Methods and Informatics for Gas-Phase Structural Biology and Drug Discovery

by

Joseph D. Eschweiler

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemistry)
in the University of Michigan
2017

Doctoral Committee:

Associate Professor Brandon T. Ruotolo, Chair
Professor Phillip C. Andrews
Professor Charles L. Brooks III
Assistant Professor Daniel R. Southworth

Joseph D. Eschweiler

joeesch@umich.edu

ORCID ID: 0000-0002-2486-9726

Dedication

I dedicate this dissertation to my dad, Damian Eschweiler, who I know would have loved nothing more than to read it in its entirety. He taught me the value of critical thinking, and without his influence this accomplishment would not have been possible.

Acknowledgements

I cannot acknowledge my advisor, Professor Brandon Ruotolo enough for his sustained support of my scientific growth. His high standards in terms of scientific rigor and quality in all forms of scientific communication have shaped me as a scientist and will continue to impact my work for many years to come. Not only did Brandon help me achieve more than I thought was possible, he did so in a way that made the doctoral experience *enjoyable*, allowing me the freedom to explore new areas and find my own internal motivation for my work.

I am also incredibly thankful to my doctoral committee members Professors Phil Andrews, Charles Brooks, and Dan Southworth. Your standards for scientific rigor and insightful questions steered my learning and the direction of my research from my candidacy exam all the way through the composition of this manuscript. I extend this acknowledgment to the entire UM Chemistry Faculty, who in my experience have been incredibly receptive to my questions, and together set an incredible standard for how a department should function as educators and scientists.

My fellow graduate students have also contributed to an excellent academic culture, both in the lab, the department and the entire campus. My Ruotolo lab colleagues, both past, and present have been extremely helpful, and give me great hope for the future of our research. I started my time in the lab by continuing the projects put forth by the first-ever graduate student from our group, Yueyang Zhong.

Yueyang's training, her preliminary data, and her writings formed the foundation for my research and I am incredibly grateful for her hard work and patience with me. The same can be said for other senior lab members, including Linjie Han, who allowed me to pick his brain many late nights in lab, and Russ Bornschein, who knew the answers to any instrumental question I could throw at him. The other lab members, whom I spent most of my time with over the last 5 years were also of great influence. Special thanks to Jessica Gibbons, who kept me on track during my first few years in the lab using Pavlovian principles and contributed to the CIUSuite platform that unfortunately only I get credit for. I'm confident in leaving the lab in the capable hands of the next most senior lab members, Yuwei Tian, Sugyan Dixit, and Dan Polasky. It's been an honor to watch these scientists grow and I know they will take future research in the lab to new levels. I also acknowledge the rotators, undergraduates, and junior lab members who I have had the privilege of working with, especially Rachel Martini, Chunyi Zhao, Sarah Fantin, and Daniel Vallejo, I know that you all will go on to do great things in your scientific careers.

Importantly, the majority of the science I did throughout graduate school would not have been possible without fantastic collaborators. Bob Hausinger and Mark Farrugia from Michigan State University set the bar for great collaborative relationships early on in my graduate career, and gave me confidence in working with other, more experienced scientists. Special thanks to Aaron Sciore, Ajitha Christie-David, Somaye Badieyan, Kyle Ferguson, Sang Joon Won, Brent Martin and Neil Marsh for giving me the opportunity to work on some of the most interesting biochemical projects I could

have asked for. I also acknowledge Aaron Frank for his support and mentorship in much of my computational modeling work.

It would be remiss to not also acknowledge some of my early scientific mentors. From UW-Milwaukee, Mary Knasinski, Henry Tomasiewicz, Andy Pacheco, and Matt Youngblut were such strong academic and scientific role models for me during my college years and I am so thankful they were there to steer me toward my current path. From the University of Michigan, I also acknowledge Raoul Kopelman for providing me with my first graduate research opportunities during my summer rotation. Brent Martin and Jaimeen Mujmadar were also fantastic mentors during my early graduate school career, and I thank them for their insights during my most formative scientific years.

Last but certainly not least, I would like to thank my family, specifically my mother Peggy Eschweiler and my fiancée Katy Robb. My mom has been so supportive of me over these last 5 years, and has always been there when I was in a pinch. If it weren't for her, I could never have made it this far in life. Perhaps the greatest thing to come out of my Ph.D. years at Michigan was my engagement to the all-around wonderful Katy Robb. Katy provided me with the emotional support and motivation to do excellent work throughout almost my entire time in graduate school, all while keeping me out of trouble. I can't wait to marry her.

Table of Contents

List of Figures

List of Tables

List of Appendices

Abstract

Methods for rapid interrogation of structure and stability attributes of proteins and protein complexes are becoming increasingly important for developing our understanding of biology and the development of pharmaceuticals. Gas-phase technologies such as mass spectrometry and ion mobility spectrometry have proven valuable in these endeavors, as they provide unique perspectives on the solution-phase equilibrium of protein complexes and their conformations. Before fully harnessing the information derived from these gas-phase techniques, new approaches for data analysis and mechanistic understanding of gas-phase protein structure are necessary. In this dissertation, we develop ion mobility mass spectrometry methods and informatics for the study of gas-phase proteins, multiprotein complexes, and protein-small molecule complexes.

In the first half of the dissertation, novel data analysis tools and experimental methodologies are outlined for the study of gas-phase protein unfolding. After providing the software tools necessary for robust analysis of gas-phase unfolding trajectories in Chapter 2, we turned our attention to understanding the mechanism of unfolding for large multidomain proteins. In Chapter 3, we focus on the factors driving changes in unfolding trajectories for a variety of serum albumin homologues, and through the use of novel unfolding experiments utilizing chemical probes and non-covalent protein constructs, a detailed mechanism for solvent-free protein unfolding is provided.

Subsequent chapters in the dissertation focus on the characterization of multiprotein complexes, especially through the use of ion mobility-mass spectrometry and coarse-grained modeling. In chapter 4, we develop and benchmark new algorithms for translating ion mobility and mass spectrometry datasets into coarse grained models. These studies outline the limits in current coarse-graining methodologies, and define the minimum restraint sets necessary to generate high confidence multiprotein models. Additionally, best practices for dealing with ambiguous models resulting from sparse datasets are described. In chapter 5, the tools developed in the previous chapter are applied to structurally characterize the urease pre-activation complex, a transient 18-subunit complex that is a target for inhibition of urease-related pathology. When our ion mobility-mass spectrometry datasets are combined with previously published chemical crosslinking and x-ray scattering data, a discrete population of conformations for the urease pre-activation complex emerges which compares favorably to previous models generated using computational techniques.

In Chapter 6, I highlight more applications of ion mobility-mass spectrometry to engineered and naturally occurring protein complexes. These applications highlight the power of ion mobility mass-mass spectrometry datasets for rapid analysis of protein oligomerization state and structure, providing a basis for further integration of the technology into pharmaceutical and structural biology workflows.

Chapter 1: Introduction

Elements of this chapter are taken from:
Eschweiler J.D.; Rabuck-Gibbons, J. N.; Kerr, R.; Ruotolo, B.T.: Sizing Up Protein–Ligand Complexes: The Rise of Structural Mass Spectrometry Approaches in the Pharmaceutical Sciences, Annu. Rev. Anal. Chem. 10 (2017)

The study of biochemical processes is imperative for continued advances in medicine, psychology, human health, food production, and environmental conservation. In an era where full genomic datasets are available for many important species, including thousands of human genomes,[1] it has become clear that the richness of biological complexity is not contained solely in nucleic acid sequences, but in the downstream chemical and physical processes that occur when gene products interact with each other and with the environment.[2-4] These gene products, proteins, are the biochemical machinery of life, each having unique structures and propensities for interaction with other biomolecules and their environments.[5] It has been the goal of structural biologists to understand the conformations and functions of proteins for nearly 100 years,[6] and these structural insights have been crucial for the discovery of pathological mechanisms and therapeutics.

Throughout history, most drugs have been small molecules that modulate the activity of a protein or protein class.[7] Though this remains true today, the fastest growing class of FDA-approved drugs are now proteins themselves, specifically monoclonal antibodies similar to those produced by the mammalian immune system.[8]

The development of new drugs, of both classical and biologic nature continues to hinge on the structural elucidation of proteins.[9,10]

Since the first studies of protein tertiary structure by X-ray crystallography in 1958,[11] over 119,785 high-resolution protein structures have been deposited in the protein data bank (PDB) to date.[12] Despite continued growth in PDB entries every year, the number of structures for new protein folds has been stagnant since 2008. This stagnation is due in part to physical and practical limitations of “gold standard” structural biology techniques for characterizing highly dynamic or unstructured proteins, as well as proteins that exist in low quantities in cells or with low solubility.[13] The ability to fully characterize the structural space of proteins is further complicated by the fact that most proteins are involved in dynamic interactions with other proteins,[14] nucleic acids,[15] or small molecules[16] that result in heterogeneous populations of protein conformation and stoichiometry that complicate analysis. In order to characterize these highly complex biological mixtures, it is necessary that we construct technologies that merge cutting-edge separation science into next-generation structure analysis tools.[17] One such intersection occurs in the gas-phase, where ion mobility-mass spectrometry (IM-MS) is emerging as a robust tool for the simultaneous separation and determination of protein stoichiometries and conformations, allowing access to highly-specific structural information within complex mixtures.[18,19] In this dissertation, we explore the application of IM-MS for gas-phase structural biology and drug discovery.

1.1 High-resolution Protein Characterization

Although the 3D structure of a protein is dictated by its amino acid sequence under a given set of conditions,[20] methods for predicting protein structure from sequence information *de novo* are still unreliable.[21] While template-based modeling is much more robust,[22] these methods are limited to proteins that share well-defined structural elements with proteins that have already been experimentally characterized at high resolution.[23] Thus, direct determination of protein structure by experiment remains the predominant approach in most structural biology and drug discovery laboratories. Of the experimental techniques for determination of protein structure, X-ray diffraction (XRD)[24] and nuclear magnetic resonance spectroscopy (NMR)[25] remain among the highest resolution approaches available currently, and thus make up the majority of structures deposited in the PDB. XRD experiments characterize crystals of individual proteins,[26] multiprotein complexes,[27,28] and protein-ligand complexes,[29] with resolution values often under 3 Å. After over 80 years of development, collecting diffraction data and appropriately processing these signals is now often routine; however, our ability to generate high-quality crystals for proteins of high mass, disordered structure, low solubility or concentration in an expedient fashion is still a limiting factor for such approaches.[30,31], Moreover, the inability of XRD to capture protein dynamics and deal with heterogeneous populations has limited its use for many complex systems.[32]

NMR, on the other hand, is well-suited to capturing protein dynamics, often at resolutions rivaling that of XRD.[33] Additionally, NMR captures proteins in solution rather than a crystalline phase, increasing confidence in structural assignments especially when studying interactions with other proteins or small molecules.[34] Despite these

advantages, NMR typically requires large amounts of soluble protein, and has limitations associated with protein mass due, in part, to line broadening and the increasing number of distance restraints needed to restrain large numbers of atoms.[13] Although many reports have overcome major challenges in NMR analysis,[33,35-39] these methods often require large amounts of optimization, purified protein, and custom data analysis techniques for each system.

To fill the gap in characterization of large proteins and protein complexes, cryo-electron microscopy (CEM) has emerged as a powerful tool capable of resolution generally around 4.5 Å, with recent examples demonstrating resolution in the 2 Å range[40-42]. Advantages of cryo-EM include its ability to capture snapshots of even the largest protein complexes in a variety of dynamic states in solution.[43] Despite great promise for this technology, reconstructing 3D images from a series of 2D micrographs remains challenging, especially when heterogeneous samples make identification of particles challenging.[44]

Although the requirements of each of these techniques vary, one common theme is the need for homogenous samples in order to achieve high resolution analysis. In the case of multiprotein or protein-ligand complexes, this requirement is severely limiting, as in many cases an equilibrium distribution of structural states and stoichiometry exists within a sample.[45] Currently, it is not possible to simultaneously characterize all of the components in such a complex sample with high-resolution technologies, as it is necessary to artificially push the equilibrium toward forming a particular species before analysis.[46] Moreover, in the context of drug discovery, crystallization or NMR analysis of large libraries of drug variants is both expensive and time consuming, making high

resolution technology amenable to very few drug leads in any development pipeline.[47,48] In many cases, however, lower resolution technologies can be utilized to answer more targeted questions about protein structure and interactions, and do so on much faster timescales than high resolution structural efforts.

1.2 Targeted Methods in Structural Biology

Targeted methods are technologies that lack the ability to produce high-resolution 3D structures of proteins and protein complexes, but have specific advantages in speed, sensitivity, or specific information content over high resolution methods. Targeted methods are often used in the context of drug discovery or integrative structural biology, as will be discussed below. Discussion of targeted methods in structural biology could conceivably cover hundreds of molecular biology, biochemistry, and analytical chemistry experiments. For brevity, only the most commonly used and integral methods with be highlighted here, with a focus on analytical technologies rather than those tools that are primarily genetic or biochemical in nature.

