enable routine top-down IM-MS analyses. The suite consists of two tools, IMTBX (“IM Toolbox”) and Grprr (“Grouper”), which together can completely process complex top-down fragmentation data into an accurate monoisotopic peak list in seconds. The list can be annotated in a sequencing program of choice. A plotting module enables flexible visualization of IM-MS data and detected peaks to evaluate the impact of peak detection parameters. The suite is modular, with a robust scripting interface to enable extension into any IM-MS data processing workflow. Both Grprr and IMTBX can be downloaded from github repositories: <https://github.com/chhh/Grprr> and <https://github.com/chhh/IMTBX> (the latter includes a combined package with both tools)

4.3 Methods

4.3.1 Experimental Section

Ubiquitin (bovine), Myoglobin (bovine), Carbonic Anhydrase (bovine), Transthyretin (human), β -lactoglobulin (bovine), Avidin (chicken), Aldolase (rabbit), and ammonium acetate were purchased from Sigma-Aldrich (St. Louis, MO). Serum amyloidp-component (human) protein was purchased from EMD Millipore (Billerica, MA). Proteins were buffer exchanged into 100 mM ammonium acetate using Micro Biospin 6 spin columns (BioRad, Hercules, CA) prior to analysis by IM-MS. All data was collected using a Synapt G2 HDMS IM-Q-ToF mass spectrometer (Waters, Milford, MA). Intact protein ions were generated using a direct infusion

nESI source in positive mode. Applied capillary voltage was 1.5 kV, and the sample cone and extraction cones were operated at 40 and 0 V, respectively. Instrument settings were optimized to preserve native-like protein ions and noncovalent complexes prior to CID, as described previously.²⁴ A single charge state protein ion was selected in the quadrupole and collisionally activated in the trap traveling-wave ion guide (voltage applied depending on protein molecular weight, from 75 V for Ubiquitin to 200 V for Aldolase), which was pressurized to $2-4 \times 10^{-2}$ mbar with argon gas. Ion mobility separation was performed at a pressure of 3 mbar with 20 V wave height and 150m/s wave velocity. The ToF mass analyzer was operated over the range of 100–6000 m/z for Ubiquitin and Myoglobin, and 100–8000 m/z for all other proteins, at a pressure of 1.5×10^{-6} mbar

4.3.2 Data Processing

Data was analyzed with IMBTX and Grppr and a “semi-automated” workflow comprised of several software packages. For IMBTX analysis, IMBTX version 2.9.1 and Grppr version 0.3.6 were used. Parameters used can be found in Supporting Information, Table 1. A wrapper script written in Python 3.5 was used to operate both IMBTX and Grppr in a batch processing mode prior to the development of a user interface. Isotopic cluster output from Grppr was annotated using ProSight Lite²⁵ or a custom peak annotation program in Python 3.5 for batch processing. For the semiautomated processing comparison workflow, regions of two-dimensional IM-MS data space corresponding to fragment charge states were extracted using TWIMExtract.¹⁵ Extracted data were smoothed (Savitsky-Golay, 0.2 m/z window size, 3 cycles), peak-picked (Intensity threshold 500, picking height 90%), and deisotoped (max charge 5, isotope mass tolerance 0.05 m/z, isotope intensity tolerance 100%) in mMass v5.5.0.²⁶⁻²⁸ Final output was annotated with the same custom annotation script as IMBTX data. Charge state comparisons

were generated by comparing the annotated charge state against the m/z and drift time of each detected cluster. ProSight Lite, TWIMExtract, and mMass are all publicly available free software packages.

Processing raw data is a two-stage process: peak picking and isotopic cluster detection. First single 2D peaks (m/z , ion mobility drift time) in raw data are detected and written to a file, followed by grouping these peaks into isotopic clusters. Two in-house built software tools perform these steps: IMTBX (Ion Mobility Toolbox), which does raw peak picking, and Grppr (Grouper), which performs isotopic grouping.

4.3.3 Raw Feature Detection (IMTBX)

IMTBX is written in C# (.NET 4.5.1), it can currently read data from Waters IM-MS or SONAR enabled instruments, such as Synapt G2-Si or Xevo systems. It can also handle regular LC-MS data without IM. An API to support custom data format adapters is under development. For assistance with unsupported data formats, please contact the authors. IMTBX can operate in two modes: scan summation/averaging followed by 2D peak detection in a single combined scan, the mode that is used throughout this manuscript, or 3D peak detection for LC-MS data where retention time is added to the model. If the data contains a lock-mass channel (that is Waters specific), it can be processed as well - time dependent mass calibration correction can be calculated and applied to spectra as they are being processed.

Each IM scan in Waters data is represented as a sparse matrix of N rows and M columns, where N is the number of IM drift bins (typically 200), and M is the number of bins for m/z axis (varies, typically 100000–300000 bins). Data from other instruments that can be represented in such a simple format can be accommodated for IMTBX analysis. IMTBX data processing steps are schematically shown in the left part of Figure 4-1. For top-down applications, all ion mobility

scans are added together or averaged (IMTBX has options for both), then a 2D filter of user-selected size and shape (typically Gaussian) is applied for smoothing. Optionally, a step of removing “lone” data points can be applied, i.e. nonzero intensity data points that do not have many neighboring nonzero points. The exact meaning of “having many neighbors” is, again, user selectable, with the default settings optimized for data observed in this and similar studies. After that, local maxima in the scan are found and used as seed points for fitting the data with 2D Gaussians. Even though the peaks are actually not ideal 2D Gaussians, especially in the IM

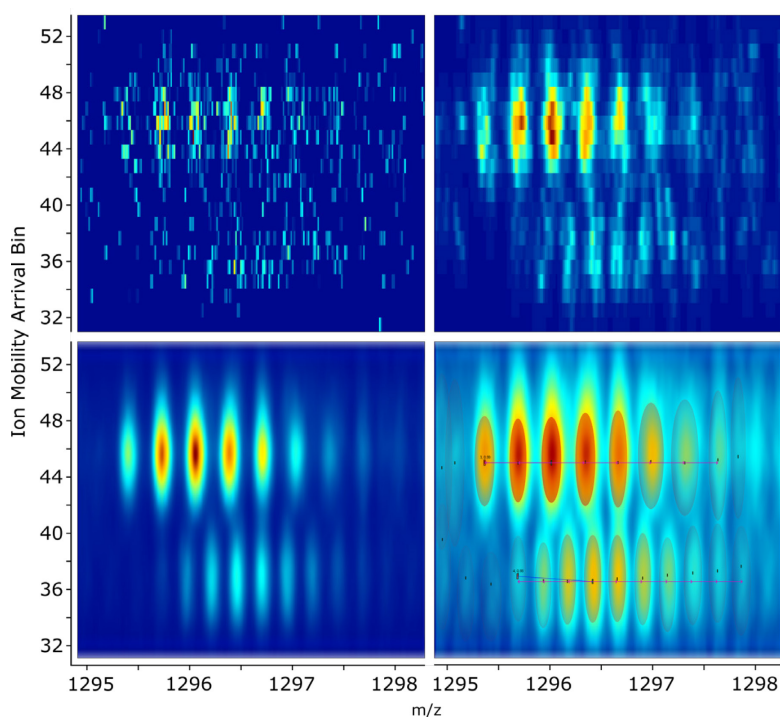


Figure 4-1 Top left: Raw unprocessed data from a single IM scan for 3+ and 4+ Ubiquitin ions from a Waters Synapt-G2 platform. A slice of data for m/z 1295–1298, drift 30–55 containing two differently charged ions is shown. Horizontal axis is m/z , vertical axis is IM drift time. Top right: Same data slice after Gaussian filtering with default parameters. Bottom left: 580 filtered scans added together, bilinear interpolation applied. Bottom right: Same as “bottom left” with intensities square rooted. Detected single 2D Gaussian peaks are marked with ovals, semiaxes corresponding to two standard deviations. Horizontal lines mark the locations and span of two detected isotopic clusters: charge 3+ at the top and 4+ at the bottom.

dimension, variations from this expectation do not negatively influence the data analysis described here to a significant extent. In addition, the actual maxima locations are also reported for each detected 2D peak, not only the means in m/z and IM dimensions. Several stages of data filtration are shown in

Figure 4-1.

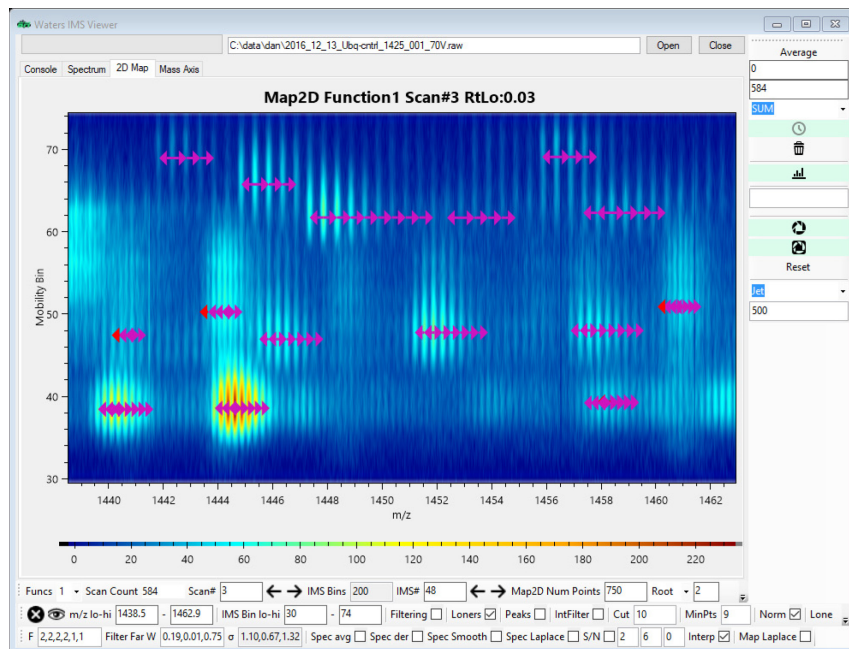


Figure 4-2 Graphical interface for viewing raw IM-MS data, processing stages and ion detection results. The software displays 1D spectra, 2D spectra, and the mass axis bin scaling. The various controls shown at the bottom allow to turn on/off and customize individual processing steps available in IMTBX. Shown in the figure is a 2D map view of a portion of an ion mobility scan (ion mobility drift time vs m/z , intensity as color) processed by IMTBX with default parameters with an overlay of isotopic clustering results by Grppr as series of purple arrows.

As top-down proteomics data often contains low SNR features that possess many local maxima even after filtration, our fitting procedure may result in multiple peaks in the proximity of the true peak apex location. Thus, the resulting list of detected peaks is further filtered to remove such spurious signals. For each peak,

multiple values are reported including: m/z , m/z standard deviation, drift time, drift time standard deviation, intensity at apex, total intensity (the sum of intensities over all data points assigned to the identified peak), and area (the number of data points used to define the peak). Several output text formats are available. For each peak, SNR is established using root-mean-square (RMS) noise estimation. Depending on user-defined parameters, peak lists that result from the above procedure may contain peaks that result from noise rather than analyte signals. In such cases, our isotopic grouping program, Grppr (described below), uses 2D isotopic clustering as a stringent filter, ignoring peaks for which other isotopologues are not reliably detected. Most of these processing steps can be visualized in the software GUI (graphical user interface) that can display 1D and 2D spectra both before and after signal processing, with options to enable or disable

individual steps. The GUI for our software can overlay detected peaks and isotopic clusters over raw data as shown in Figure 4-2.

4.3.4 Isotopic Clustering (Grppr)

Grppr is a general purpose 2D deisotoping algorithm implemented in Java 8, which is not

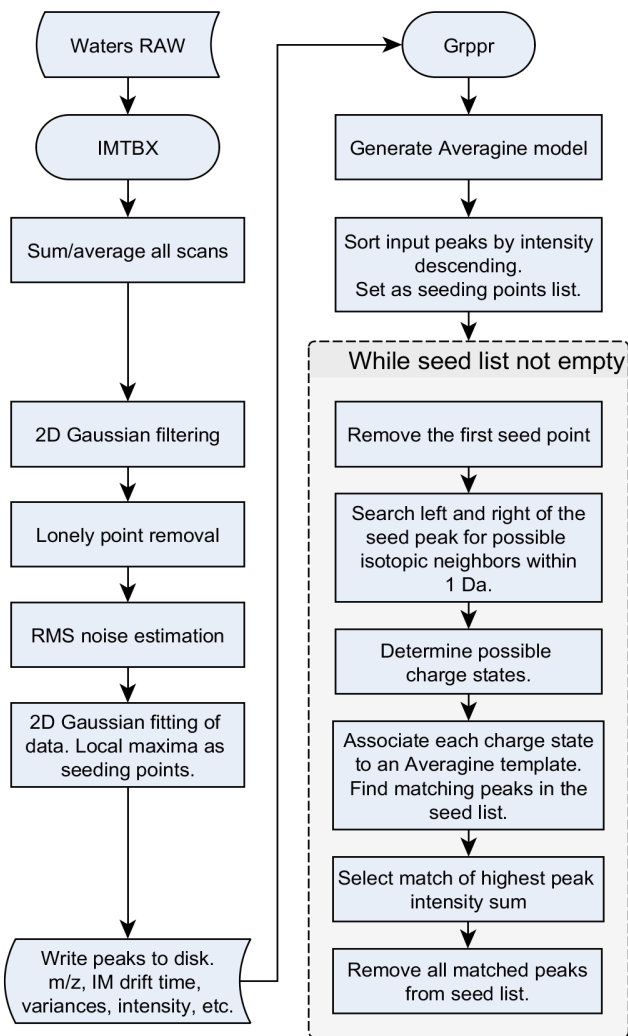


Figure 4-3 Schematic Overview of the Data Processing Steps Taken by IMTBX and Grppr for Top-down IM-MS Data Processing

tied to any particular input data format. While these two dimensions refer primarily to the IM and MS, as discussed above, Grppr can also be used for LC-MS data sets, where 2D features appear in m/z and retention time space. As input, Grppr requires a set of 2D peaks containing peak intensities as well as locations and widths in both dimensions (m/z and drift time for IM-MS data). Grppr has a simple application programming interface (API) facilitating adaptation of any input peak list format for Grppr analysis.

Grppr provides two algorithms for isotopic cluster detection: “Convex” and “Averagine”. The “Convex”

implementation searches for intensity-wise convex sets of neighboring peaks, that is, groups of peaks that are spaced in the m/z dimension in a manner consistent with a single charge state

assignment for all peaks and with intensities that consecutively decrease when moving starting from the highest peak of the set. While this approach works reasonably well for LC-MS bottom-up proteomic data, analyses of training data sets revealed systematic errors in this approach for top-down protein sequencing data sets. For example, after summation of hundreds of individual IM-MS spectra contained within a typical top-down protein sequencing data set, chemical noise from low abundance fragments results in small peaks at virtually each unit mass, which leads to incorrect isotope cluster assignments. In contrast to “Convex” isotope cluster detection, “Averagine” clustering implementation is based on the Averagine²⁹ model amino acid. For each input peak Grppr receives, a theoretical isotopic distribution is calculated and compared to the data. The advantage of this algorithm is that it can infer the correct mass of the monoisotope in low SNR data sets containing significant noise backgrounds, even when the monoisotopic peak is not detectable. For top-down proteomics data, we have found that the “Averagine” implementation is superior to the “Convex” approach described above. A high-level overview of the Averagine based deisotoping algorithm used in Grppr is presented in the right half of Figure 4-3, while a more detailed description with pseudocode can be found in Figure III-1.

4.4 Results and Discussion

Typical top-down proteomics workflows involve the separation and subsequent fragmentation of individual intact proteins in a mass spectrometer. IM offers substantial advantages for these experiments by providing a gas-phase separation technique that can separate the peptide product ions, analogous to the separation of tryptic peptides in bottom-up proteomics. Separation of such fragments using IM allows for accurate peak detection and isotopic clustering in highly complex data sets. Figure 4-4 demonstrates the utility of IM separation of fragments for a particular region of a plot of drift time versus m/z containing eight isotopic clusters derived

from the top- down collision induced dissociation (CID) of Ubiquitin. When all IM drift time spectra are combined together, two clear isotopic clusters are observed (Figure 4-4 B), along with an overlapped set of peaks that cannot be readily resolved into individual clusters. The two-

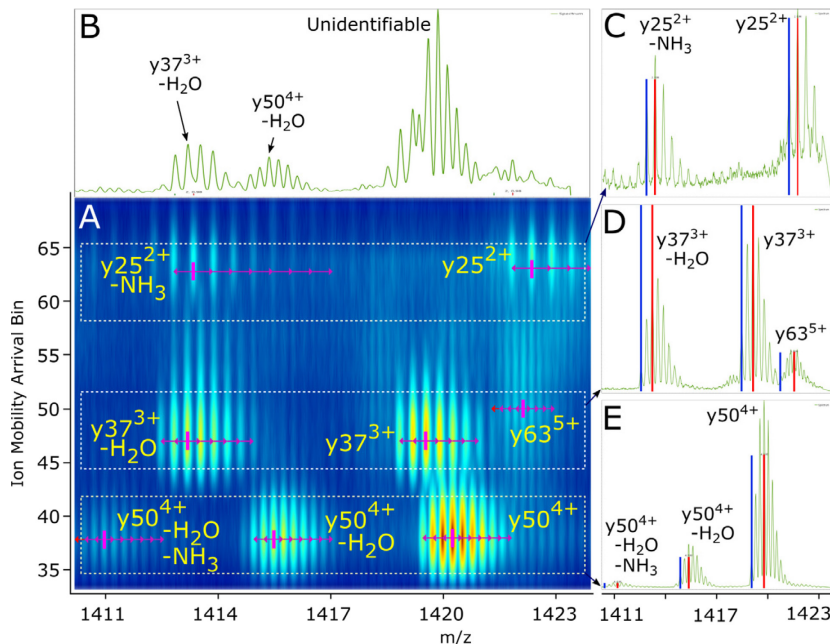


Figure 4-4 Isotopic clusters that cannot be resolved without IM separation. Two-dimensional representation of IM-MS data shows the presence of multiple isotopic clusters (A). The composite mass spectrum summed across all drift times, that is, the spectrum without IM separation, is shown at the top (B). Only the two most intense clusters at 1413 and 1415 can be identified, as the shape of the group at m/z 1419 is distorted due to interference from overlapping clusters. All 8 clusters can be identified utilizing IM separation with charge states ranging from 2+ to 5+, as shown in the mass spectra at 3 different drift times displayed on the right (C–E). Vertical red and green lines denote the most intense and monoisotopic peaks of a cluster correspondingly as reported by Grppr. Notice that fragment $y63^{5+}$ (far right ion in spectrum D) has drift time between that of 3+ and 2+ ions, following the observed trend for 5+ ions (see Figure 4).

dimensional IM-MS plot, however, clearly shows the presence of eight distinct clusters having significant relative abundance values (plotted on a square root intensity scale) due to the separation of these ions in the IM dimension (Figure 4-4 A).

Individual mass spectra can be plotted by averaging spectra from IM bins close to drift apexes of the various species (Figure 4-4 C–E), showing

accurate determination of the isotopic clusters by Grppr. By processing the data natively in two dimensions, IMTBX and Grppr are able to analyze more peaks, corresponding to more fragment ions of interest, in a given m/z range than the corresponding analysis performed with MS alone.

While the examples shown here were acquired using an IM device coupled to a time-of-flight (ToF) mass analyzer, the multi-dimensional peak detection and isotopic clustering performed by these tools can be applied to IM-MS instruments of any configuration and resolution. IM-MS

currently lacks a suitable open data format, so IMTBX is capable of reading Waters file format. Despite this, the algorithms described here can in principle be applied to any IM-MS input files, and Grppr in particular can be applied to any two-dimensional MS-based peak list.

To assess the benefits of IMTBX and Grppr for top-down IM-MS, we benchmarked our new software using top-down data acquired for a range of model proteins. Furthermore, these benchmarking results were compared to a semiautomated workflow that involves manual extraction of IM-MS regions corresponding to fragment charge state trends for analysis by existing MS data processing tools. IM and MS are only partially orthogonal, as IM drift time is a function of collision cross section (size and shape) and charge of an analyte, and mass and size tend to be generally correlated for analytes with similar overall shapes (e.g., protein fragments in

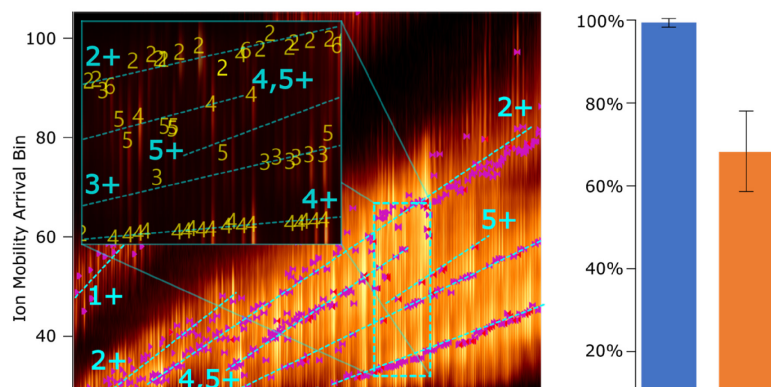


Figure 4-5 Left: Distribution of fragment ions from Ubiquitin in IM-MS space. Fragments are observed in “trend lines” due to the correlation between size and mass. Purple markers denote detected isotopic clusters. For Ubiquitin (6+ precursor) fragments, trends are observed for singly, doubly, triply, and quadruply charged fragments. However, several distinct trends can be observed for some charge state families, including 2+, 4+ and 5+, indicating the presence of multiple gas-phase conformations for these large fragments. The inset shows a zoom-in of a small IM-MS region, where yellow numbers indicate the charge state of an isotopic cluster detected at a particular location. Right: Charge assignment accuracy comparison. Charge state assignments were assessed by comparing the location of the fragment in IM-MS space and its assigned charge against the observed trend lines. For charge states 2+, 4+, and 5+, where multiple trends were observed, positioning of the signal within any of the observed trends were allowed. IMTBX + Grppr had an accuracy of 99% in assigning correct charge states vs 68% for the semiautomated data processing workflow

top-down analyses). This results in well-defined trend lines within IM-MS data, largely defined by ion charge states^{13,18} (Figure 4-5 A, cyan dashed lines). Interestingly, multiple trends were observed for fragment ions of large masses and charges (Figure 4-5, inset), likely due to the prevalence of multiple gas-phase conformations for longer amino acid sequences. Extracting each trend line separately allows for

MS analysis using existing tools while preserving some of the enhanced peak capacity provided by IM. For our semiautomated analysis, an open-source MS data viewer and peak processor, mMass,²⁶⁻²⁸ was used to accomplish signal-to-noise thresholding, peak picking, and isotopic clustering for each charge trend subset of the data. IMTBX and Grppr allow for the analysis of the complete IM-MS data set natively, by processing the raw data in two dimensions, enabling detection of off-trend peaks that are missed by the semimanual workflow, as well as greatly improving the quality of isotopic cluster detection for large fragments.

In the absence of a data set with known monoisotopic peaks for each feature, evaluating the accuracy of assignments from any computational or manual workflow is challenging. We have thus compared Grppr and the alternative workflow with respect to their relative accuracies in assigning charge states to isotopic clusters using the observed trend lines (Figure 4-5 A) as an objective comparison metric. For all isotopic clusters detected in eight analyses of Ubiquitin monomer and Avidin tetramer fragmentation data, using both IMTBX/Grppr and the semiautomated workflow, the charge state determined by the analysis was compared to the position (in m/z and drift time) of the fragment relative to the observed trend lines in order to evaluate its correctness (Figure 4-5). For higher molecular weight fragments, where multiple trends were observed, positioning of the signal within all observed trends was allowed. Grppr had an average charge state assignment accuracy of $99.0 \pm 0.6\%$ across approximately 50 data sets processed (24158 assignments) versus $68 \pm 12\%$ for the semimanual workflow (8 data sets, 4292 assignments). The natively 2D analysis of IMTBX and Grppr results in high and reproducible accuracy, allowing for unsupervised and fully automated analysis of top-down fragmentation data sets.

To demonstrate the capabilities of IMTBX and Grppr for top-down IM-MS analysis, a series of proteins and protein complex standards, ranging from 8.5 to 158 kDa, were fragmented to generate data sets for top-down analysis, collecting 300–600 IM-MS scans for each. Each data set was processed with IMTBX (requiring approximately 10–30 s for each analysis) and Grppr (requiring less than 5 s per analysis), and the resulting peak lists were exported into ProSight Lite for annotation to their respective protein sequences (Figure 4-6). IMTBX detected an average of just over 9000 peaks per data set (Figure 4-6 A), which were clustered into several hundred isotopic clusters. It is normal for many peaks not to get matched to isotopic clusters, as the peak list includes precursors and fragments at high m/z for which isotopic peaks cannot be easily resolved. Two sequence fragmentation maps are shown for reference (Figure 4-6 B): Ubiquitin, demonstrating that the annotated clusters result in excellent sequence coverage for a small protein, and Carbonic Anhydrase, showing coverage for a much larger protein, along with

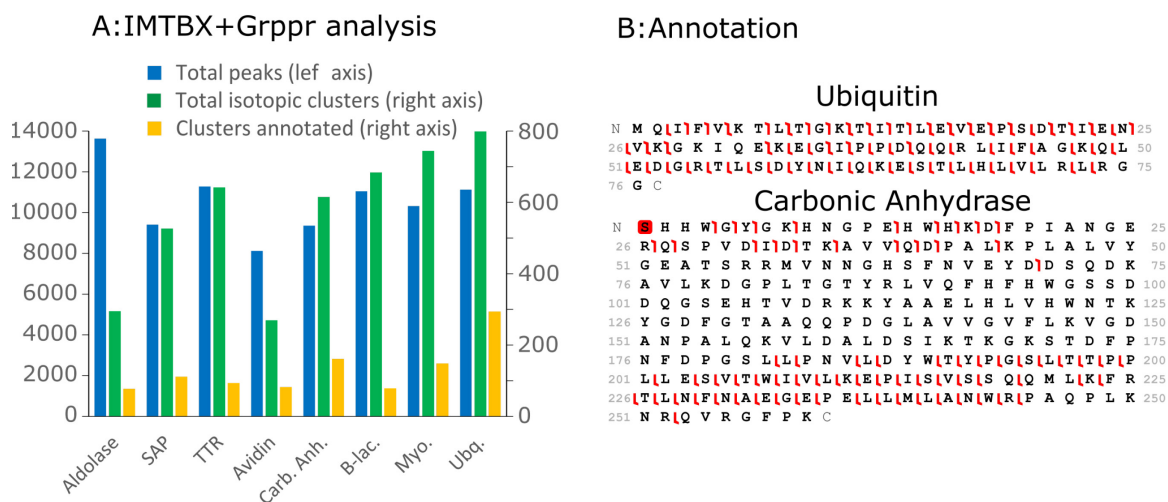


Figure 4-6 IMTBX and Grppr processing of 8 protein standards analyzed by top-down IM-MS. (A) The number of raw single peaks (2D features) detected by IMTBX for each protein (blue), the total number of isotopic clusters detected by Grppr (green), and the number of clusters matched to a fragment ion to annotate the protein sequence (yellow). Aldolase - tetramer (158 kDa), SAP = Serum Amyloid protein pentamer (125 kDa), Avidin - tetramer (64 kDa), TTR = transthyretin tetramer (56 kDa), Carb. Anh. = Carbonic Anhydrase monomer (29 kDa), B-lac. = β -lactoglobulin monomer (18 kDa), Myo. = myoglobin monomer (17 kDa), Ubq. = ubiquitin monomer (8.5 kDa). An average of 9040 features were detected for each raw data set. An average of 490 isotopic clusters were found and an average of 112 were annotated to the protein sequence. (B) Sequence maps showing sites of fragmentation for Ubiquitin (top) and Carbonic Anhydrase (bottom). The red box at the n-terminal serine of Carbonic Anhydrase indicates an acetylated serine, with the masses of all fragments containing that residue adjusted accordingly.

identification of the post- translationally acetylated serine at the n-terminus (red box). Processing even this relatively small data set manually would have required days to weeks of analysis time (Supporting Information, Table 2), given the need to manually assign thousands of peaks into hundreds of isotopic clusters to generate a peak list for annotation. IMTBX and Grpnr allow rapid (seconds to minutes), unsupervised, and accurate handling of this data to enable a wide range of top-down experiments on IM-MS instrumentation.

4.5 Conclusion

The potential benefits of IM separation for top-down proteomics have been demonstrated previously.^{14,18} Perhaps the largest obstacle preventing the realization of this potential has been a lack of software capable of automating even the initial step of translating two-dimensional IM-MS data into a list of monoisotopic peaks, which can be fed into an annotation program of choice (e.g., ProSight Lite). Fragmentation of intact proteins typically generates hundreds to thousands of fragment ions, each consisting of multiple isotopic peaks (as many as 20 for high molecular weight fragments), resulting in spectra that can contain many thousands of features. Manually annotating data of this complexity is extremely time-consuming, and the lack of automated processing tools severely limits the scope of experiments that can be performed on IM-MS platforms. With rapid and accurate determination of isotopic clusters, analysis of complex top-down fragmentation data collected on IM-MS instrumentation can be fully automated using IMTBX and Grpnr. The same software suite without modifications can also be applied to bottom-up proteomics LC-IM-MS data (not described in this manuscript). This enables experiments that require processing of hundreds or thousands of fragmentation spectra to be realistically performed on these instruments, including characterization of fragmentation mechanisms and typical top-down proteomics analyses with coupled intact protein separations.

Both Grppr and IMTBX can be downloaded from github repositories: <https://github.com/chhh/Grppr> and <https://github.com/chhh/IMTBX> (the latter includes a combined package with both tools). There is also a supporting Web site with examples and documentation at <https://chhh.github.io/IMTBX> (capitalization of IMTBX is important for the link to work properly).

4.6 References

- (1) Yates, J. R. The Revolution and Evolution of Shotgun Proteomics for Large-Scale Proteome Analysis. *J. Am. Chem. Soc.* **2013**, *135* (5), 1629–1640.
- (2) Mayne, J.; Ning, Z.; Zhang, X.; Starr, A. E.; Chen, R.; Deeke, S.; Chiang, C. K.; Xu, B.; Wen, M.; Cheng, K.; et al. Bottom-Up Proteomics (2013-2015): Keeping up in the Era of Systems Biology. *Anal. Chem.* **2016**, *88* (1), 95–121.
- (3) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M. C.; Yates, J. R. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chem. Rev.* **2013**, *113* (4), 2343–2394.
- (4) Beck, M.; Claassen, M.; Aebersold, R. Comprehensive Proteomics. *Curr. Opin. Biotechnol.* **2011**, *22* (1), 3–8.
- (5) Nesvizhskii, A. I.; Aebersold, R. Interpretation of Shotgun Proteomic Data. *Mol. Cell. Proteomics* **2005**, *4* (10), 1419–1440.
- (6) Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. Top Down Proteomics: Facts and Perspectives. *Biochem. Biophys. Res. Commun.* **2014**, *445* (4), 683–693.
- (7) Savaryn, J. P.; Catherman, A. D.; Thomas, P. M.; Abecassis, M. M.; Kelleher, N. L. The Emergence of Top-down Proteomics in Clinical Research. *Genome Med.* **2013**, *5* (6), 53.
- (8) Siuti, N.; Kelleher, N. L. Decoding Protein Modifications Using Top-down Mass Spectrometry. *Nat. Methods* **2007**, *4* (10), 817–821.
- (9) 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.
- (10) Huang, T. Y.; McLuckey, S. A. Top-down Protein Characterization Facilitated by Ion/Ion Reactions on a Quadrupole/Time of Flight Platform. *Proteomics* **2010**, *10* (20), 3577–3588.
- (11) Cannon, J. R.; Cammarata, M. B.; Robotham, S. A.; Cotham, V. C.; Shaw, J. B.; Fellers, R. T.; Early, B. P.; Thomas, P. M.; Kelleher, N. L.; Brodbelt, J. S. Ultraviolet Photodissociation for Characterization of Whole Proteins on a Chromatographic Time Scale. *Anal. Chem.* **2014**, *86* (4), 2185–2192.
- (12) 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.; et al. Mapping Intact Protein Isoforms in Discovery Mode Using Top-down Proteomics. *Nature* **2011**, *480* (7376), 254–258.
- (13) McLean, J. A.; Ruotolo, B. T.; Gillig, K. J.; Russell, D. H. Ion Mobility-Mass Spectrometry: A New Paradigm for Proteomics. *Int. J. Mass Spectrom.* **2005**, *240* (3 SPEC. ISS.), 301–315.

- (14) Baker, E. S.; Burnum-Johnson, K. E.; Ibrahim, Y. M.; Orton, D. J.; Monroe, M. E.; Kelly, R. T.; Moore, R. J.; Zhang, X.; Théberge, R.; Costello, C. E.; et al. Enhancing Bottom-up and Top-down Proteomic Measurements with Ion Mobility Separations. *Proteomics* **2015**, *15* (16), 2766–2776.
- (15) Haynes, S. E.; Polasky, D. A.; Dixit, S. M.; Majmudar, J. D.; Neeson, K.; Ruotolo, B. T.; Martin, B. R. Variable-Velocity Traveling-Wave Ion Mobility Separation Enhancing Peak Capacity for Data-Independent Acquisition Proteomics. *Anal. Chem.* **2017**, *89* (11), 5669–5672.
- (16) Meier, F.; Beck, S.; Grassl, N.; Lubeck, M.; Park, M. A.; Raether, O.; Mann, M. Parallel Accumulation-Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device. *J. Proteome Res.* **2015**, *14* (12), 5378–5387.
- (17) Halgand, F.; Habchi, J.; Cravello, L.; Martinho, M.; Guigliarelli, B.; Longhi, S. Dividing to Unveil Protein Microheterogeneities: Traveling Wave Ion Mobility Study. *Anal. Chem.* **2011**, *83* (19), 7306–7315.
- (18) Zinnel, N. F.; Pai, P. J.; Russell, D. H. Ion Mobility-Mass Spectrometry (IM-MS) for Top-down Proteomics: Increased Dynamic Range Affords Increased Sequence Coverage. *Anal. Chem.* **2012**, *84* (7), 3390–3397.
- (19) Lanucara, F.; Holman, S. W.; Gray, C. J.; Evers, C. E. The Power of Ion Mobility-Mass Spectrometry for Structural Characterization and the Study of Conformational Dynamics. *Nat. Chem.* **2014**, *6* (4), 281–294.
- (20) Paglia, G.; Astarita, G. Metabolomics and Lipidomics Using Traveling-Wave Ion Mobility Mass Spectrometry. *Nat. Protoc.* **2017**, *12* (4), 797–813.
- (21) Niu, S.; Rabuck, J. N.; Ruotolo, B. T. Ion Mobility-Mass Spectrometry of Intact Protein-Ligand Complexes for Pharmaceutical Drug Discovery and Development. *Curr. Opin. Chem. Biol.* **2013**, *17* (5), 809–817.
- (22) 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.; et al. Informed-Proteomics: Open-Source Software Package for Top-down Proteomics. *Nat. Methods* **2017**, *14* (9), 909–914.
- (23) 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.
- (24) Ruotolo, B. T.; Benesch, J. L. P.; Sandercock, A. M.; Hyung, S.-J.; Robinson, C. V. Ion Mobility-mass Spectrometry Analysis of Large Protein Complexes. *Nat. Protoc.* **2008**, *3* (7), 1139–1152.
- (25) Fellers, R. T.; Greer, J. B.; Early, B. P.; Yu, X.; Leduc, R. D.; Kelleher, N. L.; Thomas, P. M. ProSight Lite: Graphical Software to Analyze Top-down Mass Spectrometry Data. *Proteomics* **2015**, *15* (7), 1235–1238.
- (26) Niedermeyer, T. H. J.; Strohal, M. MMass as a Software Tool for the Annotation of Cyclic Peptide Tandem Mass Spectra. *PLoS One* **2012**, *7* (9), e44913.
- (27) Strohal, M.; Hassman, M.; Košata, B.; Kодиček, M. MMass Data Miner: An Open Source Alternative for Mass Spectrometric Data Analysis. *Rapid Commun. Mass Spectrom.* **2008**, *22* (6), 905–908.
- (28) Strohal, M.; Kavan, D.; Novák, P.; Volný, M.; Havlíček, V. MMass 3: A Cross-Platform Software Environment for Precise Analysis of Mass Spectrometric Data. *Anal. Chem.* **2010**, *82* (11), 4648–4651.

- (29) Senko, M. W.; Beu, S. C.; McLafferty, F. W. Automated Assignment of Charge States from Resolved Isotopic Peaks for Multiply Charged Ions. *J. Am. Soc. Mass Spectrom.* **1995**, *6* (1), 52–56.

Chapter 5 CIUSuite 2: Next-Generation Software for the Analysis of Gas-phase Protein Unfolding Data

Adapted with permission from: Daniel A. Polasky, Sugyan M. Dixit, Sarah M. Fantin, and Brandon T. Ruotolo. CIUSuite 2: Next-Generation Software for the Analysis of Gas-Phase Protein Unfolding Data. *Anal. Chem.* **2019**, *91* (4), 3147–3155.

5.1 Abstract

Ion mobility-mass spectrometry (IM-MS) has become an important addition to the structural biology toolbox, but separating closely related protein conformations remain challenging. Collision induced unfolding (CIU) has emerged as a valuable technique for distinguishing iso-crosssectional protein and protein complex ions through their distinct unfolding pathways in the gas phase. The speed and sensitivity of CIU analyses, coupled with their information-rich datasets, have resulted in the rapid growth of CIU for applications, ranging from the structural assessment of protein complexes to the characterization of biotherapeutics. This growth has occurred despite a lag in the capabilities of informatics tools available to process the complex datasets generated by CIU experiments, resulting in laborious manual analysis remaining commonplace. Here, we present CIUSuite 2, a software suite designed to enable robust, automated analysis of CIU data across the complete range of current CIU applications and to support the implementation of CIU as a true high-throughput technique. CIUSuite 2 uses statistical fitting and modeling methods to reliably quantify features of interest within CIU datasets, particularly in data with poor signal quality that cannot be interpreted with existing analysis tools. By reducing the signal-to-noise requirements for handling CIU data, we are able to demonstrate reductions in acquisition time of up to two orders of magnitude over current

workflows. CIUSuite 2 also provides the first automated system for classifying CIU fingerprints, enabling the next generation of ligand screening and structural analysis experiments to be accomplished in a high-throughput fashion.

5.2 Introduction

Native mass spectrometry (MS) has become a widespread technique in structural biology due to its ability to preserve noncovalently associated protein-protein and protein-ligand contacts and determine the stoichiometry and connectivity of these interactions.¹⁻³ The coupling of ion mobility to mass spectrometry (IM-MS) provides molecular shape information in addition to ion mass and charge, which has proven invaluable in interrogating complex biomolecular structures.^{4,5} Native IM-MS has seen dramatic growth in recent years, with applications in biotherapeutic characterization⁶ and drug discovery,⁷ joining more traditional analyses of protein complex structure and stoichiometry. A significant challenge in IM-MS is the separation of closely related protein conformations, as biologically relevant conformational variations often occur beyond the resolution limits of modern IM instrumentation. However, gas-phase activation provides a powerful approach to probe these subtle structural differences by assessing the resulting pattern of intermediate structural families produced from collisionally heating gas-phase protein ions. Early collision induced unfolding (CIU) experiments utilized this approach to differentiate charge-driven and disulfide bonding variations in small proteins.⁸ Subsequent CIU work uncovered different ligand-based stabilization mechanism in mutant TTR forms not detectable by IM-MS alone by introducing fingerprint plots that have now become a widespread analysis framework for such data.⁹

Since these early reports, CIU has seen rapid growth as such data have provided valuable approaches for a wide range of applications across the biological and pharmaceutical sciences.

The characterization of protein structure and dynamics, one of the original driving forces behind the development of CIU, remains a highly active area. Several groups have used comparative CIU of protein variants to determine the importance of specific residues, domains, and post-translational modifications on the structure and function of biomolecules.¹⁰⁻¹⁴ CIU has been used to rapidly probe the details of protein structure in response to solution and gas-phase stimuli.¹⁵⁻²⁰ CIU experiments have also been used to determine the orientation of ubiquitin non-covalent dimers through comparisons with various covalently linked dimers^{21,22} as well as assess the domain-specific unfolding of gas-phase serum albumins.²³ Many reports have described using CIU to assess ligand binding to a variety of protein targets^{9,24} in an effort to build information-rich small molecule screening platforms. CIU can be used as an analogue to stability shift assays commonly carried out in solution, as differences in gas-phase stabilities can offer similar information for unpurified samples at lower concentrations and potentially resolve intermediate transitions that may be missed in low resolution binding assays.²⁵⁻²⁷ Others have leveraged the detailed information provided by CIU to characterize the cooperativity,²⁸ binding location,²⁹ and the allostery of ligand binding events within proteins.^{30,31} CIU has also been developed into a versatile tool for the characterization of biotherapeutic antibodies.³² For example, CIU has proven to be highly sensitive to the presence of immunoglobulin isoforms,³³ differences in glycosylation and disulfide bonding patterns,³⁴ antibody-drug conjugation patterns,³⁵ and different binding epitopes.³⁶ Its relative sensitivity and speed when compared to other biophysical probes has led to the deployment of CIU in broad comparisons of biosimilar therapeutics.³⁷⁻³⁹ Finally, CIU shows promise in the context of membrane proteins,⁴⁰⁻⁴³ where such data has already proven critical in revealing some of the structural consequences of off-target drug binding.⁴⁴

Many of the CIU studies discussed above utilized manual analysis for all or part of their CIU data processing. A number of software packages are available to perform specific analytical tasks related to CIU data processing^{42,45-48}, and while they provide valuable capabilities, widespread adoption and use of software for processing of CIU data is still emerging. With CIUSuite 2, we combine additional capabilities for data processing not currently available in any software package, such as noise removal via Gaussian fitting and advanced classification of CIU data, with the integration of many capabilities into a single platform with a high degree of automation.

One of the most common outputs of CIU experiments is the accelerating voltage necessary to convert fifty percent of a compact protein form into an energetically adjacent extended state, sometimes termed a “CIU50” value. CIU50s have been used extensively to assess protein-ligand binding²⁵⁻²⁸ and the stability of domains within larger protein constructs.^{11,12,14,15,18,34} The Pulsar software package uses feature models to fit CIU50 values,⁴² but requires manual annotation of the features prior to analysis. Other packages, including the original CIUSuite⁴⁵ and Benthesisikyme,⁴⁷ annotate CIU features but lack an automated method to fit CIU50s. Another common output involves the root mean squared deviation (RMSD) analysis of CIU data, which is currently supported by several software packages^{45,48} and has proven a useful approach to detect quantitative differences in CIU fingerprints. RMSD analysis of CIU data is highly sensitive to chemical noise and to overall fluctuations in signal intensity because all differences between datasets are included in quantification. RMSD is effectively an ensemble measurement of all differences between fingerprints, which can obscure the contributions of individual changes in complex datasets. In practice, extensive signal averaging is often used to

overcome some of the noise-related limitations in RMSD analyses of CIU data, but this substantially limits the throughput of such experiments.

To address these challenges, we have developed CIUSuite 2, a software package that utilizes established fitting and statistical methods to enable the robust quantitation of CIU data across a broad range of CIU applications and analysis types, especially for enabling the analysis of low intensity CIU datasets. CIUSuite 2 extracts CIU50 values through a combination of improved feature detection and fits to logistic (sigmoid) curves that describe CIU transitions, enabling the fully automated and robust analysis of protein stabilities. By directly fitting features of interest, the signal-to-noise (S/N) ratios required for reliable analyses are reduced dramatically. These improved capabilities have, for example, enabled us to generate nearly identical output values from CIU data collected in 60-fold less time than previously published results. CIUSuite 2 significantly improves CIU fingerprint classification using linear discriminant analysis and support vector machines to enable next-generation high-throughput screening experiments. Finally, by modeling CIU data as mixtures of Gaussian functions, we are able to remove chemical noise and enable advanced feature detection within CIU datasets, producing robust analysis workflows for challenging CIU datasets. We have developed these algorithms with input from ongoing CIU projects that involve the assessment of biotherapeutic antibodies, membrane protein lipid binding events, protein-ligand interactions, and multiprotein complexes in an effort to provide a valuable set of quantitative tools for the broadest possible range of CIU applications. These capabilities are packaged into a user-friendly graphical interface that supports automated, high-throughput processing of large numbers of CIU datasets. CIUSuite 2 supports data collected on any IM-MS platform, and automated converters from vendor-specific data formats to the text file input needed for CIUSuite 2 are available. Finally,

CIUSuite 2 is a fully open-source software package and designed to be modular and readily extensible for researchers wishing to modify its capabilities for any CIU application.