1.2.1 Methods for protein structure

The first step in studying protein structure is often mass and sequence analysis by mass spectrometry.[49,50] For intact mass analysis, proteins may be denatured and analyzed most commonly by electrospray[50] or matrix assisted laser desporption ionization (MALDI)[51] and time of flight mass spectrometry instruments. Alternatively, proteins may be digested with specific proteases before the resulting peptides are

separated and analyzed by liquid chromatography and tandem mass spectrometry to recapitulate the protein sequence.[52] These techniques, although powerful, do not provide information about 3D protein structure. Optical methods are available for detecting secondary structure elements within proteins, such as circular dichroism (CD) which can monitor the relative amounts of alpha helices, beta sheets, and unstructured regions within a protein sample.[53,54] Alternatively, small angle x-ray scattering (SAXS) profiles provide information on the overall topology of a protein that can aid in structural modeling when combined with mass and secondary structure information.[55-57] If more detailed hypotheses for protein structure are available, the use of engineered Förster resonance energy transfer (FRET) tags into the amino acid sequence of a protein provide the opportunity for distance measurements to be made between the tags at given positions using fluorescence spectroscopy.[58,59]

Akin to this distance-based experiment, chemical crosslinking (CXL) of lysine and other residues within a protein is often used to provide information about the proximity of certain residues.[60,61] These chemical crosslinkers have known length, so mass spectrometric analysis of crosslinked peptides may provide information on the maximum distance between crosslinked residues. Additionally, new chemical labeling strategies such as fast photochemical oxidation of proteins (FPOP)[62,63] and hydrogen deuterium exchange (HDX)[64,65] are emerging as reliable methods for determining the solvent accessible residues of proteins.

1.2.2. Methods for Multiprotein Complexes

Genome-wide mapping of protein-protein interactions has provided the scientific community with detailed maps of modular networks of protein interactions with cells.[14,66] These maps are largely the result of yeast two hybrid experiments that utilize gene regulation machinery to report on the presence of given protein-protein interactions. Although these maps are crucial for identifying interacting networks of proteins, they do not provide information on the stoichometry, composition, or structure of individual complexes.[67]

Due to the noncovalent nature of multiprotein complexes, it is not always trivial to determine the intact mass, composition or structure of these analytes .[13] Classically, size exclusion chromatography (SEC) combined with multiangle light scattering[68] has been used for absolute determination of protein molecular weight and size, however this family of techniques only provides rough estimates of assembly mass, maximally precise to about 1 kDa. Some methodologies from the previous section have been extended to analyze multiprotein complexes, including SAXS,[56,57,69] HDX,[70-73] and FPOP,[74] but in all of these cases the presence of heterogeneous populations of protein complexes presents significant challenges in data analysis. CXL has proven to be particularly useful in mapping the topology of multiprotein complexes. When CXL is applied, even highly labile multiprotein complexes become tethered together and can survive separations in SEC or denaturing gels.[75,76] When this technology is combined with analysis of crosslinked peptides by mass spectrometry, CXL provides information on the pairwise interactions between crosslinked peptides, providing critical structural insights.[77,78]

In the last 20 years, native mass spectrometry and ion mobility mass spectrometry have also emerged as important methods for studying multiprotein complexes. The details of these technologies will be discussed in in section 1.3

1.2.3 Methods for Protein-Ligand Complexes

As the basis of modern enzymology and much of the pharmaceutical sciences, protein-ligand interactions are some of the most well-studied phenomena in biochemistry.[79] A large portion of the work in these fields revolves around measuring interaction strengths between proteins and small molecules, which can range from extremely strong nanomolar (nM)-range dissociation constants (k_D) to nearly undetectable fleeting interactions in the millimolar (mM) k_D range.[80] Technologies for the measurement of interactions strengths rely largely on optical spectroscopy, the most commonly employed being fluorescence anisotropy and surface-plasmon resonance techniques.[80,81] Although recognized as gold standard approaches for characterization of binding strength, these tools provide little to no structural information concerning protein-ligand complexes studied. Short of high-resolution 3D structural characterization, there are very few technologies capable of routinely localizing ligand binding to proteins with multiple or unknown binding sites. In some cases, hypothesis-driven mutation experiments provide the best evidence for a ligand binding pocket, however, these experiments are time consuming and require advanced molecular biology techniques.[82,83] Chemical labeling strategies coupled to MS, like HDX and FPOP, have recently been shown to provide structural insights into protein-ligand binding;[65,84-86] however these technologies are also relatively time consuming and may

not always be sensitive to all ligand binding events. As with multiprotein complexes, gas-phase technologies have recently been developed for ligand localization, especially those relying on tandem MS technology, which will be discussed in subsequent sections in this chapter.

1.3 Ion Mobility - Mass Spectrometry Instrumentation in Structural Biology

In the current structural biology landscape, mass spectrometry-based techniques have become indispensable for analysis of proteins or protein complexes that are refractory to high-resolution techniques. Structural mass spectrometry has evolved from its origins in fragmentation and sequencing of digested peptides to new areas that include intact mass measurement of noncovalent complexes,[87] ion mobility analysis of native-like ions,[17] and gas-phase calorimetry.[88] To understand these new applications, an overview of the fundamentals of native mass spectrometry and ion mobility spectrometry is necessary.

1.3.1 Ionization and Preservation of Native-like Proteins

Since 1989, electrospray ionization (ESI) has been used to ionize large biomolecules once thought to be too labile to survive the harsh transition from solution to gas-phase.[50] This method, in contrast to other techniques that involve the bombardment of samples with ions, electrons, or photons, utilizes an electrokinetic spray to more gently ionize biomolecules through charged solvent droplets . The droplets generated in this fashion undergo fission and eventual evaporation as they cross the potential and pressure gradients on their path into the mass spectrometer.[89]

This process produces a distribution of multiply charged ions that, for large proteins, depends largely on the surface area of the protein in the final stages of the electrospray process.[90,91] Further refinements to the electrospray process involved decreasing flow rates and potential gradients to yield increasingly smaller droplets which serve to increase overall ionization efficiency. The modern implementation of ESI used for structural MS operates at nL/min flow rates and utilizes a conductive capillary with an orifice on the order of 1 μm in diameter. This technology, deemed nanoelectrospray ionization (nESI) tolerates 100% aqueous buffers having a relatively wide range of ionic strengths, allowing for proteins to be electrosprayed from native buffers at room temperature.[92] In contrast to early ESI measurements of intact proteins made from acidified and partially organic solutions, nESI from native buffer results in a distribution of charge states both considerably lower and narrower, indicating that nESI is capable of generating more compact ions with memory of their solution-phase structure.[93]

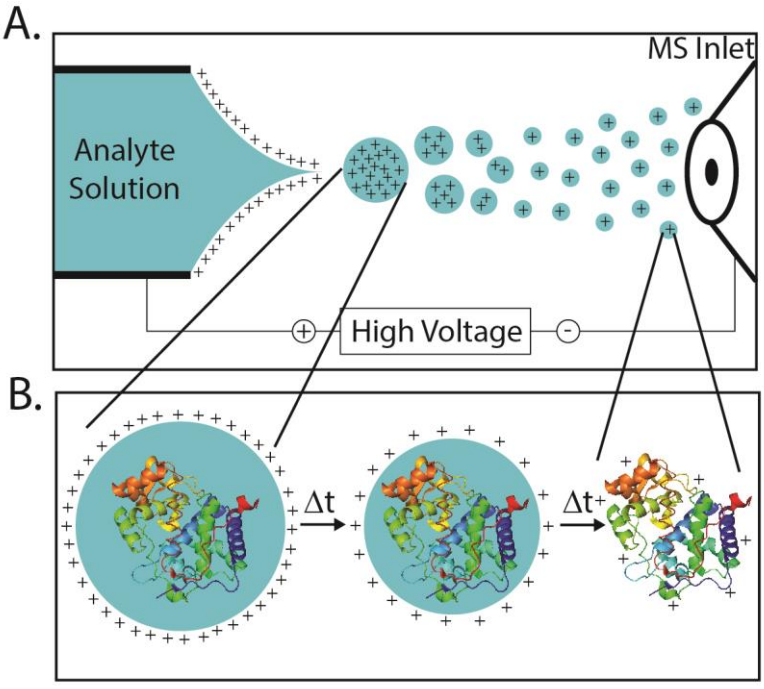


Figure 1 - 1 Electrospray Ionization. A.) Schematic of the electrospray source, including the ESI or nESI emitter on the left, and the path of analyte droplets to the MS inlet on the right. B.) depicts the evaporation of analyte-containing droplets as they travel towards and into the mass spectrometer described in the Charged Residue Mechanism of ESI.

Much research has been conducted into the mechanism of electrospray ionization,[89] and there are currently several

theories available to explain various electrospray phenomenon. Notably, the charge-residue model (CRM) explains most of the ionization behavior of proteins in nESI, accounting for charge state distributions and IM-derived collision cross sections (CCSs), as well as other phenomena such as charge reduction and amplification.[94-97] (Figure 1-1) It has been well established that evaporation of solvent molecules in ESI droplets produces droplet fission governed by the Rayleigh limit for charge density.[98,99] The CRM states that a droplet containing a macromolecule, after undergoing Rayleigh fission may continue evaporation to completion, leaving a "naked" macromolecular ion solvated only by residual, unevaporated charged particles, where the net charge generally approaches 90% of the Rayleigh limit. As a result of CRM-like behavior, the transition from solution to gas-phase is often slow and gentle enough to preserve noncovalent interactions in the gas phase.[100,101] Once in the gas-phase, noncovalent protein interactions can be maintained at least on the millisecond timescale if they are not activated by collisions with neutral gas or other ions.[102] Other models, such as the ion evaporation and the chain ejection models, account for the ionization of small molecules and highly unstructured proteins, respectively.[103-105]

1.3.2 Selection, Mass analysis, and Detection of Protein Ions

To date, the majority of structural MS research has utilized TOF mass analyzers. The TOF mass analyzer measures the mass to charge ratio (m/z) of an ion by measuring the flight time of the ion pulsed with a given amount of kinetic energy through a fixed distance to a detector, usually either a multichannel plate or a collision dynode.[106] Reflectron TOFs, the state of the art TOF technology, utilize a V shaped

flight path with a reflectron at the vertex for two reasons: the use of the reflector corrects for differences in the initial kinetic energy of ions, and the V shape effectively doubles the flight path for an ion; both factors contribute to significant increases in resolution.[107] Commercial TOF analyzers now routinely offer mass resolution values > 40,000, which allows for isotopic resolution of small proteins.[108] Compared with dedicated high resolution mass analyzers such as Fourier transform ion cyclotron resonance (FT-ICR) and Orbitraps, which routinely achieve resolutions greater than 130,000, TOF analyzers are extremely fast, able to complete scan in 50 us, at least 10 fold faster than higher resolution instruments.[109] This increase in scan time allows for greater signal accumulation and averaging, and higher duty cycle considering the continuous stream of ions generated using ESI methods.

Another mass analyzer commonly utilized in structural MS experiments is the quadrupole.[110] Although quadrupoles can function as stand-alone mass analyzers, they are generally employed in conjunction with TOF analysis (Q-TOF) in structural MS due to their ability to act as an m/z filter.[111] Quadrupoles are comprised of 4 conductive rods, of which opposing rods are electrically paired and direct (DC) and alternating current (AC) is applied. Specific combinations of DC and AC currents applied to each pair of rods act as high and low m/z filters for ions, as higher m/z ions are less sensitive to high radio frequency (RF) fields than low m/z ions. By setting the quadrupole to a specific set of DC and AC amplitudes, the quadrupole will act as a filter, usually with unit m/z resolution. In other cases, the quadrupole can act as a low mass filter, high mass filter, or allow all ions to pass through in an RF-only mode. Notably, for transmission and

selection of high m/z ions in such devices it is necessary to utilize low-frequency RF generators.[112]

1.3.3 Ion Activation and Tandem Mass Spectrometry

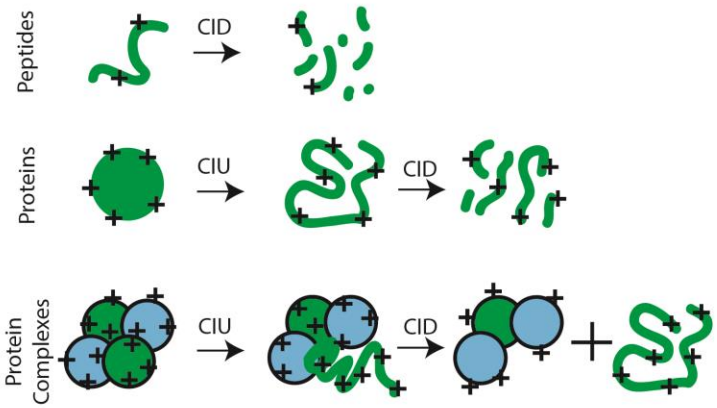


Figure 1 - 2 Collision-induced Unfolding and Dissociation of Charged Polypeptides. The fate of polypeptides at different size regimes under different amounts of collisional activation is depicted. Peptides easily fragment into smaller B and Y ions, proteins undergo unfolding to extended structures before fragmentation. In the case of multiprotein complexes, unfolding and dissociation of a subunit is the predominant mechanism, while further activation of unfolded subunits leads to fragmentation as is described above.

Despite the high accuracy and resolution of common mass analyzers, it is often necessary to perform tandem mass spectrometry experiments to confirm the identity of an unknown signal. This is especially true in the case of intact proteins analyzed by Native MS, where peak broadening can result from the combination of large isotopic clusters and heterogeneous water, buffer, or salt adducts.[113,114] Tandem mass spectrometry utilizes two mass analyzers and some mode of ion activation to dissociate an ion selected by the first analyzer into several components that can be analyzed by the second analyzer. In most structural MS applications, the configuration of choice is the Q-TOF with a collision-induced dissociation (CID) cell located between the two mass analyzers.[112] In CID, ions are accelerated from vacuum into a cell pressurized to around 10^{-2} mbar of neutral gas, resulting in ion-neutral collisions which increase the internal energy of the ion.[115] The

fate of peptides, proteins, and protein complexes in CID is depicted in Figure 1-2. At low collision energies, protein ions usually undergo unfolding to realize a series of intermediate structures prior to complete unfolding,[116] but may in some cases compact upon collisional heating.[117] If significant energy is imparted to an ion, covalent fragmentation or noncovalent dissociation may occur resulting in ejection of protein subunits, small molecule binders, or fragmented peptides.[115,116] Typical CID behavior for protein complexes involves unfolding and dissociation of a disproportionately charged monomeric subunit, leaving behind a charge-reduced protein complex which may undergo structural compaction.[118,119] This asymmetric charge partitioning phenomena has puzzled the field for many years, however recent molecular dynamics studies have provided mechanistic details for CID of protein complexes that are broadly in line with experimental evidence.[120]

CID of protein-small molecule complexes, on the other hand, proceeds by a more intuitive mechanism that involves ejection of singly-charged or neutral small molecules concomitant with local unfolding of the binding site.[121,122] In the case of peptide bond fragmentation at high CID energies, collision energy is distributed roughly evenly through the polypeptide chain, resulting in preferential fragmentation of the weakest peptide bonds giving rise to b and y ions with amine and carboxy termini resembling a new polypeptide.[49]

Other fragmentation techniques can also be utilized for interrogating proteins or protein complexes in the gas-phase. Surface-induced Dissociation (SID) is a thermal activation technique akin to CID, however it utilizes a single collision against a surface to impart energy on the ion rather than multiple collisions with gas molecules.[123] The

result of this change is that energy transfer occurs much faster, and thus can sometimes result in more symmetric charge partitioning between precursor and product ions, as well as a variety of other fragment ions not typically found in CID. Infrared multiphoton dissociation (IRMPD)[124] and blackbody infrared dissociation (BIRD)[125] are other thermal techniques which use IR radiation to directly heat the analyte ion, removing the need for collisions. Energy can also be imparted directly to ions through electron transfer using techniques like electron transfer dissociation (ETD)[126] and electron capture dissociation (ECD),[127] as well as some forms of ultraviolet photodissociation (UVPD).[128] Many of these technoloigies are capable of fragmenting peptide bonds without necessarily perturbing the global structure of the protein or protein complex ion undergoing activation. These methods, still under investigation for their utility in analysis of protein complexes, may provide unique capabilities in terms of identifying flexible regions, interfaces, and ligand binding sites in difficult to analyze protein systems.[129]

1.3.4 Ion Mobility Spectrometry

Ion mobility spectrometry (IMS) is an integral tool for gas-phase structural biology.[17] In nearly all modern implementations IMS is combined with mass spectrometry equipment to form a hybrid instrument, an ion mobility-mass spectrometer (IM-MS).[130] IMS separates ions based on their propensity to collide and interact with neutral gas molecules at a given temperature and pressure.[131] In most cases, a weak electric field is used to force ions through a pressurized drift cell, and the drift time over the fixed distance of the cell is measured and related back to the CCS of the ion.[132] In

this type of experiment, ion CCS is an orientationally averaged size parameter which describes the probability of ion interactions with the buffer gas: Ions with larger CCS will have more interactions with buffer gas, and will therefore have longer drift times compared to an ion of identical charge but lower CCS. Many IMS devices have been developed over the years which can routinely be operated with low enough field strengths to avoid activation and disruption of protein structure.[133] For brevity, the sections below focus on two commonly-used IMS devices in gas-phase structural biology research: Drift tubes and traveling wave separators.

Drift tube IMS (DTIMS) devices represent the modern incarnation of the earliest reported form of ion mobility separation .[132] For protein analysis, these devices can range from centimeters to meters in length and in general utilize a stacked-ring electrode geometry to establish a uniform potential gradient across the drift length. For continuous ion sources, an ion trap is necessary prior to the DTIMS device in order to accumulate ions into packets that can be released into the cell at a given start time. Most IMS measurements of protein ions using any device are performed in Helium, Nitrogen, or a mixture of the two gases using pressures that range from 1-4 mBar.[133] Early DTIMS devices operated without any RF focusing of ions across the drift tube, however some implementations have successfully utilized RF focusing to increase transmission efficiency in these devices without increasing the temperature of the ions substantially.[134-136] Because the physical principles surrounding of ion transport in gases are well understood,[131] the drift time of an ion in DTIMS under given conditions can be used to directly calculate an ions CCS. Importantly, IMS separations for proteins occur on the millisecond timescale, meaning that the ions exiting the drift cell can be easily

sampled by a TOF mass analyzer, which can operate at approximately 200 times the frequency of the IMS separation.[137]

Despite simplicity of the DTIMS device, the first IMS device to be commercialized and hence the most widely currently used, is the traveling-wave IMS (TWIMS) implemented in the Synapt and Vion IM-MS platforms from Waters Corporation.[130,138] Like the DTIMS, the TWIMS device utilizes a stacked-ring electrode geometry to guide ions through the mobility cell, however in this case a non-uniform, dynamic field is applied. To create this field, a DC voltage is applied to a combination of electrodes to create a waveform, which is then propagated through the device as the DC voltage is moved to adjacent electrodes at a given velocity. Since the mechanism of separation in this device depends on the ability of an ion to be propelled forward by a wave rather than be passed by the wave due to drag by the buffer gas, tuning of TWIMS wave amplitude and wave velocity allows for optimization of IMS separations for a variety of different systems. Although TWIMS is a versatile separation device, the use of a nonuniform field makes direct calculation of the CCS challenging, and therefore calibration of ion drift times based on drift tube datasets is necessary for extraction of CCS values.[136,139,140]

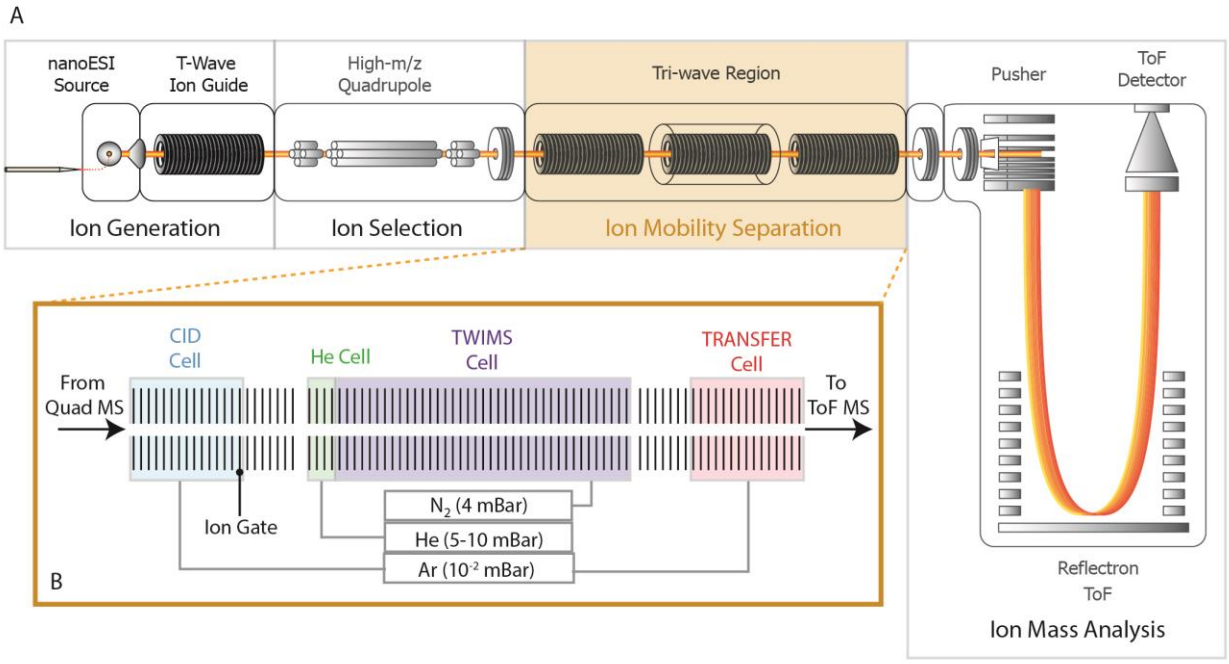


Figure 1 - 3 Schematic of the Synapt G-2 Ion Mobility-Mass Spectrometry Platform. A.) Ions are generated using nESI and can be mass-selected by the high m/z quadrupole for subsequent experiments. B.) In the Tri-wave region of the instrument, the first cell is operated as a trap where CIU or CID can be used to interrogate the ions. This trap also functions as a gate for the ion mobility separation. The TWIMS cell is operated at much higher pressure, and the transfer of ions to 4mBar nitrogen is facilitated by the helium cell. After IM separation is preformed, the pressure is rapidly dropped in the transfer cell before ions are released into the TOF mass analyzer for high resolution mass analysis.

Currently, variations of the Synapt HDMS (Waters Corp.) are by far the most commonly used instruments for studying the structure of proteins and protein complexes in the gas-phase. The Synapt platform features a nESI-Q-TWIMS-TOF configuration with two CID cells located before and after the TWIMS.[138,141] (Figure 1-3) Variants of this instrument include replacement of the TWIMS with drift tube IMS,[135] or replacement of one or more CID cells with SID,[142] ETD,[143] or UVPD.[144] Other IMS instruments have recently come to market, however these instruments have yet to find unique applications in protein structural studies.[145,146]

1.4 Ion Mobility - Mass Spectrometry Methodology in Structural Biology and Drug Discovery

Since the first proteins were observed in the gas-phase, the question of whether any elements of solution-phase structure could be maintained within the vacuum of a mass spectrometer has been investigated using a variety of techniques.[101] The first evidence for conservation of solution-phase structural elements was presented by Chowdhury in 1990,[147] when it was shown that the charge-state distribution of cytochrome C from various pH solutions correlated with known pH-dependent structural changes. Subsequent reports noted similar correlations between charge state distributions and structures perturbed by addition or organic solvent,[148] and reduction of disulfide bonds.[149] Since this early work, the field of protein structural studies using MS and IMS methods has matured substantially. Large surveys of CCS measurements by IMS have revealed strong agreement between experimental CCS measurements and CCS values calculated for solutions structures by highly accurate computational methods.[150] Further support for conservation of protein structure in the gas-phase has come from gas-phase HDX,[151] and a variety of nonergodic tandem MS experiments.[128,143,152,153]

1.4.1 IM-MS Methods for Multiprotein Complex Structure

The field of multiprotein complex studies by IM-MS is perhaps the most successful, as IM-MS offers unique advantages over other technologies. Early observations of multiprotein complexes in the gas-phase lead to a flurry of studies interrogating the stoichiometry of homomeric[154-156] and heteromeric complexes.[157-160]. In

1999, Rostom detected intact 800 kDa tetradecameric groEL,[154] and in the next 5 years methods for detection of other massive particles including ribosomes,[161] proteasomes,[162] and virus particles has been developed.[163] It wasn't until 2005, however, the first rigorous structural analysis of multiprotein complex structure in the gas-phase using IM-MS, showing that the ring structure of *trp* RNA binding attenuation protein could be maintained in the gas-phase.[117] This study was impactful not only for the observation of native-like topologies, but for the new take on coarse-grained molecular modeling that allowed the authors to visualize collapsed structures resulting from collision activation. By 2008, robust protocols for CCS analysis of multiprotein complexes had emerged,[140] and CCS analysis could be used to determine the topologies of homomeric complexes using relatively simple informatics.[164] More recent studies targeting large heteromeric complexes highlight the strengths and challenges of defining protein connectivity and topology by IM-MS.[165,166] Although IM-MS is capable