5.3 Methods

5.3.1 Experimental Section

Translocator Protein (TSPO) was purified and expressed using established protocols.⁴⁹ All lipids purchased from Avanti Polar Lipids (Alabaster, AL). Ammonium acetate and Octyl β -D-glucopyranoside (OG) were purchased from Sigma Aldrich (St. Louis, MO). All CIU data was collected using a Synapt G2 HDMS IM-Q-ToF mass spectrometer (Waters, Milford, MA). Intact protein ions were generated using a direct infusion nESI source in positive mode. Glycosylated antibody data were collected as described previously.³⁴ Briefly, enzymes were used to cleave at specific glycan residues from an antibody standard (SILuLite SigmaMAb Universal Antibody Standard human (product number: MSQC4), Sigma Aldrich, St. Louis, MO) to leave glycans of known molecular weight attached to the antibody, which was then buffer exchanged (Micro Biospin6 spin columns (BioRad, Hercules, CA)) into 100 mM ammonium acetate and analyzed by IM-MS. TSPO was buffer and detergent exchanged simultaneously from 5 mM Tris, 150 mM NaCl, 0.20% DM, pH 8.0 to 40 mM OG, 200 mM ammonium acetate, pH 8.0 using 100kDa Amicon Ultra-0.5 Centrifugal Filter Units (MilliporeSigma, Burlington, MA). Lipid binding studies were performed following established protocols.⁵⁰ Instrument settings were tuned to completely remove the micelle prior to IM separator, including source temperature (40° C), helium cell gas flow (100 mL/min), and sampling cone (120 V). All CIU analyses were performed by increasing the trap collision voltage in a stepwise manner from 5 – 200 V (antibodies) or 50 – 150 V (membrane proteins) in 5 V increments.

5.3.2 Raw Data Extraction

Raw data was converted from Waters .raw format to a text-based format (“_raw.csv”, as described previously)⁴⁵ using TWIMExtract.⁵¹ Briefly, data from the m/z range corresponding to a single protein charge state was summed across the m/z and IM drift time dimensions to generate a series of collision voltage resolved IM datasets. For those analyses that compared different instrument acquisition times, data was summed across an indicated subset of the total acquisition time at each collision voltage. Extracted profiles are concatenated by TWIMExtract into a single _raw.csv file that serves as the input to CIUSuite 2. Raw data conversion is in development for the Agilent .d IM-MS data format. Similar converters for any additional formats (for example, an open IM-MS data format if one is developed) can be added as needed.

5.3.3 CIUSuite 2 Overview

CIUSuite 2 was developed in Python 3.5 utilizing the SciPy ecosystem^{52,53}, including NumPy,⁵⁴ Matplotlib,⁵⁵ and Scikit-learn.⁵⁶ The graphical user interface was developed using the Pygubu GUI builder (<https://github.com/alejandroautalan/pygubu>). Analysis of CIU data begins by importing any number of _raw.csv (text) format files. Each file (CIU dataset) undergoes smoothing and normalization. A 2D Savitzky-Golay filter of user-specified size is the default recommended setting; however, users may also select a 1D Savitzky-Golay filter of variable size or no smoothing as options. Additional pre-processing options are available in CIUSuite 2, including cropping, interpolating (resampling) data along one or both axes, and averaging multiple datasets. Once this pre-processing is complete, a “.ciu” file (a serialized file created with Python’s pickle module to store the CIU data and the results of any processing) is created (Figure 5-1 A). An RMSD comparison module, similar to the one provided in the original CIUSuite, is included in CIUSuite 2, offering the ability to compare individual files or groups of

files to generate pairwise RMSD values. All other analysis functions described below are unique to CIUSuite 2.

5.3.4 Stability Shift (“CIU50”) Analysis

To determine quantitative stability values from CIU datasets, a series of processing and modeling steps is performed (Figure 5-1 B). First, features in the dataset are detected by grouping observed drift time peaks that are present across multiple collision voltages (Figure 5-1

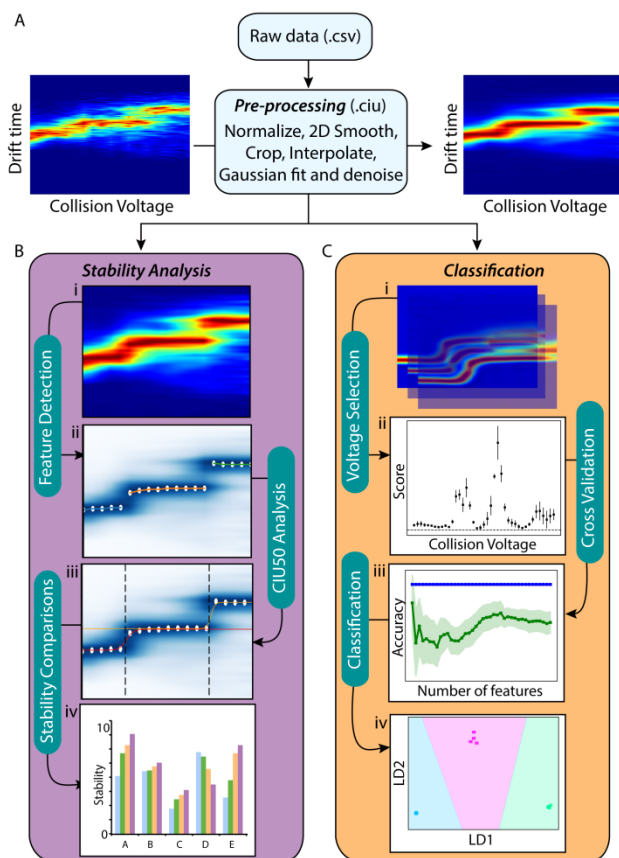


Figure 5-1 CIUSuite 2 overview. (A) Import of raw data from text (.csv) format and data preprocessing. (B) Stability analysis (“CIU50”) workflow: i) loading of preprocessed CIU datasets, ii) detection of features, iii) logistic function fitting to determine CIU50 value(s), and iv) comparison of CIU50 values among multiple datasets. (C) Classification workflow: i) Assignment of CIU training data into classes, ii) Selection of voltages most capable of differentiating classes, iii) Cross validation for model selection, and iv) Use of the selected model to perform LDA and build a SVM for classification. The capabilities illustrated here are not intended to provide a comprehensive walkthrough of CIUSuite 2 capabilities.

B, ii). The tolerances from median drift times, as well as the number of stable collision voltages required to define a feature, can be user-adjusted. Following feature detection, the transition region between features is modeled as a logistic (generalized sigmoid) function (Figure 5-1 B, iii). The logistic function parameters describe the lower and upper asymptotes (centroid drift times of the features before and after the transition), the steepness of the transition, and its midpoint voltage, which we term the “CIU50” value. Specifically, we define the CIU50 as the voltage at which 50% of a relatively compact state of the protein transitions to a more extended one, making it the effective midpoint between two

adjacent features on a CIU fingerprint. An arbitrary number of CIU datasets can be fit in a single CIUSuite 2 analysis, enabling high-throughput analyses and comparisons of stability values for all transitions detected across many CIU fingerprints (Figure 5-1 B, iv).

5.3.5 Classification

In an effort to further improve our ability to differentiate CIU fingerprint data, we developed a new classification workflow capable of sorting CIU datasets into groups using robust statistical methods. Briefly, a classification scheme is built based on training datasets from each group. First, our workflow implements a univariate feature selection (UFS) method based on an analysis of variance (ANOVA) F-test⁵⁷ to assess the significance of activation energies capable of differentiating CIU fingerprints (Figure IV-1). We iterate over all possible combinations of a training dataset in order to obtain the mean and standard deviation of $-\log_{10}(p \text{ value})$ which serves as the score for each collision voltage (Figure 5-1 C, ii). Second, we employ a “leave one out” cross-validation scheme⁵⁸ that examines the accuracy of classification, which is comprised of a linear discriminant analysis⁵⁹ (LDA) step followed by support vector machine⁶⁰ (SVM) classification of the linear discriminants, using subsets of CIU data from collision voltages in decreasing order relative to the score assigned during UFS analysis (Figure IV-2). This enables optimal selection of collision voltages to use for the resulting model and can be used to detect under- or over-fitting in the final model selected (Figure 5-1 C, iii). Finally, classification is performed on the model dataset with the optimized set of collision voltages, dividing the linear discriminant space into “decision regions” corresponding to the provided groups (Figure 5-1 C, iv). The resulting classification scheme can then be used to evaluate “unknown” CIU datasets (not used in training) to predict the class and probability for each unknown. We have also included a ‘manual’ classification mode, where users can select any

number of specific collision voltages to build a classification model. This is particularly helpful in scenarios where the accuracy observed in the cross-validation step is unacceptably low.

5.3.6 Gaussian Fitting and Automated De-noising

An optional Gaussian-fitting module enables modeling of the observed IM arrival time distribution at each collision voltage as a sum of Gaussian functions in order to provide a method for automated noise removal in complex datasets. An initial curve fitting⁵² is performed to generate high quality initial values prior to a primary analysis run by fitting a single peak to the arrival time distribution and adding peaks until the fit to the observed data (R^2) exceeds 0.99. Fitting can be performed in both “no denoising” and “denoising” modes to model a noise-free arrival time distribution or remove chemical noise, respectively. In each mode, the primary fitting run samples a range of Gaussian components (peaks) and scores each by its goodness of fit (R^2), peak width, and degree of overlap amongst its protein components. To perform a fit, Gaussian peak models for each component are assembled and fit to the arrival time data using LMFit.⁶¹ In denoising mode, the IM peak width of each Gaussian feature is used to distinguish between protein and non-protein components (Text IV-3). The highest scoring fit at each collision voltage is then taken for further evaluation, such as feature detection, CIU50 analysis, and classification workflows. The denoising workflow allows for the removal of chemical noise or other variability from CIU data prior to analysis, improving quantitative results (see below). Gaussian fit data can also be uploaded in a text format, enabling the use of other Gaussian fitting programs for CIU data as an alternative data input into CIUSuite 2. Data imported in this way can be analyzed using all of the tools described above.

5.4 Results

5.4.1 CIU Stability Shift Analysis

Assessing a shift in the gas-phase stability of a protein or protein complex in response to

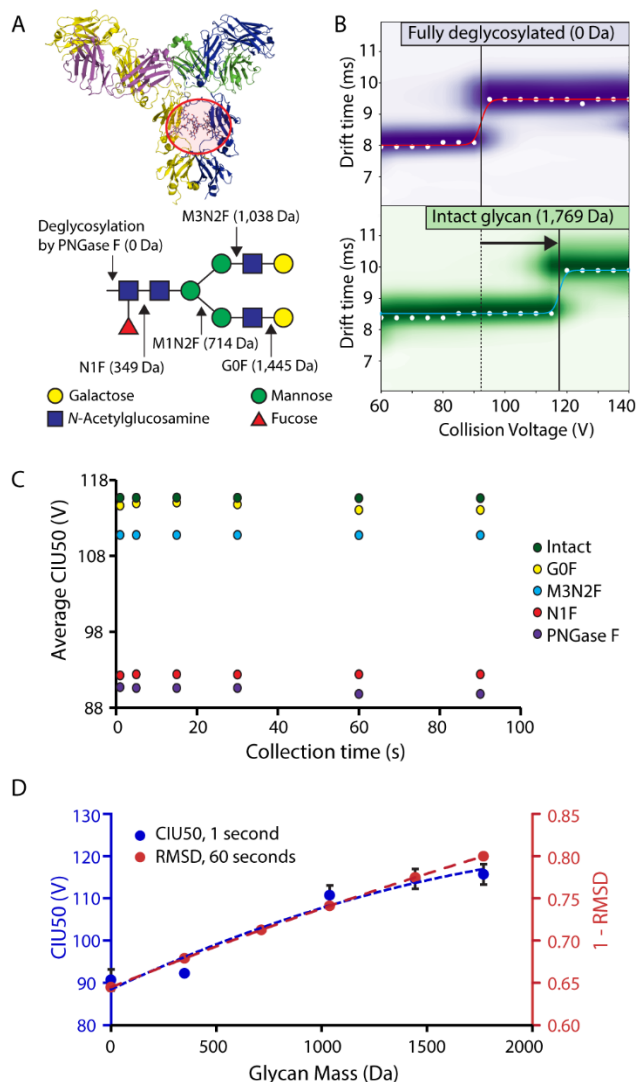


Figure 5-2 CIU50 analysis mAb glycoforms using CIUSuite 2. (A) Enzymatic reactions were used to produce glycans of varying size (cleavage site indicated by arrows), as described previously.³⁴ (B) CIU transition region shown for intact and fully deglycosylated mAbs. Larger glycans stabilize the transition between CIU features. (C) CIU50 values fit for each mAb glycoform plotted as a function of signal collection time for individual collision voltage values. Minimal variation is observed across the collection time axis, indicating that faster data acquisition is possible without affecting CIU50 values. (D) Plot of either CIU50 values from 1 s of data (blue) or previously published³⁴ RMSD using 60 s of data (red), against masses of mAb-attached glycans for each sample analyzed.

some stimulus is one of the most common applications of CIU, but current methods lack the ability to return robust quantitative values for complex or low-intensity datasets. Previous work from our laboratory has utilized RMSD analysis to detect a strong correlation between CIU fingerprint data and single-sugar changes in the glycans attached to intact monoclonal antibodies (mAbs), as generated through enzymatic reactions (Figure 5-2 A).³⁴ While the capability to detect such subtle differences in glycan structure within a 150 kDa protein is potentially enabling for mAb development, the length of time required to generate the data necessary to accurately quantify the above-described trends may make the adoption CIU technology for mAb assessments where rapidity is a requirement challenging. For example, in order to collect

our previously reported glycoform-resolved mAb CIU dataset,³⁴ 60 seconds of data was summed at each of 40 collision voltages probed (5 – 200 V) across 6 glycoforms in triplicate, for both intact mAbs and Fc fragments, resulting in a total acquisition time of approximately 24 hours. Because the quantitative comparison of each glycoform relied on a total RMSD analysis of each CIU fingerprint collected, minor fluctuations in signal intensity and chemical noise can dramatically influence the values extracted. In our previous report,³⁴ we elected to employ extended signal averaging in order to reduce the impact of such variations, at the cost of acquisition speed.

CIUSuite 2 directly models the transitions between features of CIU fingerprints, enabling the direct assessment of stability shifts without interference from other sources of variability or noise. Since our previously-reported RMSD differences largely arise from a shift in the stability of the second feature in the CIU fingerprints recorded,³⁴ we fit this transition using CIUSuite 2 in order to generate a CIU50 value for each glycoform we studied previously (Figure 5-2 B). Critically, because only the transition between the two features is used to extract correlations between sugar structures and CIU responses, the S/N ratios required for the precise assessments of such correlations are far lower than with our previous RMSD method. To demonstrate the speed improvement this affords, we extracted sub-sections of the original raw data corresponding to 1 second (a single scan collected by the instrument), 5 seconds, 15 seconds, 30 seconds, and the full 60 seconds of ion signal used in the original analysis. The S/N ratio observed in the 1-second data is approximately 60-fold lower than the full 60 seconds as expected, but the fit for the resulting logistic function that defines the CIU50 values extracted remains high quality, as shown in Figure 5-2 B. Plotting the observed CIU50 value for each glycoform as a function of signal collection time demonstrates that there is essentially no difference (less than 0.5 V)

between the results obtained using any amount of signal averaging probed here (Figure 5-2 C). As such, we can reconstruct our previous correlation between CIU response and glycan mass using only 1 second of our original data (Figure 5-2 D). The curves have similar slopes, indicating a similar glycoform sensitivity, though some variation as fundamentally different quantities different quantities are extracted from the data in the two approaches compared. Importantly, since our CIUSuite 2 method requires only 1 second of signal averaging at each collision voltage, we can reduce our total data collection time by 60-fold, resulting in a total of 24 minutes needed to quantify the same trend with equivalent precision as described in our previous report.³⁴

Furthermore, we estimate that the S/N ratio of the 1-second data is still far greater (on the order of 10^4) than is required for accurate fitting, indicating that greater reductions in acquisition time are possible with shorter (sub-second) instrument scans. By directly modeling the relevant parts of a fingerprint, CIUSuite 2 is thus able to dramatically improve the speed of CIU analyses, vastly enhancing the throughput of CIU analyses.

5.4.2 Classification of CIU data

The unfolding pathway of gas-phase protein ions has been observed to be sensitive to changes in protein structure that remain too subtle to detect using IM-MS alone.^{23,45} As such, CIU fingerprints have been deployed as means to classify protein structural states that result from changes in post-translational modifications, sequence variation, and ligand binding.^{31,45} For example, recent reports have demonstrated robust CIU classification schemes capable of differentiating allosteric and active site competitive kinase inhibitors,³¹ as well as binding event across two remote sites associated with transcriptional regulation.⁶² However, a lack statistical methods capable of sorting of unknown CIU data against known categories or of sorting between

more than two separate categories has proven to be an impediment in advancing such experiments beyond proof-of-concept demonstrations. In CIUSuite 2, we have implemented a classification workflow that uses rigorous statistical methods to generate classifying schema from known fingerprints that allows for facile evaluation of unknown samples against these schema for rapid sorting.

Data shown in Figure 5-3 displays an example implementation of our classification

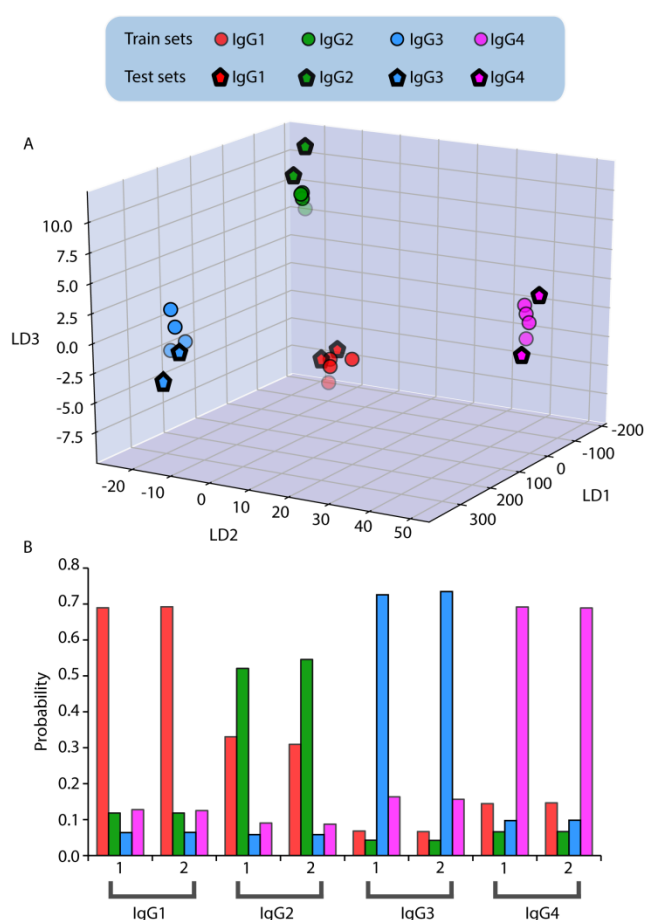


Figure 5-3 Classification of different IgG standards. (A) LDs for both training (filled circle) and test (filled pentagon) datasets corresponding to IgG1 (red), IgG2 (green), IgG3 (blue), and IgG4 (magenta) subclasses are well separated into clusters in three-dimensional space, defined by LD1, LD2, and LD3 axes. (B) Probabilities associated with each replicate (labeled as 1 and 2) in terms of categorizing the CIU data into different IgG groups. Each dataset is correctly assigned to its respective IgG subclass.

workflow, utilizing CIU data for immunoglobulin G (IgG) standards acquired across IgG1, IgG2, IgG3, and IgG4 subclasses. Each of our IgG CIU datasets contained four replicates, which we subdivided evenly into both training and test data in order to evaluate our approach. Of the forty collision voltages acquired for each CIU dataset, only a few were found to be highly differentiating between classes by UFS, with 85 V having the maximum score (Figure IV-3 E). This voltage is near the value required for the first CIU transition for each IgG subclass (Figure IV-3 A – D).

Cross-validation of UFS results revealed a classification accuracy 92.2% using only the CIU data isolated at 85 V, and decreases as

additional CIU data is added (Figure IV-3 F). Thus, our algorithm selected CIU data acquired at 85 V automatically in order to build a classification scheme. Figure 5-3 A shows the three-dimensional plot of linear discriminants (LDs) constructed using this data, which groups IgG CIU data into well-separated clusters. Furthermore, test data clustered correctly in all cases using this classification scheme (pentagons, Figure 5-3 A). We further used CIUSuite to compute the probability of each test dataset clustering into each IgG subclass (Figure 5-3 B), finding that each dataset was classified correctly with probability values ranging from 0.52 – 0.73 (Table IV-1). In general, our CIU classification workflow is generalizable, rapidly processing data in an automated fashion and accommodating any grouping scheme.

5.4.3 Classifying Noisy, Low Intensity CIU Data

While many existing tools are capable of analyzing high S/N data, low intensity CIU data containing significant amounts of chemical noise is exceptionally challenging to extract quantitative data from using current analysis paradigms. Membrane protein CIU data presents many of these challenges, as it frequently contains low-intensity protein ion signals, overlapped with intense chemical noise derived from detergents or other solubilizing agents, and is collected in a mode that thwarts typical tandem MS based CIU workflows.^{50,63} The feature detection and CIU50 functions of CIUSuite 2 were designed to extract reliable quantitative values from such datasets. Figure 5-4 illustrates the capabilities of CIUSuite 2 for such applications using data acquired for TSPO, a 36 kDa mitochondrial transmembrane protein dimer associated with benzodiazepine binding and cholesterol transport. Our feature detection workflow in CIUSuite 2 considers only the most intense IM peaks observed, and thus acts as an amplitude filter, removing such detergent and lipid based chemical noise signals from subsequent analysis steps. CIU50 values can then be fit to the observed transitions between features without interference

(Figure 5-4 A), so long as protein signals comprise the most intense peaks in the IM data analyzed. If this is not the case, automated noise removal can be employed prior to fitting (see below).

CIU has been used to characterize lipid binding to membrane proteins in order to assess stability shifts in the resulting complexes.⁴²⁻⁴⁴ Such data have been further used to distinguish between biologically relevant and nonspecifically associated lipids in membrane proteins.⁴¹

Counterintuitively, such assessments are often more straightforward to perform for larger

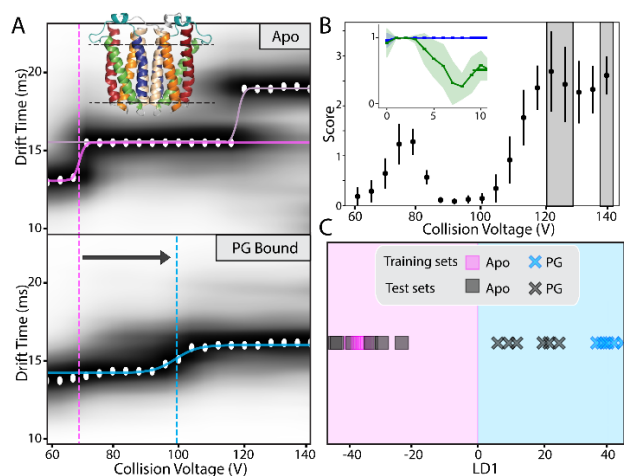


Figure 5-4 CIU50 analysis and classification TSPO-lipid complexes. (A) Feature fitting ignores low abundance chemical noise and CIU50 analysis reveals a stability shift associated with PG-bound TSPO. (B) Three voltages (120 V, 125 V, and 140 V) are used to construct a classification scheme from apo and PG-bound TSPO training sets, the inset shows a cross-validation plot indicating a high accuracy classification. (C) Additional test data sets are correctly assigned to apo (pink) or PG-bound (blue) TSPO.

proteins and complexes, as they appear at m/z and IM drift times that are frequently less contaminated by chemical noise. Since TSPO is a relatively small membrane protein complex, it is an exceptionally difficult target for CIU analysis. Preliminary screening of TSPO-lipid complexes revealed certain lipids, such as phosphatidylglycerol (PG), that significantly stabilize the protein so that CIU transitions appear distorted relative to apo protein data, making the extraction of

CIU50 values even more challenging.⁶⁴ We used the CIU50 module within CIUSuite 2 to fit these highly stabilized TSPO-PG transitions, allowing us to quantify stability imparted by lipid binding (Figure 5-4 A, lower panel). While this analysis provides high quality stability shift values, high-throughput CIU protocols require the rapid classification of ligands based on fingerprint data. To that end, we classified PG-bound and apo TSPO CIU signatures using

CIUSuite 2. By using three replicates each of apo and lipid bound data to build the classification scheme, we identified 120 V, 125 V, and 140 V as the most differentiating collision voltage values in our dataset (Figure 5-4 B). For validation of our classification scheme, four data sets that were not part of the training dataset were input as unknowns, and all were correctly classified (Figure 5-4 C). While it is clear that mass analysis alone could be used to identify PG bound and apo TSPO, these results illustrate a classification outcome that is exceptionally challenging to achieve using current CIU analysis tools.

5.4.4 Gaussian Fitting and Automated Denoising

The feature detection and CIU50 analyses presented in Figure 5-4 A enable the examination of CIU data containing a modest amount of chemical noise by employing a simple high-pass amplitude data filter and relying upon the presence of high-intensity protein signal. In many cases, however, protein signals are overlapped with chemical noise at intensity comparable to or exceeding that of the protein, rendering a high-pass filter approach ineffective. In such cases, quantified values extracted from CIU data exhibit reduced accuracy, reproducibility, and in some cases may be entirely unrecoverable. For example, the TSPO CIU data shown in Figure 5-5 A contains chemical noise that achieves greater intensity values than the detected protein ion signal at collision energies above 120 V, effectively preventing CIU50 analysis from recognizing the second protein CIU transition, which appears at 130 V. In order to surmount such signal processing difficulties, we have developed a Gaussian peak-fitting module within CIUSuite 2 that provides automated noise removal from CIU fingerprints. Other CIU analysis packages have performed Gaussian fits of CIU data for feature detection,⁴⁷ however, this approach has not been previously applied to removal of noise components from CIU data. IM peaks corresponding to protein ions exhibit a range of peak widths produced primarily by ion diffusion and

conformational heterogeneity effects^{65–67} Thus, the likely range of IM peak widths for protein signals can be utilized as a noise filter for CIU analysis, where Gaussian fits corresponding to

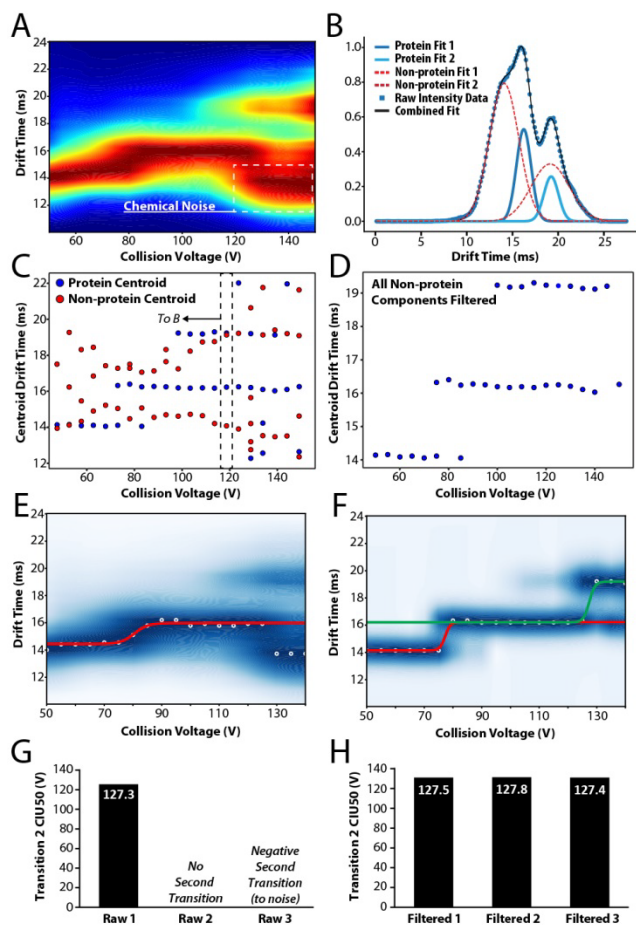


Figure 5-5 Automated de-noising of membrane protein CIU data using Gaussian fitting. (A) TSPO CIU fingerprint with chemical noise preventing analysis of the final CIU transition observed (collision voltage >120 V). (B) Example mixed-model Gaussian fit produced from selected data from A. Protein components (blue) exhibit widths within pre-defined tolerances while non-protein components (red) are broader. (C) Plot protein (blue) and non-protein (red) Gaussian fit centroids from fingerprint shown in (A). Horizontal arrays of blue dots indicate stable protein conformations (features). The region corresponding data displayed in panel B is marked. (D) Removal of non-protein components after denoising results in a centroid plot absent of the identified noise features. (E) CIU50 fitting result prior to Gaussian denoising. The second protein CIU transition is missed due to chemical noise. (F) CIU50 fitting of the same dataset as shown in E after denoising, illustrating robust recovery of both CIU transitions. (G) Histogram of CIU50 values extracted from 3 replicate datasets prior to denoising, illustrating inconsistent results. (H) Histogram of CIU50 values extracted from the same dataset as shown in panel G following denoising are highly reproducible.

features that exceed such a width tolerance are identified as noise and subtracted from the final fingerprint. Fitting is performed using two different types of components: protein components (e.g. the blue traces in Figure 5-5 B), which are Gaussian functions constrained to a narrow distribution of peak widths, and non-protein components (e.g. the red dashed lines in Figure 5-5 B), which are allowed to have any width substantially larger than the upper width limit allowed for protein components. This approach intrinsically filters resulting fit results, allowing noise components to be removed after the fitting is complete.

Figure 5-5 C shows the results of this automated fitting approach for phosphatidylcholine bound-TSPO CIU data displayed in Figure 5-5 A by plotting the peak centers determined for each protein component in blue and those determined for

each non-protein component in red. Three clear features, or sets of protein peak centers that appear at consistent arrival times, can be observed in the protein data, resembling a typical TSPO CIU fingerprint when the broader features are subtracted from the dataset (Figure 5-5 D). This denoised data can then be analyzed with any of the workflows available in CIUSuite 2, including the CIU50 determination and classification modules. CIU50 analysis of the denoised dataset allows for the recovery of the second TSPO CIU transition following the removal abundant chemical noise (Figure 5-5 F). Removing chemical noise can greatly improve the accuracy and reproducibility of CIU analyses, as in the replicate CIU50 analyses shown in Figure 5-5 G, H. Analysis of the raw TSPO CIU data results in adequate fits the second CIU transition in only one out of three datasets, missing the transition in the second replicate (pictured in Figure 5-5 E) and fitting a “negative” transition (an apparent shift from longer to shorter IM drift times) to the chemical noise observed, (Figure 5-5 G). In contrast, all three replicates can be reproducibly fit following Gaussian denoising, producing CIU50 values that vary by less than 0.4 V across all three replicates (Figure 5-5 H).

The automated removal of chemical noise or other interfering signals from CIU data thus enables the analysis CIU datasets that would be challenging to accomplish using current tools, while requiring minimal user intervention. Protein and noise components can be distinguished based on their differential peak widths, allowing such noise to be directly subtracted from CIU fingerprints. Combined with CIU50 analysis and classification workflows, Gaussian denoising represents a substantial enhancement to the CIU signal processing toolbox.

5.5 Conclusions

CIU experiments generate complex datasets containing rich protein structure information. CIUSuite 2 provides a framework, built upon established statistical methods, for extracting key

information from CIU data in a robust, automated fashion. Furthermore, CIUSuite 2 is designed to support a broad range of existing CIU applications, including the analysis of noise-contaminated membrane proteins and high-throughput screening. Our improvements to the automation CIU signal processing and acquisition speed point towards the next generation of CIU workflows, where full CIU fingerprints are collected in seconds and on-the-fly data reduction enables the rapid generation of classification schema. With support from autosampling devices, the potential exists to generate and analyze orders of magnitude more CIU data per unit time than currently possible. The ability to generate such large datasets would likely aid in answering fundamental questions regarding the relationship between solution and gas-phase stabilities and structures, while simultaneously providing a platform for rapid structural assessments of biotherapeutics and pharmaceutically relevant protein complexes. CIUSuite 2 is available for download at <https://sites.lsa.umich.edu/ruotolo/software/ciusuite-2/>, and its source code can be found at <https://github.com/RuotoloLab/CIUSuite2>.

5.6 Acknowledgements

TSPO protein was expressed and purified by the Ferguson-Miller group at Michigan State University. The authors thank M. W. Haskell for helpful discussions regarding Gaussian fitting and modeling, and the members of the Ruotolo lab for extensive feedback and beta testing. CIUSuite 2 development is supported by the National Science Foundation Division of Chemistry under Grants 1808541 and 1253384 (with co-funding from the Division of Molecular and Cellular Biosciences). Additional support for this project was provided by the Agilent Technologies Thought Leader Award and University Relations grant programs.

5.7 References

- (1) Leney, A. C.; Heck, A. J. R. Native Mass Spectrometry: What Is in the Name? *J. Am. Soc.*

- Mass Spectrom.* **2017**, *28* (1), 5–13.
- (2) Lössl, P.; van de Waterbeemd, M.; Heck, A. J. The Diverse and Expanding Role of Mass Spectrometry in Structural and Molecular Biology. *EMBO J.* **2016**, *35* (24), 2634–2657.
 - (3) Mehmood, S.; Allison, T. M.; Robinson, C. V. Mass Spectrometry of Protein Complexes: From Origins to Applications. *Annu. Rev. Phys. Chem.* **2015**, *66* (1), 453–474.
 - (4) Zhong, Y.; Hyung, S.-J.; Ruotolo, B. T. Ion Mobility–mass Spectrometry for Structural Proteomics. *Expert Rev. Proteomics* **2012**, *9* (1), 47–58.
 - (5) Lanucara, F.; Holman, S. W.; Gray, C. J.; Evers, C. E. The Power of Ion Mobility-Mass Spectrometry for Structural Characterization and the Study of Conformational Dynamics. *Nat. Chem.* **2014**, *6* (4), 281–294.
 - (6) Terral, G.; Beck, A.; Cianféroni, S. Insights from Native Mass Spectrometry and Ion Mobility-Mass Spectrometry for Antibody and Antibody-Based Product Characterization. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2016**, *1032*, 79–90.
 - (7) Bleiholder, C.; Bowers, M. T. The Solution Assembly of Biological Molecules Using Ion Mobility Methods: From Amino Acids to Amyloid β -Protein. *Annu. Rev. Anal. Chem.* **2017**, *10* (1), 365–386.
 - (8) Shelimov, K. B.; Clemmer, D. E.; Hudgins, R. R.; Jarrold, M. F. Protein Structure in Vacuo: Gas-Phase Conformations of BPTI and Cytochrome C. *J. Am. Chem. Soc.* **1997**, *119* (9), 2240–2248.
 - (9) Hyung, S. J.; Robinson, C. V.; Ruotolo, B. T. Gas-Phase Unfolding and Disassembly Reveals Stability Differences in Ligand-Bound Multiprotein Complexes. *Chem. Biol.* **2009**, *16* (4), 382–390.
 - (10) Zhang, H.; Liu, H.; Lu, Y.; Wolf, N. R.; Gross, M. L.; Blankenship, R. E. Native Mass Spectrometry and Ion Mobility Characterize the Orange Carotenoid Protein Functional Domains. *Biochim. Biophys. Acta - Bioenerg.* **2016**, *1857* (6), 734–739.
 - (11) Zhao, Y.; Singh, A.; Xu, Y.; Zong, C.; Zhang, F.; Boons, G. J.; Liu, J.; Linhardt, R. J.; Woods, R. J.; Amster, I. J. Gas-Phase Analysis of the Complex of Fibroblast GrowthFactor 1 with Heparan Sulfate: A Traveling Wave Ion Mobility Spectrometry (TWIMS) and Molecular Modeling Study. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (1), 96–109.
 - (12) Chorev, D. S.; Volberg, T.; Livne, A.; Eisenstein, M.; Martins, B.; Kam, Z.; Jockusch, B. M.; Medalia, O.; Sharon, M.; Geiger, B. Conformational States during Vinculin Unlocking Differentially Regulate Focal Adhesion Properties. *Sci. Rep.* **2018**, *8* (1), 2693.
 - (13) Byrne, D. P.; Vonderach, M.; Ferries, S.; Brownridge, P. J.; Evers, C. E.; Evers, P. A. CAMP-Dependent Protein Kinase (PKA) Complexes Probed by Complementary Differential Scanning Fluorimetry and Ion Mobility-Mass Spectrometry. *Biochem. J.* **2016**, *473* (19), 3159–3175.
 - (14) Jovcevski, B.; Kelly, M. A.; Aquilina, J. A.; Benesch, J. L. P.; Ecroyd, H. Evaluating the Effect of Phosphorylation on the Structure and Dynamics of Hsp27 Dimers by Means of Ion Mobility Mass Spectrometry. *Anal. Chem.* **2017**, *89* (24), 13275–13282.
 - (15) Chan, D. S. H.; Kavanagh, M. E.; McLean, K. J.; Munro, A. W.; Matak-Vinković, D.; Coyne, A. G.; Abell, C. Effect of DMSO on Protein Structure and Interactions Assessed by Collision-Induced Dissociation and Unfolding. *Anal. Chem.* **2017**, *89* (18), 9976–9983.
 - (16) Laszlo, K. J.; Munger, E. B.; Bush, M. F. Folding of Protein Ions in the Gas Phase after Cation-to-Anion Proton-Transfer Reactions. *J. Am. Chem. Soc.* **2016**, *138* (30), 9581–9588.