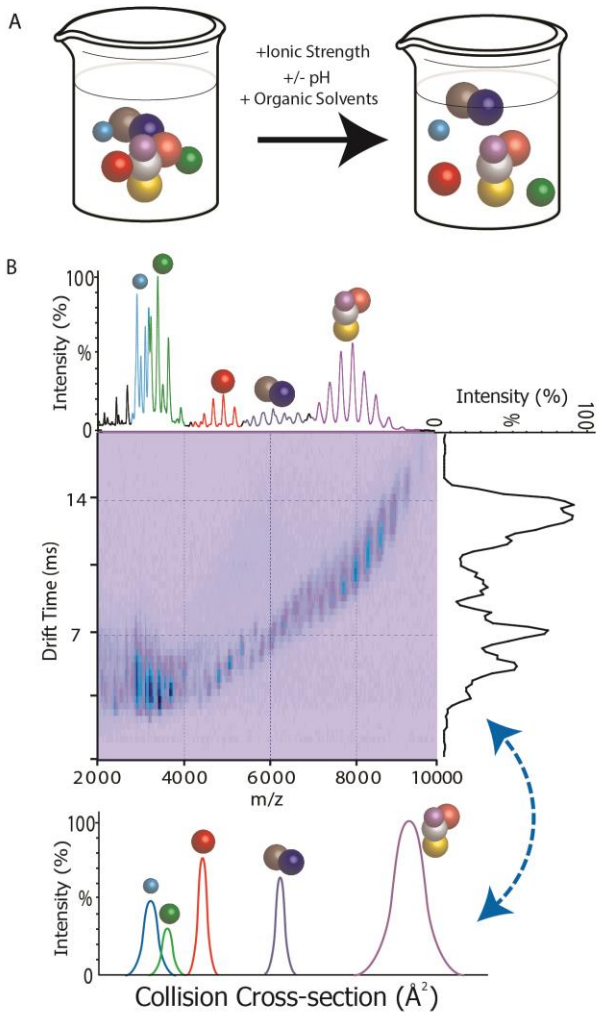


Figure 1 - 4 Information Content of IM-MS Experiments. A.) The information content of the IM-MS experiment can be greatly increased by the generation of subcomplexes in solution. B.) Simultaneous analysis of subcomplex mass and CCS allows for identification of complex connectivity and geometry. Moreover, the 3D plot of IM Drift time vs Mass to charge allows for facile deconvolution of the mass spectra.

of identifying the composition of subcomplexes resulting from intact protein species, tandem MS using CID is limited to generation of only minimal numbers of subcomplexes and is generally not informative enough for detailed connectivity assignment.[116] Thus, it is often necessary to use complementary tandem MS methods or solution-phase disruption with chaotropic agents to generate enough subcomplexes to build a useful model of the complex;[167] (Figure 1-4) New methods for generation of subcomplexes in the gas[123,168] and solution phase[169,170] are currently under development by several groups.

In addition to experimental challenges, informatics approaches for modeling heteromeric complexes from IM-MS data sets are currently underexplored. Clearly, for complexes with masses over 100 kDa, all-atom modeling is not an option due to the computing power necessary to dynamically restrain many thousands of atoms. Various coarse-graining approaches have been utilized with some success, however the utility and accuracy of highly coarse-grained models is still of some debate. One popular approach for coarse-graining large multiprotein complexes is the representation of each protein subunit as a sphere with CCS corresponding to an experimental or calculated measurement. For trimeric complexes, the CCS information necessary for restraining the topology in this case is well-defined, however for larger subunit numbers the amount of experimental CCS measurements for sub complexes needed to fully restrain a model increases linearly.[165] Hall explored this problem in 2012 and had success restraining tetrameric complexes using biophysical and limited CCS datasets, however the sample of protein complex space studied was small.[171]

Although the ability of IM-MS to accurately measure protein complex mass, composition, stoichiometry, and CCS have provided invaluable insight to the structural biology of several complexes, it is clear that IM-MS is not a stand-alone tool for those seeking atomic-level insights into multiprotein complexes. Instead, IM-MS is viewed as an emerging tool for integrative structural biology.[87] Integrative structural biology was popularized by Alber's comprehensive structure of the 456-subunit, 50 MDa nuclear pore complex in 2008,[172,173] despite the foundations for this platform being developed much earlier.[13,174] In essence, the integrative approach integrates data from multiple experimental types into a single model by way of a scoring function. The only limitation on data integration is spatial representation as a restraint on individual or groups of

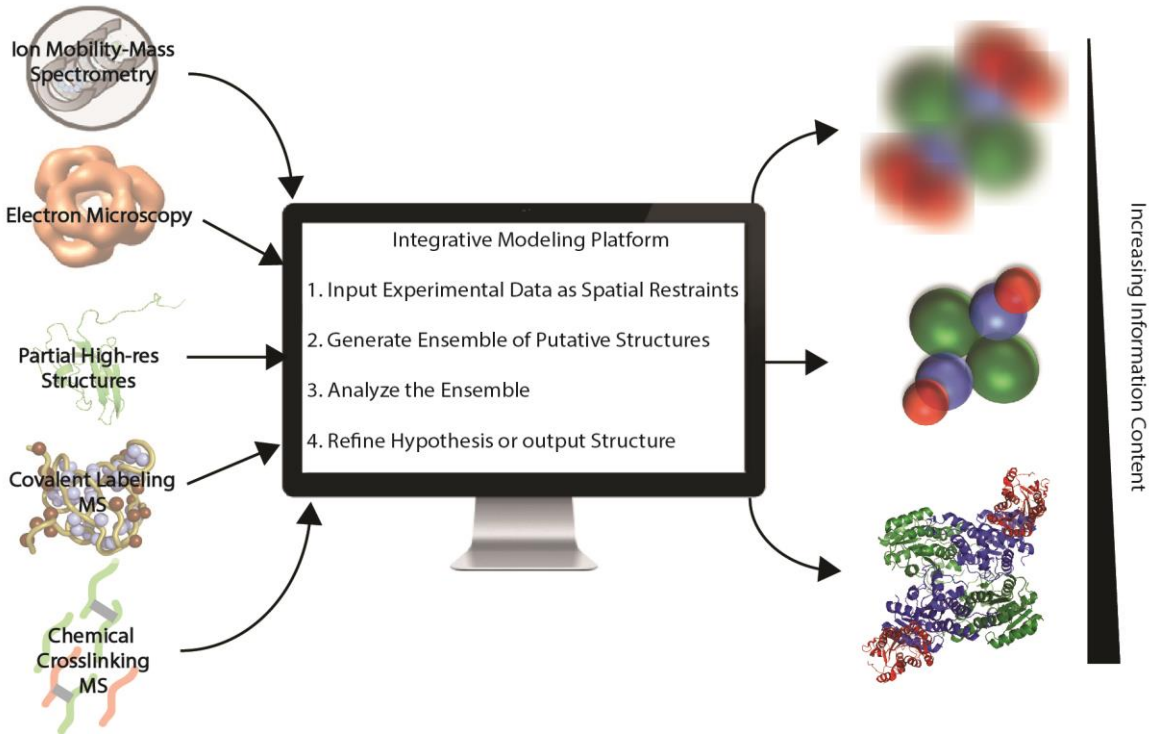


Figure 1 - 5 Integrative Structural Biology. Many challenging structural targets are refractory to high-resolution characterization by a single technique. Integrative structural biology seeks to incorporate information from multiple experimental datasets into a single model. These datasets are integrated by way of spatial restraints in a computational platform that analyzes structures consistent with all of the experimental data. The resolution of the output structures depends on the information input into the model, where new information can be added to resolve specific ambiguities.

particles within the model. This field has moved forward substantially in recent years thanks to integrative modeling software,[175,176] and the work of experimentalists to define what specific data types mean in the context of an integrative model. In Alber's model of the nuclear pore complex, datasets from ultracentrifugation, quantitative immunoblotting, affinity purification, overlay assays, electron microscopy, membrane fractionation and bioinformatics were integrated. These diverse datasets were used to restrain the shapes and sizes of each protein subunit, pairwise or subcomplex connectivities, membrane contacts, and overall shape of the complex. These restraintswere encoded into a complex scoring function that was optimized 200,000 times to generate an ensemble of 1000 structures satisfying the experimental data. The ensemble was then analyzed for commonalities and ambiguities that form the bases for new hypotheses and experiments.

Alber's integrative modeling work provides a workflow for structural biologists seeking to understand complex multiprotein systems: 1) gather experimental data 2) express data as spatial restraints in a scoring function 3) optimize the scoring function to sample all possible structures and 4) Analyze the ensemble of possible structures to form new hypotheses and repeat the process. Within this context, MS methodologies including IM-MS are emerging as powerful tools for assignment of size, shape, and connectivity information for various units within large complexes.[165,177,178] (Figure 1-5) Several reports have integrated IM-MS datasets as restraints on the connectivity and size of protein complexes alongside chemical crosslinking,[179-182] oxidative labeling,[183] and electron microscopy[184,185] datasets. Although all of these studies are foundational to our ability to use IM-MS data in conjunction with other data types, the current literature

falls short when assessing the error introduced into integrative models from the coarse-graining procedures and ambiguity in the structures refined by experimental CCS.

1.4.2 IM-MS Methods for Drug Discovery and Development

MS has proven a powerful tool for screening and measuring Protein:Ligand (P:L) interactions, including the capability to measure binding affinities, in addition to elucidating many of the structural details of such complexes. Coupled to IM, MS methods gain an enhanced structural dimension that can be leveraged to detect shifts in protein conformation and changes in protein complex stability upon ligand binding. In the early 1990s, the first reports of noncovalent complexes emerged.[186-188] Notably, the first experiments observing FK506 binding protein bound to FK506 and rapamycin also implemented control experiments with decoy ligands or denatured protein to rule out the possibility of nonspecific adduction, setting the stage for rigorous examination of protein-ligand binding constants. Since these early reports, the field of studying protein-ligand interactions with mass spectrometry has matured largely through the efforts of John Klassen and Renato Zenobi, who have developed robust methods for k_D measurement using ESI[189-194] and MALDI MS,[195,196] and provided methods for high throughput identification of small molecule binders from compound libraries from difficult classes including carbohydrates and glycolipids.[197-199] Additionally, tandem MS technologies have evolved for rapid identification of small molecule binders,[200,201] measurement of gas-phase dissociation constants,[121,202,203] and even localization of binding sites.[127,153,204]

A suite of IMS-based tools also exist for examining protein-ligand binding properties. Notably, the ion mobility shift assay has been successfully used to detect

conformational shifts as a result of ligand binding, which may be correlated with biologically relevant solution-phase processes.[205-207]

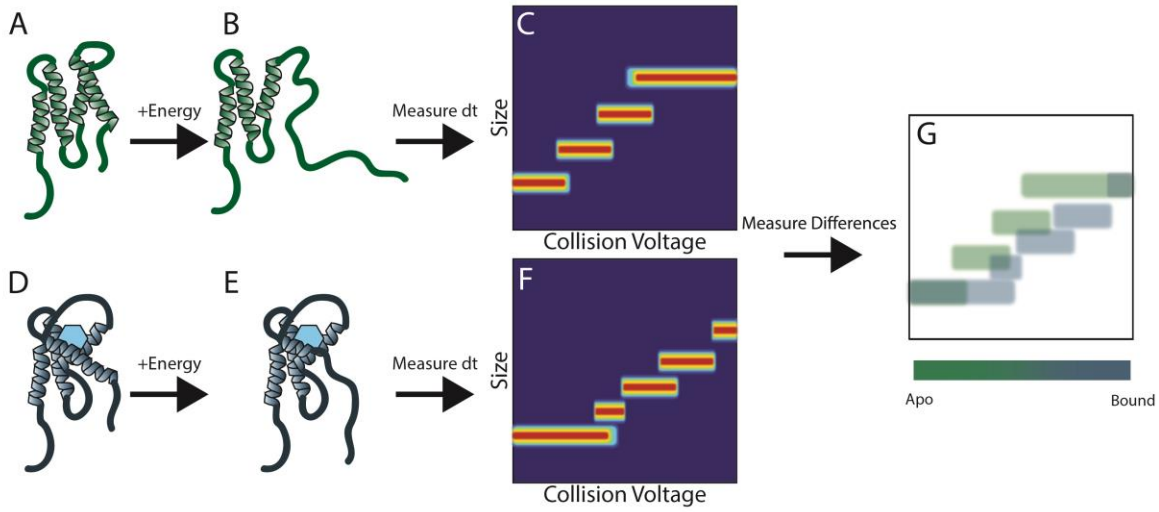


Figure 1 - 6. Collision-induced Unfolding Analysis of Protein-Ligand Complexes. A) apo-protein ions are selected and activated to produce unfolded ions (B). Measurement of ion CCS (size) as a function of Collision voltage results in a unique unfolding fingerprint for the ion (C). Unfolding fingerprints can be compared with those from proteins bound to ligands (D-F) resulting in a difference plot that represents changes in stability induced by the ligand (G).