- (17) Beynon, R. J.; Armstrong, S. D.; Claydon, A. J.; Davidson, A. J.; Evers, C. E.; Langridge, J. I.; Gómez-Baena, G.; Harman, V. M.; Hurst, J. L.; Lee, V.; et al. Mass Spectrometry for Structural Analysis and Quantification of the Major Urinary Proteins of the House Mouse. *Int. J. Mass Spectrom.* **2015**, *391*, 146–156.
- (18) Han, L.; Hyung, S. J.; Mayers, J. J. S.; Ruotolo, B. T. Bound Anions Differentially Stabilize Multiprotein Complexes in the Absence of Bulk Solvent. *J. Am. Chem. Soc.* **2011**, *133* (29), 11358–11367.
- (19) Zhong, Y.; Han, L.; Ruotolo, B. T. Collisional and Coulombic Unfolding of Gas-Phase Proteins: High Correlation to Their Domain Structures in Solution. *Angew. Chemie - Int. Ed.* **2014**, *53* (35), 9209–9212.
- (20) Samulak, B. M.; Niu, S.; Andrews, P. C.; Ruotolo, B. T. Ion Mobility-Mass Spectrometry Analysis of Cross-Linked Intact Multiprotein Complexes: Enhanced Gas-Phase Stabilities and Altered Dissociation Pathways. *Anal. Chem.* **2016**, *88* (10), 5290–5298.
- (21) Wagner, N. D.; Russell, D. H. Defining Noncovalent Ubiquitin Homodimer Interfacial Interactions through Comparisons with Covalently Linked Diubiquitin. *J. Am. Chem. Soc.* **2016**, *138* (51), 16588–16591.
- (22) Wagner, N. D.; Clemmer, D. E.; Russell, D. H. ESI-IM-MS and Collision-Induced Unfolding That Provide Insight into the Linkage-Dependent Interfacial Interactions of Covalently Linked Diubiquitin. *Anal. Chem.* **2017**, *89* (18), 10094–10103.
- (23) Eschweiler, J. D.; Martini, R. M.; Ruotolo, B. T. Chemical Probes and Engineered Constructs Reveal a Detailed Unfolding Mechanism for a Solvent-Free Multidomain Protein. *J. Am. Chem. Soc.* **2017**, *139* (1), 534–540.
- (24) Hopper, J. T. S.; Oldham, N. J. Collision Induced Unfolding of Protein Ions in the Gas Phase Studied by Ion Mobility-Mass Spectrometry: The Effect of Ligand Binding on Conformational Stability. *J. Am. Soc. Mass Spectrom.* **2009**, *20* (10), 1851–1858.
- (25) Zhao, Y.; Singh, A.; Li, L.; Linhardt, R. J.; Xu, Y.; Liu, J.; Woods, R. J.; Amster, I. J. Investigating Changes in the Gas-Phase Conformation of Antithrombin III upon Binding of Arixtra Using Traveling Wave Ion Mobility Spectrometry (TWIMS). *Analyst* **2015**, *140* (20), 6980–6989.
- (26) Nyon, M. P.; Prentice, T.; Day, J.; Kirkpatrick, J.; Sivalingam, G. N.; Levy, G.; Haq, I.; Irving, J. A.; Lomas, D. A.; Christodoulou, J.; et al. An Integrative Approach Combining Ion Mobility Mass Spectrometry, X-Ray Crystallography, and Nuclear Magnetic Resonance Spectroscopy to Study the Conformational Dynamics of A1-Antitrypsin upon Ligand Binding. *Protein Sci.* **2015**, *24* (8), 1301–1312.
- (27) Zhao, B.; Zhuang, X.; Pi, Z.; Liu, S.; Liu, Z.; Song, F. Determining the Effect of Catechins on SOD1 Conformation and Aggregation by Ion Mobility Mass Spectrometry Combined with Optical Spectroscopy. *J. Am. Soc. Mass Spectrom.* **2018**, *29* (4), 734–741.
- (28) Niu, S.; Ruotolo, B. T. Collisional Unfolding of Multiprotein Complexes Reveals Cooperative Stabilization upon Ligand Binding. *Protein Sci.* **2015**, *24* (8), 1272–1281.
- (29) Rabuck, J. N.; Hyung, S. J.; Ko, K. S.; Fox, C. C.; Soellner, M. B.; Ruotolo, B. T. Activation State-Selective Kinase Inhibitor Assay Based on Ion Mobility-Mass Spectrometry. *Anal. Chem.* **2013**, *85* (15), 6995–7002.
- (30) Beveridge, R.; Migas, L. G.; Payne, K. A. P.; Scrutton, N. S.; Leys, D.; Barran, P. E. Mass Spectrometry Locates Local and Allosteric Conformational Changes That Occur on Cofactor Binding. *Nat. Commun.* **2016**, *7*, 12163.
- (31) Rabuck-Gibbons, J. N.; Keating, J. E.; Ruotolo, B. T. Collision Induced Unfolding and

- Dissociation Differentiates ATP-Competitive from Allosteric Protein Tyrosine Kinase Inhibitors. *Int. J. Mass Spectrom.* **2018**, *427*, 151–156.
- (32) Tian, Y.; Ruotolo, B. T. The Growing Role of Structural Mass Spectrometry in the Discovery and Development of Therapeutic Antibodies. *Analyst* **2018**, *143* (11), 2459–2468.
- (33) Tian, Y.; Han, L.; Buckner, A. C.; Ruotolo, B. T. Collision Induced Unfolding of Intact Antibodies: Rapid Characterization of Disulfide Bonding Patterns, Glycosylation, and Structures. *Anal. Chem.* **2015**, *87* (22), 11509–11515.
- (34) Tian, Y.; Ruotolo, B. T. Collision Induced Unfolding Detects Subtle Differences in Intact Antibody Glycoforms and Associated Fragments. *Int. J. Mass Spectrom.* **2018**, *425*, 1–9.
- (35) Botzanowski, T.; Erb, S.; Hernandez-Alba, O.; Ehkirch, A.; Colas, O.; Wagner-Rousset, E.; Rabuka, D.; Beck, A.; Drake, P. M.; Cianfèrani, S. Insights from Native Mass Spectrometry Approaches for Top- and Middle- Level Characterization of Site-Specific Antibody-Drug Conjugates. *MAbs* **2017**, *9* (5), 801–811.
- (36) Huang, Y.; Salinas, N. D.; Chen, E.; Tolia, N. H.; Gross, M. L. Native Mass Spectrometry, Ion Mobility, and Collision-Induced Unfolding Categorize Malaria Antigen/Antibody Binding. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (11), 2515–2518.
- (37) Campuzano, I. D. G.; Larriba, C.; Bagal, D.; Schnier, P. D. Ion Mobility and Mass Spectrometry Measurements of the Humanized IgGk NIST Monoclonal Antibody. *ACS Symp. Ser.* **2015**, *1202*, 75–112.
- (38) Pisupati, K.; Tian, Y.; Okbazghi, S.; Benet, A.; Ackermann, R.; Ford, M.; Saveliev, S.; Hosfield, C. M.; Urh, M.; Carlson, E.; et al. A Multidimensional Analytical Comparison of Remicade and the Biosimilar Remsima. *Anal. Chem.* **2017**, *89* (9), 4838–4846.
- (39) Ferguson, C. N.; Gucinski-Ruth, A. C. Evaluation of Ion Mobility-Mass Spectrometry for Comparative Analysis of Monoclonal Antibodies. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (5), 822–833.
- (40) Reading, E.; Liko, I.; Allison, T. M.; Benesch, J. L. P.; Laganowsky, A.; Robinson, C. V. The Role of the Detergent Micelle in Preserving the Structure of Membrane Proteins in the Gas Phase. *Angew. Chemie - Int. Ed.* **2015**, *54* (15), 4577–4581.
- (41) Laganowsky, A.; Reading, E.; Allison, T. M.; Ulmschneider, M. B.; Degiacomi, M. T.; Baldwin, A. J.; Robinson, C. V. Membrane Proteins Bind Lipids Selectively to Modulate Their Structure and Function. *Nature* **2014**, *510* (7503), 172–175.
- (42) Allison, T. M.; Reading, E.; Liko, I.; Baldwin, A. J.; Laganowsky, A.; Robinson, C. V. Quantifying the Stabilizing Effects of Protein-Ligand Interactions in the Gas Phase. *Nat. Commun.* **2015**, *6*, 8551.
- (43) Liu, Y.; Cong, X.; Liu, W.; Laganowsky, A. Characterization of Membrane Protein–Lipid Interactions by Mass Spectrometry Ion Mobility Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (4), 579–586.
- (44) Mehmood, S.; Marcoux, J.; Gault, J.; Quigley, A.; Michaelis, S.; Young, S. G.; Carpenter, E. P.; Robinson, C. V. Mass Spectrometry Captures Off-Target Drug Binding and Provides Mechanistic Insights into the Human Metalloprotease ZMPSTE24. *Nat. Chem.* **2016**, *8* (12), 1152–1158.
- (45) Eschweiler, J. D.; Rabuck-Gibbons, J. N.; Tian, Y.; Ruotolo, B. T. CIUSuite: A Quantitative Analysis Package for Collision Induced Unfolding Measurements of Gas-Phase Protein Ions. *Anal. Chem.* **2015**, *87* (22), 11516–11522.
- (46) Sivalingam, G. N.; Yan, J.; Sahota, H.; Thalassinou, K. Amphitrite: A Program for

- Processing Travelling Wave Ion Mobility Mass Spectrometry Data. *Int. J. Mass Spectrom.* **2013**, 345–347, 54–62.
- (47) Sivalingam, G. N.; Cryar, A.; Williams, M. A.; Gooptu, B.; Thalassinos, K. Deconvolution of Ion Mobility Mass Spectrometry Arrival Time Distributions Using a Genetic Algorithm Approach: Application to A1-Antitrypsin Peptide Binding. *Int. J. Mass Spectrom.* **2018**, 426, 29–37.
- (48) Migas, L. G.; France, A. P.; Bellina, B.; Barran, P. E. ORIGAMI: A Software Suite for Activated Ion Mobility Mass Spectrometry (AIM-MS) Applied to Multimeric Protein Assemblies. *Int. J. Mass Spectrom.* **2018**, 427, 20–28.
- (49) Li, F.; Xia, Y.; Meiler, J.; Ferguson-Miller, S. Characterization and Modeling of the Oligomeric State and Ligand Binding Behavior of Purified Translocator Protein 18 KDa from *Rhodobacter Sphaeroides*. *Biochemistry* **2013**, 52 (34), 5884–5899.
- (50) Laganowsky, A.; Reading, E.; Hopper, J. T. S.; Robinson, C. V. Mass Spectrometry of Intact Membrane Protein Complexes. *Nat. Protoc.* **2013**, 8 (4), 639–651.
- (51) Haynes, S. E.; Polasky, D. A.; Dixit, S. M.; Majmudar, J. D.; Neeson, K.; Ruotolo, B. T.; Martin, B. R. Variable-Velocity Traveling-Wave Ion Mobility Separation Enhancing Peak Capacity for Data-Independent Acquisition Proteomics. *Anal. Chem.* **2017**, 89 (11), 5669–5672.
- (52) Millman, K. J.; Aivazis, M. Python for Scientists and Engineers. *Comput. Sci. Eng.* **2011**, 13 (2), 9–12.
- (53) Oliphant, T. E. Python for Scientific Computing. *Comput. Sci. Eng.* **2007**, 9 (3), 10–20.
- (54) Dubois, P. F.; Hinsen, K.; Hugunin, J. Numerical Python. *Comput. Phys.* **1996**, 10 (3), 262.
- (55) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, 9 (3), 99–104.
- (56) Pedregosa, F.; Varoquax, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O. Scikit-Learn: Machine Learning in Python. *J Mach Learn Res* **2011**, 12 (Oct), 2825–2830.
- (57) Dowdy, S. M.; Wearden, S.; Chilko, D. M. *Statistics for Research*, 3rd ed.; Wiley-Interscience: Hoboken, N.J., 2004.
- (58) Arlot, S.; Celisse, A. A Survey of Cross-Validation Procedures for Model Selection. *Stat. Surv.* **2009**, 4 (0), 40–79.
- (59) Xanthopoulos, P.; Pardalos, P. M.; Trafalis, T. B. *Linear Discriminant Analysis*; Springer, New York, NY, 2013; pp 27–33.
- (60) Mammone, A.; Turchi, M.; Cristianini, N. Support Vector Machines. In *Advanced Review*; Springer, New York, NY, 2009; Vol. 1, pp 283–289.
- (61) Newville, M.; Ingargiola, A.; Stensitzki, T.; Allen, D. B. LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python. *Zenodo* **2014**.
- (62) Rabuck-Gibbons, J. N.; Lodge, J.; Mapp, A.; Ruotolo, B. T. In Press. *J Am Soc Mass Spectrom.*
- (63) Campuzano, I. D. G.; Li, H.; Bagal, D.; Lippens, J. L.; Svitel, J.; Kurzeja, R. J. M.; Xu, H.; Schnier, P. D.; Loo, J. A. Native MS Analysis of Bacteriorhodopsin and an Empty Nanodisc by Orthogonal Acceleration Time-of-Flight, Orbitrap and Ion Cyclotron Resonance. *Anal. Chem.* **2016**, 88 (24), 12427–12436.
- (64) Fantin, S. M.; Parson, K. F.; Niu, S.; Liu, J.; Ferguson-Miller, S. M.; Ruotolo, B. T. CIU Classifies Ligand Binding Behavior of Integral Membrane Translocator Protein TSPO. *Press.*

- (65) Allen, S. J.; Giles, K.; Gilbert, T.; Bush, M. F. Ion Mobility Mass Spectrometry of Peptide, Protein, and Protein Complex Ions Using a Radio-Frequency Confining Drift Cell. *Analyst* **2016**, *141* (3), 884–891.
- (66) Zhou, M.; Politis, A.; Davies, R. B.; Liko, I.; Wu, K.-J.; Stewart, A. G.; Stock, D.; Robinson, C. V. Ion Mobility–mass Spectrometry of a Rotary ATPase Reveals ATP-Induced Reduction in Conformational Flexibility. *Nat. Chem.* **2014**, *6* (3), 208–215.
- (67) May, J. C.; McLean, J. A. Ion Mobility-Mass Spectrometry: Time-Dispersive Instrumentation. *Anal. Chem.* **2015**, *87* (3), 1422–1436.

Chapter 6 A Modified Drift Tube Ion Mobility-Mass Spectrometer for Charge Multiplexed Collision Induced Unfolding

Daniel D. Vallejo‡, Daniel A. Polasky‡, Ruwan T. Kurulugama‡, Joseph D. Eschweiler, John C. Fjeldsted, and Brandon T. Ruotolo. *Anal. Chem.* **2019**, *submitted*.
(‡ contributed equally)

6.1 Abstract

Collision induced unfolding (CIU) of protein ions and their non-covalent complexes offers relatively rapid access to a rich portfolio of biophysical information, without the need to tag or purify proteins prior to analysis. Such assays have been characterized extensively for a range of therapeutic proteins, proving exquisitely sensitive to alterations in protein sequence, structure, and post-translational modification state. Despite advantages over traditional probes of protein stability, improving the throughput and information content of gas-phase protein unfolding assays remains a challenge for current instrument platforms. In this report, we describe modifications to an Agilent 6560 drift tube ion mobility-mass spectrometer in order to perform robust, simultaneous CIU across all precursor ions detected. This approach dramatically increases the speed associated with typical CIU assays, which typically involve mass selection of narrow m/z regions prior to collisional activation, and thus their development requires a comprehensive assessment of charge-stripping reactions that can unintentionally pollute CIU data with chemical noise when more than one precursor ion is allowed to undergo simultaneous activation. By studying the unfolding and dissociation of intact antibody ions, a key analyte class associated with biotherapeutics, we reveal a predictive relationship between the precursor

charge state, the amount of buffer components bound to the ions of interest, and the amount of charge stripping detected. We then utilize our knowledge of antibody charge stripping to rapidly capture CIU data for a range of antibody subclasses and subtypes across all charge states simultaneously, demonstrating a strong charge state dependence on the information content of CIU. Finally, we demonstrate that CIU data collection times can be further reduced by scanning fewer voltage steps, enabling us to optimize the throughput of our improved CIU methods and confidently differentiate antibody variant ions using ~12.5% of the data typically collected during CIU. Taken together, our results characterize a new instrument platform for biotherapeutic stability measurements with dramatically improved throughput and information content.

6.2 Introduction

Stability is a key biophysical property of folded proteins that impacts their biological function and the ability to engineer new functions and construct effective therapeutics from existing polypeptide frameworks.^{1,2} In a practical context, biotherapeutic monoclonal antibodies (mAbs), which serve as treatments for a variety of diseases, are often screened and engineered with respect to their stabilities.³⁻⁵ Such engineering often takes on additional dimensions, as mAbs can be tailored into a number of modalities including fusion proteins,⁶⁻⁸ bispecifics,^{9,10} and antibody drug conjugates (ADCs),¹¹⁻¹⁴ that seek to improve upon the potencies of existing biotherapeutic scaffolds. The patents for many active therapeutic mAbs are due to expire,¹⁵ resulting in the rise of generic biotherapeutics, also known as biosimilars, which includes a catalog of nearly 400 protein therapeutics development.¹⁶

A range of analytical technologies, including chromatography, electrophoresis, spectroscopy, and mass spectrometry, are routinely employed to characterize therapeutic mAbs.¹⁷⁻²⁰ These techniques are often employed in the context of a multiple attribute monitoring (MAM) methodology, aimed at providing a thorough analysis of a mAb throughout its development.²¹ Many well-validated tools are currently in place to measure the stabilities of therapeutic proteins, including differential scanning calorimetry (DSC), hydrogen-deuterium exchange (HDX), and immunoassays. Despite their robust, validated history in the pharmaceutical industry, such approaches typically require large amounts of purified protein, lengthy analysis times, and homogeneous protein populations.²²⁻²⁶ Furthermore, stability assessments such as differential scanning fluorimetry (DSF) or N-glycosylation analysis require labelling chemistries that may significantly alter the protein structures they seek to measure.^{25,27} Ion mobility-mass spectrometry (IM-MS) is becoming established as a useful technology for protein structure and stability analysis. IM separates ions based on their orientationally averaged collision cross sections (CCSs).²⁸ When coupled to MS, IM drift times can be correlated with ion composition in order to reveal the influence of sequence,^{29,30} small molecule conjugation,³¹ or post-translational modification on antibody structure and stability.^{32,33} Activating protein ions in the gas phase and assessing the resulting changes in protein ion conformation during a collision induced unfolding (CIU) experiment has emerged as a useful method for rapidly assessing protein stability.³⁴ CIU can capture stability shifts associated with protein domain structure,³⁵ anion and cation adduction,^{36,37} kinase inhibitor binding,³⁸ as well as the disulfide and glycosylation patterns in intact mAbs.^{39,40}

Despite the advances in IM-MS and CIU technologies for structural biology, limitations associated with the throughput of typical CIU experiments constrain its utility for routine

antibody characterization. Most CIU workflows rely upon the MS-based isolation of single protein ion charge states prior to CIU analysis in order to prevent chemical noise related to the loss of charged adducts from protein ions having $n+1$ charges. This “charge stripping” chemical noise can pollute CIU data for adjacent signals, in extreme cases accounting for up to 50% of the observed signal intensities, preventing accurate CIU analyses. Many workflows require the collection of CIU data across several charge states in order to search for those ions that provide maximally differentiating data, but doing so serially requires long acquisition times that are incompatible with high-throughput workflows. Thus, CIU approaches that do not require such ion pre-selection would offer significant advantages in terms of analysis speed and information content.

In this report, we describe modifications to an Agilent 6560 IM-MS platform that allow for significantly increased levels of ion activation prior to IM separation. We make a series of quantitative IM-MS measurements of mAb charge stripping chemical noise under native-like and supercharged buffer conditions over a range of activation voltages. The charge stripping we observe in these experiments is used to develop a correction algorithm for CIU data acquired without prior mass selection. We then use this optimized workflow to evaluate a series of charge multiplexed CIU datasets aimed at differentiating IgG subtypes and mAb light chain variants, comparing each with previously reported, mass-selected datasets for single charge states.

6.3 Experimental Methods

6.3.1 Sample Preparation

SiLu[™]Lite SigmaMab Universal antibody standard, IgG1 κ , IgG1 λ , IgG2 κ , IgG2 λ , and IgG4 κ from human myeloma were purchased from Sigma-Aldrich and supplied as lyophilized powder

(St. Louis, MO). All samples were reconstituted using Milli-Q water (Millipore) to a concentration of 2 mg/mL unless specified otherwise. IgG1 samples were buffer exchanged into 200 mM ammonium acetate buffer using Micro Bio-spin 30 columns (Bio-Rad, Hercules, CA). Buffer exchanged samples were then diluted to a working concentration of 1 mg/mL (~6.7 μ M). Where noted, antibody samples were supercharged with 2-nitrobenzyl alcohol (m-NBA) using 1-2% (V/V %) on the Synapt G2, or using sulfolane at 5% (V/V %) on the Agilent 6560 platform, with both reagents purchased from Sigma-Aldrich (St. Louis, MO).

6.3.2 Protein Charge Stripping Analysis

Antibody standards were initially analyzed using a quadrupole-ion mobility-time-of-flight mass spectrometer (Q-IM-ToF MS) instrument (Synapt G2 HDMS, Waters, Milford, MA). Sample was transferred to a gold coated borosilicate capillary needle (prepared in-house), and ions were generated by direct infusion using a nano-electrospray ionization (nESI) source in positive mode. The electrospray capillary was operated at voltages of 1.5-1.7 kV with the sampling cone operated at 52 V. The backing pressure was set to ~7.9-8.1 mbar. The helium cell flow was operated at 200 mL/min and pressurized to 1.40×10^{-3} mbar. The trap traveling-wave ion guide was pressurized to 3.0×10^{-2} mbar of argon gas. The traveling-wave IM separator was operated at a pressure of ~2.6 mbar. IM separation was achieved with a travelling wave operated at 40 V wave height traveling at 600 m/s. The ToF-MS was operated over the m/z range of 1000–10,000 at a pressure of 1.5×10^{-6} mbar.

Antibody ions were subjected to collisions in the travelling-wave ion trap prior to the IM separation to perform antibody CIU. Tandem-MS (quadrupole selection) was used to select charge states, 20-25⁺ in native antibody samples, and 26⁺ or greater in supercharged antibody samples. The collision voltage was ramped from 5 to 200 V in 5 V increments to construct the

CIU fingerprint data. The dwell time for each 5 V step was 30 seconds. Charge stripping data were extracted after CIU was performed using the raw data for each charge state collected at each collision voltage step.

6.3.3 A Modified Agilent 6560 for Collision Induced Unfolding Experiments

IgG1 κ , IgG1 λ , IgG2 κ , IgG2 λ , and IgG4 κ antibody samples were analyzed using a modified Agilent 6560 IM-Q-TOF platform (Agilent Technologies, Santa Clara, CA). The instrument configuration for the unmodified 6560 IM-Q-TOF instrument has been described previously^{41,42}. Briefly, the ion optics design changes to the modified instrument included an additional ion lens (called fragmentor lens) at the exit of the ion transfer capillary. The DC potential on this lens can be independently controlled to adjust the ion acceleration electric field between the capillary exit and the fragmentor lens. The ambient pressure in the ion activation region is about 4.5 Torr. To increase the gas discharge potential and to improve the collision activation efficiency, we have used sulfur hexafluoride gas in the source region. SF₆ gas was added to the nitrogen drying gas line at a 10% v/v ratio and this gas mixture is carried through the ion transfer capillary to the ion activation region at the exit of the capillary. The absolute voltage difference between the capillary exit and the fragmentor lens is denoted as collision voltage hereafter.

Antibody samples were electrosprayed using a micronebulizer attached to an Agilent jet stream ion source. An external syringe pump was used to deliver the sample at 4 μ L/min flow rate. The source and drying gas temperatures were maintained at 150 °C. The mass analyzer was operated in extended mass range (10000 m/z) and ion mobility enabled mode. In ion mobility mode, maximum drift time was set to 65 ms, trap fill time was set to 50 ms and trap release time was set to 1 ms. The drift tube was operated under 3.95 Torr nitrogen gas and the electric field was set to 18.5 V/cm (drift tube entrance voltage = 1700 V and drift tube exit voltage = 250 V).

In MassHunter data acquisition software, time segment mode is used to ramp the collision voltage in 22 steps (for the full CIU scan) and the dwell time for each step was 1 min. For full CIU curve experiments, collision voltage was ramped from 220 to 500 V in 20 V steps and from 500 to 560 V in 10 V steps.

Antibody CIU data were collected across 24-29⁺ ions observed in native mAb samples and over 30-42⁺ in supercharged antibody samples. 5-voltage step fingerprints were generated using the collision voltages that were associated with both stable mAb features, and feature transitions observed in complete CIU fingerprints. The collision voltages used for such CIU experiments were 240 V, 340 V, 380 V, 420 V, and 550 V. IM data were recorded for each selected mAb ion at each collision voltage, and compiled to generate CIU fingerprints. The CIU fingerprints were recorded as CCS versus collision voltage plots. The instrument measured drift time data were converted to CCS values using the single-field CCS calculation method⁴³.

Antibody samples were collected in replicates (n=3 or higher) to generate averaged CIU fingerprints. These averaged fingerprints were compared to replicate data to determine the root-mean-square deviation (RMSD) baseline for each sample. The IgG1 κ averaged CIU fingerprints for each corresponding charge state were used to compare to the replicates of all mAb samples acquired using an identical CIU workflow, and to generate the RMSD based heat maps shown in this report. These extracted drift time data were analyzed using CIUSuite 2 to generate CIU fingerprints, CIU comparisons, and perform RMSD calculations⁴⁴.

6.3.4 Mass Spectrometry Data Analysis

Molecular masses of intact SigmaMab Universal Antibody standard and supercharged samples were calculated via the maximum entropy deconvolution method.⁴⁵⁻⁴⁷ Drift time data for

precursor and charge stripped ions were extracted at each collision voltage using TWIM Extract,⁴⁸ and manually using Driftscope for selected supercharged data (Waters, Milford, MA). Agilent MS data were viewed with MassHunter IM-MS Browser software (Agilent, Santa Clara, CA). Drift time data for each charge state for native and supercharged IgG1, IgG2, and IgG4 mAbs were extracted at each fragmentor voltage using a custom data extraction tool developed in C# that utilizes the MIDAC API for IM-MS data from Agilent (Agilent Technologies, Santa Clara, CA). The data extraction tool is available with CIUSuite 2.⁴⁹

Charge stripping data were similarly extracted for both precursor and charge-stripped ions at each collision voltage using TWIMExtract, as well as custom scripts written in Python. In order to quantify charge stripping, we measured the total intensity for each ion for each collision voltage step. The relative intensity of charge stripping was calculated as the ratio of the total ion intensity of the charge stripped product ions to the total ion intensities of the product ions and the precursor ion as in Equation 1.

$$(1) \quad CS = \frac{M^{(z-1)+}}{\sum(M^{z+} + M^{(z-1)+} + \dots + M^{(z-x)+})}$$

CS is the relative intensity of charge stripping, $M^{(z-1)+}$ is the charge stripped ion of interest, M^{z+} is the precursor ion, and $M^{(n-x)+}$ is the lowest observable charge stripped ion. For accurate comparisons of the relative charge stripping observed between each charge state, collision voltages were converted to laboratory frame energies (eV).⁵⁰ The charge stripping plots were initially fit with a logistic function using a custom python script before processing using OriginPro 2017 (Academic). The charge stripping data were mean smoothed once using a 10 point window, and a multi-data set fit mode was used such that each charge state was independently fit with a logistic function with 400 iterations. The inflection and bounds of these fitted logistic functions were used to determine the lab frame energy where charge stripping

reaches 50% of its maximum value for each charge state, which we used to estimate the initial and maximum charge stripping propensities for each mAb charge state.

The all charge state CIU analysis is represented as a heat plot generated in OriginPro 2017, in which the relative factor difference in RMSD values obtained for each charge state of IgG1 κ , or IgG2 κ in the case of light chain variants, are compared to determine the RMSD baseline (blue, bottom) and compared against the IgG1 κ , IgG1 λ , IgG2 κ , IgG2 λ , or IgG4 κ fingerprints of the same charge state. Relative difference in RMSD values are shown with a colorimetric scale, with dark blue, blue, light blue, green, yellow, and red indicating increasing levels of difference to the IgG1 κ or IgG2 κ reference relative to the highest RMSD value.

6.4 Results and Discussion

Figure 6-1 shows the modified Agilent 6560 platform and the CIU workflow used in this report. The drift tube within the 6560 typically generates higher IM resolving powers ($R\sim 60$)⁴¹ than the traveling wave IM separators typically used for CIU ($R\sim 40$), thus producing correlated advantages for CIU data analysis.⁵¹ The modifications to the 6560 ion optics shown, along with the introduction of SF₆ gas into the source region, provides higher center-of-mass frame collision energies for initiating CIU prior to IM separation (Figure 6-1 A). We found these modifications to be necessary in order to provide sufficient activation energy to completely unfold mAb ions (Figure 6-1 B). CIU was performed on the 6560 by increasing the potential of the capillary exit and fragmentor lens and collecting the IM drift time plots at each collision voltage step for all charge states. The IM drift time data was extracted for each charge state using custom software

built using the MIDAC API for IM-MS data from Agilent (Figure 6-1 C). The extracted drift times were plotted against the collision voltage (Figure 6-1 D) and used to generate a CIU fingerprint contour with drift time (ms) or CCS (nm²) on the y-axis, collision voltage (V) on the

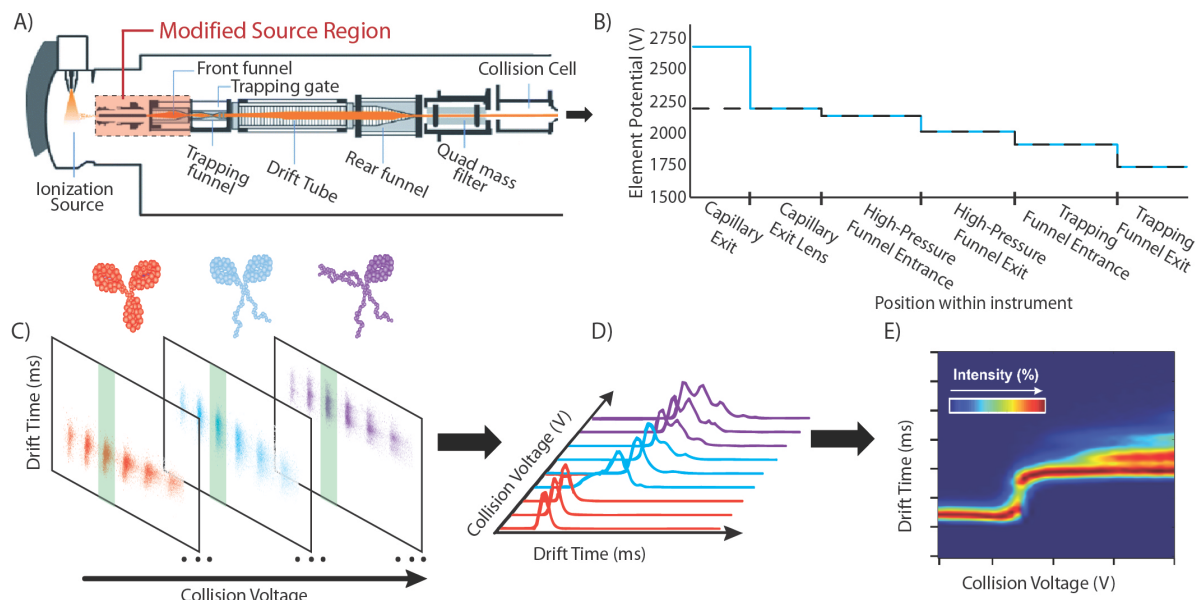


Figure 6-1 A diagram of the modified Agilent 6560 IM-MS instrument (A) with an associated maximum activation potential by instrument region (B). The blue trace indicates the potentials applied to the modified instrument, which is capable of accelerating ions into the trapping funnel at 2680 V, compared the 2200 V possible on the unmodified platform. (C) An illustration of the collision induced unfolding workflow pursued in this work for intact antibody ions without precursor selection. For CIU analysis, signals associated with individual charge states are extracted across all collision voltages. (D) Increases in IM drift times are observed and tracked at each collision voltage (E) to generate a CIU “fingerprint”.

x-axis, and the colorimetric scale formed by the relative ion intensity (Figure 6-1 E).

An all charge state CIU approach, in principle, leads to significant improvements in the information content acquired per unit of time and sample. Despite these advantages, current CIU workflows^{39,52} typically isolate individual analyte charge states prior to collisional heating in order to generate CIU fingerprints in an effort to avoid contamination from the chemical noise associated with charge stripping events. Such events typically involve the ejection of charged small molecules, or bound buffer adducts, from the precursor ions upon collision activation, often requiring low activation voltages to initiate the reaction. In order to measure mAb ion

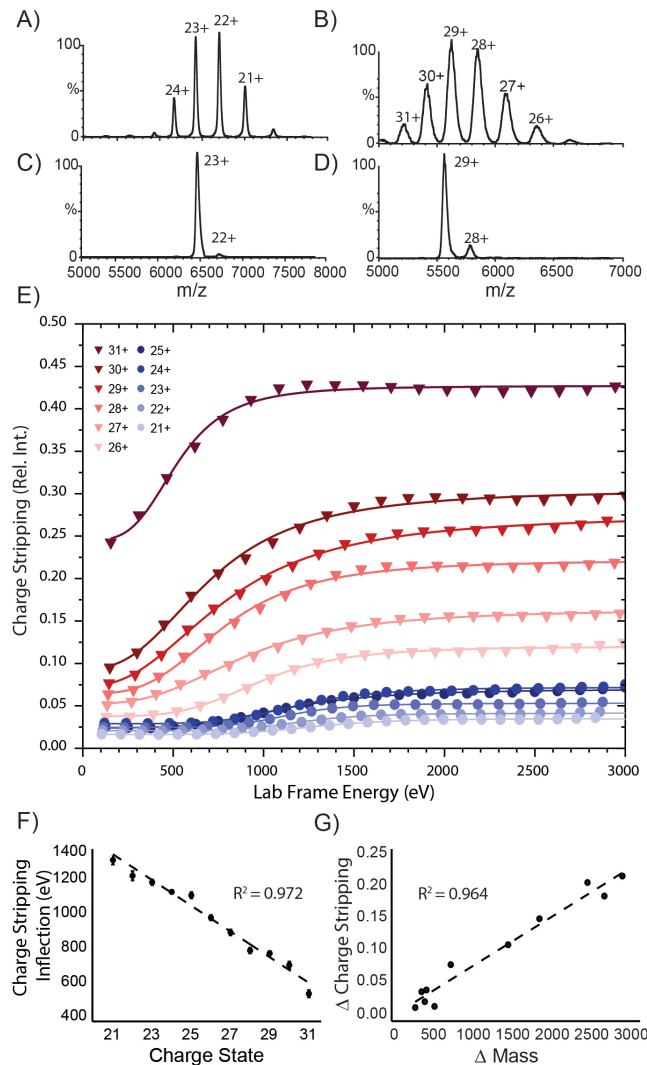


Figure 6-2 Native MS data for (A) an intact mAb standard and (C) quad-selected 23⁺ ion. The low intensity signal annotated as 22⁺ mAb ions is the result of charge stripping. Native MS data for (B) a m-NBA supercharged intact mAb standard and (D) quad-selected 29⁺ mAb ions, exhibiting increased evidence of ion charge stripping (28⁺). (E) A plot of total charge stripping observed as a function of laboratory frame collision energy for mAb charge states 21-31⁺ (acquired across both native, blue circle, and supercharging, red triangle, conditions). The amount of charge stripping observed under native MS conditions reaches a maximum of 4-6% of the total ion current, and 8-40% for supercharged mAbs. Charge stripping data were fitted with a generalized logistic function (F), and midpoint values for the fits plotted as a function of mAb charge state exhibit a strong linear correlation (R^2 value shown). The difference in average charge stripping observed relative to control data are plotted as a function of the difference in average intact mass measured at low and high activation energies, exhibiting a similarly strong linear correlation (G).

charge stripping, we isolated intact mAb charge states with the quadrupole mass analyzer under both native and supercharged conditions on a Synapt G2 IM-MS platform. Native MS data acquired for mAb standards reveal signals corresponding to the 20-25⁺ charge states of the intact antibody (Figure 6-2 A). When we then selected an m/z region corresponding to the mAb 23⁺ charge state, and applied a collision voltage of 50 V, we observed a small signal corresponding to 22⁺ charge-stripped mAb product ions, at a relative intensity of 3% when compared to the 23⁺ precursor (Figure 6-2 C).

To evaluate the effect of charge state and buffer adduct identity on the propensity of mAb ions to undergo charge stripping, we used the nESI super-charging reagent m-NBA to acquire CIU data for charge states 26-32⁺ (Figure 6-2 B).^{53,54} Such super-charged mAb ions exhibited a greater propensity to undergo charge stripping than those prepared in pure ammonium acetate

buffer. For example, isolated 29^+ mAb ions activated at 50V converts ~18% of the precursor into a charge-stripped 28^+ ion population (Figure 6-2 D). As charge stripping is due to a loss of positively charged adducts,^{55,56} it follows that the amount of charge stripping observed is likely dependent on the population of low volatility buffer components bound to the mAb, their ionizability, and any instrument settings that may lead to ion activation. In order to quantify charge stripping in mAb ions, we began by tracking the relative amount of charge stripping detected in our mass selected data as a function of the applied laboratory frame collision energies used for activation (Figure 6-2 E). Antibody ions generated under native conditions produce similar, low levels of charge stripping (2 – 5% relative abundance), and we observe this to increase at higher collision energies. Charge amplified antibody ions follow similar trends, but produce dramatically increased amounts of charge stripping at all activation energies, with higher charge states displaying lower onset energy for the charge stripping reaction. The charge stripping data shown in Figure 6-2 E was fit to logistic functions to determine the amount of charge stripping detected at the inflection point of each sigmoid, as well as the total difference in charge stripping observed for each charge state.

Ultimately, we sought to build a corrective algorithm using our mass-selected data, capable of normalizing CIU data acquired without prior mass filtering to account for any charge stripping signals present. Our data revealed a strong linear correlation between the laboratory-frame activation energy required to reach the inflection points in the charge stripping reaction and charge state analyzed (Figure 6-2 F), indicating that mAb ions of increased charge undergo more facile charge stripping in a highly predictable manner. We then computed the average difference in total charge stripping across all laboratory-frame activation energies probed using our fitted data from Figure 6-2 E, and plotted these values against the average total difference in

observed mAb ion mass collected at low and high activation conditions (Figure V-1), 5-10 V and 100-150 V respectively (Figure 6-2 G). Again, we observe a strong linear correlation, allowing us to quantitatively relate the observed mass of adducts bound to mAb ions with the amount of charge stripped product ions produced. Taken together, the trends observed in Figure 6-2 E and G create a means to calculate the amount of charge stripping expected across a wide range of mAb ion masses, energies, and charge states in CIU assays where pre-IM mass filtering is either not available or not desirable.

In order to test our charge multiplexed CIU approach, we began by analyzing IgG subclasses using our modified 6560 IM-MS platform that were previously differentiated using CIU in a mass-selected mode.³⁹ For example, we observe three CIU features for 26⁺ ions across data acquired for the IgG1, IgG2 and IgG4 subclasses studied here (Figure 6-3), similar to our previous analyses of mass-selected mAb CIU data.³⁹ The first structural state we detect centers on ~80 nm² (Table V-1), a value significantly greater than previous mAb CCS measurements.^{30,57} Given that mAb ions have previously been observed to adopt a wide range of gas-phase structures,⁵⁸ we interpret these increased CCS values as evidence that such ions undergo a lesser degree of structural compaction within our modified drift tube IM-MS platform than observed previously. This initial feature transitions to occupy a range of CCS values spanning ~86 nm² to 95 nm² from 400-450 V, producing a final stable set of conformers centered on ~100 nm² (Figure 6-3 A-C). CIU difference plots reveal low RMSD baseline values and a general ability to confidently differentiate IgG subclasses. The baseline RMSD is a measure of the difference between CIU replicates, and can be used to evaluate reproducibility, providing a control for downstream comparisons between samples. Specifically, we compute the CIU RMSD baseline for IgG1κ 26⁺ to be 1.0±0.2 (Figure 6-3 D), a factor of ~3 lower than was achieved with

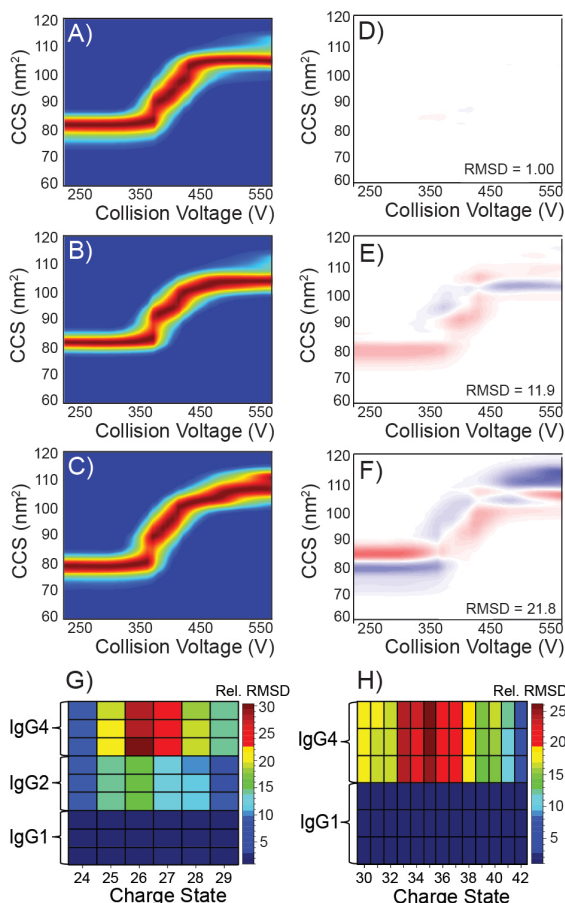


Figure 6-3 CIU fingerprints of (A-C) IgG1 κ , IgG2 κ , and IgG4 κ , respectively. (D) To determine the reproducibility and sensitivity of our CIU RMSD baselines on our modified drift tube platform, IgG1 CIU data were averaged ($n=3$), and a difference analysis conducted, yielding an RMSD baseline value of 2.93%. CIU comparison plots of IgG1 versus (E) IgG2 and (F) IgG4 were generated and produced RMSD values comparable to previous reports. Comprehensive charge state heat maps for (G) native and (H) supercharged mAbs enable a rapid assessment of optimal charge states for CIU-base differentiation.

population improves IM and MS resolution, leading to the enhanced differentiating capabilities we observe in CIU fingerprints on the modified 6560. Specifically, we previously achieved factor increases in CIU RMSD analyses for IgG subclasses ranging between 6 and 8,³⁹ whereas CIU data acquired on our modified 6560 IM-MS platform produces CIU RMSD factor increases of 12 and 22 for differentiating the same IgG subclasses.

these samples on our traveling-wave IM-MS platform (Figure V-2 D), resulting in a concomitant improvement in our ability to differentiate IgG subclasses. Difference plots comparing IgG1 κ to both IgG2 κ and IgG4 κ reveal RMSD values of 11.9 ± 0.2 and $21.8 \pm 0.9\%$, respectively (Figure 6-3 E-F), with comparison values of $23 \pm 2\%$ and $19 \pm 2\%$ produced using our traveling-wave IM-MS instrument (Figure V-2 E-F). We attribute the improvement noted above in the baseline RMSD values obtained for these samples to the modifications we implemented on the 6560 IM-MS platform. The tunable optics for ion activation prevent variation in adduction mass resulting in average predicted charge stripping values of 2-3%. This homogenous ion

The results discussed above leverage CIU data extracted from only a single charge state to differentiate IgG subclasses. Charge multiplexed CIU, however, enables us to rapidly identify those charge states that maximally differentiate analytes of interest (Figure 6-3 G). We chose to analyze such charge multiplexed CIU data as a heat map, in which CIU RMSD values (Table V-3 and Table V-4) are compared between IgG1 κ , IgG2 κ and IgG4 κ datasets across all the mAb charge states observed in our experiments. By comparing CIU data in this way, we readily identify optimal charge states for differentiating IgG subclasses. Specifically, our data indicates that 26⁺ ions provide maximal differentiation for IgG1 κ and IgG2 κ samples, and either 26⁺ or 27⁺ charge states appear optimal for similarly-structured IgG4 κ comparisons. Supercharged IgG CIU data also reveals RMSD variations when comparisons are made between IgG1 κ and IgG4 κ samples (Figure 6-3 H). Our heat map analysis indicates that 35⁺ (25.5 \pm 0.5%) provides CIU data that maximizes our ability to differentiate these IgG subclasses. Additionally, the 33⁺, 34⁺, 36⁺, and 37⁺ provide CIU that is highly differentiating (20-25%). Taken together, Figure 6-3 G and H demonstrate the potential utility of a charge multiplexed CIU data analysis, where a group of optimal charge states can be selected for comparisons in order to maximize the detection of structural differences within mAb ion populations.

We analyzed light chain variants of IgG1 and IgG2 antibodies. These light chain variants are distinguished by only subtle differences in sequence and epitope binding, but are responsible for a broad range of phenotypic differences such as conformational flexibility, mAb half-life, and alterations in antibody specificity.² Several methods have been developed to determine the light

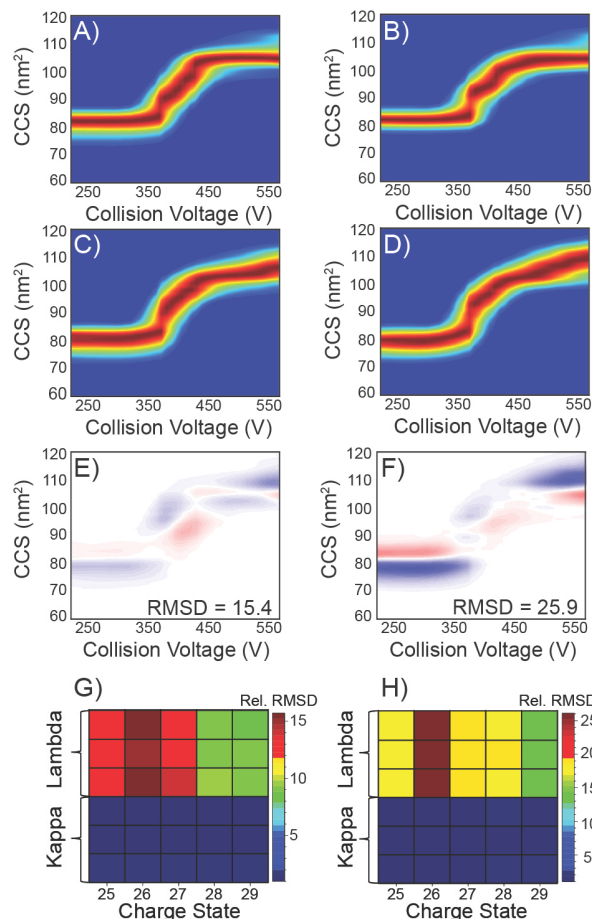


Figure 6-4 CIU fingerprints of A) IgG1 κ , B) IgG1 λ , C) IgG2 κ , and D) IgG2 λ . Subtle differences in (E) IgG1 and (F) IgG2 light chain variant unfolding patterns can be observed, with the greatest differences being observed in the beginning of the CIU process, and minor differences observed in later conditions. RMSD heatmap comparison of (G) IgG1 and (H) IgG2 light chain variants. These heatmaps provide a comprehensive charge state assessment that allows for increased discriminatory capabilities when compared to single charge state CIU, while providing diagnostic information on which charge states can provide the greatest difference in future analysis.

chain type and ratios from cell lines in mAbs and in serum, but the resulting alterations in mAb higher order structure and stability have remained largely unexplored.² As such, we sought to test the ability of our charge multiplexed CIU approach to differentiate such IgG variants. As above, we began by comparing average CIU fingerprint data for the 26⁺ charge state of IgG1 κ (Figure 6-4 A), IgG1 λ (Figure 6-4 C), IgG2 κ (Figure 6-4 B), and IgG2 λ (Figure 6-4 D). Notably, the fingerprints recorded for κ -containing IgG1 and IgG2 mAbs possess intermediate CIU features of enhanced stability when compared to equivalent λ -containing mAb data. CIU difference plots for IgG1 (Figure 6-4 E) and IgG2 (Figure 6-4 F) reveal RMSD values of $15.4 \pm 0.2\%$ and $25.9 \pm 0.1\%$, respectively, driven by CCS

differences observed at both high and low collision voltage values. In addition, λ -containing mAbs produce a broader range of CCS values during CIU when compared to equivalent κ -containing constructs. The subtle differentiation of these fingerprints indicates the sensitivity of all charge state CIU to detect higher order differences between light chain variants of monoclonal antibodies. RMSD heat maps were generated for the light chain variants using the

RMSD values obtained from replicates (Table V-5 and Table V-6), and similarly identifies the 26⁺ charge state as optimally differentiating for both the IgG light chain variants probed here (Figure 6-4 G and H).

The charge multiplexed CIU approach demonstrated in Figure 6-3 and Figure 6-4 results in an increase in throughput over mass-selective CIU modes of operation, while maintaining the same information content. In order to further increase the throughput of such CIU methods, we probed the ability of our modified drift tube IM-MS platform to generate CIU data using a significantly fewer number of voltage steps, strategically chosen to retain key information (*i.e.* transitions between features) from the complete CIU fingerprint.