In many cases, conformational shifts resulting from ligand binding may be too small to resolve via CCS measurement alone, even with high resolution IMS instruments. For these cases, collision induced unfolding (CIU) technologies are being developed by multiple laboratories, which are highly sensitive to subtle conformational shifts unobservable by traditional IMS.[208] (Figure 1-6) Despite success in extracting useful empirical information from these experiments, the physics involved in gas-phase protein unfolding, ligand stabilization, and ejection during these processes are largely unknown. In order for the field to move toward high-throughput drug discovery applications of CIU, our understanding of the fundamental processes in gas-phase activation needs to be improved.

1.5 Summary of the Dissertation

This dissertation represents my work in the areas of gas-phase structural biology and drug discovery. The 2nd and 3rd chapters focus on collision induced unfolding technology for protein structure and drug discovery applications. In chapter 2, I discuss my contributions to an informatics platform for analysis of CIU of proteins. This work has been previously published as **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), pp 11516–11522** In chapter 3, I presentmechanistic studies of CIU using a series of albumins as a model system. In this chapter, I also discuss future applications of CIU technology in drug discovery. This work has been previously published as **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), pp 534–540.**

The 4th and 5th and 6th chapters focus on my work developing IM-MS and informatics methods for building structures of multiprotein complexes. In Chapter 4, I present data relating to the information content of IM-MS datasets in modeling multiprotein complexes, and outlines best practices for reporting ambiguities in modeling results. This paper is under review as **Coming to Grips with Ambiguity: Ion Mobility-Mass Spectrometry for Protein Quaternary Structure Assignment.** In Chapter 5, I use IM-MS and the informatics tools outlined in chapter 4 to characterize the structure of the urease activation complex, a 610 kDa octadecamer. This manuscript is awaiting submission. In Chapter 6, I highlight my contributions to several studies focused on structural determination of protein complexes. Stoichiometry assessment of

engineered multiprotein complexes in collaboration with the Neil Marsh group, including results from **Flexible, symmetry-directed approach to assembling protein cages**[209] and **Evaluation of *de novo*-designed coiled coils as off-the-shelf components for protein assembly**[210]. I also present further application of IM-MS based modeling to the ApoE Tetramer, for which I present the first structural model of this species.

To conclude, I provide a comprehensive list of my published work for the interested reader, which includes as its first entry a recent review article.

1.5 Publications

1. Eschweiler J.D.; Rabuck-Gibbons, J. N.; Kerr, R.; Ruotolo, B.T.: Sizing Up Protein–Ligand Complexes: The Rise of Structural Mass Spectrometry Approaches in the Pharmaceutical Sciences, Annu. Rev. Anal. Chem. 10 (2017)

2. Soper-Hopper, M.T.; Eschweiler, J.D.; Ruotolo, B.T.: Ion Mobility-Mass Spectrometry Reveals a Dipeptide That Acts as a Molecular Chaperone for Amyloid β. ACS Chem. Biol. 12 (4), 1113-1120 (2017)

3. Cristie-David, A.S., Sciore, A., Badieyan, S., Escheweiler, J.D., Koldewey, P., Bardwell, J.C.A., Ruotolo, B.T., Marsh, E.N.G.: Evaluation of de novo-designed coiled coils as off-the-shelf components for protein assembly. Mol. Syst. Design & Engin. (2017).

4. 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.,, 139 (1), pp 534–540. (2016)

5. Sciore, A., Su, M., Koldewey, P., Eschweiler, J.D., Diffley, K.A., Linhares, B.M., Ruotolo, B.T., Bardwell, J.C.A., Skiniotis, G., Marsh, E.N.G.: Flexible, symmetry-directed approach to assembling protein cages. Proc. Natl. Acad. Sci. U. S. A. 113(31), 8681-8686 (2016)

6. Won, S.J.; Eschweiler, J.D.; Majmudar, J.D.; Chong, F.S.; Hwang, S.Y.; Ruotolo, B.T.; Martin B.R. Affinity-Based Selectivity Profiling of an In-Class Selective Competitive Inhibitor of Acyl Protein Thioesterase 2. ACS Med. Chem. Lett., 8 (2), pp 215–220 (2016)

7. Bornschein, R.E.; Niu, S.; Eschweiler, J.D; Ruotolo, B.T.: Ion Mobility-Mass Spectrometry Reveals Highly-Compact Intermediates in the Collision Induced Dissociation of Charge-Reduced Protein Complexes. J. Am. Soc. Mass Spectrom. 27: 41 (2016)

8. 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., 87 (22), pp 11516–11522 (2015)

1.6 References

(1) Sabeti, P. C.; Varilly, P.; Fry, B.; Lohmueller, J.; Hostetter, E.; Cotsapas, C.; Xie, X.; Byrne, E. H.; McCarroll, S. A.; Gaudet, R.; Schaffner, S. F.; Lander, E. S. *Nature* **2007**, *449*, 913.

(2) Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; Wu, C.; Sweet, S. M. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L. *Nature* **2011**, *480*, 254.

(3) Prabakaran, S.; Lippens, G.; Steen, H.; Gunawardena, J. *Wiley Interdiscip. Rev.: Syst. Biol. Med.* **2012**, *4*, 565.

(4) Perkins, J. R.; Diboun, I.; Dessailly, B. H.; Lees, J. G.; Orengo, C. *Structure* **2010**, *18*, 1233.

(5) Tompa, P.; Davey, Norman E.; Gibson, Toby J.; Babu, M. M. *Mol. Cell* **2014**, *55*, 161.

(6) Campbell, I. D. *Nat Rev Mol Cell Biol* **2002**, *3*, 377.

(7) Jones, A. W. *Drug Test. Anal.* **2011**, *3*, 337.

(8) Conroy, P. J.; Law, R. H. P.; Caradoc-Davies, T. T.; Whisstock, J. C. *Methods* **2017**, *116*, 12.

(9) Congreve, M.; Murray, C. W.; Blundell, T. L. *Drug Discovery Today* **2005**, *10*, 895.

(10) Surade, S.; Blundell, Tom L. *Chem. & Biol.* **2012**, *19*, 42.

(11) Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C. *Nature* **1958**, *181*, 662.

(12) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235.

(13) Sali, A.; Glaeser, R.; Earnest, T.; Baumeister, W. *Nature* **2003**, *422*, 216.

(14) Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M.; Hattori, M.; Sakaki, Y. *Proc. Nat. Academ. Sci. U.S.A* **2001**, *98*, 4569.

(15) Luger, K.; Phillips, S. E. V. *Curr. Opin. Struct. Biol.* **2010**, *20*, 70.

(16) McFedries, A.; Schwaid, A.; Saghatelian, A. *Chem. and Biol.*, *20*, 667.

(17) Zhong, Y.; Hyung, S.-J.; Ruotolo, B. T. Exp. Rev. Proteomics **2012**, *9*, 47.

(18) Hopper, J. T. S.; Robinson, C. V. *Angew. Chem. Int. Ed.* **2014**, *53*, 14002.

(19) Maurer, M. M.; Donohoe, G. C.; Valentine, S. J. *Analyst* **2015**, *140*, 6782.

(20) Alberts B, J. A., Lewis J, et al. In *Molecular Biology of the Cell*; 4 ed.; Garland Science.: New York, 2002.

(21) Kinch, L. N.; Li, W.; Monastyrskyy, B.; Kryshtafovych, A.; Grishin, N. V. *Proteins: Struct., Funct., Bioinf.* **2016**, *84*, 51.
(22) Modi, V.; Xu, Q.; Adhikari, S.; Dunbrack, R. L. *Proteins: Struct., Funct., Bioinf.***2016**, *84*, 200.
(23) Fiser, A. Template-Based Protein Structure Modeling. In *Computational Biology*; Fenyö, D., Ed.; Humana Press: Totowa, NJ, 2010, p 73.
(24) Hull, A. W. *J. Am. Chem. Soc.* **1919**, *41*, 1168.
(25) Rabi, I. I.; Zacharias, J. R.; Millman, S.; Kusch, P. *Phys. Rev.* **1938**, *53*, 318.
(26) Majorek, K. A.; Porebski, P. J.; Dayal, A.; Zimmerman, M. D.; Jablonska, K.; Stewart, A. J.; Chruszcz, M.; Minor, W. *Mol. Immunol.* **2012**, *52*, 174.
(27) Wimberly, B. T.; Brodersen, D. E.; Clemons, W. M.; Morgan-Warren, R. J.; Carter, A. P.; Vonrhein, C.; Hartsch, T.; Ramakrishnan, V. *Nature* **2000**, *407*, 327.
(28) Robinson, R. C.; Turbedsky, K.; Kaiser, D. A.; Marchand, J.-B.; Higgs, H. N.; Choe, S.; Pollard, T. D. *Science* **2001**, *294*, 1679.
(29) Pearson, M. A.; Michel, L. O.; Hausinger, R. P.; Karplus, P. A. *Biochemistry* **1997**, *36*, 8164.
(30) Bhowmick, A.; Brookes, D. H.; Yost, S. R.; Dyson, H. J.; Forman-Kay, J. D.; Gunter, D.; Head-Gordon, M.; Hura, G. L.; Pande, V. S.; Wemmer, D. E.; Wright, P. E.; Head-Gordon, T. *J. Am. Chem. Soc.* **2016**, *138*, 9730.
(31) von Heijne, G. *Nat Rev Mol Cell Biol* **2006**, *7*, 909.
(32) DePristo, M. A.; de Bakker, P. I. W.; Blundell, T. L. *Structure* **2004**, *12*, 831.
(33) Lisi, G. P.; Loria, J. P. *Prog. Nucl. Magn. Reson. Spectrosc.* **2016**, *92-93*, 1.
(34) Lisi, G. P.; Loria, J. P. *Chem. Rev.* **2016**, *116*, 6323.
(35) Hansen, M. R.; Graf, R.; Spiess, H. W. *Accounts Chem. Res.* **2013**, *46*, 1996.
(36) Merloni, A.; Dobrovolska, O.; Zambelli, B.; Agostini, F.; Bazzani, M.; Musiani, F.; Ciurli, S. *Biochim. Biophys. Acta, Proteins Proteomics* **2014**, *1844*, 1662.
(37) Chiliveri, S. C.; Deshmukh, M. V. *J. Biosci.* **2016**, *41*, 787.
(38) Xu, G. H.; Li, C. G.; Liu, M. L. *Prog. Chem.* **2017**, *29*, 75.
(39) Quinn, C. M.; Polenova, T. *Q. Rev. Biophys.* **2017**, *50*, 1.
(40) Glaeser, R. M.; Hall, R. J. *Biophys.J.***2011**, *100*.
(41) Skiniotis, G.; Southworth, D. R. *Microscopy* **2016**, *65*, 9.
(42) Bartesaghi, A.; Merk, A.; Banerjee, S.; Matthies, D.; Wu, X.; Milne, J. L. S.; Subramaniam, S. *Science* **2015**, *348*, 1147.
(43) Spahn, C. M. T.; Penczek, P. A. *Curr.Opin. Chem. Biol.* **2009**, *19*, 623.
(44) Mackay, J. P.; Landsberg, M. J.; Whitten, A. E.; Bond, C. S. *Trends Biochem Sci*, *42*, 155.
(45) Marsh, J. A.; Teichmann, S. A. In *Annu. Rev. Biochemistry* 2015; Vol. 84, p 551.
(46) Hassell, A. M.; An, G.; Bledsoe, R. K.; Bynum, J. M.; Carter, H. L.; Deng, S.-J. J.; Gampe, R. T.; Grisard, T. E.; Madauss, K. P.; Nolte, R. T.; Rocque, W. J.; Wang, L.; Weaver, K. L.; Williams, S. P.; Wisely, G. B.; Xu, R.; Shewchuk, L. M. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2007**, *63*, 72.
(47) Tautermann, C. S. *Bioorg. Med. Chem. Lett.* **2014**, *24*, 4073.
(48) Blundell, T. L.; Patel, S. *Curr. Opin. Pharmacol.* **2004**, *4*, 490.
(49) Hunt, D. F.; Yates, J. R.; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Nat. Academ. Sci. U.S.A* **1986**, *83*, 6233.
(50) Fenn, J.; Mann, M.; Meng, C.; Wong, S.; Whitehouse, C. *Science* **1989**, *246*, 64.
(51) Karas, M.; Hillenkamp, F. *Anal. Chem.* **1988**, *60*, 2299.
(52) Eng, J. K.; McCormack, A. L.; Yates, J. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976.
(53) Brahms, S.; Brahms, J. *J Mol Biol***1980**, *138*, 149.
(54) Kelly, S. M.; Jess, T. J.; Price, N. C. *Biochim. Biophys. Act. Proteins and Proteomics* **2005**, *1751*, 119.
(55) Dmitri, I. S.; Michel, H. J. K. *Rep. Progress in Phys.* **2003**, *66*, 1735.
(56) Mertens, H. D. T.; Svergun, D. I. *J Struct Biol* **2010**, *172*, 128.
(57) Schneidman-Duhovny, D.; Hammel, M.; Tainer, J. A.; Sali, A. *Nucleic Acids Res.* **2016**.
(58) Förster, T. *Annalen der Physik* **1948**, *437*, 55.