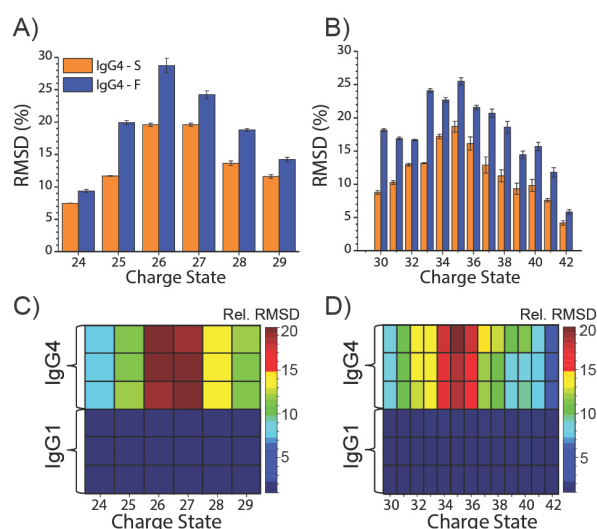


Figure 6-5 (A) Native and (B) supercharged mAb samples demonstrate that RMSD based CIU data generated by “snapshot” CIU (orange) enable faster analysis time with minimal loss to differentiating capabilities afforded by “full” CIU datasets (blue). 5-step RMSD heat maps of (C) Native and (D) supercharged IgG1 and IgG4 antibodies.

To demonstrate our CIU methodology utilizing a reduced number of voltage steps, we generated RMSD comparisons for mAb CIU fingerprints consisting of only 5 collision voltage steps in a charge multiplexed mode, where the voltage values were selected to provide maximum coverage of known IgG transitions in “snapshot” CIU fingerprints. Bar graphs of relative RMSD difference values (Table V-7 and Table V-8) produced for IgG4κ

“snapshots” (Orange) and “full” (Blue) CIU fingerprints on the 6560 for native (Figure 6-5 A), and supercharged conditions (Figure 6-5 B). For IgG1κ, the RMSD baselines computed for both conditions are comparable, despite a 4.4-fold (22 minutes to 5 minutes) reduction in acquisition time for the snapshot CIU. In general, our 5-step data often produced increased standard

deviations, which we associate with an increased sensitivity to noise inherent in using fewer datasets to construct fingerprints.

A CIU RMSD heat map analysis of our charge multiplexed 5-step fingerprints (Figure 6-5 C and D) reveals strong similarities to equivalent analyses performed for CIU data comprised of a larger number of voltage steps, as shown in Figure 6-3. For example, 5-step CIU data identifies the 26⁺ as maximally differentiating for IgG1/IgG4 comparisons, similar to the S/N corrected data from Figure 6-3. In addition, supercharged 5-step mAb CIU data identifies 35-38⁺ ions as maximally differentiating for IgG1/IgG4 comparisons, and a similar group as is highlighted in Figure 6-3. Despite minor shifts in the distribution of RMSD values, the overall difference magnitudes observed in our analyses remain similar between the snapshot and full scan CIU data. As such, the data shown in Figure 6-5 strongly indicates that a 5-step CIU protocol retains the ability to confidently differentiate IgG sub-classes, while delivering a 4.4-fold improvement in throughput over the CIU workflows discussed in Figure 6-3 and Figure 6-4.

6.5 Conclusions

Here, we describe modifications to a drift tube IM-MS platform that enables high quality, charge multiplexed CIU and characterize this new platform in terms of its capacity to sensitively differentiate mAb structural variants. In order to create a high-quality charge multiplexed CIU workflow, we carefully analyzed mAb ion charge stripping reactions, evaluated their potential to contaminate CIU data for neighboring charge states, and generated a correction algorithm capable of normalizing CIU data containing significant amounts of such chemical noise. Our findings illustrate that higher charge states, particularly those formed under supercharging conditions, produce increased charge stripping. We observe improved CIU variability for our drift tube IM-MS data when compared to previous results and attribute this improved

performance to our modified ion source, which serves to more completely remove adduct populations from protein ions prior to CIU.

Furthermore, we demonstrate the capabilities of charge multiplexed CIU in the context of IgG sub-class differentiation, compare our results to previous IM-MS literature, and move forward to sensitively differentiate between mAb light chain variants using CIU for the first time. Charge multiplexed CIU offers a significant reduction in data collection time along while maintaining information content, allowing for the rapid identification of charge states optimally suited for mAb differentiation. It is important to note that the charge stripping trends reported here are likely only valid for mAbs or proteins with similar properties prepared under the two general buffer conditions explored here. Finally, we test a CIU workflow that utilizes ~4-fold fewer voltage steps than typical CIU assays but retains most of the expected differentiating capabilities expected for mAb variants. We envision that a detailed understanding of mAb charge stripping, combined with the modified drift tube IM-MS instrument described here, and the charge multiplexed assay structures tested in this study, will drive the development of CIU toward a high-throughput, validated assay for future biotherapeutic development.

6.6 References

- (1) Beck, A.; Goetsch, L.; Dumontet, C.; Corvaia, N. Strategies and Challenges for the next Generation of Antibody–drug Conjugates. *Nat. Rev. Drug Discov.* **2017**, *16* (5), 315–337.
- (2) Brinkmann, U.; Kontermann, R. E. The Making of Bispecific Antibodies. *MAbs* **2017**, *9* (2), 182–212.
- (3) Ecker, D. M.; Jones, S. D.; Levine, H. L. The Therapeutic Monoclonal Antibody Market. *MAbs* **2015**, *7* (1), 9–14.
- (4) Chames, P.; Van Regenmortel, M.; Weiss, E.; Baty, D. Therapeutic Antibodies: Successes, Limitations and Hopes for the Future. *Br. J. Pharmacol.* **2009**, *157* (2), 220–233.
- (5) Urquhart, L. Market Watch: Top Drugs and Companies by Sales in 2017. *Nat. Rev. Drug Discov.* **2018**, *17* (4), 232.
- (6) Lipsky, P. E.; van der Heijde, D. M. F. M.; St. Clair, E. W.; Furst, D. E.; Breedveld, F. C.; Kalden, J. R.; Smolen, J. S.; Weisman, M.; Emery, P.; Feldmann, M.; et al. Infliximab and

- Methotrexate in the Treatment of Rheumatoid Arthritis. *N. Engl. J. Med.* **2000**, *343* (22), 1594–1602.
- (7) Moreland, L. W.; Baumgartner, S. W.; Schiff, M. H.; Tindall, E. A.; Fleischmann, R. M.; Weaver, A. L.; Ettlinger, R. E.; Cohen, S.; Koopman, W. J.; Mohler, K.; et al. Treatment of Rheumatoid Arthritis with a Recombinant Human Tumor Necrosis Factor Receptor (P75)–Fc Fusion Protein. *N. Engl. J. Med.* **1997**, *337* (3), 141–147.
 - (8) Feldmann, M.; Maini, R. N. Anti-TNF α Therapy of Rheumatoid Arthritis: What Have We Learned? *Annu. Rev. Immunol.* **2001**, *19* (1), 163–196.
 - (9) Holliger, P.; Hudson, P. J. Engineered Antibody Fragments and the Rise of Single Domains. *Nat. Biotechnol.* **2005**, *23* (9), 1126–1136.
 - (10) Rodgers, K. R.; Chou, R. C. Therapeutic Monoclonal Antibodies and Derivatives: Historical Perspectives and Future Directions. *Biotechnol. Adv.* **2016**, *34* (6), 1149–1158.
 - (11) Cho, K.; Wang, X.; Nie, S.; Chen, Z.; Shin, D. M. Therapeutic Nanoparticles for Drug Delivery in Cancer. *Clin. Cancer Res.* **2008**, *14* (5), 1310–1316.
 - (12) Sievers, E. L.; Senter, P. D. Antibody-Drug Conjugates in Cancer Therapy. *Annu. Rev. Med.* **2013**, *64* (1), 15–29.
 - (13) Beck, A.; Reichert, J. M. *Antibody-Drug Conjugates Present and Future*; 2014; Vol. 6.
 - (14) Xu, K.; Liu, L.; Dere, R.; Mai, E.; Erickson, R.; Hendricks, A.; Lin, K.; Junutula, J. R.; Kaur, S. Characterization of the Drug-to-Antibody Ratio Distribution for Antibody-Drug Conjugates in Plasma/Serum. *Bioanalysis* **2013**, *5* (9), 1057–1071.
 - (15) Konara, C. S.; Barnard, R. T.; Hine, D.; Siegel, E.; Ferro, V. The Tortoise and the Hare: Evolving Regulatory Landscapes for Biosimilars. *Trends Biotechnol.* **2016**, *34* (1), 70–83.
 - (16) Fox, J. L. Debate over Details of US Biosimilar Pathway Continues to Rage. *Nat. Biotechnol.* **2012**, *30* (7), 577.
 - (17) Masson, G. R.; Jenkins, M. L.; Burke, J. E. An Overview of Hydrogen Deuterium Exchange Mass Spectrometry (HDX-MS) in Drug Discovery. *Expert Opin. Drug Discov.* **2017**, *12* (10), 981–994.
 - (18) Tabrizi, M.; Bornstein, G. G.; Klakamp, S. L. Development of Antibody-Based Therapeutics: Translational Considerations. *Dev. Antibody-Based Ther. Transl. Considerations* **2012**, *9781441959*, 1–425.
 - (19) Beck, A.; Wagner-Rousset, E.; Ayoub, D.; Van Dorsselaer, A.; Sanglier-Cianfèrani, S. Characterization of Therapeutic Antibodies and Related Products. *Anal. Chem.* **2013**, *85* (2), 715–736.
 - (20) Rathore, D.; Faustino, A.; Schiel, J.; Pang, E.; Boyne, M.; Rogstad, S. The Role of Mass Spectrometry in the Characterization of Biologic Protein Products. *Expert Rev. Proteomics* **2018**, *15* (5), 431–449.
 - (21) Harris, R. J.; Shire, R. J.; Winter, C. Commercial Manufacturing Scale Formulation and Analytical Characterization of Therapeutic Recombinant Antibodies. *Drug Dev. Res.* **2004**, *61* (3), 137–154.
 - (22) Johnson, C. M. Differential Scanning Calorimetry as a Tool for Protein Folding and Stability. *Arch. Biochem. Biophys.* **2013**, *531* (1–2), 100–109.
 - (23) Clas, S.; Dalton, C.; Hancock, B. Differential Scanning Calorimetry: Applications in Drug Development. *Pharm. Sci. Technol. Today* **1999**, *2* (8), 311–320.
 - (24) Moreno, M. R.; Tabitha, T. S.; Nirmal, J.; Radhakrishnan, K.; Yee, C. H.; Lim, S.; Venkatraman, S.; Agrawal, R. Study of Stability and Biophysical Characterization of Ranibizumab and Aflibercept. *Eur. J. Pharm. Biopharm.* **2016**, *108*, 156–167.

- (25) Niesen, F. H.; Berglund, H.; Vedadi, M. The Use of Differential Scanning Fluorimetry to Detect Ligand Interactions That Promote Protein Stability. *Nat. Protoc.* **2007**, *2* (9), 2212–2221.
- (26) Bruylants, G.; Wouters, J.; Michaux, C. Differential Scanning Calorimetry in Life Science: Thermodynamics, Stability, Molecular Recognition and Application in Drug Design. *Curr. Med. Chem.* **2005**, *12* (17), 2011–2020.
- (27) Higel, F.; Seidl, A.; Sörgel, F.; Friess, W. N-Glycosylation Heterogeneity and the Influence on Structure, Function and Pharmacokinetics of Monoclonal Antibodies and Fc Fusion Proteins. *Eur. J. Pharm. Biopharm.* **2016**, *100*, 94–100.
- (28) Ruotolo, B. T.; Benesch, J. L. P.; Sandercock, A. M.; Hyung, S.-J.; Robinson, C. V. Ion Mobility–mass Spectrometry Analysis of Large Protein Complexes. *Nat. Protoc.* **2008**, *3* (7), 1139–1152.
- (29) Watanabe, Y.; Vasiljevic, S.; Allen, J. D.; Seabright, G. E.; Duyvesteyn, H. M. E.; Doores, K. J.; Crispin, M.; Struwe, W. B. Signature of Antibody Domain Exchange by Native Mass Spectrometry and Collision-Induced Unfolding. *Anal. Chem.* **2018**, *90* (12), 7325–7331.
- (30) Hernandez-Alba, O.; Wagner-Rousset, E.; Beck, A.; Cianfèrani, S. Native Mass Spectrometry, Ion Mobility, and Collision-Induced Unfolding for Conformational Characterization of IgG4 Monoclonal Antibodies. *Anal. Chem.* **2018**, *90* (15), 8865–8872.
- (31) Debaene, F.; Bœuf, A.; Wagner-Rousset, E.; Colas, O.; Ayoub, D.; Corvaia, N.; Van Dorselaer, A.; Beck, A.; Cianfèrani, S. Innovative Native MS Methodologies for Antibody Drug Conjugate Characterization: High Resolution Native MS and IM-MS for Average DAR and DAR Distribution Assessment. *Anal. Chem.* **2014**, *86* (21), 10674–10683.
- (32) Thompson, N. J.; Rosati, S.; Rose, R. J.; Heck, A. J. R. The Impact of Mass Spectrometry on the Study of Intact Antibodies: From Post-Translational Modifications to Structural Analysis. *Chem. Commun.* **2013**, *49* (6), 538–548.
- (33) Zhang, H.; Cui, W.; Gross, M. L. Mass Spectrometry for the Biophysical Characterization of Therapeutic Monoclonal Antibodies. *FEBS Lett.* **2014**, *588* (2), 308–317.
- (34) Dixit, S. M.; Polasky, D. A.; Ruotolo, B. T. Collision Induced Unfolding of Isolated Proteins in the Gas Phase: Past, Present, and Future. *Curr. Opin. Chem. Biol.* **2018**, *42*, 93–100.
- (35) Zhong, Y.; Han, L.; Ruotolo, B. T. Collisional and Coulombic Unfolding of Gas-Phase Proteins: High Correlation to Their Domain Structures in Solution. *Angew. Chemie - Int. Ed.* **2014**, *53* (35), 9209–9212.
- (36) Han, L.; Hyung, S. J.; Ruotolo, B. T. Bound Cations Significantly Stabilize the Structure of Multiprotein Complexes in the Gas Phase. *Angew. Chemie - Int. Ed.* **2012**, *51* (23), 5692–5695.
- (37) Han, L.; Hyung, S. J.; Mayers, J. J. S.; Ruotolo, B. T. Bound Anions Differentially Stabilize Multiprotein Complexes in the Absence of Bulk Solvent. *J. Am. Chem. Soc.* **2011**, *133* (29), 11358–11367.
- (38) Rabuck, J. N.; Hyung, S. J.; Ko, K. S.; Fox, C. C.; Soellner, M. B.; Ruotolo, B. T. Activation State-Selective Kinase Inhibitor Assay Based on Ion Mobility-Mass Spectrometry. *Anal. Chem.* **2013**, *85* (15), 6995–7002.
- (39) Tian, Y.; Han, L.; Buckner, A. C.; Ruotolo, B. T. Collision Induced Unfolding of Intact Antibodies: Rapid Characterization of Disulfide Bonding Patterns, Glycosylation, and

- Structures. *Anal. Chem.* **2015**, *87* (22), 11509–11515.
- (40) Pisupati, K.; Tian, Y.; Okbazghi, S.; Benet, A.; Ackermann, R.; Ford, M.; Saveliev, S.; Hosfield, C. M.; Urh, M.; Carlson, E.; et al. A Multidimensional Analytical Comparison of Remicade and the Biosimilar Remsima. *Anal. Chem.* **2017**, *89* (9), 4838–4846.
- (41) Kurulugama, R. T.; Darland, E.; Kuhlmann, F.; Stafford, G.; Fjeld-sted, J. C. Evaluation of Drift Gas Selection in Complex Sample Analyses Using a High Performance Drift Tube Ion Mobility-QTOF Mass Spectrometer. *Analyst* **2015**, *140*, 6834–6844.
- (42) May, J. C.; Goodwin, C. R.; Lareau, N. M.; Leaptrot, K. L.; Morris, C. B.; Kurulugama, R. T.; Mordehai, A.; Klein, C.; Barry, W.; Darland, E.; et al. Conformational Ordering of Biomolecules in the Gas Phase: Nitrogen Collision Cross Sections Measured on a Prototype High Resolution Drift Tube Ion Mobility-Mass Spectrometer. **2014**.
- (43) Stow, S. M.; Causon, T. J.; Zheng, X.; Kurulugama, R. T.; Mairinger, T.; May, J. C.; Rennie, E. E.; Baker, E. S.; Smith, R. D.; McLean, J. A.; et al. An Interlaboratory Evaluation of Drift Tube Ion Mobility-Mass Spectrometry Collision Cross Section Measurements. *Anal. Chem.* **2017**, *89* (17), 9048–9055.
- (44) Daniel A. Polasky, Sugyan M. Dixit, Sarah M. Fantin, and B. T. R. CIUSuite2_v6_BTR. *Anal. Chem.* **2019**.
- (45) Tito, M. A.; Tars, K.; Valegard, K.; Hajdu, J.; Robinson, C. V. Electrospray Time-of-Flight Mass Spectrometry of the Intact MS2 Virus Capsid [15]. *J. Am. Chem. Soc.* **2000**, *122* (14), 3550–3551.
- (46) McKay, A. R.; Ruotolo, B. T.; Ilag, L. L.; Robinson, C. V. Mass Measurements of Increased Accuracy Resolve Heterogeneous Populations of Intact Ribosomes. *J. Am. Chem. Soc.* **2006**, *128* (35), 11433–11442.
- (47) Winkler, R. ESIprot: A Universal Tool for Charge State Determination and Molecular Weight Calculation of Proteins from Electrospray Ionization Mass Spectrometry Data. *Rapid Commun. Mass Spectrom.* **2010**, *24* (3), 285–294.
- (48) Haynes, S. E.; Polasky, D. A.; Dixit, S. M.; Majmudar, J. D.; Neeson, K.; Ruotolo, B. T.; Martin, B. R. Variable-Velocity Traveling-Wave Ion Mobility Separation Enhancing Peak Capacity for Data-Independent Acquisition Proteomics. *Anal. Chem.* **2017**, *89* (11), 5669–5672.
- (49) Polasky, D. A.; Dixit, S. M.; Fantin, S. M.; Ruotolo, B. T. CIUSuite 2: Next-Generation Software for the Analysis of Gas-Phase Protein Unfolding Data. *Anal. Chem.* **2019**, *91* (4), 3147–3155.
- (50) Hall, Z.; Politis, A.; Bush, M. F.; Smith, L. J.; Robinson, C. V. Charge-State Dependent Compaction and Dissociation of Protein Complexes: Insights from Ion Mobility and Molecular Dynamics. *J. Am. Chem. Soc.* **2012**, *134* (7), 3429–3438.
- (51) Zhong, Y.; Hyung, S.-J.; Ruotolo, B. T. Characterizing the Resolution and Accuracy of a Second-Generation Traveling-Wave Ion Mobility Separator for Biomolecular Ions. *Analyst* **2011**, *136* (17), 3534.
- (52) Eschweiler, J. D.; Rabuck-Gibbons, J. N.; Tian, Y.; Ruotolo, B. T. CIUSuite: A Quantitative Analysis Package for Collision Induced Unfolding Measurements of Gas-Phase Protein Ions. *Anal. Chem.* **2015**, *87* (22), 11516–11522.
- (53) Metwally, H.; McAllister, R. G.; Popa, V.; Konermann, L. Mechanism of Protein Supercharging by Sulfolane and M-Nitrobenzyl Alcohol: Molecular Dynamics Simulations of the Electrospray Process. *Anal. Chem.* **2016**, *88* (10), 5345–5354.
- (54) Going, C. C.; Williams, E. R. Supercharging with M-Nitrobenzyl Alcohol and Propylene

- Carbonate: Forming Highly Charged Ions with Extended, Near-Linear Conformations. *Anal. Chem.* **2015**, *87* (7), 3973–3980.
- (55) Van Duijn, E.; Simmons, D. A.; Van Den Heuvel, R. H. H.; Bakkes, P. J.; Van Heerikhuizen, H.; Heeren, R. M. A.; Robinson, C. V.; Van Der Vies, S. M.; Heck, A. J. R. Tandem Mass Spectrometry of Intact GroEL - Substrate Complexes Reveals Substrate-Specific Conformational Changes in the Trans Ring. *J. Am. Chem. Soc.* **2006**, *128* (14), 4694–4702.
- (56) Sobott, F.; Robinson, C. V. Characterising Electrosprayed Biomolecules Using Tandem-MS - The Noncovalent GroEL Chaperonin Assembly. *Int. J. Mass Spectrom.* **2004**, *236* (1–3), 25–32.
- (57) Campuzano, I. D. G.; Larriba, C.; Bagal, D.; Schnier, P. D. Ion Mobility and Mass Spectrometry Measurements of the Humanized IgGk NIST Monoclonal Antibody. *ACS Symp. Ser.* **2015**, *1202*, 75–112.
- (58) Pacholarz, K. J.; Porrini, M.; Garlish, R. A.; Burnley, R. J.; Taylor, R. J.; Henry, A. J.; Barran, P. E. Dynamics of Intact Immunoglobulin G Explored by Drift-Tube Ion-Mobility Mass Spectrometry and Molecular Modeling. *Angew. Chemie - Int. Ed.* **2014**, *53* (30), 7765–7769.

Chapter 7 An Algorithm for Building Multi-State Classifiers Based on Collision Induced Unfolding Data

Daniel A. Polasky, Sugyan M. Dixit, Daniel D. Vallejo, Kathryn D. Kulju, and Brandon T. Ruotolo.

7.1 Introduction

Native mass spectrometry (MS) and ion mobility-mass spectrometry (IM-MS) have been increasingly adopted techniques for the determination of protein-protein and protein-ligand contacts, stoichiometry, and shape.¹⁻³ Native IM-MS has seen rapid growth in the characterization of proteins^{4,5} protein-ligand complexes,^{6,7} and multi-protein complexes.⁸ A significant challenge in these analyses remains the relatively low resolution of IM in the context of protein structure, limiting the ability of IM-MS to distinguish subtle, but biologically relevant, conformational variations that occur below the resolution limits of modern instrumentation. The activation of protein ions in the gas phase prior to IM separation in an effort to follow their subsequent structural transitions represents a useful method to distinguish such structural differences. This approach, termed collision-induced unfolding (CIU) when the ion activation is accomplished using collisions with an inert gas, has a rich history in the IM-MS analysis of protein structure⁹⁻¹¹ and has seen rapid growth for drug discovery¹²⁻¹⁴ and biotherapeutic characterization.¹⁵⁻²⁰ The relative speed of CIU, combined with detailed comparative structure information, make it a promising technique for the development of structure-sensitive screening methods at medium to high throughput.

A number of reports have demonstrated proof-of-principle methods using CIU to distinguish ligand binding sites for kinases inhibitors,^{13,14,21} quantifying cooperative binding of ligands within a protein complex,²² and detecting protein allostery upon ligand attachment.²³ Screening approaches sensitive to these structural parameters are in great demand for a wide range of applications associated with protein biophysics. The relative comparison of CIU fingerprints under different conditions, for example following ligand binding to a target protein or after applying heat stress to a biotherapeutic, enables the determination of useful information about the structure of a protein and its response to perturbations.

Converting the complex datasets generated in CIU experiments into this structural information requires robust statistical methods. Several recent reports have developed quantitative methods to compare CIU fingerprints in support of these analyses.²⁴⁻²⁹ For screening workflows in particular, supervised learning approaches show great promise. In these methods, “training” CIU data is acquired using known standards and used to generate a classifier that can then distinguish unknown CIU data. We recently developed CIUSuite 2, a software package that includes an automated workflow to construct classifiers for CIU data.²⁶ This approach was used to differentiate ligand and lipid binding modes in a membrane protein system³⁰ and shows promise for high-throughput screening and characterization of biotherapeutics.

Despite these successes, the current method is limited to the comparison of a single charge state of CIU data and relatively small quantities of training data. Native IM-MS experiments typically generate multiple charge states, each with a unique CIU fingerprint. Recent work has demonstrated the benefits of including CIU information from multiple charge states in distinguishing the structures of monoclonal antibodies.³¹ The incorporation of all information available from multiple charge states provides, in principle, great potential for

improving CIU classification and screening methods without increasing data acquisition time. In this report, we describe the creation of a supervised classification algorithm that can accommodate CIU data from multiple protein ‘states,’ improve processing speed to enable processing of large datasets, and expand the scope of the classification workflow to include comparative analyses that move beyond the concept of using a single group of charge states alone. We demonstrate the utility of these approaches to characterize ligand binding modes in a protein-inhibitor context and in distinguishing a highly similar innovator/biosimilar pair of biotherapeutic monoclonal antibodies.

7.2 Methods

7.2.1 Sample Preparation

SiLuLite SigmaMab Universal antibody standard, IgG1 λ , and IgG4 λ from human myeloma were purchased from Sigma-Aldrich and supplied as lyophilized powder (St. Louis, MO). Samples were reconstituted using Milli-Q water (Millipore) to a concentration of 2 mg/mL unless specified otherwise. Avastin® (Genentech, 25 mg/mL) and Avegra® (Biocad, 25 mg/mL) were purchased and supplied in solution formulation (158.6 mM Trehalose dehydrate, 40.9mM Sodium Phosphate, 0.16% Polysorbate 80, pH 6.2). Biotherapeutic samples were diluted to 1mg/mL using 0.9% bacteriostatic sodium chloride injection, USP. (Pfizer Inc. New York City, NY). Stressed samples were incubated at 40 °C with 250 RPM orbital shaking for 4 weeks. All antibody samples were buffer exchanged into 200 mM ammonium acetate buffer using Micro Bio-spin 30 columns (Bio-Rad, Hercules, CA). Buffer exchanged samples were then diluted to a working concentration of 1 mg/mL (~6.7 μ M).

Src kinase domain DNA was synthesized by GeneArt (Life Technologies, Grand Island, NY) using E. coli modified codons and subcloned into pET28a with a modified TEV-protease

cleavable N-terminal 6x-His tag. The plasmid was transformed by electroporation into BL21 DE3 electrochemically competent cells with a YopH in pCDFDuet-1. Cell growth, protein expression, and purification were adapted from protocols previously developed for the c-Src kinase domain³² without cleavage of the His-tag. Dasatinib, staurosporine, foretinib, and ponatinib were purchased from LC Laboratories (Woburn, MA). Protein was reconstituted and buffer exchanged into 200 mM ammonium acetate (Sigma-Aldrich, St. Louis, MO) at pH 7.0 using Micro Bio-Spin 6 columns (BioRad, Hercules, CA) to a final concentration of 10 μ M. Samples were incubated at a ratio of 3:1 inhibitor:protein, on ice for 15 minutes prior to analysis by IM-MS.

7.2.2 CIU Acquisition

All CIU data were acquired using a Synapt G2 quadrupole-ion mobility-time-of-flight mass spectrometer (Q-IM-ToF MS) instrument (Waters, Milford, MA). Sample was transferred to a gold-coated borosilicate capillary needle (prepared in-house), and ions were generated by direct infusion using a nano-electrospray ionization (nESI) in positive mode. The electrospray capillary was operated at voltages of 1.5-1.7 kV with the sampling cone at 40 V. The backing pressure was set to 7.9-8.1 mbar for antibody samples or 5.0 mbar for kinase samples. The trap collision cell was pressurized to $4-5 \times 10^{-2}$ mbar of argon gas, helium cell flow to 1.4×10^3 mbar, traveling-wave IM separator to 3.4 mbar, and ToF MS to 1.5×10^{-6} mbar. CIU experiments were performed by ramping the collision voltage in the trap cell from 5 to 200 V (antibodies) or 10 to 125 V (Src kinase) in 5 V increments with a dwell time of 6 s at each collision voltage.

7.2.3 Classification

IM arrival time data was extracted from raw data for each charge state using TWIMExtract³³ and smoothed with CIUSuite 2²⁶ (Savitzky-Golay 2D smoothing, window 5, 2 iterations). An

updated version of the CIUSuite 2 classification interface that recognizes the labels across multiple states (e.g. charge states) was used to assemble the training data for each classifier. Classifiers were generated in 'all_data' mode with cross validation test sizes of 6, 1, and 3 for data presented in Figure 7-1, Figure 7-2, and Figure 7-3, respectively. Input data for Figure 7-1, Figure 7-2 was normalized but not standardized; input data for Figure 7-3 was both normalized and standardized. The classification algorithm presented here is based on the original CIUSuite 2 algorithm, utilizing the scikit-learn Python library,³⁴ with the following key differences: support for division of the input data into subclasses throughout the classification, addition of data standardization to improve classifier performance, and implementation of random sampling cross validation to allow large input training datasets to be used without prohibitive memory and computation costs. Input training data is standardized within each subclass and collision voltage by scaling to zero mean and unit variance. For input Gaussian data, each attribute of each Gaussian peak is considered separately so that centroids are only standardized with centroids, widths with widths, and so forth. Standardized and labeled training data for each subclass is assessed separately by the univariate feature selection (UFS) method in CIUSuite 2, which uses ANOVA F-value to assess the variation within and between classes at each collision voltage. The highest scoring collision voltages are then chosen from amongst all subclasses for cross validation and final classifier construction, meaning that a classifier can contain data from multiple subclasses.

Cross validation is performed by holding back a portion of the training data (of configurable size), constructing a classifier with the remaining training data, then testing the withheld data (the "test" data) to see if it is classified correctly. As in CIUSuite 2, cross validation involves adding "features" in decreasing order of UFS score to determine the number

of features that results in the most accurate classifier. In the workflow describe here, the features represent a single collision voltage from one of the subclasses, so a particular voltage can be included multiple times if it scores highly in multiple subclasses. The original CIUSuite 2 cross validation method tested all possible permutations of training and test data from an input dataset, which resulted in exponential time and memory cost with increasing dataset size and proved prohibitive for the larger datasets evaluated in this work. Random sampling from the possible input permutations was implemented to reduce this to a linear increase in performance cost by sampling only a user-specified number of the possible permutations, chosen at random. Following determination of the optimal number of features to include, final classifiers are generated as in CIUSuite 2.

7.3 Results and Discussion

Each charge state observed in a native IM-MS experiment undergoes a substantially different unfolding trajectory during CIU, providing potentially complementary information for a multi-state CIU-based classifier. To evaluate the utility of combining data from multiple charge states for CIU classification, we compared monoclonal antibodies IgG1 and IgG4, which differ only slightly in disulfide bonding pattern (Figure 7-1 A). The native mass spectrum of IgG1 shows charge states from 22-26⁺, with 24⁺ being the most abundant (Figure 7-1 B). The CIU fingerprints of IgG1 and IgG4 at the 24⁺ charge state are quite similar, aside from minor differences in the second CIU feature in the 60-80 V activation range (Figure 7-1 A, bottom). Performing a single charge state comparison using the 24⁺ charge state only, as would be done in the original CIUSuite 2 workflow, results in a feature selection plot showing minor differences in the 60-80 V region as expected, with minimal difference outside that region (Figure 7-1 C). The classifier that can be trained from this data is of relatively low quality, achieving a maximum

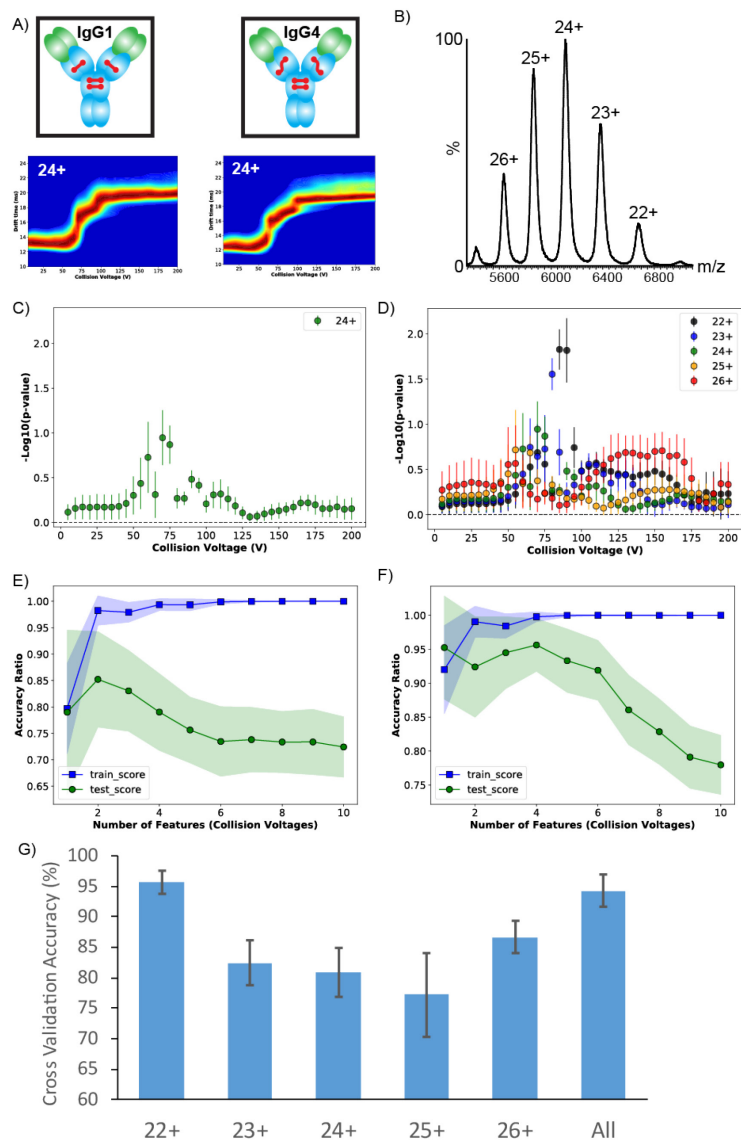


Figure 7-1 Multiple charge state classification of IgGs. A) IgG1 and IgG4 subtypes differ primarily in disulfide bond linkage, resulting in slightly different CIU fingerprints. B) Native mass spectrum of IgG1 with 22-26⁺ charge states. C) UFS score plot distinguishing IgG1 and IgG4 at the 24⁺ charge state only. D) UFS plot for all charge states of IgG1 and IgG4. E) Cross validation accuracies for 1-10 features from the 24⁺ charge state alone and F) for all charge states incorporated into one classifier. G) Optimal cross validation accuracy from each individual charge state and all charge states combined.

cross validation accuracy of 85% when using two features (70 and 75 V) (Figure 7-1 E). Assessing all charge states with the classification workflow, however, reveals that the 24⁺ charge state, despite being the highest signal in the mass spectrum, is not the optimal CIU data to differentiate these two antibodies. To examine all charge states, we perform feature selection sequentially for each, meaning that the 22⁺ charge state of IgG1 is compared to the 22⁺ of IgG4, the 23⁺ to 23⁺, and so on. This results in five feature selection plots, which can be overlaid to evaluate each charge state (Figure 7-1 D). The 22⁺ charge state has the two highest scoring individual voltages (black, 85 and 90 V), followed by 80

V in the 23⁺ charge state (blue), then 75 and 70 V in the 24⁺ charge state (green) (Figure 7-1 F).

As in our standard single charge state classification mode, cross validation is performed by incorporating the data into classifiers in decreasing order of feature selection score; but in

classifiers derived from multiple charge states, the input data can originate in any of the charge states included in the analysis. The cross validation indicates that the optimal classifier in this case uses four collision voltages, two from the 22⁺ charge state and one each from the 23⁺ and 24⁺ charge states, to achieve an accuracy of 95%, significantly improved over the 85% accuracy achieved by the classifier using just the 24⁺ charge state.

To complete the comparison, we generated single charge state classifiers for all five charge states and compared the cross validation accuracy at the optimal number of collision voltages for each classifier (Figure 7-1 G). Given the pair of very high scores from the 22⁺ charge state in the feature selection, it is not surprising that it results in the best single charge state classifier, and indeed achieves nearly identical accuracy to the combined classifier that considered all charge states (95%). The 23-26⁺ charge states each individually achieve accuracies in the 85-90% range, lower than the 22⁺ or combined classifiers. In this case, because one charge state is substantially better at differentiating the classes than the other charge states, its data drives the performance of the combined classifier, resulting in very similar output accuracies. Performing the classification with all charge states is, in this case, primarily a means to rapidly identify the optimal charge state and ensure it is incorporated into the final classifier. Indeed, the 22⁺ ions are the lowest intensity signals included in the analysis, and would not be an obvious choice if using only IM-MS precursor data. In cases where several charge states achieve similar feature selection scores, however, combining data from multiple charge states can generate a superior classifier to any individual charge state.

We applied our multi-state classification workflow to a number of challenging proteins and complexes that had previously confounded CIU classification efforts using data from a single charge state. Src, a non-receptor protein tyrosine kinase, plays a key role in several cell

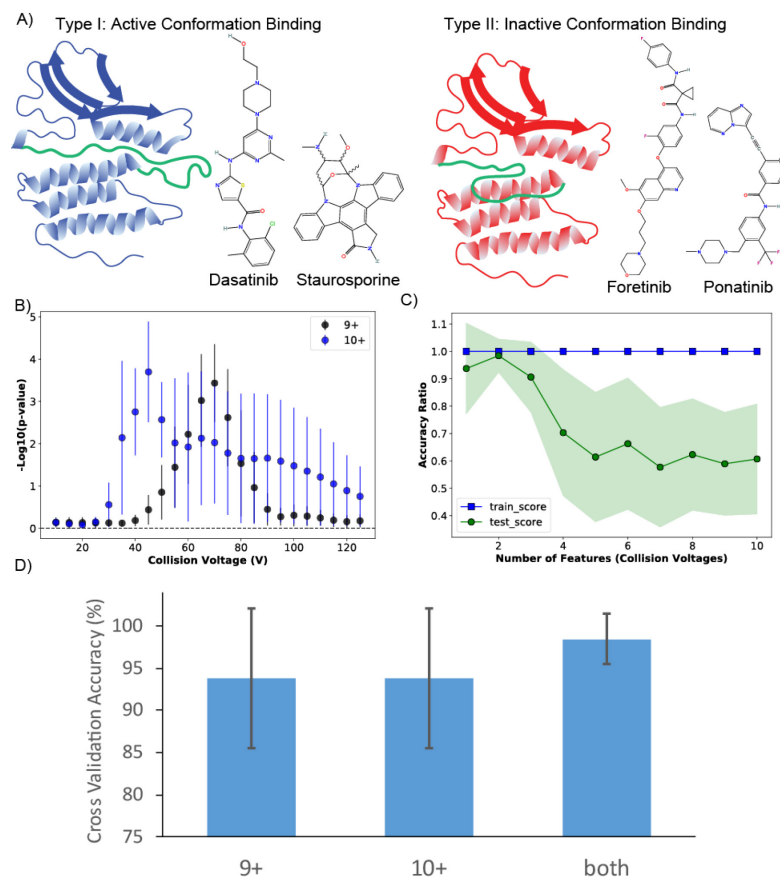


Figure 7-2 Multiple charge state classification of Src kinase. A) Type I and II kinase inhibitors target the active (left) or inactive (right) conformations of the kinase. B) UFS plot comparing Src CIU fingerprints with bound Type I (Dasatinib and Staurosporine) against Type II (Foretinib and Ponatinib) inhibitors at all charge states. C) Cross validation accuracy as a function of number of features included in the classifier for the combined classifier. D) Optimal cross validation accuracy for individual charge state and combined classifiers.

the DFG loop is in the “out” conformation (green loop, Figure 7-2 A, right). While single charge state classifiers and analogous methods have been successful in differentiating such tertiary structures within Abl,^{13,25} a related kinase, differentiating these binding modes within Src using our previous single charge state classification method has proven challenging. Using the multi-state workflow developed here, we observe similar feature scores that distinguish Type I from Type II kinase inhibitors for both the 9⁺ and 10⁺ charge states (Figure 7-2 B). As a result, the optimal classifier uses a single collision voltage each from 9⁺ and 10⁺, resulting in a cross

signaling processes^{35,36} and has been observed to be overexpressed in certain carcinomas and glioblastomas.³⁷ Several classes of inhibitors to kinases like Src are known to target different conformations of the kinase. Type I inhibitors like Dasatinib and Staurosporine bind to the active state, in which the DFG loop is in the “in” conformation, wrapping around the helices (green loop, Figure 7-2 A, left). Type II inhibitors like Foretinib and Ponatinib bind the protein in the inactive conformation, in which

validation accuracy of 98% (Figure 7-2 C). The optimal classifiers for the 9⁺ and 10⁺ charge states individually utilized only the highest scoring single voltage in each case, but achieved accuracies of only 92-93% (Figure 7-2 D). The large error bars for the individual charge state classifier accuracies also indicate substantial uncertainty in their performance, with lower accuracy possible for external validation. Thus, the combined classifier using multiple charge states is superior in this case to any of the individual charge state classifiers, and enabled robust classification of ligand binding modes in a system that had proven challenging to classify with a single charge state alone.

Finally, we examined a biotherapeutic innovator/biosimilar pair, Avastin and Avegra,

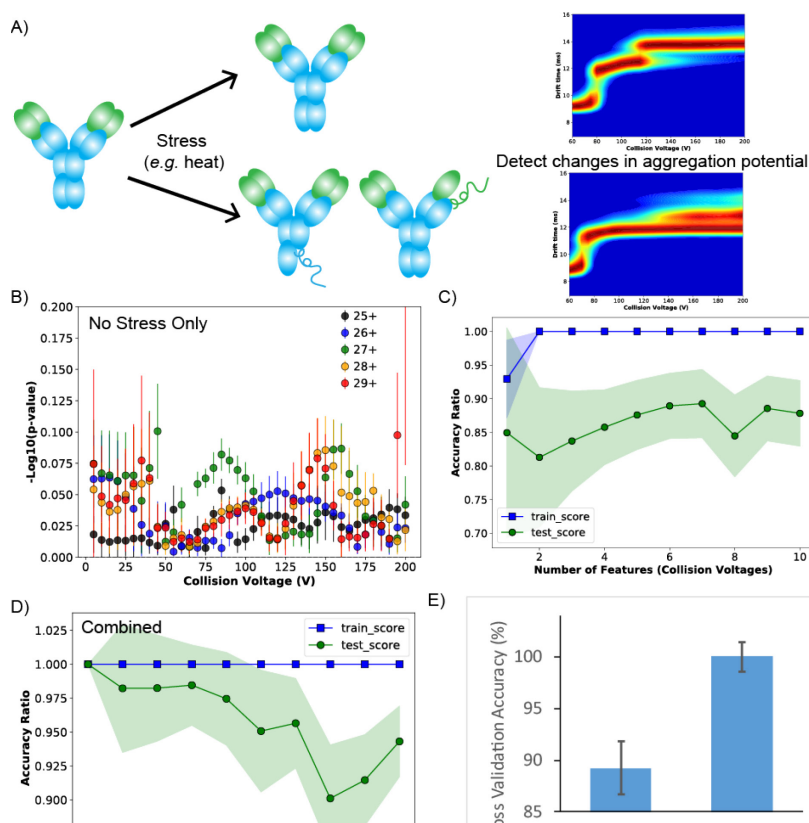


Figure 7-3 Stress subclasses distinguish Avastin and Avegra. A) Stressed antibodies can undergo structural transitions that increase their propensity to aggregate, potentially resulting in differences that can be identified using CIU. B) UFS plot for all charge states of unstressed Avastin and Avegra showing minimal differences. C) Cross validation of unstressed Avastin and Avegra. D) Cross validation from a combined classifier using stressed and unstressed Avastin and Avegra. E) Best cross validation accuracy for unstressed alone and both conditions (stressed and unstressed) Avastin and Avegra.

incorporating both multiple charge states and stress conditions into a multi-state classifier. Assessing a biosimilar, or generic form of an innovator protein therapeutic, presents significant analytical challenges due to the typical size and complexity of monoclonal antibodies. Comparing higher order structure (HOS) information is particularly challenging without resorting to low-throughput,

high-resolution structural biology techniques. As biosimilars, Avastin and Avegra are highly similar, and classification using CIU data across all charge states of the antibodies resulted in a low degree of differentiation. The feature selection scores were low (Figure 7-3 B), and despite some minor regions of difference, the optimal classifier achieved accuracy of only 87% (Figure 7-3 C). Charge states are not the only states that can be examined using our multi-state CIU data analysis algorithm. Our approach considers data acquired across any state that results in a different CIU pathway, so long as it can be applied equally across the classes being compared. A key attribute monitored in biotherapeutics is the propensity to aggregate during transport and storage, which can be challenging to assess in the laboratory. Early warning methods for aggregation that detect structural changes following various types of stress (for example, heat or oxidation) are thus highly useful (Figure 7-3 A). Avastin and Avegra were stressed by heating to 40 °C and applying orbital shaking at 250 RPM for 4 weeks. CIU data from the stressed samples was incorporated into a multi-state classifier, along with all the observed charge states, for a total of 10 states (5 charge states each from stressed and unstressed conditions). The combined classifier achieved cross validation accuracy above 99% (Figure 7-3 D), indicating very robust differentiation between Avastin and Avegra, significantly outperforming the classifier that used data from all charge states but only compared the unstressed antibodies (Figure 7-3 E). Our analysis indicates that Avastin and Avegra have different structural responses to the stress employed in this study, which can be utilized to develop a classifier capable of robustly distinguishing between them using our multi-state classification method.