(59) Clegg, R. M. In *Laboratory Techniques in Biochemistry and Molecular Biology*; Elsevier: 2009; Vol. Volume 33, p 1.
(60) Jin Lee, Y. *Mol. BioSystems* **2008**, *4*, 816.
(61) Jensen, O. N.; Barofsky, D. F.; Young, M. C.; Von Hippel, P. H.; Swenson, S.; Seifried, S. E. *Rapid Comm. Mass Spectrom* **1993**, *7*, 496.
(62) Sheshberadaran, H.; Payne, L. G. *Proc. Nat. Academ. Sci. U. S. A.* **1988**, *85*, 1.
(63) Yan, Y. T.; Chen, G. D.; Wei, H.; Huang, R. Y. C.; Mo, J. J.; Rempel, D. L.; Tymiak, A. A.; Gross, M. L. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 2084.
(64) Zhang, Z.; Smith, D. L. *Protein Science* **1993**, *2*, 522.
(65) Yang, L.; Broderick, D.; Jiang, Y.; Hsu, V.; Maier, C. S. *Biochim. Biophys. Act. Proteins and Proteomics* **2014**, *1844*, 1684.
(66) Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T. A.; Judson, R. S.; Knight, J. R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; Qureshi-Emili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.; Johnston, M.; Fields, S.; Rothberg, J. M. *Nature* **2000**, *403*, 623.
(67) Aloy, P.; Russell, R. B. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 5896.
(68) Wen, J.; Arakawa, T.; Philo, J. S. *Anal. Biochem.* **1996**, *240*, 155.
(69) Quiroz-Valenzuela, S.; Sukuru, S. C. K.; Hausinger, R. P.; Kuhn, L. A.; Heller, W. T. *Arch. Biochem. Biophys.* **2008**, *480*, 51.
(70) Konermann, L.; Tong, X.; Pan, Y. *J. Mass Spectrom.* **2008**, *43*, 1021.
(71) Pan, L. Y.; Salas-Solano, O.; Valliere-Douglass, J. F. *Anal. Chem.* **2014**, *86*, 2657.
(72) Politis, A.; Borysik, A. J. *Proteomics* **2015**, *15*, 2792.
(73) Marcoux, J.; Cianferani, S. *Methods* **2015**, *89*, 4.
(74) Yan, Y.; Chen, G.; Wei, H.; Huang, R. Y.-C.; Mo, J.; Rempel, D. L.; Tymiak, A. A.; Gross, M. L. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 2084.
(75) Politis, A.; Schmidt, C.; Tjioe, E.; Sandercock, A. M.; Lasker, K.; Gordiyenko, Y.; Russel, D.; Sali, A.; Robinson, C. V. *Chem. Biol.* **2014**.
(76) Shallan, M.; Radau, B.; Salnikow, J.; Vater, J. *Biochim. Biophys act.* **1991**, *1057*, 64.
(77) Bennett, K. L.; Kussmann, M.; Bjork, P.; Godzwon, M.; Mikkelsen, M.; Sorensen, P.; Roepstorff, P. *Protein Science* **2000**, *9*, 1503.
(78) Back, J. W.; Sanz, M. A.; De Jong, L.; De Koning, L. J.; Nijtmans, L. G. J.; De Koster, C. G.; Grivell, L. A.; Van der Spek, H.; Muijsers, A. O. *Protein Science* **2002**, *11*, 2471.
(79) Du, X.; Li, Y.; Xia, Y.-L.; Ai, S.-M.; Liang, J.; Sang, P.; Ji, X.-L.; Liu, S.-Q. *Int. J. of Mol. Sci.s* **2016**, *17*, 144.
(80) Rossi, A. M.; Taylor, C. W. *Nat. Protocols* **2011**, *6*, 365.
(81) Szabo, A.; Stolz, L.; Granzow, R. *Curr. Opin. Struct. Biol.* **1995**, *5*, 699.
(82) Saggio, I.; Gloaguen, I.; Poiana, G.; Laufer, R. *EMBO J.* **1995**, *14*, 3045.
(83) Kim, P.; Zhao, J.; Lu, P.; Zhao, Z. *Nucleic Acids Res.h* **2017**, *45*, D256.
(84) DeArmond, P. D.; West, G. M.; Anbalagan, V.; Campa, M. J.; Patz, E. F.; Fitzgerald, M. C. *J. biomol. screen.* **2010**, *15*, 1051.
(85) Strickland, E. C.; Geer, M. A.; Tran, D. T.; Adhikari, J.; West, G. M.; DeArmond, P. D.; Xu, Y.; Fitzgerald, M. C. *Nat.protocols* **2013**, *8*, 148.
(86) Mondal, T.; Wang, H.; DeKoster, G. T.; Baban, B.; Gross, M. L.; Frieden, C. *Biochemistry* **2016**, *55*, 2613.
(87) Hyung, S. J.; Ruotolo, B. T. *Proteomics* **2012**, *12*, 1547.
(88) Hyung, S.-J.; Robinson, C. V.; Ruotolo, B. T. *Chem. Biol.* **2009**, *16*, 382.
(89) Konermann, L.; Ahadi, E.; Rodriguez, A. D.; Vahidi, S. *Anal. Chem.* **2013**, *85*, 2.
(90) Kaltashov, I. A.; Abzalimov, R. R. *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 1239.
(91) Kaltashov, I. A.; Mohimen, A. *Anal. Chem.* **2005**, *77*, 5370.
(92) Wilm, M.; Mann, M. *Anal. Chem.* **1996**, *68*, 1.
(93) Guevremont, R.; Siu, K. W. M.; Le Blanc, J. C. Y.; Berman, S. S. *J. Am. Soc. Mass Spectrom.* **1992**, *3*, 216.
(94) Dole, M.; Mack, L. L.; Hines, R. L.; Mobley, R. C.; Ferguson, L. D.; Alice, M. B. *J. Chem. Phys.* **1968**, *49*, 2240.
(95) Kebarle, P.; Verkerk, U. H. *Mass Spectrom .Rev* **2009**, *28*, 898.
(96) Iavarone, A. T.; Williams, E. R. *J. Am. Chem. Soc.* **2003**, *125*, 2319.

(97) Fernandez de la Mora, J. *Anal.Chim. Acta* **2000**, *406*, 93.
(98) Rayleigh, L. *Philosophical Magazine Series 5* **1882**, *14*, 184.
(99) Dawson, G. A. *J. Geophys.Res.h* **1970**, *75*, 701.
(100) Silveira, J. A.; Fort, K. L.; Kim, D.; Servage, K. A.; Pierson, N. A.; Clemmer, D. E.; Russell, D. H. *J. Am. Chem. Soc.* **2013**, *135*, 19147.
(101) Hoaglund-Hyzer, C. S.; Counterman, A. E.; Clemmer, D. E. *Chem. Rev.* **1999**, *99*, 3037.
(102) Breuker, K.; McLafferty, F. W. *Proc Natl Acad Sci U S A* **2008**, *105*, 18145.
(103) Wilm, M. *Mol. Cell. Proteomics* **2011**, *10*.
(104) Iribarne, J. V.; Thomson, B. A. *J. Chem Phys* **1976**, *64*, 2287.
(105) Konermann, L.; Rodriguez, A. D.; Liu, J. *Anal. Chem.* **2012**, *84*, 6798.
(106) Chernushevich, I. V.; Loboda, A. V.; Thomson, B. A. *J. Mass Spectrom.* **2001**, *36*, 849.
(107) B.A. Mamyrin, V. I. K., D.V. Shmikk, V.A. Zagulin *J. Exp. Theor. Phys.* **1973**, *37*, 82.
(108) Wu, X.; Li, X.; Miller, C.; Waddell, K.; Tang, N. *J. of Biomol. Techniques : JBT* **2010**, *21*, S48.
(109) Domon, B.; Aebersold, R. *Science* **2006**, *312*, 212.
(110) W. Paul, H. Z. S. *Z. Naturforsch* **1953**, *8a*, 448.
(111) Howard R. Morris; Thanai Paxton; Anne Dell; Jean Langhorne; Matthias Berg; Robert S. Bordoli; John Hoyes; Bateman, R. H. *Rapid Commun Mass Spectrom* **1996**, *10*, 889.
(112) Sobott, F.; Hernández, H.; McCammon, M. G.; Tito, M. A.; Robinson, C. V. *Anal. Chem.* **2002**, *74*, 1402.
(113) McKay, A. R.; Ruotolo, B. T.; Ilag, L. L.; Robinson, C. V. *J. Am. Chem. Soc.* **2006**, *128*, 11433.
(114) Loo, J. A.; Udseth, H. R.; Smith, R. D. *Anal. Biochem.* **1989**, *179*, 404.
(115) Mitchell Wells, J.; McLuckey, S. A. In *Methods in Enzymology*; Academic Press: 2005; Vol. Volume 402, p 148.
(116) Benesch, J. L. P. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 341.
(117) Ruotolo, B. T.; Giles, K.; Campuzano, I.; Sandercock, A. M.; Bateman, R. H.; Robinson, C. V. *Science* **2005**, *310*, 1658.
(118) Bornschein, R. E.; Ruotolo, B. T. *Analyst* **2015**, *140*, 7020.
(119) Bornschein, R. E.; Niu, S.; Eschweiler, J.; Ruotolo, B. T. *J. Am. Soc. Mass Spectrom.* **2016**, *27*, 41.
(120) Popa, V.; Trecroce, D. A.; McAllister, R. G.; Konermann, L. *The J. Phys. Chem. B* **2016**, *120*, 5114.
(121) Hunter, C. L.; Mauk, A. G.; Douglas, D. J. *Biochem.***1997**, *36*, 1018.
(122) Mayer, P. M.; Martineau, E. *Phys. Chem. Chem. Phys* **2011**, *13*, 5178.
(123) Zhou, M. W.; Wysocki, V. H. *Accounts Chem. Res.* **2014**, *47*, 1010.
(124) Zhang, X.; Li, H.; Moore, B.; Wongkongkathep, P.; Ogorzalek Loo, R. R.; Loo, J. A.; Julian, R. R. *Rapid Comm. Mass Spectrom. : RCM* **2014**, *28*, 2729.
(125) Deng, L.; Kitova, E. N.; Klassen, J. S. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 988.
(126) Syka, J. E. P.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 9528.
(127) Xie, Y.; Zhang, J.; Yin, S.; Loo, J. A. *J. Am. Chem. Soc.* **2006**, *128*, 14432.
(128) O'Brien, J. P.; Li, W.; Zhang, Y.; Brodbelt, J. S. *J. Am. Chem. Soc.* **2014**, *136*, 12920.
(129) Brodbelt, J. S. *Anal. Chem.* **2016**, *88*, 30.
(130) Giles, K.; Pringle, S. D.; Worthington, K. R.; Little, D.; Wildgoose, J. L.; Bateman, R. H. *Rapid Commun Mass Spectrom* **2004**, *18*, 2401.
(131) Mason, E. A.; McDaniel, E. W. In *Transport Properties of Ions in Gases*; Wiley-VCH Verlag GmbH & Co. KGaA: 1988.
(132) Wyttenbach, T.; Kemper, P. R.; Bowers, M. T. *Int. J. Mass Spectrom.* **2001**, *212*, 13.
(133) May, J. C.; McLean, J. A. *Anal. Chem.* **2015**, *87*, 1422.
(134) Allen, S. J.; Giles, K.; Gilbert, T.; Bush, M. F. *Analyst* **2016**, *141*, 884.
(135) Allen, S. J.; Bush, M. F. *J. Am. Soc. Mass Spectrom.* **2016**, *27*, 2054.
(136) Bush, M. F.; Hall, Z.; Giles, K.; Hoyes, J.; Robinson, C. V.; Ruotolo, B. T. *Anal. Chem.* **2010**, *82*, 9557.
(137) Hoaglund, C. S.; Valentine, S. J.; Sporleder, C. R.; Reilly, J. P.; Clemmer, D. E. *Anal. Chem.* **1998**, *70*, 2236.