7.4 Conclusions

CIU experiments generate rich datasets that have proven capable of distinguishing subtle differences in protein structures. Applying our multi-state classification workflow presented here to analyze all charge states observed within in a CIU experiment maximizes the detection of these subtle differences by incorporating more of the experimental data into the statistical framework for classification. Improvements to the core algorithm have increased the accuracy of the classifiers developed through data standardization and have dramatically reduced the computational requirements for large datasets, enabling the extension of these algorithms to much larger training sets than analyzed previously. Finally, we demonstrate incorporating states other than protein charge states by generating a robust classifier to distinguish Avastin from its biosimilar Avegra. This work indicates the potential of the multi-state classification workflow to be used with a wide range of conditions or perturbations, as any change that causes differences in CIU for an analyte of interest can be incorporated into a classifier using this method. Incorporating differential responses to stimulus into CIU classification has the potential to make CIU sensitive to even more subtle structural differences and provide a rapid and informative workflow for evaluating protein structures.

7.5 References

- (1) Leney, A. C.; Heck, A. J. R. Native Mass Spectrometry: What Is in the Name? *J. Am. Soc. Mass Spectrom.* **2017**, *28* (1), 5–13.
- (2) Lössl, P.; van de Waterbeemd, M.; Heck, A. J. The Diverse and Expanding Role of Mass Spectrometry in Structural and Molecular Biology. *EMBO J.* **2016**, *35* (24), 2634–2657.
- (3) Zhong, Y.; Hyung, S.-J.; Ruotolo, B. T. Ion Mobility–mass Spectrometry for Structural Proteomics. *Expert Rev. Proteomics* **2012**, *9* (1), 47–58.
- (4) Tian, Y.; Ruotolo, B. T. The Growing Role of Structural Mass Spectrometry in the Discovery and Development of Therapeutic Antibodies. *Analyst* **2018**, *143* (11), 2459–2468.
- (5) Terral, G.; Beck, A.; Cianféroni, S. Insights from Native Mass Spectrometry and Ion Mobility-Mass Spectrometry for Antibody and Antibody-Based Product Characterization.

- J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2016**, *1032*, 79–90.
- (6) Bleiholder, C.; Bowers, M. T. The Solution Assembly of Biological Molecules Using Ion Mobility Methods: From Amino Acids to Amyloid β -Protein. *Annu. Rev. Anal. Chem.* **2017**, *10* (1), 365–386.
 - (7) Liko, I.; Allison, T. M.; Yen, H.-Y.; Hopper, J. S. T.; Robinson, C. V. Using Native Mass Spectrometry to Inform Drug Discovery. *Drug Target Rev.* **2017**, *4* (2), 44–47.
 - (8) Mehmood, S.; Allison, T. M.; Robinson, C. V. Mass Spectrometry of Protein Complexes: From Origins to Applications. *Annu. Rev. Phys. Chem.* **2015**, *66* (1), 453–474.
 - (9) Clemmer, D. E.; Hudgins, R. R.; Jarrold, M. F. Naked Protein Conformations: Cytochrome c in the Gas Phase. *J. Am. Chem. Soc.* **1995**, *117* (40), 10141–10142.
 - (10) Dixit, S. M.; Polasky, D. A.; Ruotolo, B. T. Collision Induced Unfolding of Isolated Proteins in the Gas Phase: Past, Present, and Future. *Curr. Opin. Chem. Biol.* **2018**, *42*, 93–100.
 - (11) Hopper, J. T. S.; Oldham, N. J. Collision Induced Unfolding of Protein Ions in the Gas Phase Studied by Ion Mobility-Mass Spectrometry: The Effect of Ligand Binding on Conformational Stability. *J. Am. Soc. Mass Spectrom.* **2009**, *20* (10), 1851–1858.
 - (12) Ruotolo, B. Searching for Conformationally-Selective Small Molecule Therapeutics Using Ion Mobility-Mass Spectrometry (227.1). *FASEB J.* **2014**, *28* (1 Supplement).
 - (13) Rabuck, J. N.; Hyung, S. J.; Ko, K. S.; Fox, C. C.; Soellner, M. B.; Ruotolo, B. T. Activation State-Selective Kinase Inhibitor Assay Based on Ion Mobility-Mass Spectrometry. *Anal. Chem.* **2013**, *85* (15), 6995–7002.
 - (14) Rabuck-Gibbons, J. N.; Keating, J. E.; Ruotolo, B. T. Collision Induced Unfolding and Dissociation Differentiates ATP-Competitive from Allosteric Protein Tyrosine Kinase Inhibitors. *Int. J. Mass Spectrom.* **2018**, *427*, 151–156.
 - (15) Tian, Y.; Han, L.; Buckner, A. C.; Ruotolo, B. T. Collision Induced Unfolding of Intact Antibodies: Rapid Characterization of Disulfide Bonding Patterns, Glycosylation, and Structures. *Anal. Chem.* **2015**, *87* (22), 11509–11515.
 - (16) Tian, Y.; Ruotolo, B. T. Collision Induced Unfolding Detects Subtle Differences in Intact Antibody Glycoforms and Associated Fragments. *Int. J. Mass Spectrom.* **2018**, *425*, 1–9.
 - (17) Pisupati, K.; Tian, Y.; Okbazghi, S.; Benet, A.; Ackermann, R.; Ford, M.; Saveliev, S.; Hosfield, C. M.; Urh, M.; Carlson, E.; et al. A Multidimensional Analytical Comparison of Remicade and the Biosimilar Remsima. *Anal. Chem.* **2017**, *89* (9), 4838–4846.
 - (18) Hernandez-Alba, O.; Wagner-Rousset, E.; Beck, A.; Cianfèrani, S. Native Mass Spectrometry, Ion Mobility, and Collision-Induced Unfolding for Conformational Characterization of IgG4 Monoclonal Antibodies. *Anal. Chem.* **2018**, *90* (15), 8865–8872.
 - (19) Watanabe, Y.; Vasiljevic, S.; Allen, J. D.; Seabright, G. E.; Duyvesteyn, H. M. E.; Doores, K. J.; Crispin, M.; Struwe, W. B. Signature of Antibody Domain Exchange by Native Mass Spectrometry and Collision-Induced Unfolding. *Anal. Chem.* **2018**, *90* (12), 7325–7331.
 - (20) Tian, Y.; Lippens, J. L.; Netirojjanakul, C.; Campuzano, I. D. G.; Ruotolo, B. T. Quantitative Collision-Induced Unfolding Differentiates Model Antibody–drug Conjugates. *Protein Sci.* **2019**, *28* (3), 598–608.
 - (21) Rabuck-Gibbons, J. N.; Lodge, J. M.; Mapp, A. K.; Ruotolo, B. T. Collision-Induced Unfolding Reveals Unique Fingerprints for Remote Protein Interaction Sites in the KIX Regulation Domain. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (1), 94–102.
 - (22) Niu, S.; Ruotolo, B. T. Collisional Unfolding of Multiprotein Complexes Reveals

- Cooperative Stabilization upon Ligand Binding. *Protein Sci.* **2015**, *24* (8), 1272–1281.
- (23) Beveridge, R.; Migas, L. G.; Payne, K. A. P.; Scrutton, N. S.; Leys, D.; Barran, P. E. Mass Spectrometry Locates Local and Allosteric Conformational Changes That Occur on Cofactor Binding. *Nat. Commun.* **2016**, *7*, 12163.
- (24) Migas, L. G.; France, A. P.; Bellina, B.; Barran, P. E. ORIGAMI: A Software Suite for Activated Ion Mobility Mass Spectrometry (AIM-MS) Applied to Multimeric Protein Assemblies. *Int. J. Mass Spectrom.* **2018**, *427*, 20–28.
- (25) Eschweiler, J. D.; Rabuck-Gibbons, J. N.; Tian, Y.; Ruotolo, B. T. CIUSuite: A Quantitative Analysis Package for Collision Induced Unfolding Measurements of Gas-Phase Protein Ions. *Anal. Chem.* **2015**, *87* (22), 11516–11522.
- (26) Polasky, D. A.; Dixit, S. M.; Fantin, S. M.; Ruotolo, B. T.; Daniel A. Polasky, Sugyan M. Dixit, Sarah M. Fantin, and B. T. R. CIUSuite 2: Next-Generation Software for the Analysis of Gas-Phase Protein Unfolding Data. *Anal. Chem.* **2019**, *91* (4), 3147–3155.
- (27) Allison, T. M.; Reading, E.; Liko, I.; Baldwin, A. J.; Laganowsky, A.; Robinson, C. V. Quantifying the Stabilizing Effects of Protein-Ligand Interactions in the Gas Phase. *Nat. Commun.* **2015**, *6*, 8551.
- (28) Sivalingam, G. N.; Yan, J.; Sahota, H.; Thalassinos, K. Amphitrite: A Program for Processing Travelling Wave Ion Mobility Mass Spectrometry Data. *Int. J. Mass Spectrom.* **2013**, *345–347*, 54–62.
- (29) Sivalingam, G. N.; Cryar, A.; Williams, M. A.; Gooptu, B.; Thalassinos, K. Deconvolution of Ion Mobility Mass Spectrometry Arrival Time Distributions Using a Genetic Algorithm Approach: Application to A1-Antitrypsin Peptide Binding. *Int. J. Mass Spectrom.* **2018**, *426*, 29–37.
- (30) Fantin, S. M.; Parson, K. F.; Niu, S.; Liu, J.; Ferguson-Miller, S. M.; Ruotolo, B. T. CIU Classifies Ligand Binding Behavior of Integral Membrane Translocator Protein TSPO. *Under Rev.*
- (31) Vallejo, D. D.; Polasky, D. A.; Kurulugama, R. T.; Eschweiler, J. D.; Fjeldsted, J. C.; Ruotolo, B. T. A Modified Drift Tube Ion Mobility-Mass Spectrometer for Charge Multiplexed Collision Induced Unfolding. *Anal. Chem.* **2019**.
- (32) Seeliger, M. A.; Young, M.; Henderson, M. N.; Pellicena, P.; King, D. S.; Falick, A. M.; Kuriyan, J. High Yield Bacterial Expression of Active C-Abl and c-Src Tyrosine Kinases. *Protein Sci.* **2005**, *14* (12), 3135–3139.
- (33) Haynes, S. E.; Polasky, D. A.; Dixit, S. M.; Majmudar, J. D.; Neeson, K.; Ruotolo, B. T.; Martin, B. R. Variable-Velocity Traveling-Wave Ion Mobility Separation Enhancing Peak Capacity for Data-Independent Acquisition Proteomics. *Anal. Chem.* **2017**, *89* (11), 5669–5672.
- (34) Pedregosa, F.; Varoquax, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O. Scikit-Learn: Machine Learning in Python. *J Mach Learn Res* **2011**, *12* (Oct), 2825–2830.
- (35) Brown, M. T.; Cooper, J. A. Regulation, Substrates and Functions of Src. *Biochim. Biophys. Acta - Rev. Cancer* **1996**, *1287* (2–3), 121–149.
- (36) Parsons, S. J.; Parsons, J. T. Src Family Kinases, Key Regulators of Signal Transduction. *Oncogene* **2004**, *23* (48 REV. ISS. 7), 7906–7909.
- (37) Creedon, H.; Brunton, V. G. . Src Kinase Inhibitors: Promising Cancer Therapeutics? *Crit. Rev. Oncog.* **2012**, *17* (2), 145–159.

Chapter 8 Conclusions and Future Directions

8.1 Conclusions

Full characterization of proteins and the macromolecular complexes into which they associate remains a lofty goal, requiring knowledge of amino acid sequence, post-translational modifications, three-dimensional structure, and interactions and associations into complexes. In the push for increasingly individualized treatments for disease, all of these levels of information will be critical for understanding disease with increased precision and developing new treatments. Mass spectrometry and IM-MS are already utilized at each level of protein information, but many technological gaps remain, particularly in detecting and localizing PTMs and rapidly obtaining high-resolution structures. This thesis develops methods aimed at improving both sequencing and structural characterization experiments using IM-MS.

In Chapters 2 and 3, we evaluate the impact of chemical modification on the fragmentation pathways of intact proteins. Intact proteins fragment through several pathways with sufficiently similar energetics for factors like protein charge state and side chain chemistry to influence which pathway is dominant.^{1,2} In Chapter 2, we show that altering the side chain chemistry by fixing stable, intrinsic charges can improve the degree of fragmentation obtained from protein complexes.³ Despite the utility of this approach, it did not work for all proteins surveyed; in part due to imperfect modification chemistry resulting in splitting a single starting protein peak into many differentially modified states, diluting the useful signal obtained from fragmentation. In some cases, however, it was clear that signal dilution was not the primary

cause of reduced fragmentation following chemical modification. This prompted the work in Chapter 3, in which detailed studies with two different modification reagents demonstrated the factors influencing which fragmentation pathways would be favored. Proteins with low charge mobility (from having more basic residues, particularly Arg, and fewer charging protons) predominately fragment at acidic sites through charge-remote pathways, while proteins with higher charge mobility fragment through charge directed “mobile proton” pathways. Using our chemical modifications to either replace protons with fixed charges or block fragmentation at acidic sites, we were able to push proteins to fragment via either pathway with appropriately tuned chemistry. In addition to predicting and altering the fragmentation pathways of proteins, we found evidence for extensive charge solvation of positive charges in gas-phase protein ions, a finding with implications for gas-phase structural assessments and an indication of the physical limitations facing further improvements top-down sequencing.

A major limitation of IM-MS methods for both sequence and structural analyses is a lack of tools and methods to process the complex data acquired in these methods. A sequencing experiment on a single, purified protein can generate more than 10,000 unique signals in a combined IM-MS spectrum, resulting in prohibitively lengthy data processing times. In Chapter 4, we developed software to process top-down IM-MS data in collaboration with the Nesvizhskii lab.⁴ The ability to process complex top-down IM-MS datasets in seconds was fundamental to the work performed in Chapters 2 and 3, and is now available for other researchers to use in their own IM-MS work.

Data processing capabilities also limited structural analyses with CIU. Development of a suite of software tools to analyze CIU data is discussed in Chapter 5. This software, CIUSuite 2, aimed to address several key limitations in existing analysis methods, including low acquisition

speed from signal averaging required to prevent noise from impacting the analysis, inability to handle datasets contaminated by chemical noise (for example, in membrane protein analyses), and a need for additional automation of stability shift and screening methods.⁵ Data pre-processing methods, including two-dimensional smoothing and interpolation, combined with the stability shift tools to reduce the acquisition time needed for reproducible analysis of glycosylation state of biotherapeutic proteins by a factor of 60, reducing data acquisition time from hours to minutes. Gaussian fitting and noise removal enabled stability shift and classification analyses of membrane protein data that had proven extremely difficult to analyze with previous tools.

Chapters 6 and 7 investigate improvements to CIU by incorporating data from multiple charge states and hardware and software developments required to implement that capability. Typical CIU experiments select only a single charge state of a protein ion for analysis, as simultaneous analysis of multiple charge states can result in data contaminated by charge stripping, in which a protein ion at one charge state loses a charged adduct during the experiment, altering its charge state and contaminating the data for the new state it has become. In Chapter 6, we investigate this phenomenon in detail and develop a predictive algorithm that uses the degree of adduct loss and propensity for adducts to leave as a charged group to estimate the expected amount of charge stripping and potentially correct for it. This enables simultaneous collection of CIU data for all observed charge states, increasing the amount of data collected per unit time. These methods were then employed on a modified Agilent 6560 IM-MS instrument and used to characterize monoclonal antibody samples by CIU. Chapter 7 explores improved data processing methods to take advantage of the ability to acquire data from multiple charge states for enhanced classification and screening methods. By combining data from multiple

charge states, we generated classifiers with substantially improved accuracy over data from a single charge state, without increasing the time required to acquire or process the data.

8.2 Future Directions

8.2.1 Charge manipulation for optimized fragmentation of intact proteins

The chemical modification methods developed in Chapters 2 and 3 make it clear that the charge mobility (the ratio of charging protons to charge-sequestering groups) is the primary factor in determining the fragmentation pathway of proteins in slow heating activation methods. Moving fragmentation away from charge remote pathways, which tend to fragment the protein specifically at a few residues, can result in significant improvements in sequence coverage.

There appears to be an optimum level of charge mobility, as too little results in fragmentation primarily at acidic residues, while too much can result in fragmentation primarily at Proline.^{2,6}

Using the charge mobility approximations developed in Chapter 3, it could be possible to predict optimal charge states for fragmentation of given protein sequences and use charge manipulation to achieve those states. Chemical modification offers one pathway to altering charge mobility, for example, by affixing stable charges to groups that are typically not charged in solution (*i.e.* acidic or neutral residues).⁷ Development of stable, intrinsically-charged reagents targeting acidic or neutral residues may prove challenging, in which case, solution modifiers for charge manipulation may prove more effective.

Protein charge states have been manipulated with a range of solution modifiers,⁸⁻¹⁰ as well as gas-phase ion-ion and ion-neutral reactions.¹¹⁻¹³ Solution modifiers would require the starting and desired charge states of a protein to be known in advance, making that approach suitable for targeted analyses only. However, gas-phase charge manipulation could be used online with sufficiently advanced instrument control software. An initial sequencing event under

default conditions could yield the protein sequence and charge state, determining the optimal charge manipulation for a rapid follow-up sequencing experiment and then using gas phase ion chemistry to achieve the desired charge state. Current approaches to maximize sequence coverage typically employ multiple activation methods and/or energies to generate complementary fragmentation; adding a varying charge manipulation step could improve the contribution of collisional activation methods in these experiments and help increase overall sequence coverage for top-down proteomics.

Ultimately, the future of intact protein sequencing is likely to require multiple stages of activation, first breaking large proteins or protein complexes into medium sized fragments, then performing a second stage of fragmentation on those ions to generate comprehensive sequence coverage. Intact protein fragmentation can generate immensely complex product ion spectra, which could prevent a subsequent fragmentation step from being effective, due to dilution of signal into too many product ion channels. Utilizing low charge state CID as the first stage of fragmentation, in which proteins fragment relatively specifically at acidic residues, could be a strategy analogous to enzymatic cleavage with trypsin in bottom-up proteomics, essentially performing a middle-down experiment entirely within the instrument. Intelligent strategies for choosing the initial fragment ions for secondary sequencing will likely still be critical to coupling such an approach to separations methods, potentially requiring online annotation of the initial fragments to protein sequence to determine the regions requiring additional coverage and/or containing PTMs to be localized.

8.2.2 Development of top-down sequence annotation software for PTM localization

Analysis of top-down proteomics data involves two primary steps: “pre-processing,” or converting raw MS data into a list of isotopic clusters (a peak list), and sequence annotation, or matching the detected peaks to fragments of the amino acid sequence of a protein. The software developed in Chapter 4 provides pre-processing for IM-MS data but not sequence annotation, since existing software packages¹⁴⁻²⁰ were capable of annotating peak lists generated after IM-MS analysis, just not generating a peak list from IM-MS data directly. However, to our knowledge, existing software tools for top-down sequence annotation focus on protein and PTM identification, leaving PTM localization primarily to manual analysis and post-processing. Identifying the protein(s), sequence variant(s), and PTM(s) present in a sample is crucial for any proteomics analysis, but can generally be accomplished using bottom-up proteomics for protein sequence and variants, or high resolution intact mass analysis for PTM presence. The key information that top-down proteomics provides in comparison to bottom-up proteomics and high-resolution intact mass is proteoform identification, through the location of PTMs along the amino acid sequence, but this capability is currently underserved by existing software.

To annotate the sequences of chemically modified proteins, I developed a custom software package, which was used to annotate the data presented in Chapters 2 and 3. Because chemical modification sites occur throughout the protein sequence, and their presence or absence was a key factor in assessing fragmentation, the software package was developed with location of PTMs in mind. To generalize the current capabilities to any PTM, a localization algorithm could be developed that uses the presence or absence of a modification from fragment ions along a region of protein sequence to generate a location score for the modification. As technology for top-down proteomics continues to improve, generation of sufficient fragment density to

accurately localize PTMs (and chemical modifications) is likely to become increasingly common, creating a need for software to accurately assign modifications to specific sites in a protein sequence automatically.

8.2.3 Advanced CIU software: width analysis and direct incorporation of mass information

The Gaussian fitting and noise removal modules discussed in Chapter 5 have proven highly useful in cleaning up membrane protein data for analysis, but have the potential to provide additional information about CIU data that is currently not explored. Previous work has used deviations from the expected width of peaks in the mobility dimension in IM-MS experiments (based on the resolution of the device) can be related to conformational dynamics on the timescale of the mobility separation.²¹ Protein structural dynamics are often crucial to understanding function, but remain understudied due, in part, to a lack of tools capable of assessing dynamics, particularly in a rapid fashion. Gaussian fitting of CIU data provides an estimated peak width for all components detected, which is used in Chapter 5 to filter chemical noise from protein signal of interest. These peak widths could be combined with the width to dynamics relationship developed previously²¹ to build a conformational dynamics analysis module in CIUSuite 2. The inclusion of peak width in CIU analysis could provide additional detail on protein stability and structural changes in response to perturbations, showing, for example, increased dynamics as a protein structure begins to be disrupted by heat stress.

Improvements to the Gaussian fitting algorithm are also possible through improved statistics for estimating the correct number of components to fit, or through the inclusion of the *m/z* information collected in CIU data, which is currently ignored in data analysis by CIUSuite 2. Determining the optimal number of Gaussian components to fit to a given arrival time

distribution is a challenging task accomplished in the current software using an empirically-derived penalty function for overlapping peaks. This may not be sufficient for all data, as differences in peak characteristics (*e.g.* from data collected with different IM resolution) may result in poor estimation. Bayesian or Akaike information criteria are a common approach to estimating the optimal number of parameters for model fitting, and could be employed for improved component number estimation. In some cases, chemical noise can be filtered by peak width alone, but this is not always sufficient, particularly when the chemical “noise” is a contaminant protein possessing similar width characteristics to the analyte of interest. This has proven particularly problematic in the analysis of membrane proteins encapsulated in protein nanodiscs, as the scaffolding proteins cannot be filtered by width alone when they overlap in m/z with the membrane protein of interest. Performing the Gaussian fit on the raw data in 2D (including both IM and MS information, as opposed to IM alone in the current software) could provide greatly increased capability to distinguish peaks of interest from contaminants and improve the effectiveness of noise removal. Fitting peaks in the m/z dimension would also provide automated peak detection, removing the need for manual peak selection and extraction of mobility spectra from the raw data prior to CIU analysis. Manual selection of peaks is one of the few remaining manual steps in current CIU analyses, and introduces significant challenges in complex datasets where adduct loss can result in a changing mass across collision voltages. Finally, including the mass dimension in CIU analyses would enable tracking of dissociation products (for example, of a ligand in ligand-bound CIU experiments) automatically, a common data analysis method that is currently performed manually.

8.2.4 High-throughput CIU with microfluidics for rapid sample introduction and automated acquisition

Improvements in CIU data processing discussed in Chapter 5 have made very rapid acquisition of CIU data feasible. For the glycosylated monoclonal antibody data examined in Chapter 5, for example, one second of data acquisition resulted in an S/N ratio in excess of 10,000, indicating that sub-second acquisitions should still generate data of sufficient quality for robust analysis. Each mobility separation takes on the order of 20-30 ms, meaning that 30-50 mobility separations are summed in a 1 s acquisition. Reducing acquisition time at each collision voltage by an order of magnitude, to 3-5 mobility separations and 0.1 s total scan time, would result in complete CIU fingerprints in less than 5 s. At these speeds, our current sample introduction method (samples manually loaded into capillary tips for static infusion) would be the limiting factor in throughput, as it can take up to a minute to switch to a new sample following completion of a fingerprint. Automating the instrument acquisition method would also be essential, as it would not be possible for a human operator to change voltage settings every 0.1 s.

Recent work in collaboration with the Kennedy lab (UM Chemistry) has explored the idea of coupling a segmented flow microfluidic sample introduction method with CIU for high-throughput analyses. In segmented flow, aqueous droplets containing a sample are separated by a chemically inert oil phase, typically using perfluorinated oils. Droplet trains can be rapidly generated from well plates by covering the plate in the oil and aspirating the contents of each well in succession with an oil layer in between. When coupled to an ESI source, the aqueous droplets ionize while the oil phase does not, carrying only the samples into the mass spectrometer. Each droplet is typically only analyzed for a few seconds, so using droplets for CIU sample introduction would require the extremely rapid CIU enabled by the new data

processing methods. To acquire CIU data at this speed, we updated the “method editor” script used to generate methods for CIU data collection to enable sub-second acquisition times and with a variable delay to account for the oil phase between segmented droplets.

With automated sample introduction via segmented flow and automated instrument control through the method editor, true high-throughput CIU analyzing up to 10 samples per minute appears achievable. Combined with the consideration of all charge states explored in Chapters 6 and 7, sensitive screening methods could be developed to operate on this timescale. Classification methods, in which training CIU data is used to determine a few key voltages that most distinguish the analytes of interest, are particularly suited to this high-throughput format, as individual samples could be screened using a reduced voltage set in less than a second each. The ultimate limit on the speed of CIU experiments is the time required to perform the IM separation, which is on the order of 10-100 ms with current IM methods. Using a single IM separation for each step of a fingerprint could enable full CIU acquisitions to be accomplished at a rate of 1-2 per second, or reduced datasets for input to a classifier at up to 10 per second. This level of throughput could make CIU competitive with classical high-throughput screening techniques while providing increased information content and the capability to evaluate polydisperse samples.

8.3 References

- (1) Palzs, B.; Suhal, S. Fragmentation Pathways of Protonated Peptides. *Mass Spectrom. Rev.* **2005**, *24* (4), 508–548.
- (2) 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.
- (3) Polasky, D. A.; Lermyte, F.; Nshanian, M.; Sobott, F.; Andrews, P. C.; Loo, J. A.; Ruotolo, B. T. Fixed-Charge Trimethyl Pyridinium Modification for Enabling Enhanced

- Top-Down Mass Spectrometry Sequencing of Intact Protein Complexes. *Anal. Chem.* **2018**, *90* (4), 2756–2764.
- (4) Avtonomov, D. M.; Polasky, D. A.; Ruotolo, B. T.; Nesvizhskii, A. I. IMTBX and Grppr: Software for Top-Down Proteomics Utilizing Ion Mobility-Mass Spectrometry. *Anal. Chem.* **2018**, *90* (3), 2369–2375.
 - (5) Polasky, D. A.; Dixit, S. M.; Fantin, S. M.; Ruotolo, B. T. CIUSuite 2: Next-Generation Software for the Analysis of Gas-Phase Protein Unfolding Data. *Anal. Chem.* **2019**, *91* (4), 3147–3155.
 - (6) Haverland, N. A.; Skinner, O. S.; Fellers, R. T.; Tariq, A. A.; Early, B. P.; LeDuc, R. D.; Fornelli, L.; Compton, P. D.; Kelleher, N. L. Defining Gas-Phase Fragmentation Propensities of Intact Proteins During Native Top-Down Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (6), 1203–1215.
 - (7) Krusemark, C. J.; Frey, B. L.; Belshaw, P. J.; Smith, L. M. Modifying the Charge State Distribution of Proteins in Electrospray Ionization Mass Spectrometry by Chemical Derivatization. *J. Am. Soc. Mass Spectrom.* **2009**, *20* (9), 1617–1625.
 - (8) Catalina, M. I.; van den Heuvel, R. H. H.; van Duijn, E.; Heck, A. J. R. Decharging of Globular Proteins and Protein Complexes in Electrospray. *Chem. - A Eur. J.* **2005**, *11* (3), 960–968.
 - (9) Lomeli, S. H.; Yin, S.; Ogorzalek Loo, R. R.; Loo, J. A. Increasing Charge While Preserving Noncovalent Protein Complexes for ESI-MS. *J. Am. Soc. Mass Spectrom.* **2009**, *20* (4), 593–596.
 - (10) Going, C. C.; Williams, E. R. Supercharging with M-Nitrobenzyl Alcohol and Propylene Carbonate: Forming Highly Charged Ions with Extended, Near-Linear Conformations. *Anal. Chem.* **2015**, *87* (7), 3973–3980.
 - (11) Frey, B. L.; Krusemark, C. J.; Ledvina, A. R.; Coon, J. J.; Belshaw, P. J.; Smith, L. M. Ion-Ion Reactions with Fixed-Charge Modified Proteins to Produce Ions in a Single, Very High Charge State. *Int. J. Mass Spectrom.* **2008**, *276* (2–3), 136–143.
 - (12) Laszlo, K. J.; Munger, E. B.; Bush, M. F. Folding of Protein Ions in the Gas Phase after Cation-to-Anion Proton-Transfer Reactions. *J. Am. Chem. Soc.* **2016**, *138* (30), 9581–9588.
 - (13) Stephenson, J. L.; McLuckey, S. A. Ion/Ion Reactions in the Gas Phase: Proton Transfer Reactions Involving Multiply-Charged Proteins. *J. Am. Chem. Soc.* **1996**, *118* (31), 7390–7397.
 - (14) 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), 703–714.
 - (15) 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. *Mol. Cell. Proteomics* **2010**, *9* (12), 2772–2782.
 - (16) 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), M111.008524.
 - (17) 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.
 - (18) D. LeDuc, R.; L. Kelleher, N. Using ProSight PTM and Related Tools for Targeted

Protein Identification and Characterization with High Mass Accuracy Tandem MS Data. In *Current Protocols in Bioinformatics*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2007; Vol. 19, p 13.6.1-13.6.28.

- (19) Leduc, R. D.; Fellers, R. T.; Early, B. P.; Greer, J. B.; Thomas, P. M.; Kelleher, N. L. The C-Score: A Bayesian Framework to Sharply Improve Proteoform Scoring in High-Throughput Top down Proteomics. *J. Proteome Res.* **2014**, *13* (7), 3231–3240.
- (20) 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.; et al. Informed-Proteomics: Open-Source Software Package for Top-down Proteomics. *Nat. Methods* **2017**, *14* (9), 909–914.
- (21) Dixit, S. M.; Ruotolo, B. T. A Semi-Empirical Framework for Interpreting Traveling Wave Ion Mobility Arrival Time Distributions. *J. Am. Soc. Mass Spectrom.* **2019**, 1–11.

Appendices

I. Chapter 2 Supporting Information

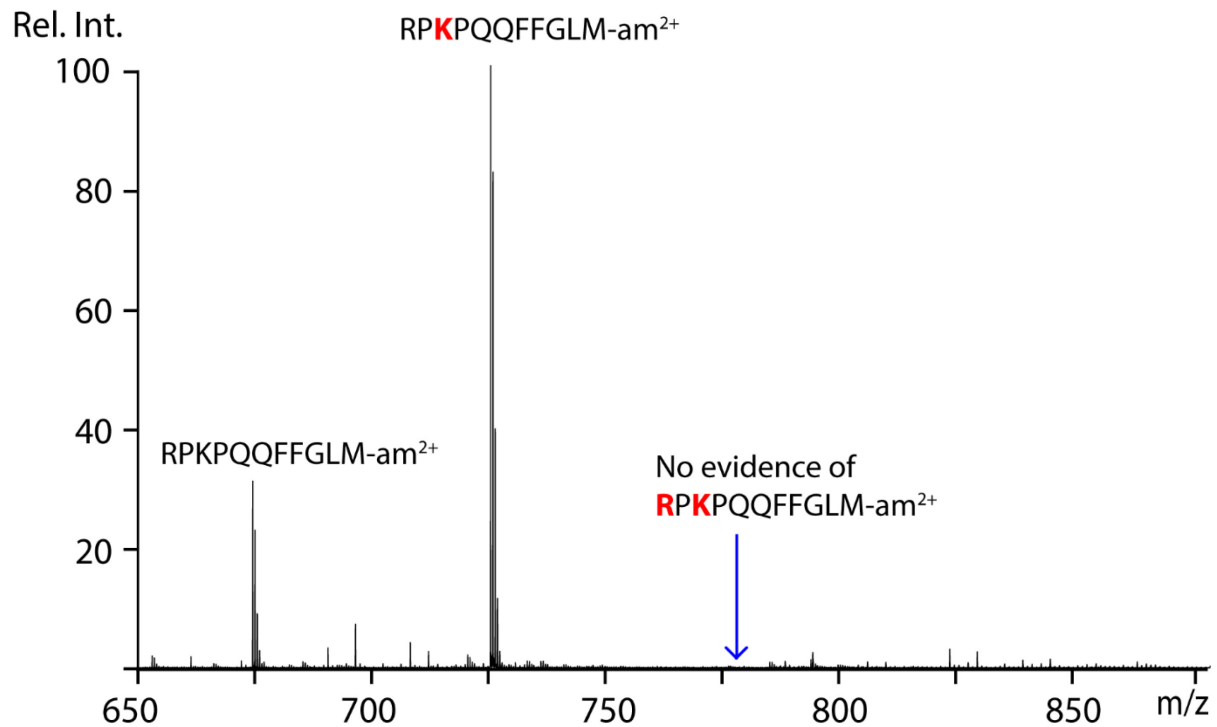


Figure I-1 TMP modification of Substance P peptide (RPKPQQFFGLM-amide). Modification was performed under the same conditions as for proteins (see experimental), with TMP to Lysine molar ratio adjusted accordingly. The majority of the Substance P is modified at Lysine (red highlighted, center), but a significant minority remains unmodified after 24 hours. No evidence of modification at the n-terminal Arginine is observed, despite the availability of a primary amine at the terminus. No evidence of any other side reactions is observed.

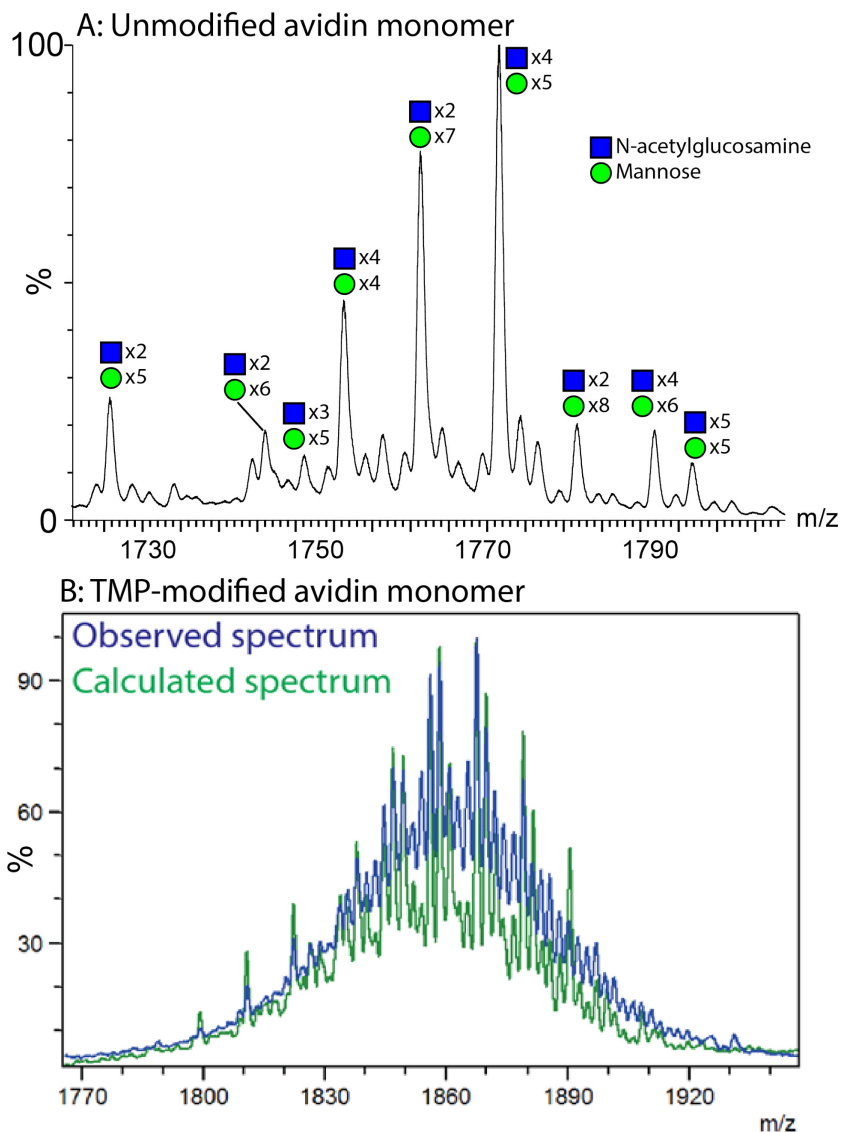


Figure I-2 A) Mass spectrum of a monomer of unmodified avidin, displaying the heterogeneity of its glycosylation. This heterogeneity, combined with some variation in the number of TMP modifications observed on various monomers results in the spectrum shown in B. B) Modeled (calculated, green) and experimentally observed (blue) mass spectra for TMP-modified avidin monomer. The model was obtained by sequentially adding the m/z of a TMP modification to the spectrum of an unmodified avidin monomer to generate predicted spectra for each possible number of TMP modifications considered (0 to 15). A linear system of equations representing each possible modification state was solved for the optimal distribution of intensity (i.e. how much intensity in each modification state best recapitulates the observed mass spectrum) using the partial conjugate gradient method, resulting in the bar graph in Figure I-3 and the calculated spectrum above. The major peaks in the observed spectrum are well represented in the calculated spectrum, though there is some discrepancy in the intensities. This is likely due to differences in peak width due to increased adduction in the TMP-modified data, a distinction which is not of great importance for this simple estimation of modification efficiency.

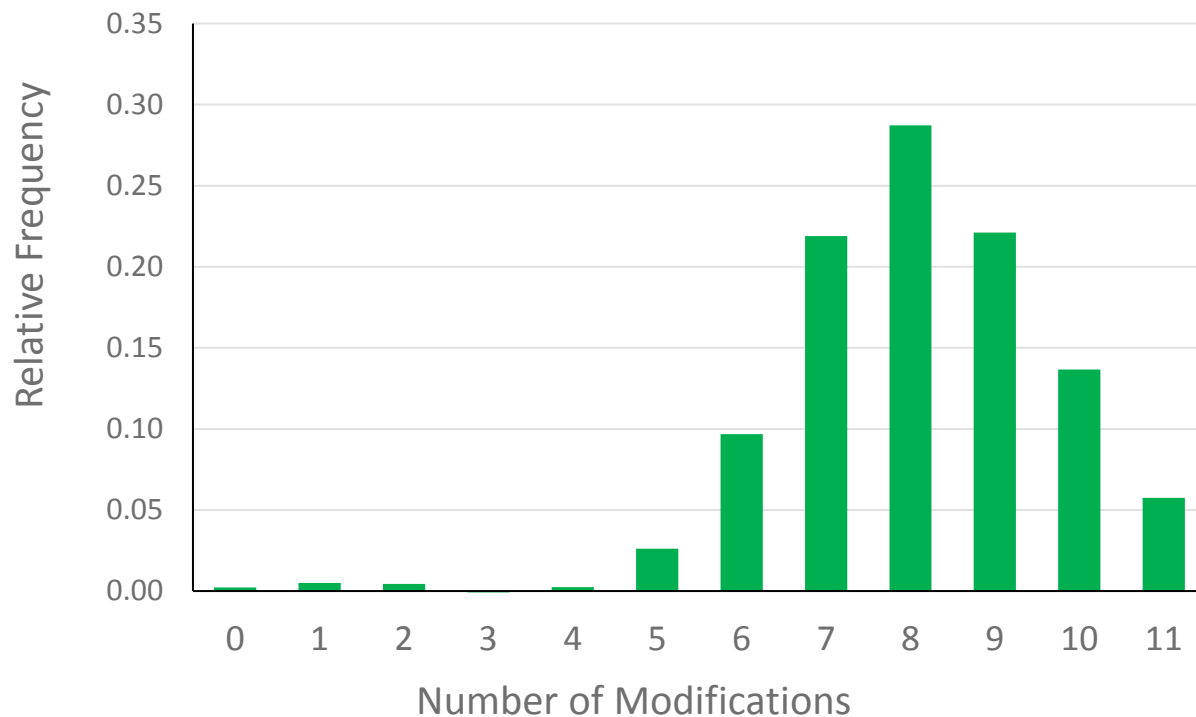


Figure I-3 Number of TMP modifications to a single avidin monomer from model analysis described in Figure I-2. There are 10 canonical sites (primary amines) available for modification in the form of nine lysine residues and the N-terminus, as well as at least one site on the glycosylated side chain of Asn-17 available for TMP modification. Data indicates generally good incorporation of the TMP tag with some variability in the final number of modifications. Achieving complete incorporation of tags is highly challenging in native, folded protein complexes as not all reaction sites are exposed to solvent, likely the primary cause of the distribution of modification states observed here. Aside from the glycan, no side reactions are observed with any non-Lysine amino acid residues.

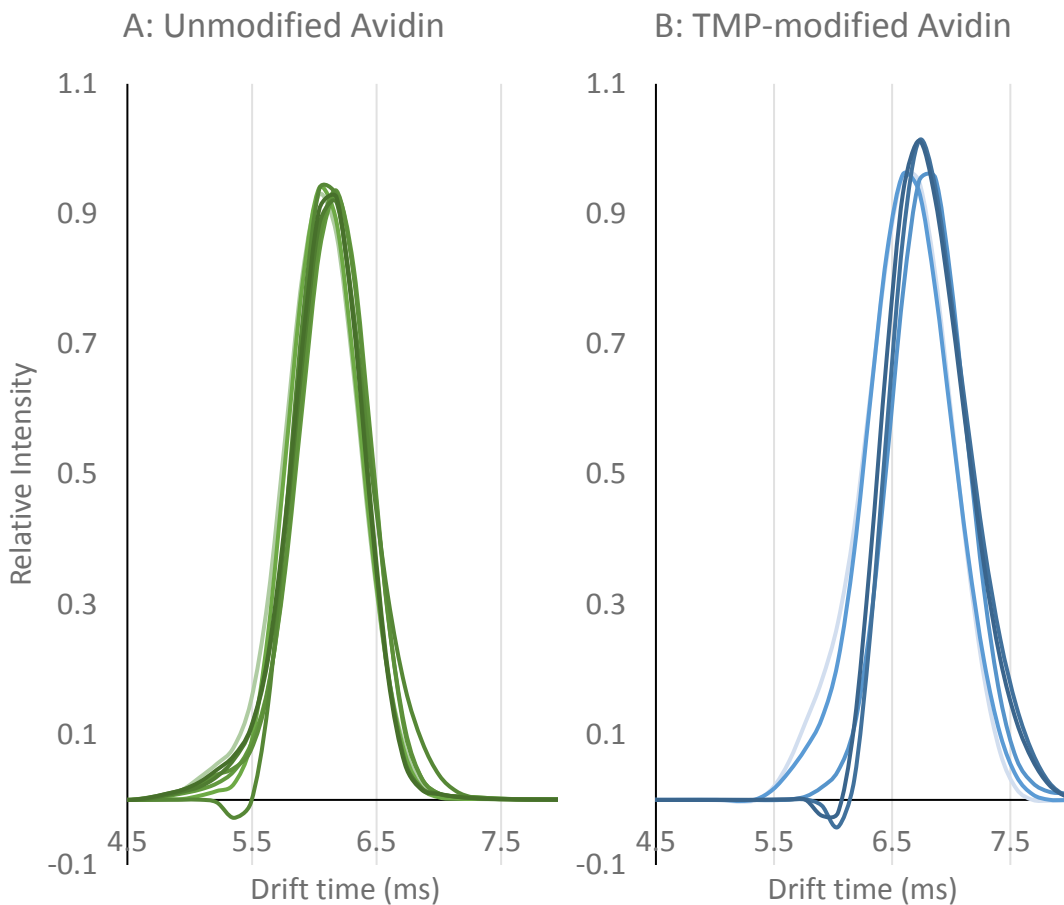


Figure I-4 Ion mobility arrival time distributions of replicate analyses of the 17^+ charge state of A) the unmodified (green lines, m/z 3750) and B) TMP-modified (blue lines, m/z 3960) avidin tetramer with no excess collisional activation energy. Both unmodified and TMP-modified distributions display a single peak, generally indicating the presence of a single, native-like global structure. The measured mass of unmodified Avidin tetramer was 63.9 kDa, while the measured mass of the TMP-modified Avidin was 67.5 kDa, with variation of approx. ± 0.5 kDa between labeling reactions. The average centroid drift time for unmodified Avidin was 6.2 ms and the average for TMP-modified Avidin was 6.7 ms. The resulting 7% increase in drift time is in line with expectations from previous work in which Avidin was crosslinked with various reagents, resulting in 5-10% increases in CCS without perturbing structure¹. We attribute the slight variations in drift time in TMP-modified Avidin to differences in final mass as a result of differences in labeling efficiency across several reaction trials.

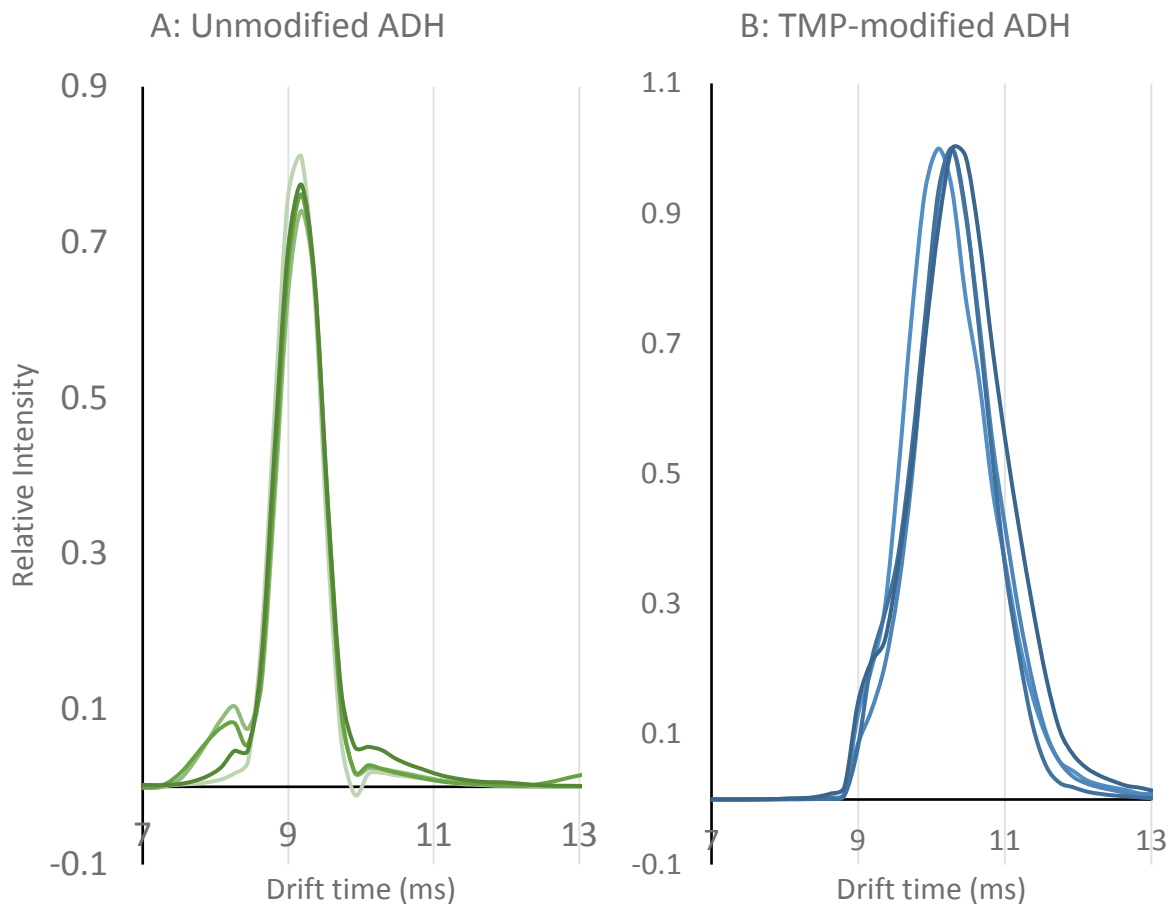


Figure I-5 Ion mobility arrival time distributions of replicate analyses of the 26^+ charge state of the A) unmodified (green lines, m/z 5675) and B) TMP-modified (blue lines, m/z 6100) ADH tetramer with no collisional activation energy applied. Both unmodified and TMP-modified distributions display a single peak, generally indicating the presence of a single, native-like global structure. The measured mass of unmodified ADH tetramer was 147.8 kDa, while the measured mass of the TMP-modified ADH was 158.3 kDa, with variation of approx. ± 1 kDa between labeling reactions. The average centroid drift time for unmodified ADH was 9.2 ms and the average for TMP-modified ADH was 10.3 ms. The resulting 11% increase in drift time is in line with expectations from previous work in which several globular proteins were crosslinked with various reagents, resulting in 5-10% increases in CCS without perturbing structure¹. We attribute the slight variations in centroid drift time and increased peak broadness in TMP-modified ADH to differences in final mass as a result of differences in labeling efficiency.

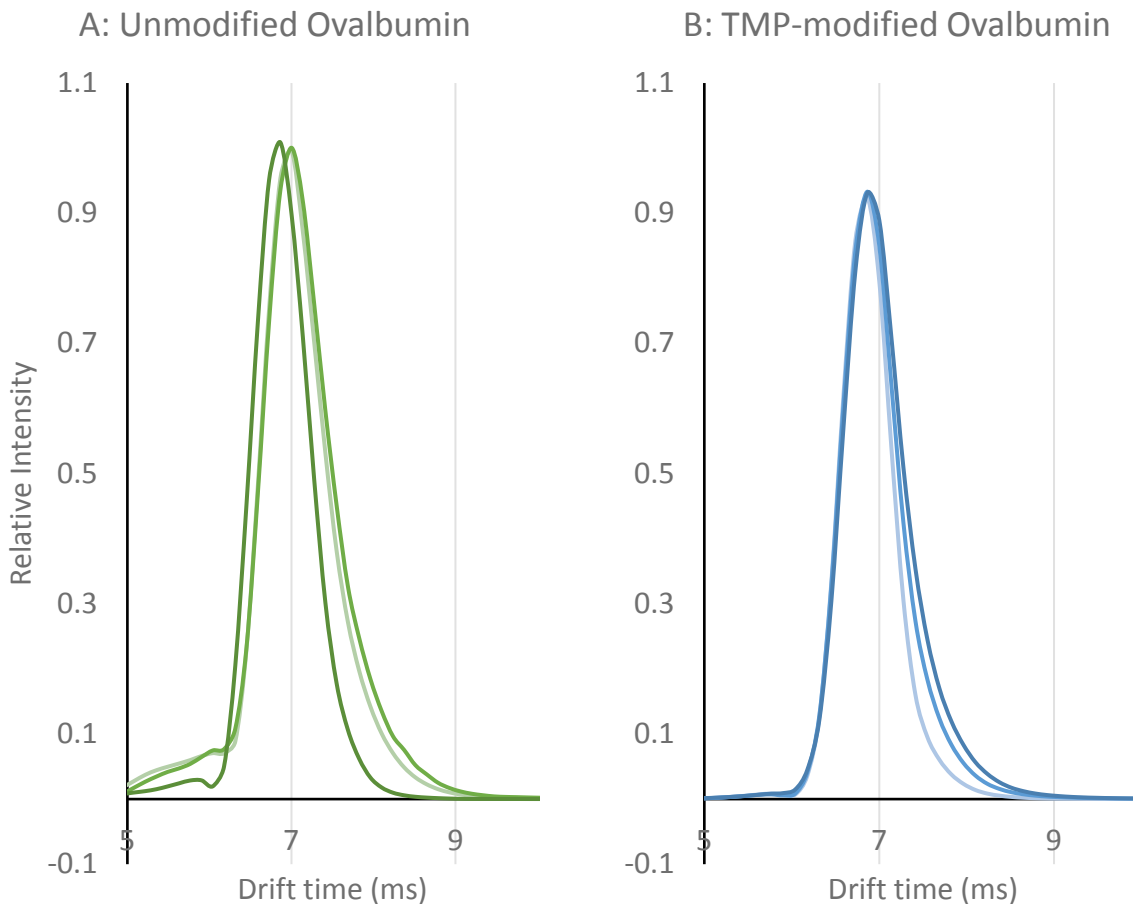


Figure I-6 Ion mobility arrival time distributions of replicate analyses of the A) unmodified 12^+ (green lines, m/z 3720) and B) TMP-modified 13^+ (blue lines, m/z 3780) Ovalbumin monomer with no collisional activation energy applied. Note that different charge states were used for this analysis as the modified and unmodified charge state distributions were very narrow and did not share any high abundance peaks. Both unmodified and TMP-modified distributions display a single peak, generally indicating the presence of a single, native-like global structure. The measured mass of unmodified Ovalbumin monomer was 44.7 kDa, while the measured mass of the TMP-modified Ovalbumin monomer was 49.0 kDa, with variation of approx. ± 0.5 kDa between labeling reactions. The average centroid drift time for unmodified Ovalbumin was 6.9 ms and the average for TMP-modified Ovalbumin was 6.9 ms. As the TMP-modified Ovalbumin here is at a higher charge state than the unmodified, it will have a larger collision cross section value despite the similarity in drift time, but, as in the case of ADH and Avidin, the increase in cross section is on par with the 5-10% observed following labeling of protein complexes in previous work¹.

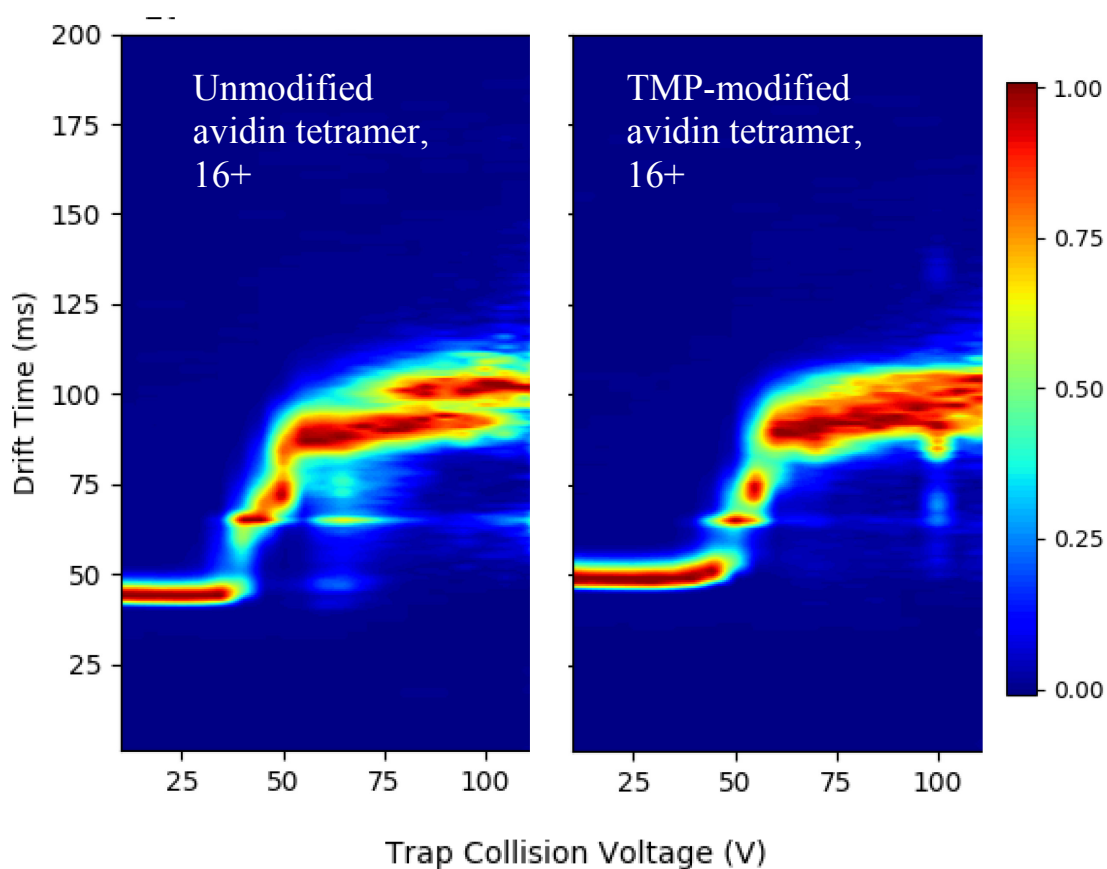


Figure I-7 Collision induced unfolding (CIU) profiles of the same charge state of unmodified (left) and TMP-modified (right) avidin tetramer. The profiles show the increases in drift time (unfolding) as a function of applied collisional energy (trap collision voltage, x-axis). Notably, both modified and unmodified profiles show the same five features, albeit at slightly different intensities and onset voltages. This broad similarity is indicative of similar global structure (with slight changes possible to local structure), as we would expect following chemical modification with TMP. The similarity of the CIU profiles observed, combined with the same initial drift times (Figure I-4 - Figure I-6), indicates that TMP modification does not cause measurable structural changes in the protein complexes analyzed in this work.

II. Chapter 3 Supporting Information

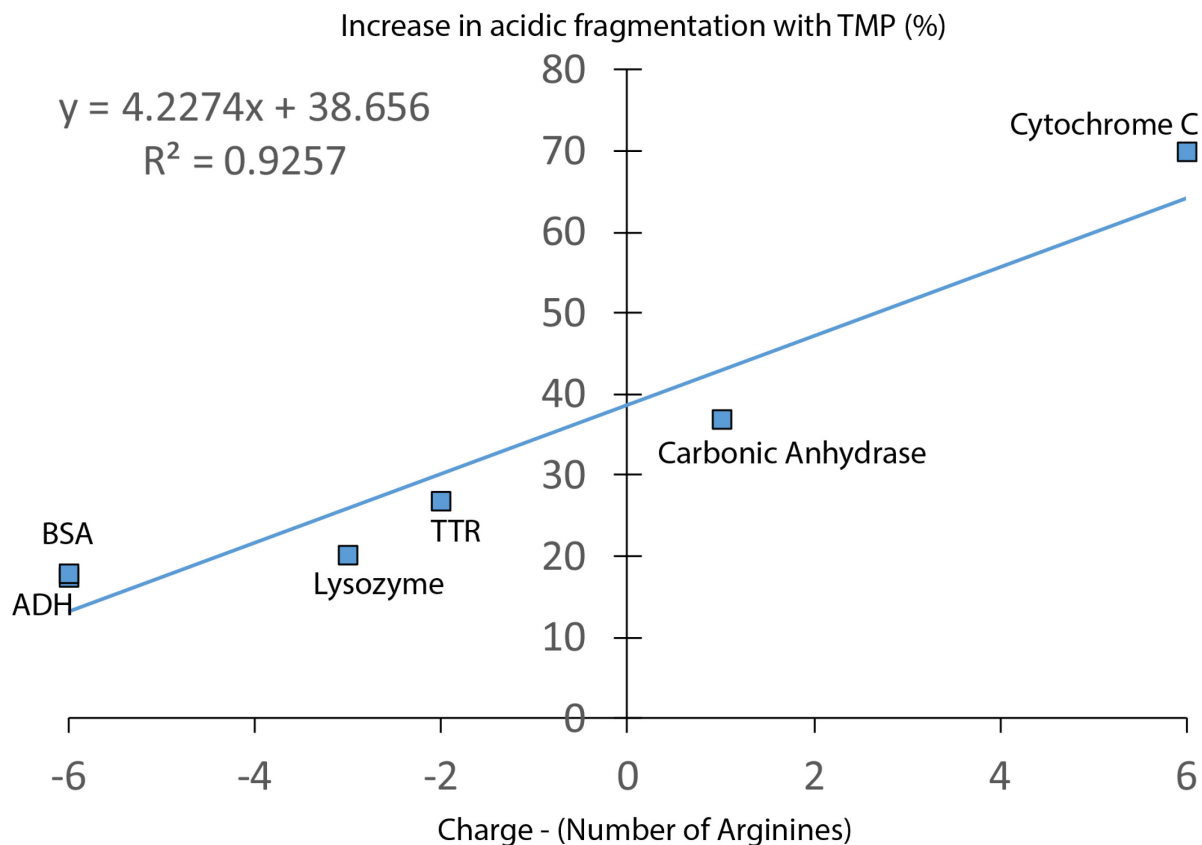
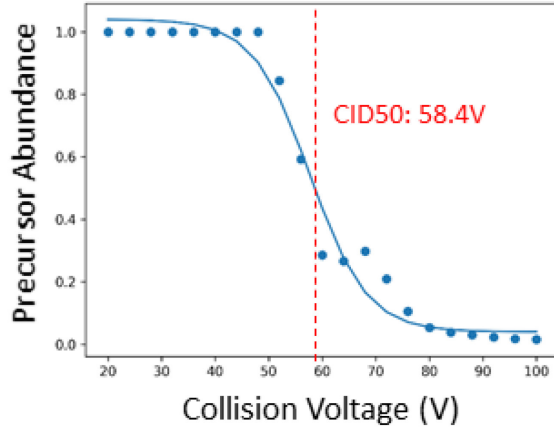


Figure II-1 Linear trend fit (alternative fit to Figure 3-1G). As in Figure 3-1G, the increase in fragmentation at acidic residues following derivatization by TMP is shown as a function of a crude charge mobility “score,” derived by subtracting the number of arginine residues present from the observed charge of the protein. Proteins with extremely low charge mobility scores (SAP, -21 and aldolase, -35) are excluded from this analysis, as the linear trend fitted would place them well below 0%.

A) Unmodified



B) EDC-modified

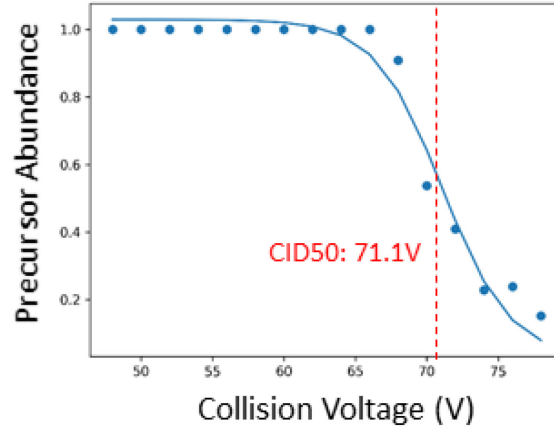


Figure II-2 CID-50 values for EDC-modified SERF vs unmodified SERF. Precursor abundance (height of precursor peak vs height of base peak) as a function of collision voltage for unmodified (A) and EDC-modified (B) SERF. The 8+ charge state is used in both cases. EDC-modified SERF requires substantially higher activating voltage to cause dissociation, with a 50% dissociation point for the intact protein precursor (CID-50) occurring at 71V, as opposed to 58V for the unmodified protein.

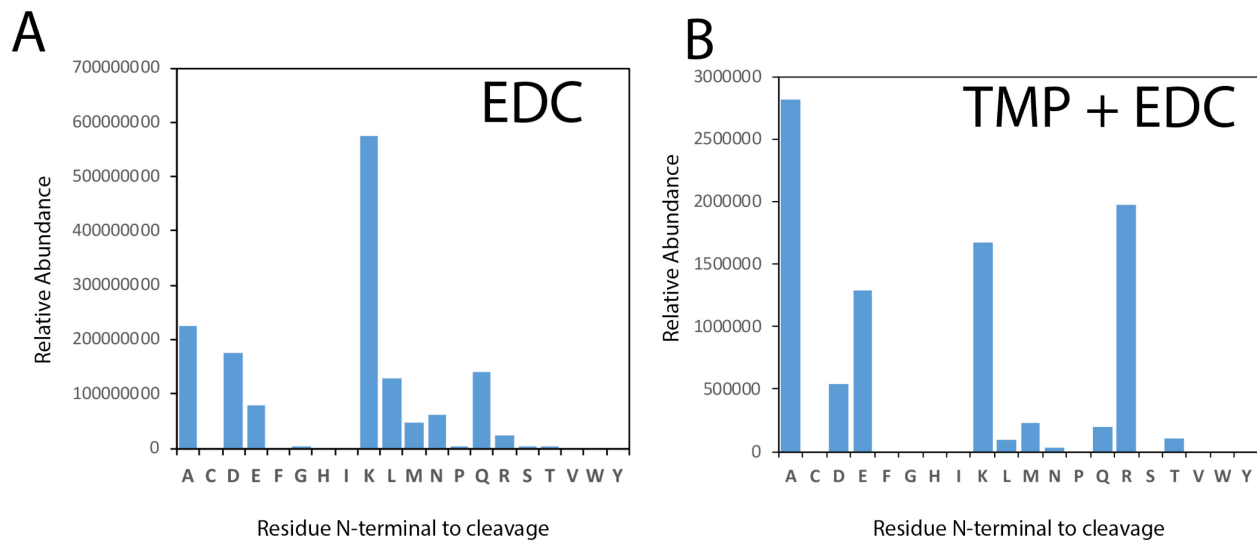


Figure II-3 Fragmentation propensity maps by amino acid summed across all charge states of chemically modified SERF. A) Summed fragmentation propensity for SERF with acidic residues capped by glycinamide coupled with EDC. B) Summed fragmentation propensity across all charge states of SERF with both acidic residues capped by glycinamide/EDC and lysine residues derivatized with TMP.

Table II-1 Mass assignments for EDC-modified 6+ SERF (Figure 3-3B). Mass assignments and error for all peaks annotated in Figure 3B. Masses are calculated as average rather than monoisotopic masses, as peaks are not fully isotopically resolved. Mass errors are thus significantly higher than would be expected for monoisotopic peak assignments. Theoretical m/z values are calculated from the mass of SERF and the added glycinamides, crosslinks, and/or adducts present as detailed in the table.

Observed m/z	Theoretical m/z (average mass)	Mass Error (Da)	Number of Glycinames added	Number of crosslinks	Adducts/other modifications present
1403.8	1403.965	0.17	7	2	
1406.7	1406.968	0.27	8	1	
1413.3	1413.310	0.01	8	2	
1416.2	1416.312	0.11	9	1	
1419.0	1419.315	0.32	10	0	
1425.7	1425.657	0.01	10	1	
1428.5	1428.659	0.13	11	0	
1431.3	1431.326	0.05	11	0	Oxidation
1432.6	1432.491	-0.11	11	0	Sodium adduct
1435.3	1435.001	-0.26	11	1	
1438.0	1438.004	0.01	12	0	
1440.7	1440.670	0.01	12	0	Oxidation
1442.0	1441.836	-0.16	12	0	Sodium adduct
1445.0	1445.667	0.67	12	0	2 Sodium adducts

Table II-2 Mass assignments for TMP + EDC-modified SERF (Figure 3-3G). Unlike Table II-1 Mass assignments for EDC-modified 6+ SERF (Figure 3-3B). Mass assignments and error for all peaks annotated in Figure 3B. Masses are calculated as average rather than monoisotopic masses, as peaks are not fully isotopically resolved. Mass errors are thus significantly higher than would be expected for monoisotopic peak assignments. Theoretical m/z values are calculated from the mass of SERF and the added glycinamides, crosslinks, and/or adducts present as detailed in the table., data presented in this table were collected on an Orbitrap Fusion Lumos instrument at high resolution, enabling monoisotopic peak assignments. Intact mass assignments were determined from monoisotopic peaks chosen by deconvolution in BioPharmaFinder 3.0 (see methods for details). In some cases, incorrect monoisotopic peaks were chosen and correct manually, as indicated in the final column of the table. Following correction, all mass assignments are within 5 ppm of each other.

Observed m/z	Theoretical m/z (monoisotopic mass)	Mass Error (Da)	Number of TMPs added	Number of Glycinames added	Number of crosslinks	Isotope off by X corrected
872.2469	872.2206	-0.026	10	10	1	
874.9774	874.9492	-0.028	11	8	2	
879.1615	879.1341	-0.027	10	12	0	2
881.7096	881.6808	-0.029	11	10	1	
883.2545	883.2272	-0.027	11	11	0	-1
884.4385	884.4094	-0.029	12	8	2	
885.9842	885.9558	-0.028	12	9	1	-1
888.4412	888.4124	-0.029	11	12	0	
891.1715	891.1410	-0.031	12	10	1	
892.7155	892.6874	-0.028	12	11	0	-1
893.9000	893.8695	-0.030	13	8	2	
895.4452	895.4160	-0.029	13	9	1	-1
897.9029	897.8726	-0.030	12	12	0	
899.2655	899.2366	-0.029	13	9	2	3
900.6326	900.6012	-0.031	13	10	1	
902.1765	902.1476	-0.029	13	11	0	-1
903.3603	903.3297	-0.031	14	8	2	
904.9056	904.8761	-0.029	14	9	1	-1
907.3638	907.3328	-0.031	13	12	0	
908.2707	908.2422	-0.029	14	9	2	-2
910.0929	910.0614	-0.032	14	10	1	
911.6373	911.6078	-0.029	14	11	0	-1
913.3645	913.3365	-0.028	14	10	2	-2
916.2755	916.2464	-0.029	15	10	0	3
916.8236	916.7930	-0.031	14	12	0	
918.0027	917.9751	-0.028	15	9	2	1
919.5500	919.5216	-0.028	15	10	1	
921.0955	921.0680	-0.027	15	11	0	-1
922.8257	922.7967	-0.029	15	10	2	-2
926.3718	926.3441	-0.028	15	12	0	1

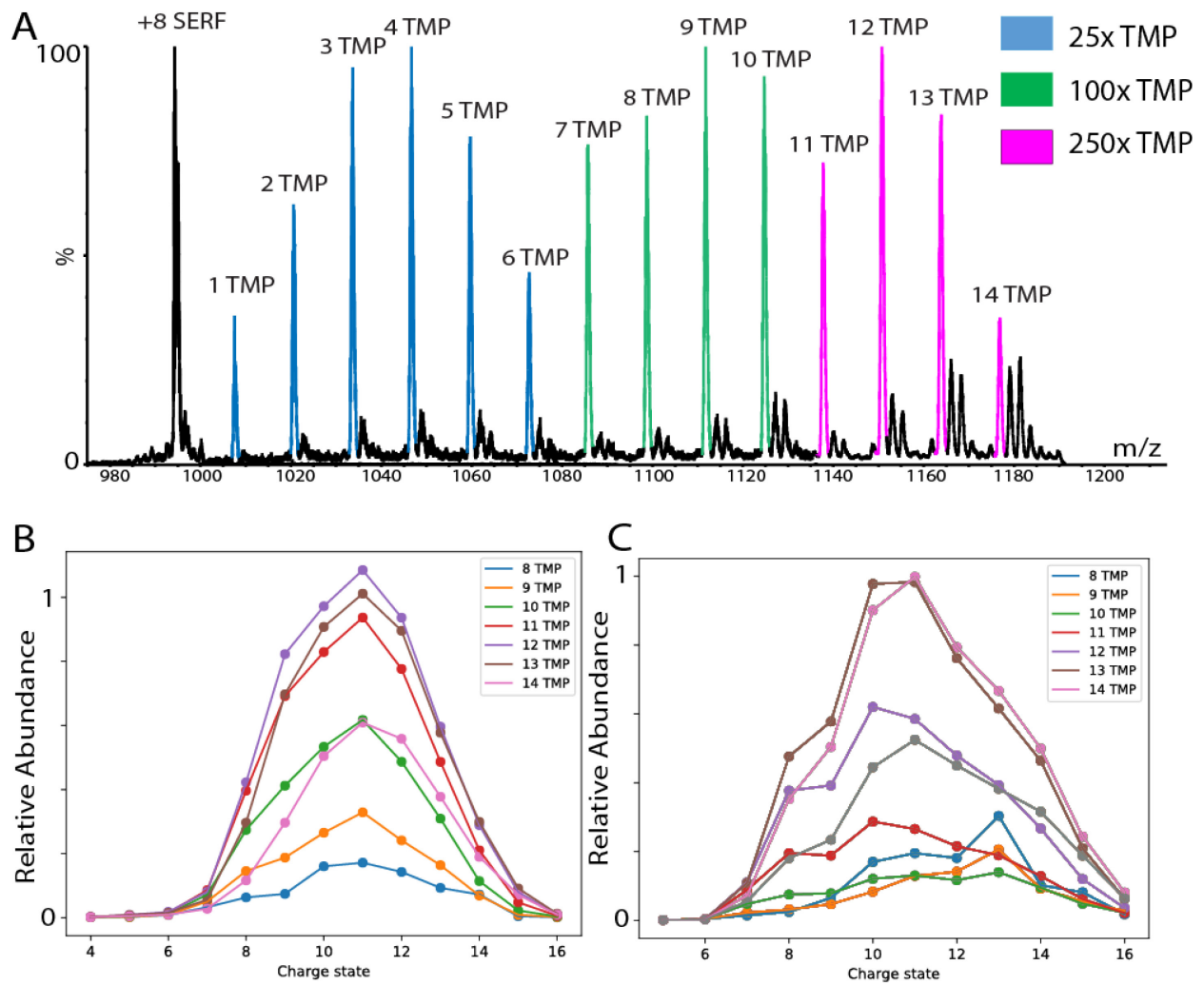


Figure II-4 Fixed charge modification reveals extensive charge solvation by intact protein ions. A) Composite mass spectrum from three reaction conditions of SERF with TMP (25, 100, and 250:1 ratios of TMP to lysine residues), demonstrating up to 14 intrinsically-charged TMP modifications on 8+ SERF. B) Charge state distribution for SERF with 8-14 TMP modifications and SERF modified with EDC and 8-14 TMP modifications.

Text II-1. Supporting Methods for Model Structure Analysis

Topology and Parameter files

Residue topology and parameter file of lysine modified with TMP, and EDC capped aspartic acid and glutamic acid residues are provided as separate files. Part of residue topology file and parameter file of lysine modified with TMP was made using ParamChem²⁻⁴. The penalties for charges, bond lengths, and bond angles were less than 50 in most cases. Set of dihedral angles connecting lysine to TMP moiety had penalties above 50. However, that is most likely due to absence of such angles in the training set^{3,4}. Since our models agree well with the experimental data, we believe that the parameters generated using ParamChem do not deviate from true values.

Hierarchical Clustering

Hierarchical clustering method^{5,6} from scipy^{7,8} was used to classify structural families extracted from the simulations. 2000 structures were selected at regular intervals in the last 10 ns for 300 K replica. Pairwise RMSD values were calculated for all combinations of structures. Pairwise Euclidean distance matrix was generated using RMSD matrix. The resulting distance matrix was then used for hierarchical clustering using average method. The number of clusters was set by performing an F test and analyzing the pairwise RMSD distribution histogram.

Secondary Structure Assignment

Secondary structure assignment was performed using dssp program^{9,10}. Fraction of secondary structure for each model was calculated as

$$Frac_{secstruct} = \frac{\sum_i^N i_{secstruct}}{N} \quad (1)$$

where *secstruct* is the secondary structure assignment as α helix, π helix, 3_{10} helix, beta bridge, beta buldge, turns, curve, and other types of loop, $i_{secstruct}$ is the residue assigned to the

secstruct, and N is the total number of residue. In addition, the following was used to calculate fraction helix, beta strand, and loop

$$Frac_{helix} = \frac{\sum_i^N i_{alpha_helix} + i_{pi_helix} + i_{310_helix}}{N} \quad (2)$$

$$Frac_{beta_strand} = \frac{\sum_i^N i_{beta_bridge} + i_{beta_buldge}}{N} \quad (3)$$

$$Frac_{loop} = \frac{\sum_i^N i_{turns} + i_{curve} + i_{other_loop}}{N} \quad (4)$$

Fraction of a residue in a specific secondary structure in a cluster was calculated as

$$Frac_{res} = \frac{\sum_{res}^{N_c} res_{secstruct}}{N_c} \quad (5)$$

where *res* is the residue and N_c is the total number of structures in the cluster.

Investigating interactions between fixed charge in lysine to all backbone carbonyl oxygen atoms

For all the structures in a cluster Euclidean distance between the positively charged nitrogen atom in Lysine with or without TMP modification and peptide backbone carbonyl oxygen atoms was calculated. Distance matrix was created with Lys residues against all other residues in the structure. Contact matrix with distance (d) $> 5 \text{ \AA}$ and $10 \text{ \AA} < d < 5 \text{ \AA}$ was created using the distance matrix. Z-score was calculated on the summed contact matrix in a cluster to assess the pairwise interactions.

Theoretical CCS Calculations

IMPACT¹¹ and IMOS^{12,13} were used for CCS calculation for model structures. IMPACT was used for CCS calculations on all structures resulting from MD simulations. The IMOS diffusive trajectory method (TM), which accounts for diffuse scattering in momentum transfer calculation, was used with He gas at 300 K with 50,000 total gas molecules. CHARMM non-

integer partial charges were included in the IMOS calculation. Overall, IMOS TM calculations were employed on a random subset of 80 structures from each SERF model. Linear regression between IMPACT TM and IMOS TM was used to convert the IMPACT TM values to IMOS TM values. Error in the final CCS values were calculated using: $error = \sqrt{\sigma^2 + rmse^2}$ where σ is standard deviation of CCS values and $rmse$ is root mean square error from the linear regression between IMPACT and IMOS TM CCS values.

Table II-3 . Experimental CCS of unmodified SERF ions. Conf shows different CCS distribution seen in a charge state. CCS calibration RMSE was 1.23 % and database error was set as 3 %.

conf	charge	CCS (nm ²)	total_error (nm ²)
conf1	5	8.93	0.29
conf2	5	9.24	0.30
conf1	6	11.42	0.38
conf2	6	9.45	0.31
conf1	7	14.20	0.46
conf2	7	13.38	0.44
conf3	7	12.37	0.40
conf1	8	14.69	0.48
conf1	9	15.34	0.50
conf1	10	15.86	0.51
conf1	11	16.38	0.53
conf1	12	17.50	0.57
conf1	13	18.74	0.61
conf1	14	19.32	0.63
conf1	15	19.68	0.64

Table II-4 Experimental CCS of modified SERF ions. Conf shows different CCS distribution seen in a charge state. CCS calibration RMSE was 0.54 % and database error was set as 3 %.

conf	charge	CCS (nm ²)	total_error (nm ²)
conf1	8	11.69	0.49
conf2	8	13.06	0.43
conf3	8	13.64	0.49
conf4	8	14.40	0.46
conf5	8	15.21	0.47
conf6	8	16.45	0.64
conf7	8	17.65	0.66
conf8	8	19.03	0.65
conf1	9	12.87	0.47
conf2	9	13.75	0.45
conf3	9	14.70	0.52
conf4	9	15.24	0.55
conf5	9	17.08	0.53
conf1	10	14.38	0.93
conf2	10	16.68	0.55
conf3	10	17.44	0.56
conf1	11	16.98	0.71
conf2	11	18.08	0.55
conf1	12	17.74	0.67
conf2	12	19.36	0.60
conf1	13	18.04	0.72
conf2	13	20.12	0.64
conf1	14	18.56	0.65
conf2	14	20.61	0.64
conf1	15	19.89	0.70
conf2	15	22.09	0.69

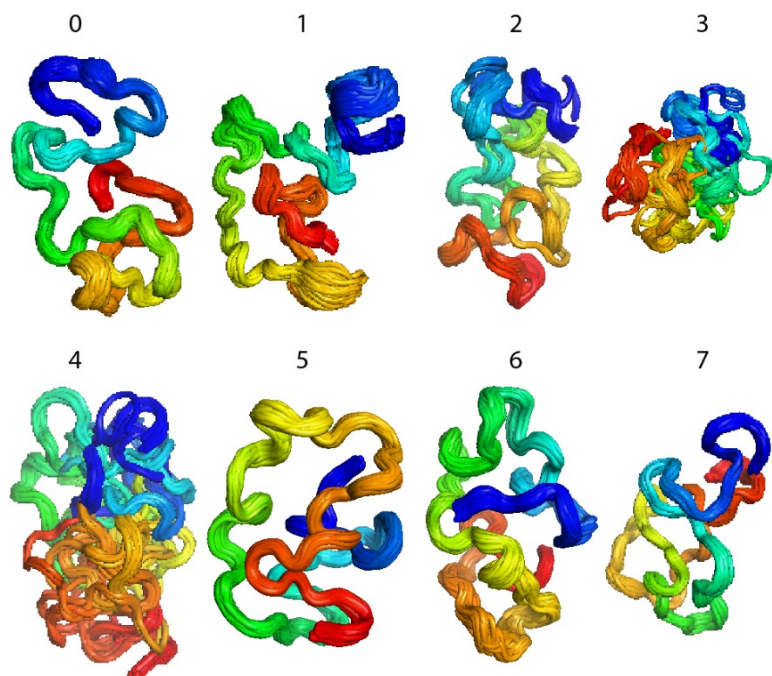


Figure II-5 Numbered cluster structures for unmodified SERF.

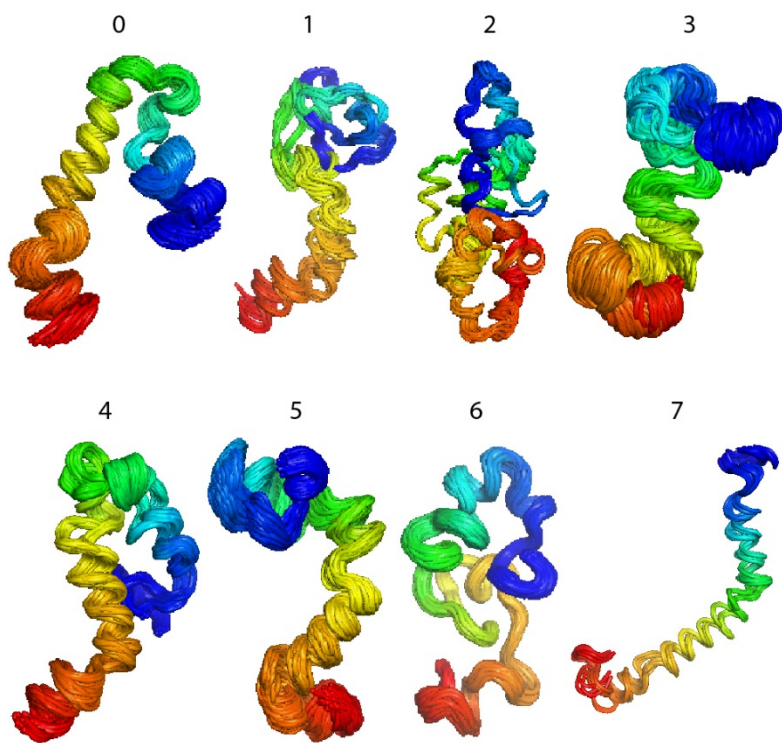


Figure II-6 Numbered cluster structures for modified SERF.

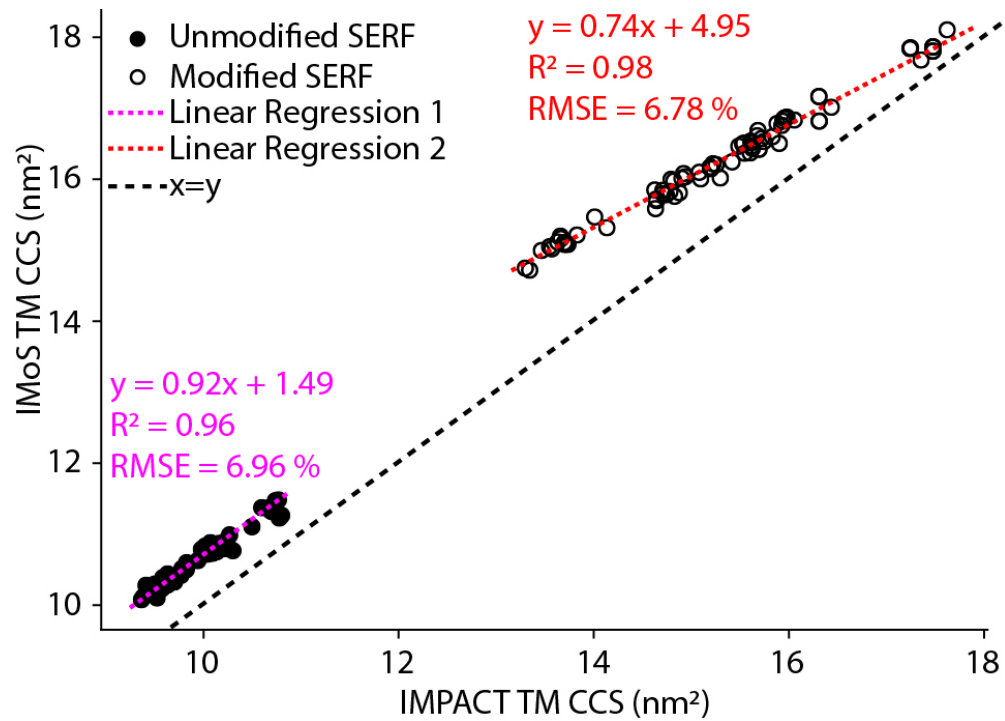


Figure II-7 Plot of TM CCS values determined via IMPACT against IMoS. Separate linear regression models are required to fit the data for unmodified (Linear Regression 1) and modified (Linear Regression 2) SERF. The equation provided for each model is used to convert the IMPACT TM CCS values to IMoS TM CCS values. RMSE is used in the computation of total error.

Table II-5 Theoretical CCS of models from each cluster in the unmodified version of SERF. μ and σ are the mean and standard deviation of CCS values, respectively.

Cluster number	Number of structures	Size of cluster (%)	μ (nm ²)	σ (nm ²)	error (nm ²)
0	1087	54.38	10.82	0.07	0.76
1	110	5.50	11.33	0.10	0.80
2	256	12.81	10.78	0.08	0.76
3	114	5.70	10.46	0.38	0.82
4	43	2.15	10.60	0.32	0.80
5	172	8.60	10.33	0.07	0.72
6	165	8.25	10.52	0.09	0.74
7	52	2.60	10.78	0.05	0.75

Table II-6 Theoretical CCS of models from each cluster in the modified version of SERF. μ and σ are the mean and standard deviation of CCS values, respectively.