(138) Giles, K.; Williams, J. P.; Campuzano, I. *Rapid Commun Mass Spectrom* **2011**, *25*, 1559.
(139) Salbo, R.; Bush, M. F.; Naver, H.; Campuzano, I.; Robinson, C. V.; Pettersson, I.; Jorgensen, T. J.; Haselmann, K. F. *Rapid Commun Mass Spectrom* **2012**, *26*, 1181.
(140) Ruotolo, B. T.; Benesch, J. L.; Sandercock, A. M.; Hyung, S.-J.; Robinson, C. V. *Nature protoc.***2008**, *3*, 1139.
(141) Pringle, S. D.; Giles, K.; Wildgoose, J. L.; Williams, J. P.; Slade, S. E.; Thalassinos, K.; Bateman, R. H.; Bowers, M. T.; Scrivens, J. H. *Int. J. Mass Spectrom.* **2007**, *261*, 1.
(142) Zhou, M.; Dagan, S.; Wysocki, V. H. *Angew. Chem. Int. Ed.* **2012**, *51*, 4336.
(143) Lermyte, F.; Konijnenberg, A.; Williams, J. P.; Brown, J. M.; Valkenborg, D.; Sobott, F. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 343.
(144) Theisen, A.; Yan, B.; Brown, J. M.; Morris, M.; Bellina, B.; Barran, P. E. *Anal. Chem.* **2016**, *88*, 9964.
(145) Michelmann, K.; Silveira, J. A.; Ridgeway, M. E.; Park, M. A. *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 14.
(146) Kurulugama, R. T.; Darland, E.; Kuhlmann, F.; Stafford, G.; Fjeldsted, J. *Analyst* **2015**, *140*, 6834.
(147) Chowdhury, S. K.; Katta, V.; Chait, B. T. *J. Am. Chem. Soc.* **1990**, *112*, 9012.
(148) Loo, J. A.; Loo, R. R. O.; Udseth, H. R.; Edmonds, C. G.; Smith, R. D. *Rapid Comm. Mass Spectrom.* **1991**, *5*, 101.
(149) Loo, J. A.; Edmonds, C. G.; Udseth, H. R.; Smith, R. D. *Anal. Chem.* **1990**, *62*, 693.
(150) Benesch, J. L. P.; Ruotolo, B. T. *Curr. Opin. Struct. Biol.* **2011**, *21*, 641.
(151) Robinson, E. W.; Williams, E. R. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 1427.
(152) Lermyte, F.; Sobott, F. *Proteomics* **2015**, *15*, 2813.
(153) Cammarata, M. B.; Thyer, R.; Rosenberg, J.; Ellington, A.; Brodbelt, J. S. *J. Am. Chem. Soc.* **2015**, *137*, 9128.
(154) Rostom, A. A.; Robinson, C. V. *J. Am. Chem. Soc.* **1999**, *121*, 4718.
(155) Tito, M. A.; Miller, J.; Griffin, K. F.; Williamson, E. D.; Titball, R. W.; Robinson, C. V. *Protein Sci.* **2001**, *10*, 2408.
(156) Van Berkel, W. J. H.; Van Den Heuvel, R. H. H.; Versluis, C.; Heck, A. J. R. *Protein Sci.* **2000**, *9*, 435.
(157) Tito, M. A.; Miller, J.; Walker, N.; Griffin, K. F.; Williamson, E. D.; Despeyroux-Hill, D.; Titball, R. W.; Robinson, C. V. *Biophys. J.* **2001**, *81*, 3503.
(158) Pinkse, M. W. H.; Maier, C. S.; Kim, J.-I.; Oh, B.-H.; Heck, A. J. R. *J. Mass Spectrom.* **2003**, *38*, 315.
(159) Harmer, N. J.; Ilag, L. L.; Mulloy, B.; Pellegrini, L.; Robinson, C. V.; Blundell, T. L. *J. Mol. Biol.***2004**, *339*, 821.
(160) Sharon M.; Robinson, C. V. *Annu. Rev. Biochemistry* **2007**, *76*, 167.
(161) Videler, H.; Ilag, L. L.; McKay, A. R. C.; Hanson, C. L.; Robinson, C. V. *FEBS Lett.* **2005**, *579*, 943.
(162) Loo, J. A.; Berhane, B.; Kaddis, C. S.; Wooding, K. M.; Xie, Y.; Kaufman, S. L.; Chernushevich, I. V. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 998.
(163) Uetrecht, C.; Barbu, I. M.; Shoemaker, G. K.; van Duijn, E.; Heck, A. J. *Nature Chem.* **2011**, *3*, 126.
(164) Pukala, T. L.; Ruotolo, B. T.; Zhou, M.; Politis, A.; Stefanescu, R.; Leary, J. A.; Robinson, C. V. *Structure* **2009**, *17*, 1235.
(165) Politis, A.; Park, A.; Hyung, S.-J.; Barsky, D.; Ruotolo, B. T.; Robinson, C. V. *PLoS ONE* **2010**, *5*.
(166) Marsh, J. A.; Hernández, H.; Hall, Z.; Ahnert, S. E.; Perica, T.; Robinson, C. V.; Teichmann, S. A. *Cell* **2013**, *153*.
(167) Hernández, H.; Dziembowski, A.; Taverner, T.; Séraphin, B.; Robinson, C. V. *EMBO reports* **2006**, *7*, 605.
(168) Quintyn, Royston S.; Yan, J.; Wysocki, Vicki H. *Chem. Biol.* **2015**, *22*, 583.
(169) Zhong, Y.; Feng, J.; Ruotolo, B. T. *Anal. Chem.* **2013**, *85*, 11360.
(170) Samulak, B. M.; Niu, S.; Andrews, P. C.; Ruotolo, B. T. *Anal. Chem.* **2016**, *88*, 5290.
(171) Hall, Z.; Politis, A.; Robinson, C. V. *Structure* **2012**, *20*, 1596.

(172) Alber, F.; Dokudovskaya, S.; Veenhoff, L. M.; Zhang, W.; Kipper, J.; Devos, D.; Suprapto, A.; Karni-Schmidt, O.; Williams, R.; Chait, B. T.; Rout, M. P.; Sali, A. *Nature* **2007**, *450*, 683.
(174) Alber, F.; Kim, M. F.; Sali, A. *Structure* **2005**, *13*.
(175) Alber, F.; Förster, F.; Korkin, D.; Topf, M.; Sali, A. *Biochemistry* **2008**, *77*, 443.
(176) Russel, D.; Lasker, K.; Webb, B.; Velázquez-Muriel, J.; Tjioe, E.; Schneidman-Duhovny, D.; Peterson, B.; Sali, A. *PLoS Biol.***2012**, *10*.
(177) Politis, A.; Park, A.; Hall, Z.; Ruotolo, B. T.; Robinson, C. V. *J. Mol. Biol.***2013**, *425*.
(178) Marcoux, J.; Politis, A.; Rinehart, D.; Marshall, D. P.; Wallace, M. I.; Tamm, L. K.; Robinson, C. V. *Structure* **2014**, *22*, 781.
(179) Sinz, A.; Arlt, C.; Chorev, D.; Sharon, M. *Protein Sci.* **2015**, *24*, 1193.
(180) Bernstein, S. L.; Dupuis, N. F.; Lazo, N. D.; Wyttenbach, T.; Condron, M. M.; Bitan, G.; Teplow, D. B.; Shea, J.-E.; Ruotolo, B. T.; Robinson, C. V. *Nat.Chem.***2009**, *1*, 326.
(181) Marcoux, J.; Politis, A.; Rinehart, D.; Marshall, D. P.; Wallace, M. I.; Tamm, L. K.; Robinson, C. V. *Structure* **2014**, *22*.
(182) Politis, A.; Stengel, F.; Hall, Z.; Hernandez, H.; Leitner, A.; Walzthoeni, T.; Robinson, C. V.; Aebersold, R. *Nat Meth* **2014**, *11*, 403.
(183) Schmidt, C.; Macpherson, J. A.; Lau, A. M.; Tan, K. W.; Fraternali, F.; Politis, A. *Anal. Chem.* **2017**, *89*, 1459.
(184) van Duijn, E.; Barbu, I. M.; Barendregt, A.; Jore, M. M.; Wiedenheft, B.; Lundgren, M.; Westra, E. R.; Brouns, S. J. J.; Doudna, J. A.; van der Oost, J.; Heck, A. J. *Mol. Cell. proteomics :***2012**, *11*, 1430.
(185) Rouillon, C.; Zhou, M.; Zhang, J.; Politis, A.; Beilsten-Edmands, V.; Cannone, G.; Graham, S.; Robinson, C. V.; Spagnolo, L.; White, M. F. *Mol. Cell* **2013**, *52*.
(186) Ganem, B.; Li, Y. T.; Henion, J. D. *J. Am. Chem. Soc.* **1991**, *113*, 6294.
(187) Ganem, B.; Li, Y. T.; Henion, J. D. *J. Am. Chem. Soc.* **1991**, *113*, 7818.
(188) Loo, R. R. O.; Goodlett, D. R.; Smith, R. D.; Loo, J. A. *J. Am. Chem. Soc.* **1993**, *115*, 4391.
(189) El-Hawiet, A.; Kitova, E. N.; Liu, L.; Klassen, J. S. *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 1893.
(190) Liu, L.; Kitova, E. N.; Klassen, J. S. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 310.
(191) Lin, H.; Kitova, E. N.; Klassen, J. S. *Anal. Chem.* **2013**, *85*, 8919.
(192) Mathur, S.; Badertscher, M.; Scott, M.; Zenobi, R. *Phys. Chem. Chem. Phys.* **2007**, *9*, 6187.
(193) Jecklin, M. C.; Touboul, D.; Jain, R.; Toole, E. N.; Tallarico, J.; Drueckes, P.; Ramage, P.; Zenobi, R. *Anal. Chem.* **2009**, *81*, 408.
(194) Gavriilidou, A. F. M.; Gülbakan, B.; Zenobi, R. *Anal. Chem.* **2015**, *87*, 10378.
(195) Daniel, J. M.; Friess, S. D.; Rajagopalan, S.; Wendt, S.; Zenobi, R. *Int. J. Mass Spectrom.* **2002**, *216*, 1.
(196) Madler, S.; Erba, E. B.; Zenobi, R. In *Applications of Maldi-Tof Spectroscopy*; Cai, Z., Liu, S., Eds.; Springer-Verlag Berlin: Berlin, 2013; Vol. 331, p 1.
(197) Kitova, E. N.; El-Hawiet, A.; Klassen, J. S. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 1908.
(198) Han, L.; Kitova, E. N.; Li, J.; Nikjah, S.; Lin, H.; Pluvinage, B.; Boraston, A. B.; Klassen, J. S. *Anal. Chem.* **2015**, *87*, 4888.
(199) Han, L.; Shams-Ud-Doha, K.; Kitova, E. N.; Klassen, J. S. *Anal. Chem.* **2016**.
(200) McCammon, M. G.; Scott, D. J.; Keetch, C. A.; Greene, L. H.; Purkey, H. E.; Petrassi, H. M.; Kelly, J. W.; Robinson, C. V. *Structure* **2002**, *10*, 851.
(201) McCammon, M. G.; Hernández, H.; Sobott, F.; Robinson, C. V. *J. Am. Chem. Soc.* **2004**, *126*, 5950.
(202) Gulbakan, B.; Barylyuk, K.; Zenobi, R. *Curr. Opin. Biotechnol.* **2015**, *31*, 65.
(203) Cong, X.; Liu, Y.; Liu, W.; Liang, X.; Russell, D. H.; Laganowsky, A. *J. Am. Chem. Soc.* **2016**, *138*, 4346.
(204) Heath, B. L.; Jockusch, R. A. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 1911.
(205) Kudryashova, E.; Quintyn, R.; Seveau, S.; Lu, W. Y.; Wysocki, V. H.; Kudryashov, D. S. *Immunity* **2014**, *41*, 709.
(206) Zhao, Y. J.; Singh, A.; Li, L. Y.; Linhardt, R. J.; Xu, Y. M.; Liu, J.; Woods, R. J.; Amster, I. J. *Analyst* **2015**, *14*, 6980.

(207) Harvey, S. R.; Porrini, M.; Stachl, C.; MacMillan, D.; Zinzalla, G.; Barran, P. E. *J Am Chem Soc* **2012**, *134*, 19384.

(208) Rabuck, J. N.; Hyung, S.-J.; Ko, K. S.; Fox, C. C.; Soellner, M. B.; Ruotolo, B. T. *Anal. Chem.* **2013**, *85*, 6995.

(209) Sciore, A.; Su, M.; Koldewey, P.; Eschweiler, J. D.; Diffley, K. A.; Linhares, B. M.; Ruotolo, B. T.; Bardwell, J. C. A.; Skiniotis, G.; Marsh, E. N. G. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 8681.

(210) Cristie-David, A. S.; Sciore, A.; Badieyan, S.; Escheweiler, J. D.; Koldewey, P.; Bardwell, J. C. A.; Ruotolo, B. T.; Marsh, E. N. G. *Mol. Systems Design & Engineering* **2017**.

Chapter 2: CIUSuite: A Quantitative Analysis Package for Collision Induced Unfolding Measurements of Gas-Phase Protein Ions

Joseph D. Eschweiler, Jessica N. Rabuck-Gibbons, Yuwei Tian, and Brandon T. Ruotolo
***Anal. Chem.*, 2015, *87* (22), pp 11516–11522**
DOI: 10.1021/acs.analchem.5b03292
References in this chapter are formatted according to ACS Anal. Chem. standards

2.1 Abstract

Ion mobility-mass spectrometry (IM-MS) is a technology of growing importance for structural biology, providing complementary 3D structure information for biomolecules within samples that are difficult to analyze using conventional analytical tools through the near-simultaneous acquisition of ion collision cross sections (CCSs) and masses. Despite recent advances in IM-MS instrumentation, the resolution of closely related protein conformations remains challenging. Collision induced unfolding (CIU) has been demonstrated as a useful tool for resolving isocrossectional protein ions, as they often follow distinct unfolding pathways when subjected to collisional heating in the gas phase. CIU has been used for a variety of applications, from differentiating binding modes of activation state-selective kinase inhibitors to characterizing the domain structure of multidomain proteins. With the growing utilization of CIU as a tool for structural biology, significant challenges have emerged in data analysis and interpretation, specifically the normalization and comparison of CIU data

sets. Here, we present CIUSuite, a suite of software modules designed for the rapid processing, analysis, comparison, and classification of CIU data. We demonstrate these tools as part of a series of workflows for applications in comparative structural biology, biotherapeutic analysis, and high throughput screening of kinase inhibitors. These examples illustrate both the potential for CIU in general protein analysis as well as a demonstration of best practices in the interpretation of CIU data.