Cluster number	Number of structures	Size of cluster (%)	μ (nm ²)	σ (nm ²)	error (nm ²)
0	733	36.67	16.47	0.14	1.13
1	27	1.35	16.60	0.27	1.16
2	62	3.10	15.58	0.85	1.35
3	185	9.25	15.84	0.09	1.08
4	89	4.45	16.02	0.12	1.09
5	514	25.71	16.46	0.18	1.13
6	375	18.76	15.10	0.13	1.03
7	14	0.70	17.82	0.11	1.21

Table II-7 Table of average and standard deviation of fraction of secondary structure elements in each cluster for unmodified SERF.

cluster	helix_avg	helix_std	beta_strand_avg	beta_strand_std	loop_avg	loop_std	310helix_avg	310helix_std	alphahelix_avg	alphahelix_std	pihelix_avg	pihelix_std
0	0.09	0.02	0.02	0.01	0.88	0.03	0.02	0.02	0.07	0.01	0.00	0.00
1	0.01	0.02	0.00	0.00	0.99	0.02	0.01	0.02	0.00	0.00	0.00	0.00
2	0.07	0.04	0.01	0.01	0.92	0.04	0.03	0.03	0.04	0.03	0.00	0.00
3	0.06	0.05	0.01	0.02	0.94	0.04	0.05	0.04	0.01	0.02	0.00	0.00
4	0.02	0.02	0.01	0.01	0.98	0.02	0.02	0.02	0.00	0.00	0.00	0.00
5	0.05	0.02	0.03	0.00	0.92	0.02	0.05	0.02	0.00	0.00	0.00	0.00
6	0.05	0.04	0.00	0.01	0.95	0.04	0.02	0.03	0.03	0.03	0.00	0.00
7	0.05	0.03	0.03	0.01	0.92	0.03	0.05	0.03	0.00	0.00	0.00	0.00

Table II-8 Table of average and standard deviation of fraction of secondary structure elements in each cluster for unmodified SERF.

cluster	helix_avg	helix_std	beta_strand_avg	beta_strand_std	loop_avg	loop_std	310helix_avg	310helix_std	alphahelix_avg	alphahelix_std	pihelix_avg	pihelix_std
0	0.35	0.03	0.00	0.00	0.65	0.03	0.00	0.00	0.18	0.03	0.17	0.03
1	0.33	0.05	0.00	0.00	0.67	0.05	0.02	0.03	0.19	0.05	0.12	0.03
2	0.47	0.07	0.02	0.01	0.51	0.07	0.02	0.03	0.18	0.16	0.27	0.15
3	0.33	0.03	0.00	0.00	0.67	0.03	0.02	0.02	0.22	0.04	0.09	0.02
4	0.33	0.04	0.00	0.00	0.67	0.04	0.03	0.02	0.28	0.03	0.03	0.04
5	0.41	0.04	0.00	0.00	0.59	0.04	0.07	0.01	0.34	0.04	0.00	0.00
6	0.21	0.03	0.00	0.00	0.79	0.03	0.03	0.03	0.09	0.03	0.09	0.02
7	0.61	0.03	0.00	0.00	0.39	0.03	0.00	0.00	0.26	0.04	0.35	0.04

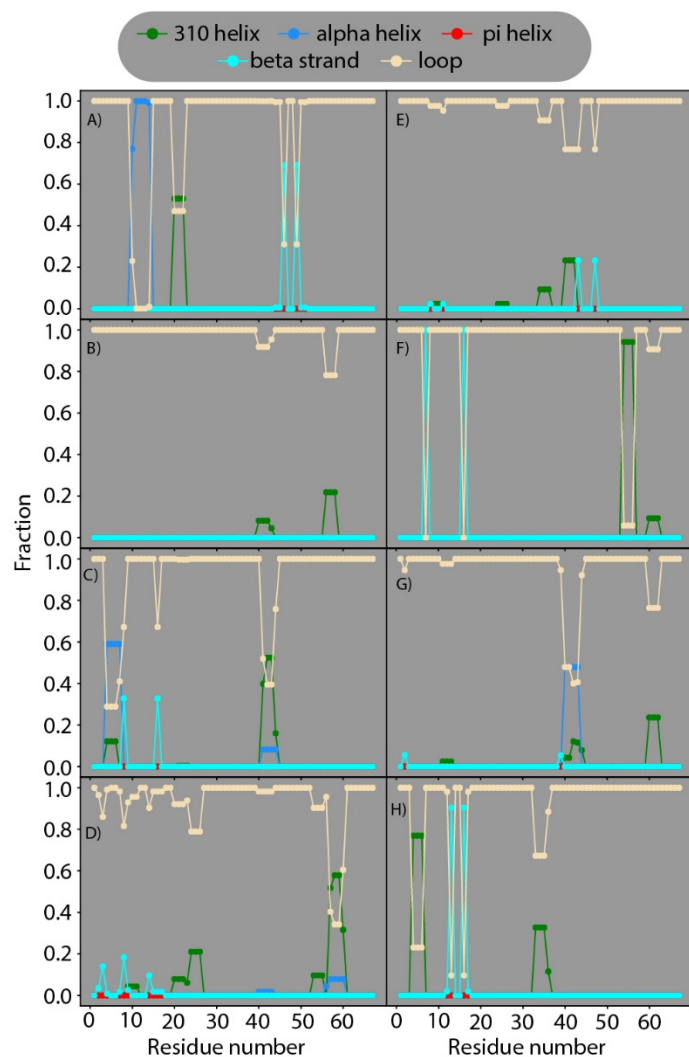


Figure II-8 Plots of fraction of secondary structure vs. residue number for unmodified SERF clusters 0, 1, 2, 3, 4, 5, 6, 7, in A), B), C), D), E), F), G), and H), respectively.

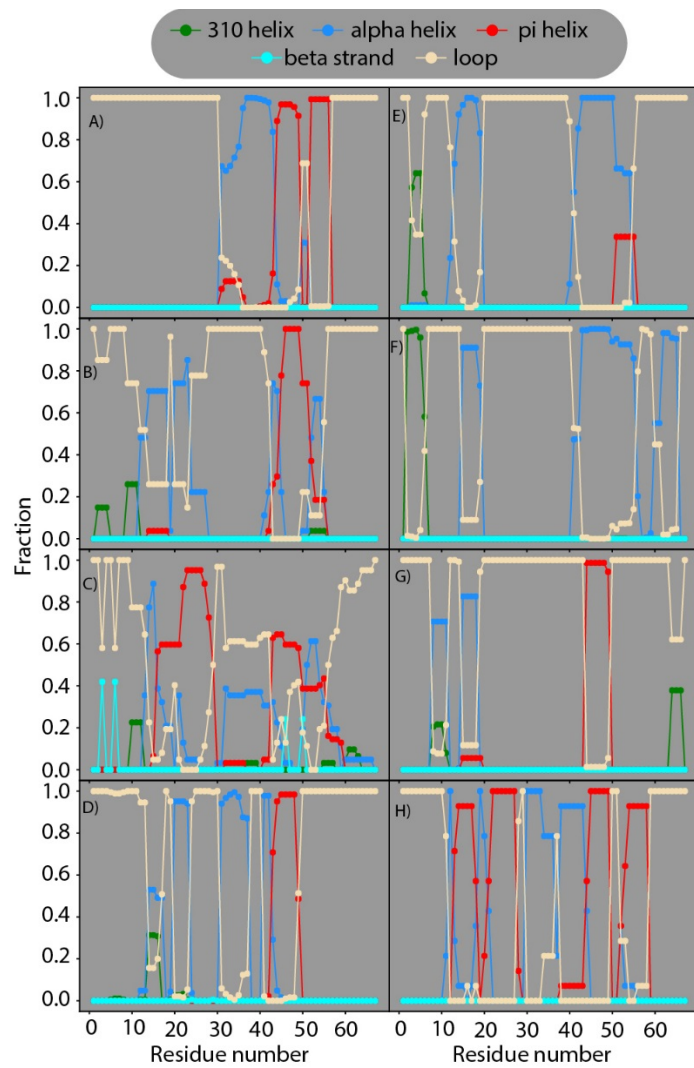


Figure II-9 Plots of fraction of secondary structure vs. residue number for modified SERF clusters 0, 1, 2, 3, 4, 5, 6, 7, in A), B), C), D), E), F), G), and H), respectively.

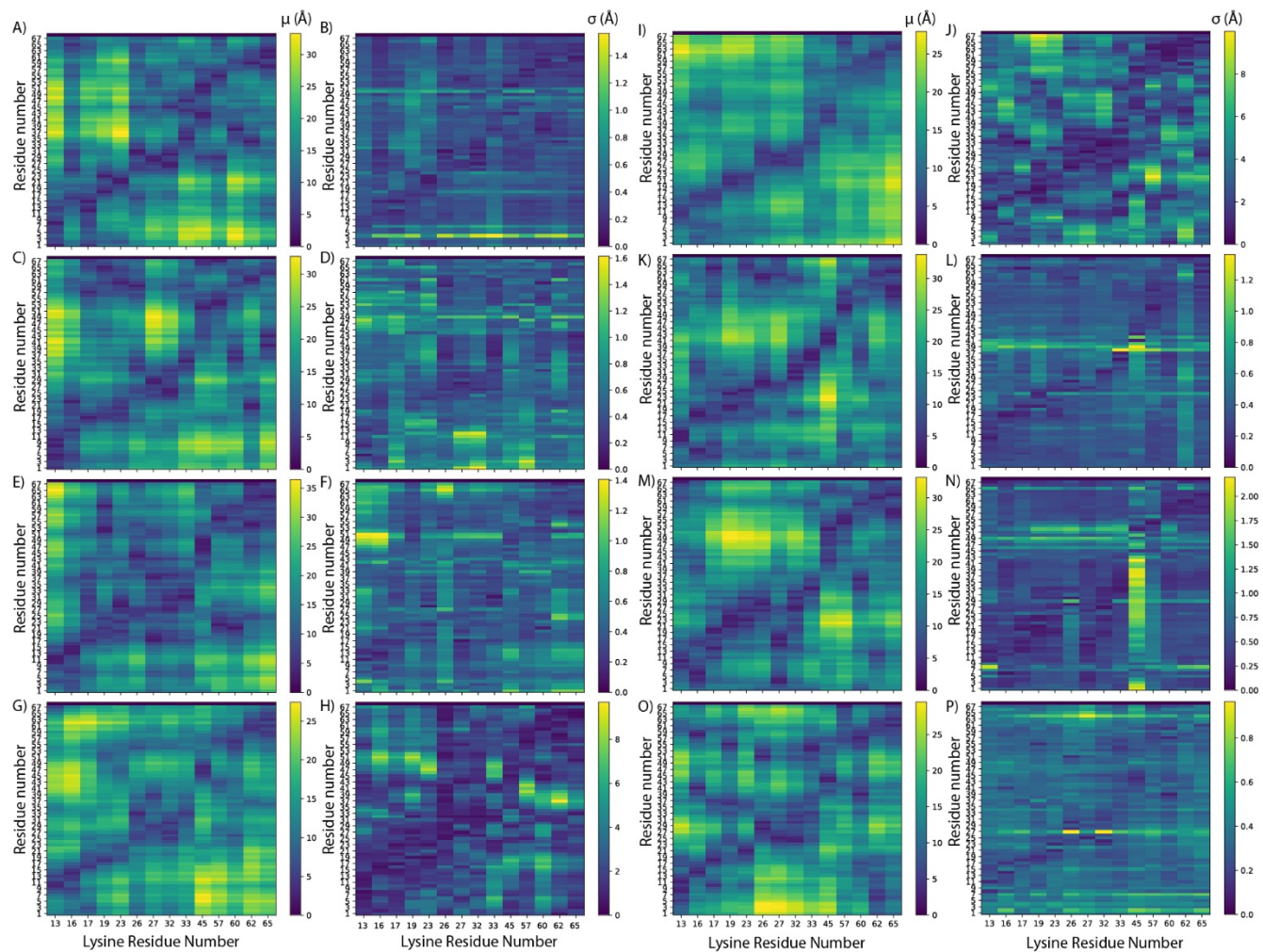


Figure II-10 Contour plots of mean (μ) distance between the positive charge carrying nitrogen atom in lysine and all the carbonyl oxygen atoms in unmodified SERF for clusters 0, 1, 2, 3, 4, 5, 6, 7 in A), C), E), G), I), K), M), and O), respectively. Contour plots of standard deviation (σ) of distance between the positive charge carrying nitrogen atom in lysine and all the carbonyl oxygen atoms in unmodified SERF for clusters 0, 1, 2, 3, 4, 5, 6, 7 in B), D), F), H), J), L), N), and P), respectively.

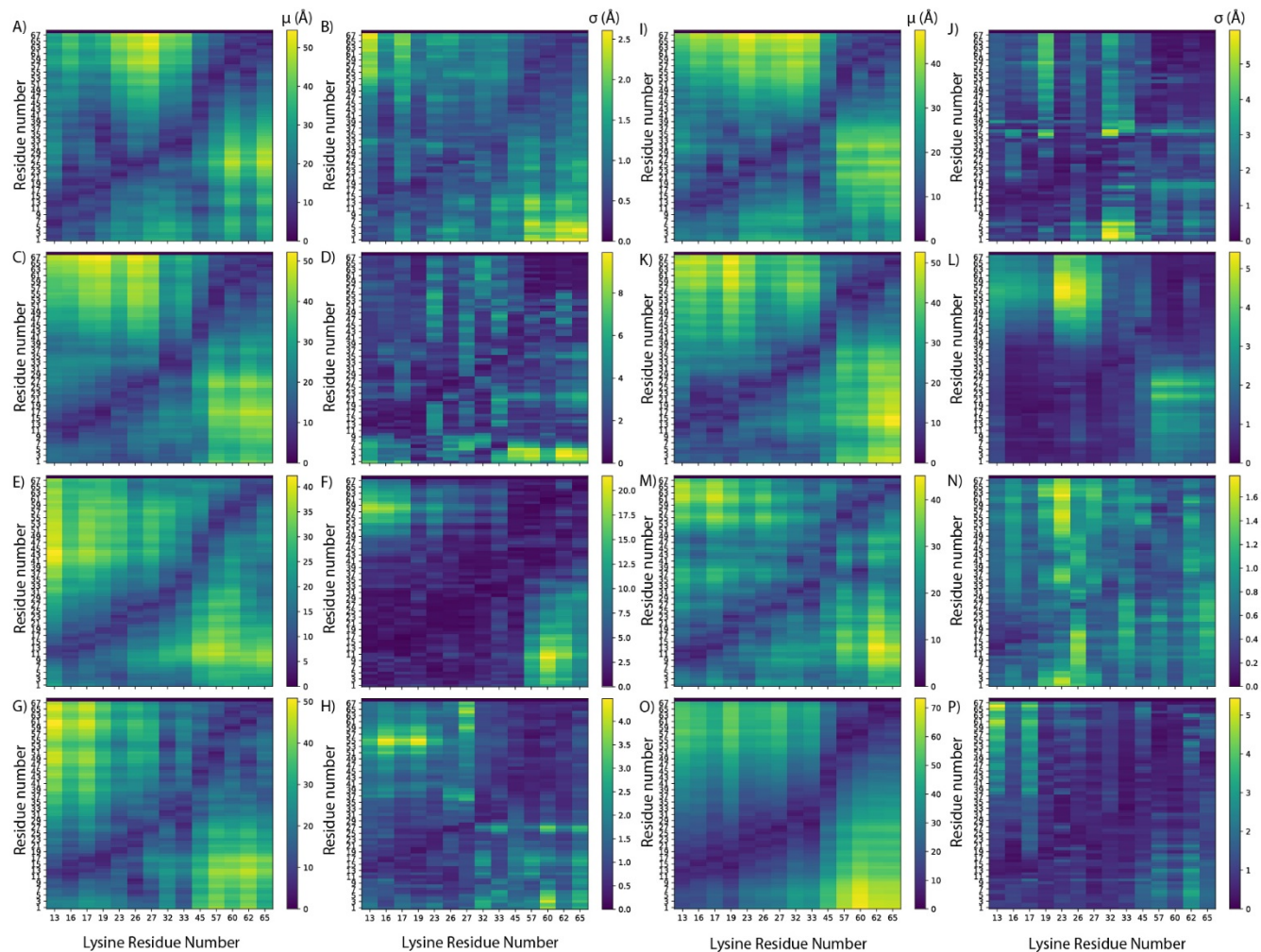


Figure II-11 Contour plots of mean (μ) distance between the positive charge carrying nitrogen atom in lysine and all the carbonyl oxygen atoms in modified SERF for clusters 0, 1, 2, 3, 4, 5, 6, 7 in A), C), E), G), I), K), M), and O), respectively. Contour plots of standard deviation (σ) of distance between the positive charge carrying nitrogen atom in lysine and all the carbonyl oxygen atoms in unmodified SERF for clusters 0, 1, 2, 3, 4, 5, 6, 7 in B), D), F), H), J), L), N), and P), respectively.

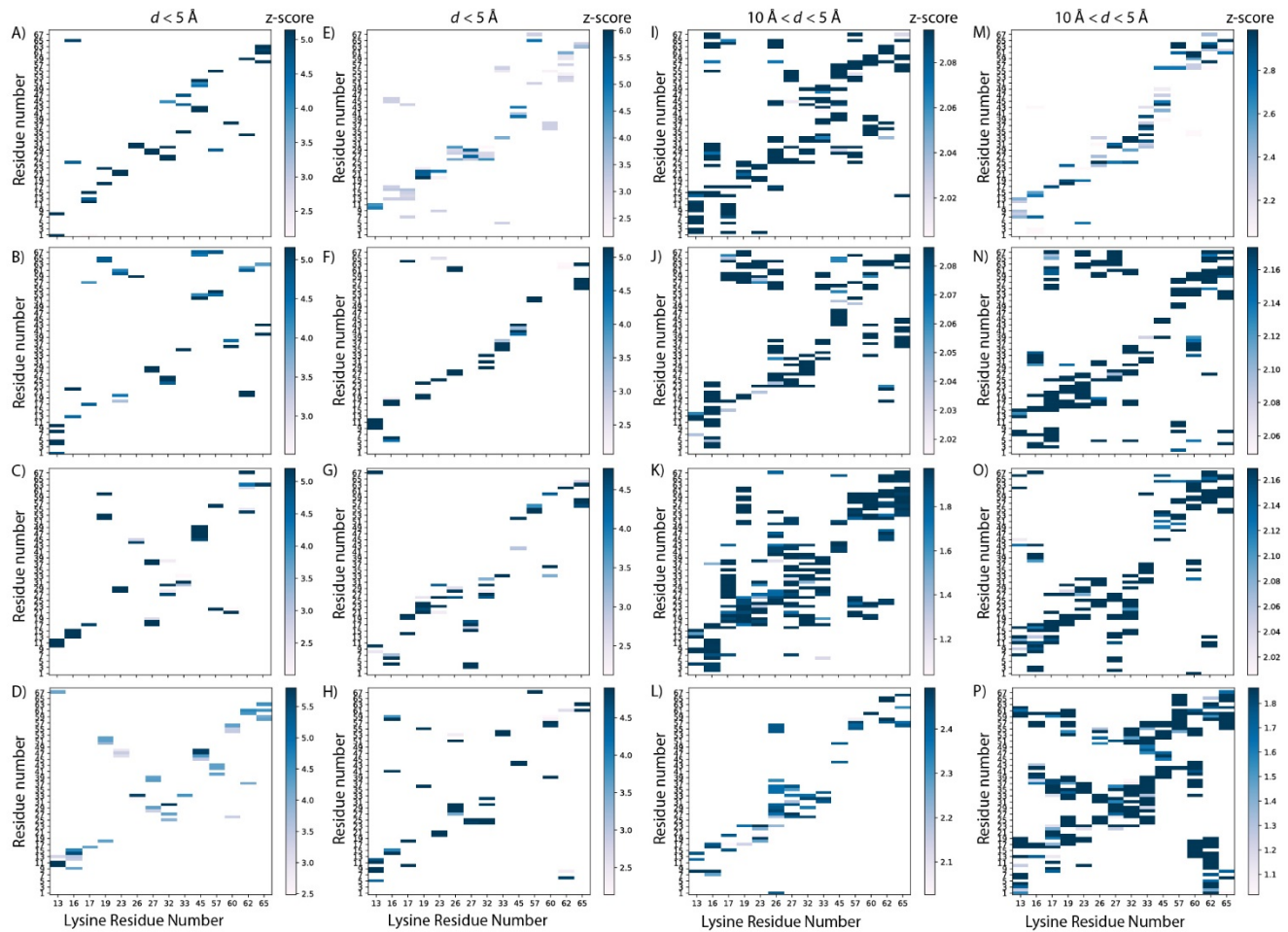


Figure II-12 Contour plots of z-score of pairwise contact matrix for distance (d) $< 5 \text{ \AA}$ between positive charge carrying nitrogen atom in lysine residues versus peptide backbone carbonyl oxygen atoms in all residues for unmodified SERF clusters 0, 1, 2, 3, 4, 5, 6, and 7 in A), B), C), D), E), F), G), and H), respectively. Similar plots for $10 \text{ \AA} < d < 5 \text{ \AA}$ for clusters 0, 1, 2, 3, 4, 5, 6, and 7 in I), J), K), L), M), N), O), and P), respectively. Z-score values greater than 2 are considered in all contour plots, except for clusters 2 and 7 in $10 \text{ \AA} < d < 5 \text{ \AA}$. Since there were no values greater than 2 for those clusters, values greater than 1 are plotted in the contour plots.

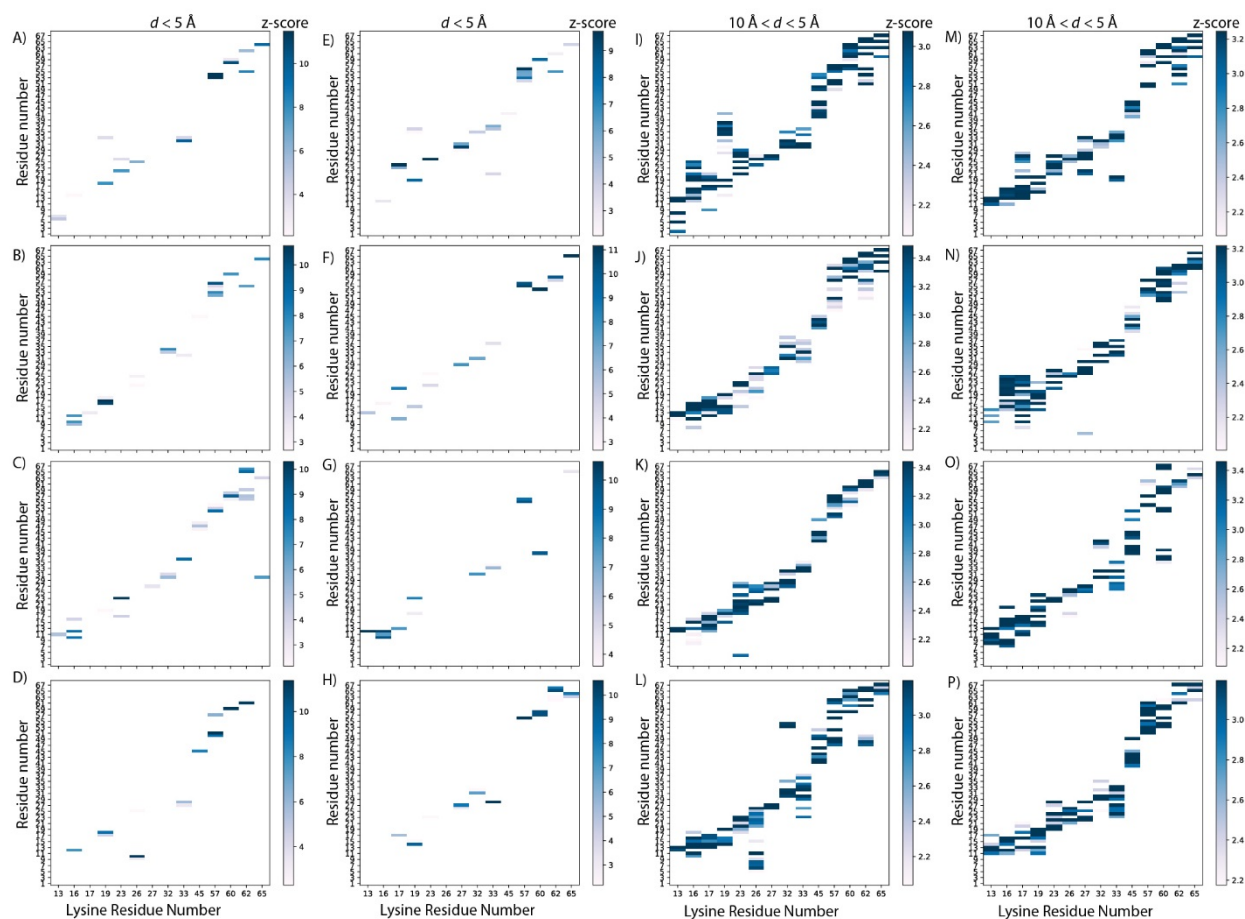


Figure II-13 Contour plots of z-score of pairwise contact matrix for distance (d) $< 5 \text{ \AA}$ between positive charge carrying nitrogen atom in lysine residues versus peptide backbone carbonyl oxygen atoms in all residues for modified SERF clusters 0, 1, 2, 3, 4, 5, 6, and 7 in A), B), C), D), E), F), G), and H), respectively. Similar plots for $10 \text{ \AA} < d < 5 \text{ \AA}$ for clusters 0, 1, 2, 3, 4, 5, 6, and 7 in I), J), K), L), M), N), O), and P), respectively. Z-score values greater than 2 are considered in all contour plots.

III. Chapter 4 Supporting Information

Text III-1. Averagine based isotopic grouping algorithm used in Grppr

Briefly, the list of input peaks is first sorted by intensity in descending order. The first peak is removed from the list and its possible isotopic neighbors are searched for within +1 m/z. For each possible determined charge state, the potential neutral mass is computed and a “template” is generated based on it. A template is a set of peaks with intensities normalized to unity and inter-peak mass differences defined; as such, it represents an isotopic envelope through mass differences to neighboring peaks rather than the exact S1 absolute mass of each peak. To generate a template for a seed peak with an m/z value of MZ_s , IM drift time of IM_s and a potential charge state z_s , the neutral mass M_s is calculated as $M_s = z_s(MZ_s - M_H)$, where M_H is the mass of a proton, and the corresponding elemental composition according to the Averagine model is computed. For this study we have used the standard definition of Averagine as an amino acid of mass 111.1254 Da, having an elemental composition of $C_{4.93}H_{7.75}N_{1.35}O_{1.47}S_{0.04}$. We have found that the exact mass and composition values are not critical in identifying isotopic envelopes, as slight changes in those values introduce only minor variations to the resulting shape of the isotopic envelope. The most intense peak within the template is aligned to the seed peak, i.e. its m/z and IM drift time are set to MZ_s and IM_s respectively. The template’s peaks are then traversed, calculating theoretical m/z value for each and searching for possible matching peaks in the input peak list. If multiple candidates are found in the proximity of a theoretical peak, the one with intensity closest to that of the template peak is retained. To account for potential errors in peak intensity measurements that might lead to the most intense peak being identified incorrectly, the template is further varied in m/z direction, shifting it by one peak position left and right. Kullback-Leibler divergence (KLD) is used as a goodness-of-fit value to

select the best matching location for the template. All the peaks assigned to an isotopic cluster are then blacklisted, disallowing them to serve as seed points for other clusters. For a typical top-down sequencing experiment we also set a threshold on the minimum number of peaks required to call an isotopic cluster to at least 3-4 peaks, which is a user-defined parameter. This clustering workflow is summarized in pseudo-code in Algorithm 1.

Algorithm 1: Isotopic grouping of 2D peaks lists using Averagine model.

```
input : List of 2D peaks (Peaks)
output: List of isotopic clusters

1 Seeds ← Peaks
2 Blacklist, Clusters, Results ← empty list
3 sort (Seeds) by intensity descending
4 foreach Seed in Seeds do
5   if Seed is in Blacklist then
6     | continue
7   ClusterCandidates ← empty list
8   Search Peaks for Neighbors up +/-1 Da away from Seed
9   foreach Neighbor in Neighbors do
10    Z = 1/abs (mz (Neighbor) - mz (Seed) )
11    if Z is not close enough to an integer then
12      | continue
13    Cluster ← new isotopic cluster with Seed in it, charge int (Z)
14    Build Template from Seed's neutral mass using Averagine model
15    Align Template's most intense peak to Seed
16    foreach TPeak in Template do
17      | Search Peaks for all IPeak that are near TPeak
18      | Select one IPeak that is closest in intensity to TPeak
19      | Add IPeak to Cluster
20    end
21    if length (Cluster) < threshold then
22      | continue
23    Extend Cluster by extra N lower intensity peaks from Peaks to the left and
      right
24    foreach possible Shift of Template relative to Cluster do
25      | Shift.KLD ← KLD (Shift, Template, Cluster)
26    end
27    Trim Cluster to leave only peaks that intersect with the best Shifted
      Template
28    Add Cluster to ClusterCandidates
29  end
30  BestCluster ← most intense Cluster from ClusterCandidates
31  Add BestCluster to Results
32  Add all of BestCluster's peaks to Blacklist
33 end
34 return Results
```

Figure III-1 Algorithm 1

Table III-1 User-definable parameters used for IMTBX processing.

Parameters for WatersIMSPeakDetector		
Parameter name	Flag	Value
Analysis mode	N/A	peaks
Smoothing/filtering	--filter	4 4 2 2 0 0
Absolute intensity cutoff	--cut	30
Scan combination mode	--mode	Sum
Minimum number of points for a peak	--npts	16
Valley depth between peaks in IM dimension	--drop	0.1
m/z bounds on data to consider	--bounds_mz	100 5000
RMS noise estimation moving window size	--noise -- noiseWnd	4 8
Minimum S/N ratio for a peak to be detected	--snr	2
all other parameters left at default values		

Table III-2 User-definable parameters used for Grppr processing.

Parameters for Isotopic Clustering tool		
Parameter name	Flag	Value
Function and Min number of peaks in a cluster	--deisotope	1 4
Grouping in retention time (for LC-coupled analyses) true/false	--group	false
Minimum S/N for most abundant and all other peaks in isotopic cluster	--dsnr	3 2.5
Min charge to consider	--zLo	1
Max charge to consider	--zHi	6
Isotopic grouping scoring algorithm	--isoAlg	AVERAGINE
Mass error tolerated when searching for members of a cluster	--isoMzV	100
Units for mass error when searching for members of a cluster	--IsoMzT	PPM

Table III-3 Estimated processing time for top-down data by manual, semi-automated, and IMTBX+Grppr workflows.

A	Estimated total processing time (minutes)		
Number of analyses	Manual annotation	Semi-automated (TWIMExtract + mMass)	IMTBX + Grppr
1	225	7.5	0.81
10	2250	48	3.6
100	22500	453	31.5
B	Estimated total user ("active") time (minutes)		
Number of analyses	Manual annotation	Semi-automated (TWIMExtract + mMass)	IMTBX + Grppr
1	225	7.5	0.5
10	2250	48	0.5
100	22500	453	0.5

IV. Chapter 5 Supporting Information

Text IV-1 Univariate Feature Selection (UFS) for Classification

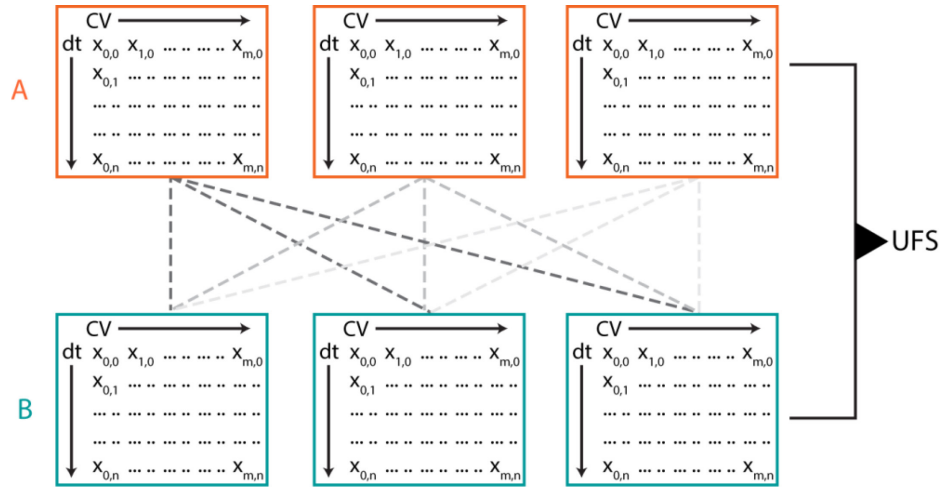


Figure IV-1 Depiction of querying each possible combination of input data between classes. The variation within and between classes for each combination of input data is used to perform feature selection across the input dataset.

Univariate feature selection, comprised of performing F-test statistics¹⁴ to evaluate the significance of each collision voltage in differentiating the training data classes, is performed as the first step in classification workflow. Scheme S1 shows UFS for two groups A and B, with F statistics calculated for each data set in each group, as shown by the dotted lines. F ratio is calculated as described below, and F ratio is then converted to p-value using the F-distribution.

$$F = \frac{\text{Between group variability}}{\text{Within group variability}}$$

$$\text{Between group variability} = \sum_{i=1}^K n_i (\bar{Y}_i - \bar{Y})^2 / (K - 1)$$

where K is the number of groups, n_i and \bar{Y}_i are the number of observations and the mean of the i -th group, respectively, and \bar{Y} is the overall mean of the data.

$$\text{Within group variability} = \sum_{i=1}^K \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 / (N - K)$$

where Y_{ij} is the j -th observation in the i -th group and N is the overall sample size.

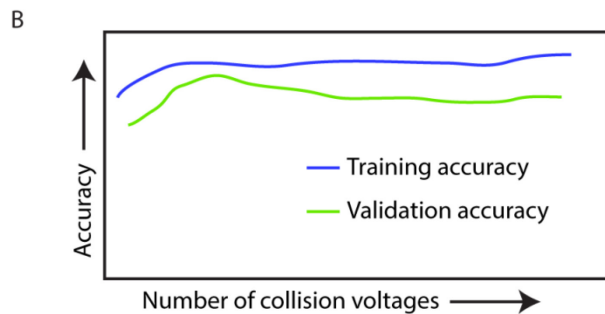
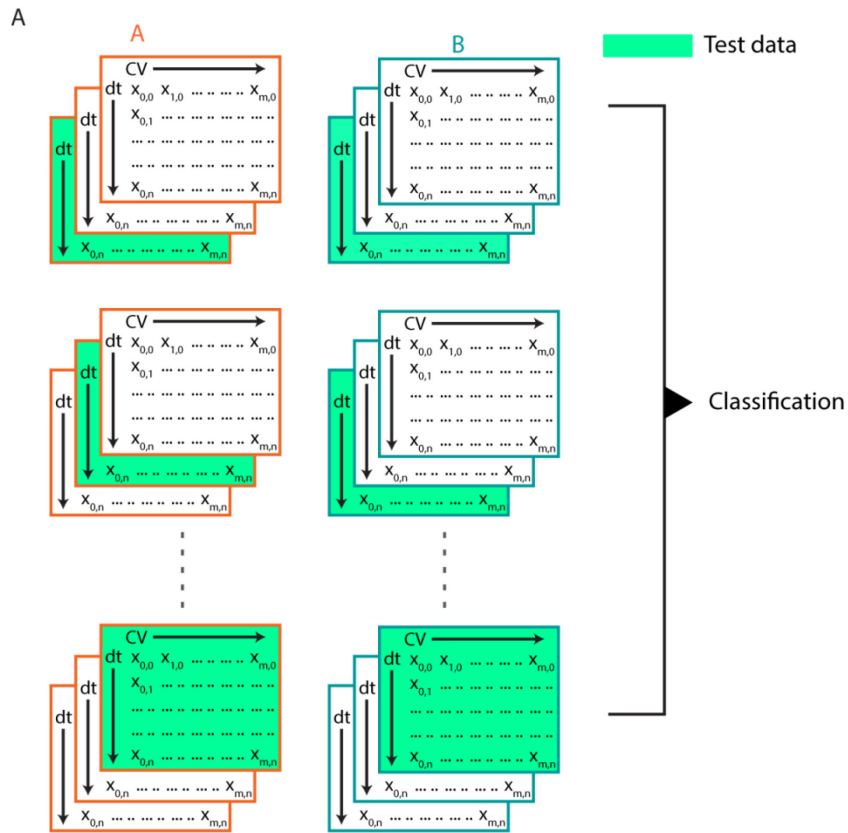


Figure IV-2 Cross-validation workflow. A “leave one out” cross-validation¹⁵ is performed by implementing the classification workflow, which consists of linear discriminant analysis (LDA)¹⁶ followed by construction of a support vector machine (SVM)¹⁷, across a range of sub-sections of CIU datasets corresponding to sequential addition of arrival time distributions from individual collision voltages in decreasing order of UFS score. Classification accuracy of training and test data sets are then plotted as a function of the number of collision voltages to determine the optimal model selection. A) Scheme showing “Leave one out” cross validation approach where a single CIU dataset is treated as test data in each group and the remaining datasets are used as training data. All possible combinations of training and test data set are created. The training set is used to build the classification model and the test data is used to validate the accuracy of the classification model. B) Accuracy of training and test data sets are plotted as a function of a range of CIU datasets that has sequential addition of collision voltages. The number of collision voltages resulting in greatest validation accuracy is used as the final classification scheme.

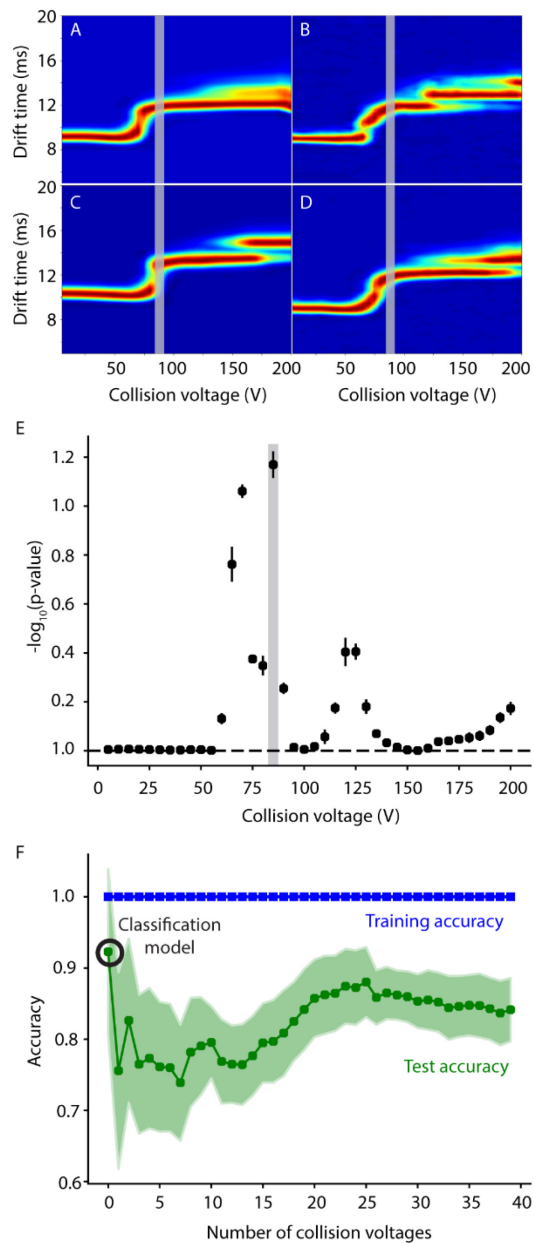


Figure IV-3 Feature selection and cross-validation for IgG1, IgG2, IgG3, and IgG4 classification from primary text Figure 3. A, B, C, and D are representative CIU fingerprints of IgG1, IgG2, IgG3 and IgG4, respectively. E) UFS results. Plot of $-\log_{10}(\text{p-value})$ against collision voltage assessing the significance of each collision voltage in differentiating the groups. 85 V, which has the highest score, is highlighted in grey and is also show in CIU fingerprints (A, B, C, and D). F) Cross-validation plot showing classification accuracy for training (blue) and test (green) data set with accuracy on y-axis and number of collision voltages on x-axis. The shaded region indicates the standard deviation from all the combinations of training and test datasets created. In this example, a single collision voltage (85V, circle marker) had the highest validation accuracy and was chosen as the final classification scheme to generate the data in primary text Figure 5-3.

Table IV-1 Test data probability values for classification of IgG subclasses. Each subclass has higher probability values when it is classified into its respective type, as highlighted in green.

IgG Subclass	Replicate	Probability for IgG1	Probability for IgG2	Probability for IgG3	Probability for IgG4
IgG1	1	0.69	0.12	0.06	0.13
	2	0.69	0.12	0.06	0.13
IgG2	1	0.33	0.52	0.06	0.09
	2	0.31	0.55	0.06	0.09
IgG3	1	0.07	0.04	0.73	0.16
	2	0.07	0.04	0.73	0.16
IgG4	1	0.14	0.07	0.10	0.69
	2	0.15	0.07	0.10	0.69

Table IV-2 Benchmarks for Gaussian fitting and Classification analysis time. All testing data was generated on a Dell Optiplex 990 workstation with an Intel Intel Core i7-2600 CPU (4 cores, hyperthreaded) at 3.40 GHz and 16 GB of RAM. For Gaussian fitting, two datasets were tested: the “complex” set, including both several protein and several noise signals and a “simple” set containing only high abundance protein signals. For “complex” data, 5 signal components and 7 noise components were allowed, while for “simple” data, 4 signal components and 0 noise components were allowed. Both analyses were performed using multithreading with 8 cores. Times to complete analysis are presented both as average time per individual collision voltage (top) and across a complete CIU dataset with 20 collision voltages. Average and standard deviation are from 10 separate runs on different datasets. For classification, default parameters were used to classify between two classes from CIU datasets containing 40 collision voltages. Classifying schemes were generated from 3, 6, and 10 replicates per class with the resulting times to complete the entire classification procedure.

Gaussian Fitting		
Dataset	Complex	Simple
	Per individual collision voltage (CV)	
Avg (s)	1.69	0.16
Std Dev (s)	0.27	0.02
	Per Complete Dataset (20 CVs)	
Avg (s)	33.80	3.20
Std Dev (s)	5.31	0.48
Classification		
		2-way, 40 features
3 replicates	Time (s)	2.2
6 replicates	Time (s)	16.3
10 replicates	Time (s)	68.0

Text IV-2: Additional details regarding fitting stability shift (CIU50) values

Stability shift (“CIU50”) analysis fits a generalized logistic function, of the form below, to the observed data in one of several modes specified by the user. In the default (and generally recommended) mode, the arrival time distribution at each collision voltage is centroided by simply taking the location (drift time) of the most intense signal. This is called “max” mode by the software. The resulting dataset is a centroid drift time at each collision voltage. To perform the logistic fitting, a transition region is set up by considering a pair of consecutive features from feature detection. The transition region is defined by any space between the two features (collision voltages not included in either feature but present between the max collision voltage of

the first feature and before the min collision voltage of the second feature) plus a “padding” amount (adjustable by the user) that includes a small amount of each feature in the transition region. The remaining portions of the features outside the transition region are set to the median centroid value of each feature to provide a floor and ceiling (min and max) for the logistic function. A least-squares minimization is performed (using `optimize.curve_fit` in Scipy⁸) between the logistic function and transition data to determine the optimal fit.

$$\text{Logistic function: } y = y_{min} + \frac{y_{max} - y_{min}}{1 + e^{-k(x - x_c)}}$$

y_{min} is the minimum function value (corresponding to the median drift time of the first feature)

y_{max} is the maximum function value (corresponding to the median drift time of the second feature)

k is a steepness factor (how quickly the transition occurs across collision voltages)

x_c is the center of the transition, also called the CIU50 value.

In some cases, particularly if features coexist at similar intensities across a range of collision voltages, it can be detrimental to centroid to the maximum intensity without considering other features. In these cases, spectral average and median modes are available that centroid the arrival time distribution by average and median, respectively. However, average and median centroided data do not always resemble a sigmoid function, and can result in poor fitting. In addition, any chemical or other noise or additional features beyond the two being considered can influence the fits, making these modes much less robust than the max mode centroiding.

Text IV-3: Additional details regarding Gaussian fitting and scoring

Gaussian fitting is performed using a least-squares minimization, implemented in the LMFit¹⁸ library in Python, of a sum of Gaussian functions (see below) to the observed arrival time distribution at each collision voltage. Expected width (full width at half max, FWHM) values and a maximum number of components (individual Gaussian functions) are provided by the user. Automated peak width prediction is not done as it requires detailed knowledge of both the analyte in question and the instrument settings (IM pressures, fields, etc.) used to collect the data, making it challenging to generalize in this context.

$$\text{Gaussian function: } y = A * e^{-\frac{(x-c)^2}{2\sigma^2}}$$

A = amplitude (peak height), c = peak center, σ = standard deviation (peak width)

Given a maximum number of components, CIUSuite 2 performs a least-squares minimization for each possible number of components, then scores the resulting fit to determine the best result. For example, if no chemical noise components are considered and the user allows a maximum of 4 components, fits utilizing 1, 2, 3, and 4 Gaussian functions are performed and the highest scoring fit is saved as the best result. If chemical noise components are considered, all combinations of signal and noise peaks are modeled. For example, if 3 signal components and 3 noise components are allowed by the user, a total of 9 fits are performed corresponding to 1 signal + 1 noise, 1 signal + 2 noise, ... , 3 signal + 3 noise. Each component is constrained to sample only width values with the ranges specified by the user.

To evaluate the results of each fitting procedure, a scoring function is employed. The score for a given fit is the r^2 value of the fit to the observed data minus any penalties incurred. Thus, the maximum score is 1, corresponding to a perfect correlation with the observed data and no penalties. Fits are penalized in 3 ways. First, a width penalty is assessed on any peaks that

deviate from the allowed ranges specified by the user. This penalty is computed as the difference between the observed width and the edge of the allowed window with no adjustment, and is rarely used as the fitted widths are constrained to be within these boundaries. Second, a peak overlap penalty is computed for signal components to prevent highly overlapping components from being considered. This penalty is computed according to the formulas below, allowing the user to specify strict penalizing of overlaps (>50% overlap is penalized), relaxed penalizing (>75% overlap penalized) or no penalty for overlap. These options allow prevention of component over-fitting into a single observed peak, particularly when a broad range of peak widths are allowed.

$$\text{[Strict penalty] Shared area penalty} = (1.25 * \text{sharedAreaRatio} - 0.25)^4$$

$$\text{[Relaxed penalty] Shared area penalty} = (\text{sharedAreaRatio} - 0.40)^4$$

“Shared area ratio” is computed as the area under a given Gaussian component that is also under another Gaussian component in the fit result divided by the total area of the Gaussian component in question. Thus, if a peak is almost entirely underneath another peak, the shared area ratio will approach 1 and the penalty will approach 1 in strict mode or 0.13 in relaxed mode. Finally, in fits considering both signal and chemical noise components, a fit is penalized if there are no signal components above a user-defined intensity threshold (default: 20% relative intensity). The penalty is computed as the difference between the threshold (e.g. 20%, or 0.20) and the amplitude of the most intense signal component (so a fit with a most intense signal component of 0.15 relative intensity would incur a penalty of $0.20 - 0.15 = 0.05$) to be subtracted from the r^2 along with any other penalties.

V. Chapter 6 Supporting Information

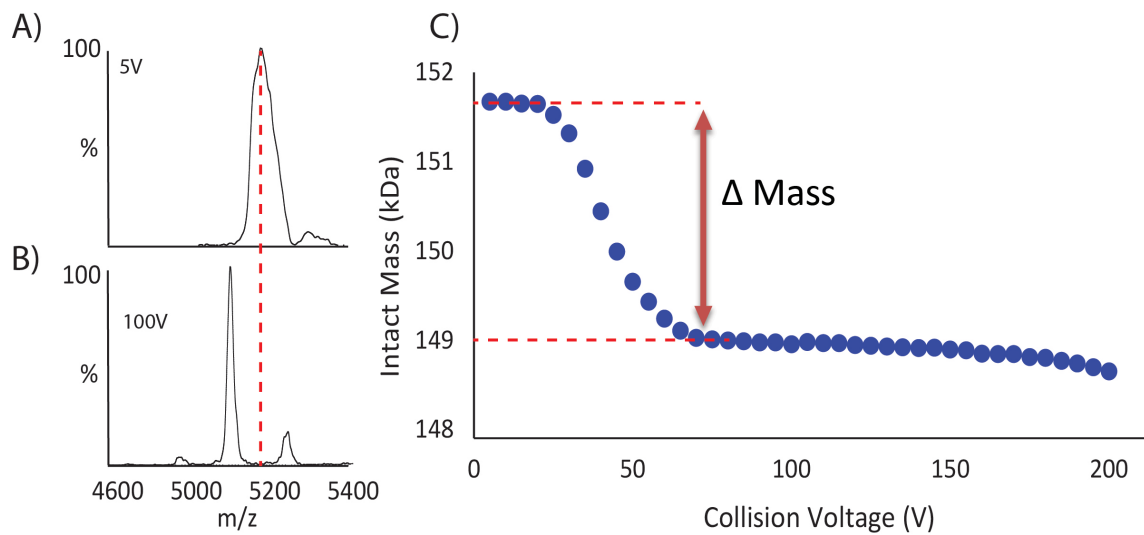


Figure V-1 Activation of supercharged 30+ mAb ion at (A) 5 V and (B) 100 V. We observe the selected 30+ ion peak shifting to lower m/z as adducts dissociate upon activation. (C) The intact mass plotted versus activation energy. The average mass at low and high activation energies is taken, and the difference is used to determine the Δ mass value used in Figure 6-2 G.

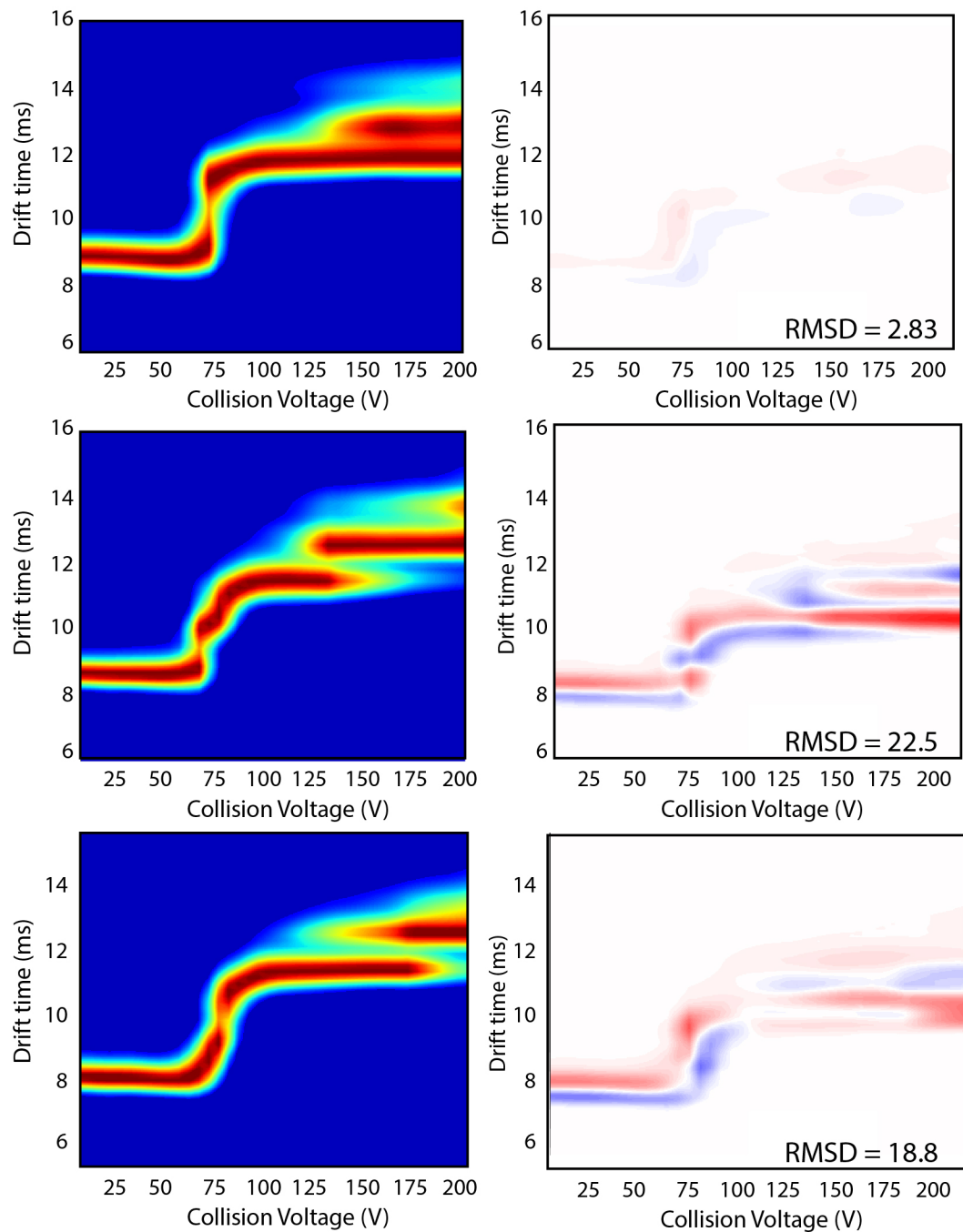


Figure V-2 CIU fingerprints of (A-C) IgG1k, IgG2k, and IgG4k, respectively on the Synapt G2 system generated using CIUSuite 2. (D) To determine the reproducibility and sensitivity of the RMSD baseline on the IgG1 standard fingerprints were average ($n=3$), and the average fingerprint was compared to all of the replicates producing an RMSD baseline value of nominally 3%. Comparison plots of IgG1 versus (E) IgG2 and (F) IgG4 were generated and produced RMSD of 22.5% and 18.8%, and a difference factor of 8 and 6, respectively.

Table V-1 CCS measurements extracted from IgG subclass CIU.

Charge State	IgG1 CCS (Å²)	IgG2 CCS (Å²)	IgG4 CCS (Å²)
24+	7870±50	8090±25	7750±43
25+	8070±45	8170±26	7800±45
26+	8260±27	8250±27	7980±27
27+	8340±28	8330±28	8030±28
28+	8470±29	8460±29	8210±58
29+	8640±30	8650±30	8500±30

Table V-2 Calculated charge stripping for select charge states in 6560 mass spectra. The low and high collision energies for m/z values were obtained at 220 V and 500 V, respectively. The equation from figure 2G was: Charge Stripping = $0.000075 * (\Delta \text{ Adduct Mass}) + 0.006675$

Condition	Sample	Low m/z	High m/z	$\Delta m/z$	z	$\Delta \text{ mass (Da)}$	CS	CS %
Native	IgG1_1	5302	5298	4	28	112	0.015	1.5
Native	IgG1_2	5316	5295	21	28	588	0.051	5.1
Native	IgG1_3	5300	5295	5	28	140	0.017	1.7
Native	IgG2_1	5309	5306	3	28	84	0.013	1.3
Native	IgG2_2	5313	5304	9	28	252	0.026	2.6
Native	IgG2_3	5310	5303	7	28	196	0.021	2.1
Native	IgG4_1	5334	5328	6	28	168	0.019	1.9
Native	IgG4_2	5336	5330	6	28	168	0.019	1.9
Native	IgG4_3	5331	5326	5	28	140	0.017	1.7

Supercharged	IgG1_1	3618	3607	11	40	440	0.040	4.0
Supercharged	IgG1_2	3617	3611	6	40	240	0.025	2.5
Supercharged	IgG1_3	3622	3615	7	40	280	0.028	2.8
Supercharged	IgG1_1	4612	4606	6	35	210	0.022	2.2
Supercharged	IgG1_2	4117	4112	5	35	175	0.020	2.0
Supercharged	IgG1_3	4014	4010	4	35	140	0.017	1.7
Supercharged	IgG1_1	4805	4791	14	30	420	0.038	3.8
Supercharged	IgG1_2	4806	4792	14	30	420	0.038	3.8
Supercharged	IgG1_3	4809	4792	17	30	510	0.045	4.5
Supercharged	IgG4_1	3679	3671	8	40	320	0.031	3.1
Supercharged	IgG4_2	3684	3678	6	40	240	0.025	2.5
Supercharged	IgG4_3	3688	3678	10	40	400	0.037	3.7
Supercharged	IgG4_1	4072	4068	4	35	140	0.017	1.7
Supercharged	IgG4_2	4075	4070	5	35	175	0.020	2.0
Supercharged	IgG4_3	4080	4070	10	35	350	0.033	3.3
Supercharged	IgG4_1	4847	4843	4	30	120	0.016	1.6
Supercharged	IgG4_2	4851	4842	9	30	270	0.027	2.7
Supercharged	IgG4_3	4859	4844	15	30	450	0.040	4.0

Table V-3 RMSD values for Figure 6-3 G

Antibody	IgG1			IgG2			IgG4		
Replicate	1	2	3	1	2	3	1	2	3
29+	2.1	1.6	1.2	10.7	10.1	9.6	16.8	16.4	16.0
28+	1.8	1.3	0.9	10.0	9.5	8.9	16.8	16.5	16.3
27+	1.6	1.2	0.8	10.6	10.5	10.2	20.1	19.1	19.0
26+	1.3	1.0	0.8	11.5	11.9	12.1	23.1	21.3	21.2
25+	1.7	1.5	1.0	14.6	14.8	14.9	20.5	20.2	19.7
24+	2.2	2.0	2.0	19.3	19.3	19.0	19.2	18.8	18.1

Table V-4 RMSD values for Figure 6-3 H

Antibody	IgG1			IgG4		
	1	2	3	1	2	3
42+	2.00	1.98	1.86	10.0	10.8	11.6
41+	1.1	1.3	1.6	12.3	13.4	14.3
40+	0.95	0.86	1.03	12.8	13.7	14.0
39+	0.96	0.90	0.98	12.3	13.1	13.5
38+	0.81	0.70	0.83	12.2	13.3	13.6
37+	0.76	0.60	0.73	11.9	12.6	12.8
36+	0.66	0.57	0.63	12.1	12.3	12.5
35+	0.56	0.51	0.64	12.7	13.0	13.4
34+	0.74	0.62	0.74	13.8	14.0	14.4
33+	0.76	0.60	0.69	14.2	14.5	14.6
32+	0.90	0.88	0.98	14.6	14.7	14.7
31+	1.01	0.90	1.08	15.0	15.1	15.5
30+	1.01	0.88	1.08	15.7	16.0	16.2

Table V-5 RMSD values for Figure 6-4 G

Antibody	Kappa			Lambda			
	Replicate	1	2	3	1	2	3
29+		2.1	1.6	1.2	10.7	10.2	10.1
28+		1.8	1.3	0.9	8.2	7.9	8.0
27+		1.6	1.2	0.8	11.0	10.3	10.4
26+		1.3	1.0	0.8	11.9	11.5	11.6
25+		1.7	1.5	1.0	12.2	12.3	12.4

Table V-6 RMSD values for Figure 6-4 H

Antibody	Kappa			Lambda			
	Replicate	1	2	3	1	2	3
29+		2.1	1.1	1.6	14.3	15.3	15.4
28+		1.6	0.8	1.6	14.7	15.2	14.7
27+		1.4	1.1	1.3	20.2	20.3	19.9
26+		1.3	1.0	1.2	26.1	25.9	25.8
25+		1.4	1.8	1.9	24.9	26.0	25.9

Table V-7 RMSD values for Figure 6-5 C

Antibody	IgG1			IgG4		
Replicate	1	2	3	1	2	3
29+	1.0	1.50	1.1	11.3	11.8	12.0
28+	1.0	0.9	1.2	11.5	11.3	12.0
27+	0.85	0.68	0.73	13.5	13.4	13.1
26+	0.93	0.94	1.11	17.9	18.4	18.4
25+	1.5	1.4	1.8	16.8	16.7	16.6
24+	2.5	2.3	1.7	13.2	13.4	13.2

Table V-8 RMSD values for Figure 6-5 D

Antibody	IgG1			IgG4		
Replicate	1	2	3	1	2	3
42+	2.7	4.1	4.4	8.9	9.4	8.4
41+	3.8	3.7	5.7	16.0	15.1	15.6
40+	2.4	3.0	3.7	10.1	9.2	11.2
39+	1.6	1.9	2.9	12.0	12.2	12.9
38+	1.3	2.2	2.5	11.9	11.7	14.2
37+	1.4	1.6	2.2	12.4	12.4	14.9
36+	1.2	1.5	2.0	12.5	12.0	14.4
35+	1.1	1.0	1.6	11.6	11.7	14.2
34+	0.9	0.8	1.4	13.0	13.0	14.7
33+	0.8	1.1	1.5	14.5	14.5	15.9
32+	0.9	1.4	1.9	15.4	15.9	16.2
31+	1.7	1.2	2.4	16.0	15.9	16.0
30+	1.5	1.0	1.5	12.6	13.0	12.9

VI. Compounds Synthesized for Optimizing Intrinsically-charged Labeling of Proteins

VI-I. Overview

The effectiveness of the trimethyl pyrylium (TMP) reagent used for fixed-charge labeling of proteins in Chapters 2 and 3 for improved sequencing is limited by the low solubility and reactivity of the compound. The modification procedure employed in Chapters 2 and 3 employs a 24 hour reaction at room temperature, pH 9, at a molar excess of reagent to lysine residues ranging from 50-200:1. Increasing the ratio of reagent proved problematic, particularly for large proteins containing many lysine residues, due to the limited aqueous solubility of the reagent. In collaboration with the Andrews lab (Michigan Biological Chemistry) and Dr. Hollis Showalter and Susan Hagen (Vahlteich Medicinal Chemistry Core, Michigan Medicinal Chemistry), we attempted to develop reagents based on TMP with improved aqueous solubility and reactivity towards protein lysine residues under near-physiological pH conditions.

VI-II. Methods

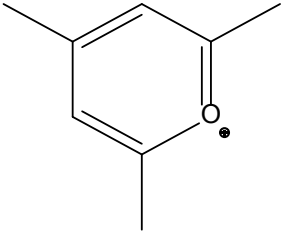
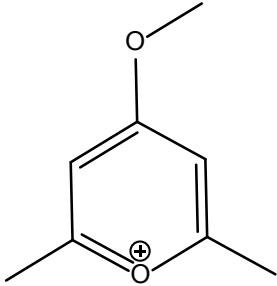
Synthesized compounds from the Medicinal Chemistry Core group were reacted with a test substrate, Substance P, to provide a comparative basis for reactivity in a system with a single lysine residue available. Aqueous solubility at the necessary concentrations for the modification reactions in proteins was evaluated by observing if a precipitate formed during upon resuspension to the desired concentration. A 200:1 ratio of reagent was mixed with 50uM Substance P at pH 9 in 100mM triethylammonium bicarbonate (TEAB) buffer for 1 hour. Reactions were quenched by addition of 1M ammonium acetate solution before cleanup. Excess reagent removal and buffer exchange were accomplished using C18 spin columns (Pierce #89870) according to manufacturer's instructions, with 4 aqueous washes employed instead of 2. Substance P was eluted in 50% acetonitrile with 0.1% formic acid for analysis by MS. Samples

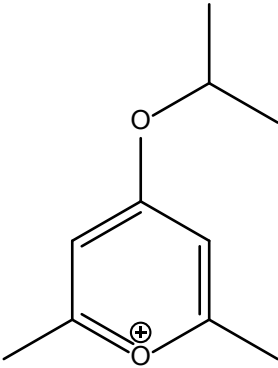
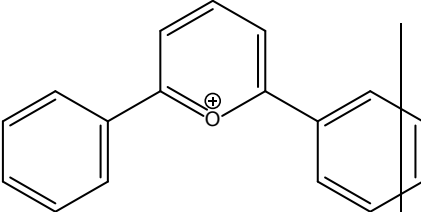
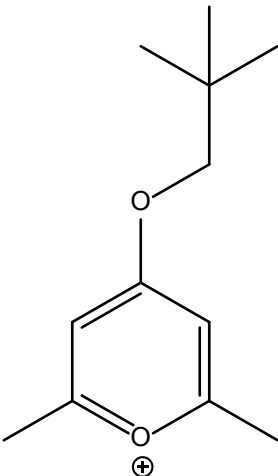
were analyzed by nanoESI-IM-MS on the Synapt G1 or G2 and the height of characteristic peaks for unmodified and modified forms of substance P were used to evaluate reaction completion. Activation of modified peptides in the trap region was used to evaluate the gas-phase stability of the modification.

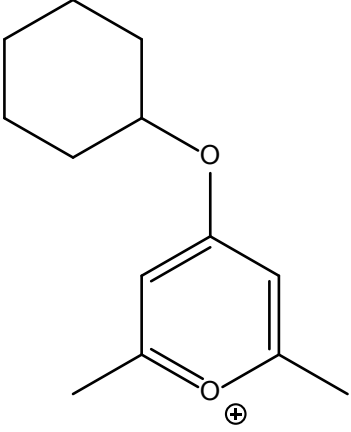
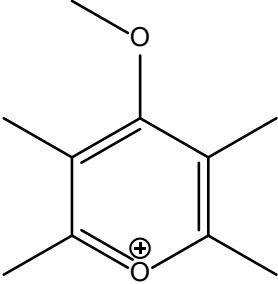
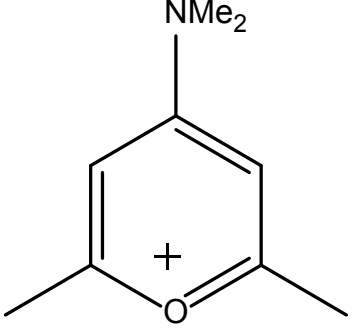
VI-III. Results

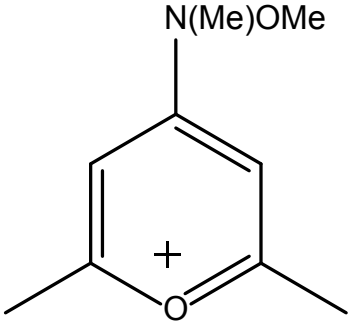
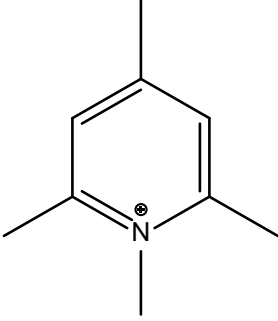
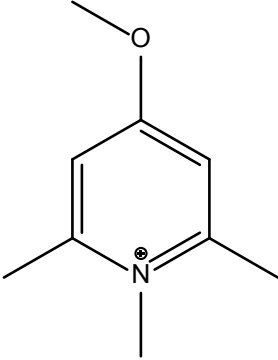
All compounds tested are compared in the table below. Many of the reagents exhibited an undesired side reaction product, labeled “B”.

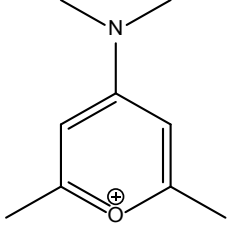
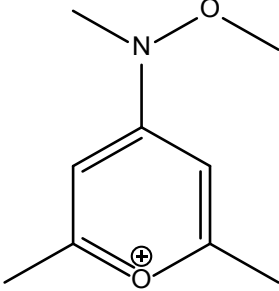
Table VI-1 Synthesized compounds tested and results of testing. “A” is the desired product of a pyrylium-like reaction, while “B” is an undesired side reaction introduced by -oxy substituents at the para position to the oxonium.

Name	Structure	% A formed	% B formed	Comments
2,4,6-trimethylpyrylium, BF ₄ salt	 $C_8H_{11}O^+$	75	0	Original compound - commercially available. Single reaction pathway. Limited reactivity and solubility
4-methoxy-2,6-dimethylpyrylium, triflate	 4-methoxy-2,6-dimethylpyrylium $C_8H_{11}O_2^+$	75	25	Much improved solubility and reactivity over initial compound, but also formed alternate product

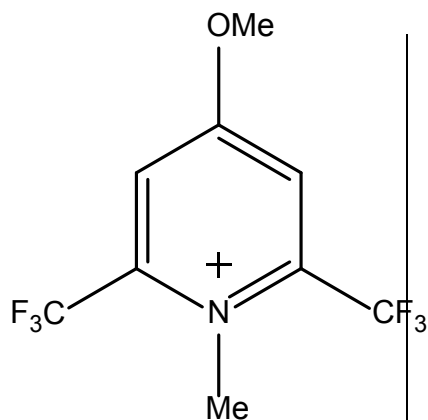
<p>4-isopropoxy-2,6-dimethylpyrylium, triflate</p>	 <p>$C_{10}H_{15}O_2^+$</p>	<p>75</p>	<p>10</p>	<p>Most selective of compounds tested. Required pH 10 OR 24 hours at pH 9 to get significant yield. Small yield of alternate product still caused problems for protein modification.</p>
<p>2,6-diphenylpyrylium, BF₄ salt</p>	 <p>$C_{17}H_{13}O^+$</p>	<p>0</p>	<p>0</p>	<p>Nearly completely insoluble in water. No reaction observed in water, DMSO, or mixed water/DMSO solutions.</p>
<p>2,6-dimethyl-4-(neopentyloxy)pyrylium, triflate</p>	 <p>$C_{12}H_{19}O_2^+$</p>	<p>33</p>	<p>66</p>	<p>Soluble and highly reactive, very similar to the 4-methoxy dimethyl pyrylium. Increasing pH to 10 improved selectivity, but still favored the alternate over primary product.</p>

<p>4-(cyclohexyloxy)-2,6-dimethylpyrylium, triflate</p>	 <p>$C_{13}H_{19}O_2^+$</p>	<p>20</p>	<p>60</p>	<p>Less reactive and less selective (or more selective for the alternate product rather than the primary)</p>
<p>4-methoxy-2,3,5,6-tetramethylpyrylium, triflate</p>	 <p>$C_{10}H_{15}O_2^+$</p>	<p>20</p>	<p>30</p>	<p>**Mix of compound and starting material, so reactivity analysis not perfect**. Challenging to analyze, but appears to be less reactive than most and generally favor the alternate product</p>
<p>4-(dimethylamino)-2,6-dimethylpyrylium</p>	 <p>$C_9H_{14}NO^+$</p>	<p>0</p>	<p>0</p>	<p>No Reaction</p>

<p>4-(methoxy(methyl)amino)-2,6-dimethylpyrylium</p>	 <p>$C_9H_{14}NO_2^+$</p>	<p>20</p>	<p>0</p>	<p>Selective for the correct form, but limited reactivity. Also shown to be readily cleavable in CID</p>
<p>1,2,4,6-tetramethylpyridin-1-ium, triflate</p>	 <p>$C_9H_{14}N^+$</p>	<p>0</p>	<p>0</p>	<p>No Reaction</p>
<p>4-methoxy-1,2,6-trimethylpyridin-1-ium, triflate</p>	 <p>$C_9H_{14}NO^+$</p>	<p>0</p>	<p>0</p>	<p>No Reaction</p>

4-(dimethylamino)-2,6-dimethylpyrylium	 4-(dimethylamino)-2,6-dimethylpyrylium $C_9H_{14}NO^+$	0	0	No Reaction
4-methoxy(methyl)amino-2,6-dimethylpyrylium	 4-(methoxy(methyl)amino)-2,6-dimethylpyrylium $C_9H_{14}NO_2^+$	25	0	Low reactivity, but with desired selectivity. *NOTE: loss of HOCH3 can occur under CID conditions, but wasn't a major dissociation path for modified peptide*
Unsuccessful Syntheses	Structure			

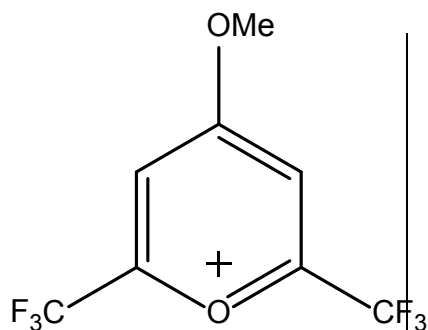
4-methoxy-1-methyl-2,6-bis(trifluoromethyl)pyridin-1-ium, triflate



$C_9H_8F_6NO^+$

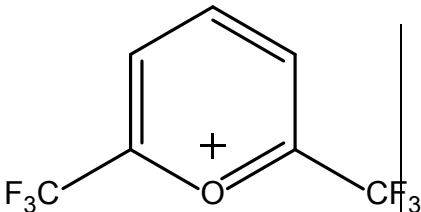
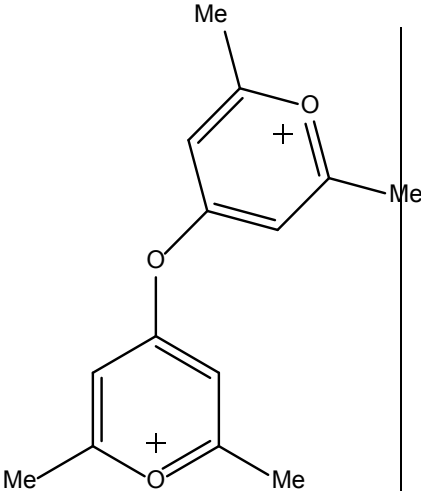
Precursor 4-OMe pyridine made. CF₃ groups make pyridine nitrogen extremely electron deficient, thus not reactive to further alkylating agents, such as methyl triflate even at high temp. Any product that forms likely unstable, reverting to starting material. Not enough e- density in the ring to do the chemistry required to form these compounds

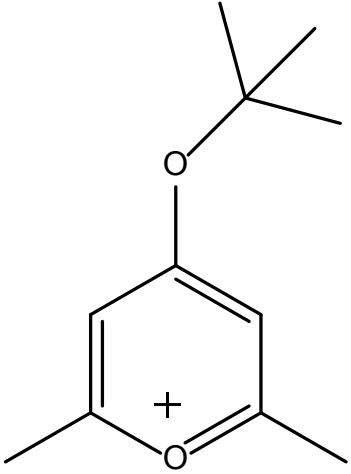
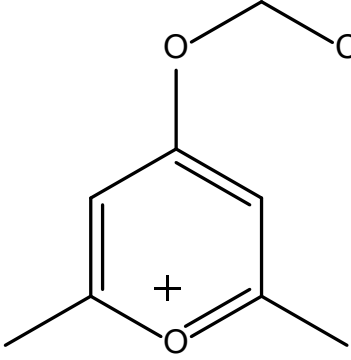
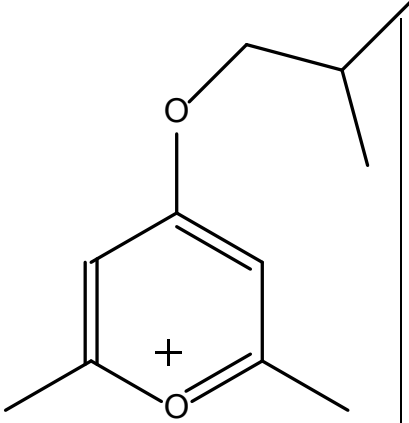
4-methoxy-2,6-bis(trifluoromethyl)pyrylium, triflate

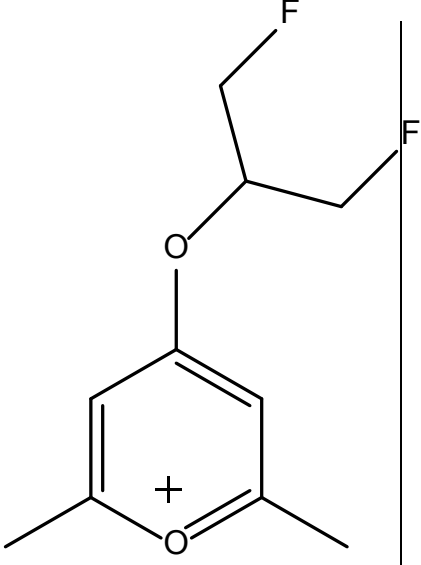
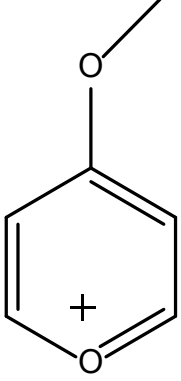


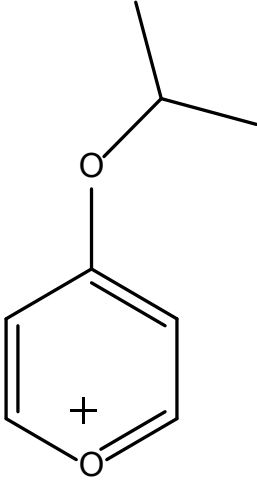
$C_8H_5F_6O_2^+$

Methyl triflate at 120 deg gives no reaction. Also no reaction by triflic anhydride procedure. Difficult to make due to deactivating effect of CF₃ groups; O-Me bond likely quite weak so if

			formed would readily revert to starting ketone
2,6- bis(trifluoromethyl)pyryliu m, BF4 or perchlorate salt	 $C_7H_3F_6O^+$		Several attempts made using literature procedures analogous to those for the 2,6-diphenyl congeners. Saw no product formation by TLC
4,4'-oxybis(2,6- dimethylpyrylium), bis triflate	 $C_{14}H_{16}O_3^{2+}$		Made according to literature procedure. Very hygroscopic and limited stability

<p>4-(tert-butoxy)-2,6-dimethylpyrylium, triflate</p>	 <p>$C_{11}H_{17}O_2^+$</p>		<p>No reaction using the triflic anhydride procedure</p>
<p>2,6-dimethyl-4-(2,2,2-trifluoroethoxy)pyrylium, triflate</p>	 <p>$C_9H_{10}F_3O_2^+$</p>		<p>Triflic anhydride procedure. Likely product formation, but appears to revert to methyl ether upon purification eluting with 3% MeOH in DCM</p>
<p>4-isobutoxy-2,6-dimethylpyrylium, triflate</p>	 <p>$C_{11}H_{17}O_2^+$</p>		<p>Triflic anhydride procedure. Likely product formation, but appears to revert to methyl ether upon purification eluting with 3% MeOH in DCM</p>

<p>4-((1,3-difluoropropan-2-yl)oxy)-2,6-dimethylpyrylium, triflate</p>	 <p>$C_{10}H_{13}F_2O_2^+$</p>	<p>Triflic anhydride procedure. Likely product formation, but reverts to 4:1:1 mixture of product, 4-pyrone, and 4-i-Pr ether upon purification eluting with 4% IPA in DCM</p>
<p>4-methoxypyrylium, triflate</p>	 <p>$C_6H_7O_2^+$</p>	<p>Attempted to make by both methyl triflate and triflic anhydride procedures. Both appeared to give 1:1 mixture of sm:pdt by TLC. But HPLC (sm only) and NMR (1:1 sm:prd) show product is unstable</p>

<p>4-isopropoxyppyrylium, triflate</p>	 <p>$C_8H_{11}O_2^+$</p>	<p>Similar results to Me ether. Crude product by NMR appears to revert to pyrone upon purification</p>
--	--	--

VI-IV. Summary

Despite trying many new compounds, we were unable to find any that provided improved reactivity and solubility without either also introducing competing, off-target pathways or becoming CID-cleavable. Reactivity of the initial TMP reagent is much improved in small and/or unstructured systems than globular proteins and protein complexes, indicating that some of the reactivity challenges may be a result of surface inaccessibility of lysine residues. The most promising reagent was 4-methoxy-2,6-dimethylpyrylium, which improved reactivity and solubility dramatically, but introduced a competing reaction pathway at the methoxy group. Introduction of other electron-withdrawing substituents at the 4-position that cannot provide a competing pathway could provide a path to the desired improvements in a reagent.

VII. Protamine Analysis Procedure

VII-I. Overview

Protamines have traditionally been considered inert nuclear proteins serving as passive structural elements that condense the paternal genome. However, emerging evidence calls for a need to revisit protamine protein's presumed biological function.¹⁹ A mass spectrometry analysis revealed that mouse protamines bear a number of post-translational modifications (PTMs),²⁰ indicating that these modifications may bear a "protamine code" analogous to the "histone code." So far, however, none of these novel modifications have been explored, nor have the corresponding human protamine proteins. In collaboration with Dr. Sue Hammoud and Dr. Samantha Schon (Michigan Medicine), as well as Dr. Philip Andrews (Michigan Biological Chemistry), we developed a protocol to evaluate the proteoforms of human protamine samples using top-down mass spectrometry.

VII-II. Methods

Protamine proteins were obtained by Dr. Samantha Schon from discarded semen samples from men undergoing routine semen analysis and the University of Michigan Center for Reproductive Medicine using an acid extraction protocol. Extracted samples were dialyzed for 24-48 hours at 4C in a 2 kDa MWCO cassette against 200mM ammonium acetate, pH 7. Note that dialysis against lower concentrations of ammonium acetate resulted in adduction of substantial amounts of phosphoric acid during MS analysis, likely to do the highly basic nature of the protamine proteins. 5 μ L of dialyzed protein solution was transferred to a gold-coated borosilicate capillary (0.78 mm i.d., Harvard Apparatus, Holliston, MA) for direct infusion using a Nanospray Flex ion source (Thermo Scientific, Waltham, MA). Capillary voltage was 1.6 kV, transfer capillary temperature was 275C. Intact mass analysis was conducted in the Orbitrap analyzer with a

resolution of 120,000 at m/z 400 with an AGC target of 1e6 and 5 microscans and data was accumulated for several minutes to ensure sufficient signal-to-noise ratio for deconvolution. Data was analyzed using BioPharmaFinder v3.0 software. Intact mass deconvolution using BioPharmaFinder was performed in “average over selected retention time” mode with a mass range of 3000-20000.

VII-III. Results

Protamine exhibits several isoforms and the potential for several PTMs, including phosphorylation. The isoforms present in Uniprot and associated abbreviations are displayed in Figure 1. MS analysis of the intact masses of protamine peaks indicates the abundances shown in


Names & Taxonomy ¹		Chain	Name
Protein names ¹ Recommended name: Protamine-2 Alternative name(s): <ul style="list-style-type: none"> Sperm histone P2 Sperm protamine P2 Cleaved into the following 7 chains: <ul style="list-style-type: none"> Basic nuclear protein HPI1 Basic nuclear protein HPI2 Basic nuclear protein HPS1 Basic nuclear protein HPS2 Sperm histone HP4 Alternative name(s): "PRM2_HP4" <ul style="list-style-type: none"> Sperm protamine P4 Sperm histone HP2 Alternative name(s): "PRM2_HP2" <ul style="list-style-type: none"> Sperm protamine P2 Short name: P2' Sperm histone HP3 Alternative name(s): "PRM2_HP3" <ul style="list-style-type: none"> p2'' Sperm protamine P3 	2 - 102	Basic nuclear protein HPI1	
	22 - 102	Basic nuclear protein HPI2	
	34 - 102	Basic nuclear protein HPS1	
	37 - 102	Basic nuclear protein HPS2	
	45 - 102	Sperm histone HP4	
	46 - 102	Sperm histone HP2	
	49 - 102	Sperm histone HP3	
Gene names ¹	Name: PRM2		
Organism ¹	Homo sapiens (Human)		
Taxonomic identifier ²	9606 [NCBI]		
Taxonomic lineage ¹	Eukaryota > Metazoa > Chordata > Craniata > Vertebrata > Euteleostomi > Mammalia > Eutheria > Euarchontoglires > Primates > Haplorrhini > Catarrhini > Hominidae > Homo 		
Proteomes ¹	UP000005640 Component ¹ : Chromosome 16		

Figure VII-1 Naming scheme used to describe protamine isoforms. There are two known protamine sequences: protamine 1 (PRM1), which has only one isoform, and protamine 2 (PRM2), which has the 7 isoforms listed in this figure.

Figure 2 following deconvolution. The most abundant isoform is protamine 2 sperm histone HP3, followed by protamine 1, then protamine 2 sperm histone HP4. A small amount of single

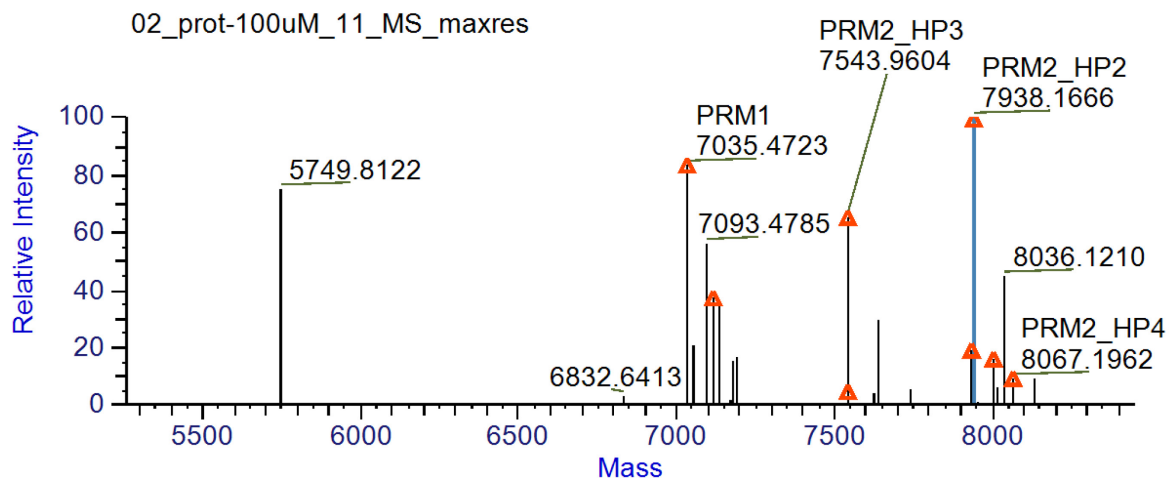


Figure VII-2 Intact masses of protamine proteoforms observed from pooled samples. A number of unlabeled peaks correspond to adducts of phosphoric acid or mass shifts with unknown annotation. The peak at m/z 5749 is observed in some samples, but has not been annotated to a protamine (or other) sequence.

and double phosphorylation was consistently observed on protamine 1, but no PTMs were detected on protamine 2.

VIII. References

- (1) Samulak, B. M.; Niu, S.; Andrews, P. C.; Ruotolo, B. T. Ion Mobility-Mass Spectrometry Analysis of Cross-Linked Intact Multiprotein Complexes: Enhanced Gas-Phase Stabilities and Altered Dissociation Pathways. *Anal. Chem.* 2016, 88 (10), 5290–5298.
- (2) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; et al. CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* 2009, 31 (4), NA-NA.
- (3) Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* 2012, 52 (12), 3144–3154.
- (4) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* 2012, 52 (12), 3155–3168.
- (5) Mullner, D. Modern Hierarchical, Agglomerative Clustering Algorithms. eprint arXiv:1109.2378 2011.
- (6) Zepeda-Mendoza, M. L.; Resendis-Antonio, O. Hierarchical Agglomerative Clustering. In *Encyclopedia of Systems Biology*; Dubitzky, W., Wolkenhauer, O., Cho, K.-H., Yokota, H., Eds.; Springer New York: New York, NY, 2013; pp 886–887.
- (7) Walt, S. van der; Colbert, S. C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* 2011, 13 (2), 22–30.
- (8) Oliphant, T. E. Python for Scientific Computing. *Comput. Sci. Eng.* 2007, 9 (3), 10–20.
- (9) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* 1983, 22 (12), 2577–2637.
- (10) Touw, W. G.; Baakman, C.; Black, J.; te Beek, T. A. H.; Krieger, E.; Joosten, R. P.; Vriend, G. A Series of PDB-Related Databanks for Everyday Needs. *Nucleic Acids Res.* 2015, 43 (Database issue), D364-8.
- (11) Marklund, E. G.; Degiacomi, M. T.; Robinson, C. V.; Baldwin, A. J.; Benesch, J. L. P. Collision Cross Sections for Structural Proteomics. *Structure* 2015, 23 (4), 791–799.
- (12) Larriba, C.; Hogan, C. J. Free Molecular Collision Cross Section Calculation Methods for Nanoparticles and Complex Ions with Energy Accommodation. *J. Comput. Phys.* 2013, 251, 344–363.
- (13) Larriba, C.; Hogan, C. J. Ion Mobilities in Diatomic Gases: Measurement versus Prediction with Non-Specular Scattering Models. *J. Phys. Chem. A* 2013, 117 (19), 3887–3901.
- (14) Dowdy, S. M.; Wearden, S.; Chilko, D. M. *Statistics for Research*, 3rd ed.; Wiley-Interscience: Hoboken, N.J., 2004.
- (15) Arlot, S.; Celisse, A. A Survey of Cross-Validation Procedures for Model Selection. *Stat. Surv.* 2009, 4 (0), 40–79.
- (16) Xanthopoulos, P.; Pardalos, P. M.; Trafalis, T. B. *Linear Discriminant Analysis*; Springer, New York, NY, 2013; pp 27–33.
- (17) Izenman, A. J. *Support Vector Machines*; Springer, New York, NY, 2013; pp 369–406.
- (18) Newville, M.; Ingargiola, A.; Stensitzki, T.; Allen, D. B. *LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python*. Zenodo 2014.

- (19) Björndahl, L.; Kvist, U. Human Sperm Chromatin Stabilization: A Proposed Model Including Zinc Bridges. *Mol. Hum. Reprod.* 2009, 16 (1), 23–29.
- (20) Brunner, A. M.; Nanni, P.; Mansuy, I. M. Epigenetic Marking of Sperm by Post-Translational Modification of Histones and Protamines. *Epigenetics Chromatin* 2014, 7 (1), 2.