2.2 Introduction

Native mass spectrometry (MS) is now a widespread technique in the structural biology community due to its ability to study the stoichiometry and connectivity of heterogeneous biomolecules while having lesser requirements on concentration and purity of such samples than other common techniques.[1,2] Coupling native MS with ion mobility spectrometry (IM-MS) allows for simultaneous interrogation of the mass, charge, and size of biological macromolecules, which has proven invaluable in the structural analysis of complex biological systems.[3] Recently, IM-MS has been successfully utilized to solve the structures of important macromolecular complexes,[4-7] probe structural changes upon ligand binding,[8,9] examine the polydispersity of protein complexes,[10,11] and study the effects of small molecules on amyloid formation in disease models.[12-14]

A key feature of IM-MS for structural and pharmaceutical applications is the ability to measure the orientationally averaged collision cross section (CCS) of an ion in addition to its mass and charge. The CCS is a coarse-grained size parameter that is limited in information content when viewed alone, but can become information rich when measured as a function of stoichiometry,[15] ligand binding,[16] or ion activation.[2,17]

Additionally, the experimental CCS is an extremely important scoring metric for modeling complex systems, as it can be compared to CCSs calculated from other known or inferred structures.[18]

A long-term challenge for IM-MS has been the resolution of closely-related protein conformations, commonly observed by X-ray and NMR analyses. Despite recent enhancements to IM resolving power, IMS still faces significant challenges when attempting separating protein conformations that differ by less than 2% in CCS. The information content of an IM-MS experiment can be greatly enriched by the addition of gas-phase ion activation, as some structural differences in protein structure are too subtle to be detected by classical IM separations. Early experiments that utilized gas-phase protein unfolding to both study and differentiate protein structures focused on small, single domain proteins and detected stability difference for proteins as a function of charge state, and for those with intact disulfide bonds.[19] Subsequent experiments extended these observations to the ligand-bound forms of wild-type (WT) and disease-associated variants of tetrameric transthyretin (TTR).[2] In this study, a 3D contour plot of ion intensity as a function of activation voltage and drift time, termed a collision induced unfolding (CIU) fingerprint, was used to perform an in-depth analysis of subtle differences in the unfolding and dissociation pathways of TTR variants, identifying additional ligand-based protein stabilization in mutant TTR forms not detectable by IM-MS alone.

Since these earlier experiments, CIU fingerprints have been used in the context of various applications. These efforts include: studying the influence of bound anions and cations on gas-phase protein stability,[20,21] distinguishing between inhibitors that

stabilize either the active or the inactive form of the Abelson protein tyrosine kinase (Abl),[22] measuring stability enhancements and cooperativity effects in proteins upon ligand-binding,[17,23] probing the selectivity of lipid binding in membrane proteins,[24] determining the domain structures for 16 proteins with varying molecular weights and domain structures,[25] and differentiating between disulfide binding isoforms in antibodies.[26]

Despite these varied and potentially impactful applications, CIU has not reached its full potential as a tool for structural biology and drug discovery. Key challenges for the technique include the general underutilization of structural information content of CIU data, as well as a lack of high throughput experimental frameworks and data analysis tools. Recent advances have been made in analysis and interpretation of other IM-MS data types, including deconvolution algorithms,[26,27] and an array of methods for prediction of CCSs from experimental or model structures,[28-30], but thus far CIU data has not been the focus of any such data analysis packages.

In order to move forward in the use of CIU as a tool for general structural biology, as well as for high-throughput pharmaceutical applications specifically, data analysis tools and strategies for handling the large amount of data that is produced by CIU fingerprints must be developed and implemented. In this report, we describe such software tools, collectively named CIUSuite, designed to ameliorate many of the challenges described above. Additionally, through detailed discussions of three diverse applications of the CIU technique, we illustrate workflows and best practices for extracting maximal information content from CIU data.

2.3 CIUSuite Overview

To facilitate interpretation of CIU data for a variety of applications, we designed CIUSuite, a series of Python[31] modules for generation and manipulation of CIU fingerprints. CIUSuite consists of six modules that allow the user to readily access statistical and structural information from CIU experiments by designing user-defined CIUSuite workflows. The main data structure in CIUSuite is the three dimensional size, activation energy, and intensity matrix that forms the CIU fingerprint. The size axis is IM-MS drift time (ms) by default, however the user is able to substitute CCS values when available. The activation energy axis can be expressed in volts, laboratory–frame energy, or center-of-mass frame energy. In an attempt to obviate potential problems arising from signal intensity variation between samples, the intensities for each activation energy are normalized to a maximum value of 1 and smoothed using a Savitsky-Golay filter with a window length of three and polynomial order of two, however these parameters can also be easily adjusted by the user.

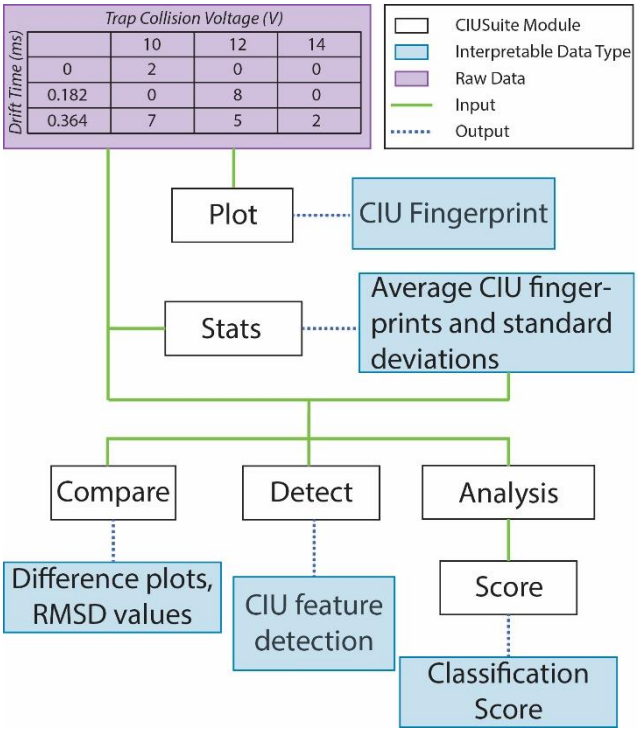


Figure 2-1 Schematic representation of CIUSuite modules. All modules (shown as white boxes) take as input raw data in the form of a 2D matrix (purple) formatted such that ion intensity is collected a function of drift time and trap collision voltage. Additionally, modules can accept outputs from CIU_stats for groupwise comparisons using average and standard deviation measurements. Example outputs from each of these modules are also shown (blue). Both the modules and their outputs are discussed in detail in the text.

CIUSuite_plot forms the basis of the CIUSuite. CIUSuite_plot batch processes any CIU data in its working directory tagged with the suffix "_raw.csv" and writes the corresponding contour plot to a .png file.

CIUSuite_stats outputs both visual information that can be interpreted by the user and numerical matrices that can be used for downstream analysis. The current implementation of CIU_stats calculates the average and standard deviation of all of the CIU data in a directory with the "_raw.csv" tag. This calculation is performed for every data point in the CIU fingerprint, and both the average and standard deviation matrices are output as a .csv matrix as well as a .png figure in the same fashion as CIU_plot. This module requires at least 3 datasets to calculate the standard deviation fingerprint

CIUSuite_compare allows for facile comparison of CIU fingerprints by matrix subtraction and visualization of the difference matrix. Inputs for CIUSuite_compare can be raw data matrices or the average matrices output from CIUSuite_stats. CIUSuite_compare also utilizes the root mean square deviation (RMSD) parameter to report the absolute .

difference between two matrices and prints the RMSD on the difference plot. Here, RMSD is defined in Equation 1 as:

$$RMSD = \sqrt{\frac{\sum(A-B)^2}{m \times n}} \times 100\% \qquad (1)$$

where *A* and *B* are both *m* x *n* CIU matrices. The module operates in three modes:

Basic Mode: Outputs the difference plot and RSMD for two user defined inputs.

Batch Mode: Outputs the difference plot and RMSD for a reference dataset compared to all other _raw.csv files in the directory. Here, RMSDs are also saved to a .csv file for future reference.

Cluster Mode: Calculates pairwise RMSDs for all _raw.csv files in the directory and utilizes K-medoids clustering to form a user-defined number of clusters. Cluster mode outputs a .csv file with the optimal clustering, where each file is assigned a cluster number corresponding to the file that is the medoid of the cluster. Additionally, the pairwise RMSD matrix (distance matrix) is output as a .csv if the user desires to utilize other clustering algorithms.

CIUSuite_detect is a simple feature detection algorithm that allows for quantitative analysis of CIU data. The algorithm utilizes the first derivative test to identify local maxima in the data, before refining the shape of the feature using user-defined data scaling and intensity thresholds. After features are identified and refined, their stabilities in collision voltage space as well as their centroid drift times (or CCSs) are output to a file summarizing the dataset. CIUSuite_detect works in a batch processing mode that allows for detailed quantitative comparisons of CIU datasets beyond the absolute difference output from CIUSuite_compare. Because CIUSuite_detect may require tuning of data scaling and intensity thresholding parameters, we also output reconstructed CIU fingerprint plots showing only the features used in the analysis.

CIUSuite_analysis was developed as a tool for adaptation of CIU fingerprinting for high throughput ligand screening and structural biology. CIU_analysis allows the user to identify areas within the CIU fingerprint that are useful for categorizing datasets into groups, such as type I or type II kinase inhibitors (*vide infra*). The current

implementation of CIU_analysis takes as input a training dataset, where each file is annotated as either a type I (_typeI_raw.csv) or type II (_typeII_raw.csv) fingerprint. CIU_analysis utilizes a scaled deviation score where each fingerprint in the dataset is compared to the average fingerprints for both the type I and type II groups. Here, the type I scaled deviation score (SDS) is defined in Equation 2:

$$SDS_i = \sum_{j=0}^{j=m} \frac{(X_{ij} - A_{ij}^1) \times A_{ij}^1}{S_{ij}^1} \quad (2)$$

Where X is a CIU matrix, A^1 is the average type I matrix, S^1 is the type I standard deviation matrix, i is a given collision energy, j a given drift time, and m is the total number of drift time bins. The primary outputs are two plots of SDS vs Collision Voltage, one corresponding to the type I average SDS value, the other corresponding to the type II average SDS value. These plots display the average SDS value (with 2 standard deviations as the error bars) for both type I and type II fingerprints compared to the corresponding average value. The information contained in these two plots is extremely valuable for the accurate classification of unknown fingerprints as well as targeting CIU workflows toward optimal regions of dissimilarity between the two data classes, increasing the throughput of the experiment. CIUSuite_analysis also outputs a plot of SDS vs Collision Voltage for each component of the training dataset, allowing the user to identify outliers or other anomalies that may bias the analysis.

CIUSuite_score is predicated on data from CIUSuite_analysis. CIUSuite_score accepts as input a training data set and "unknown" data that is tagged with "_uk_raw.csv." Previous analyses of CIU fingerprint data[22] have shown that focusing on specific collision voltages, rather than using the entire CIU fingerprint, can increase the

throughput and robustness of the resulting screen. After identifying the voltage ranges or drift times in the CIU fingerprints that yield significant group-wise deviation values using CIUSuite_analysis, the user can enter these values into the scoring module to calculate classification scores based only on these regions. Each training data set is grouped according to a user-defined tag, and corresponding SDS values are calculated. SDS values are then summed over all of the collision energies to be scored, assigning a single scaled deviation value for each fingerprint. For example, type I fingerprints should have low overall deviation relative to the type I average, whereas they should have higher deviation scores relative to the type II average. The type I z-score with respect to the type I training data for an unknown fingerprint is simply the z-score of its SDS compared to the average SDS for a type I compared to the type I average, as described in Equation 3:

$$z\ score^{1} = \frac{SDS^{1} - \bar{x}^{1}}{s^{1}} \qquad (3)$$

Where SDS^1 is the SDS of the unknown compared to the type I average, $\bar{x}^1$ is the average SDS of a type I compared to the type I average, and s^1 is the standard deviation of SDS values around $\bar{x}^1$.The output for CIUSuite_score comprises of a .csv file that contains the type I and type II z-scores for each dataset and a graph showing the type II classification-score vs type I classification-score. The resulting plot displays type I training data in blue, type II training data in red, and unknown scores in cyan. A blue and red box around the data sets indicates two standard deviations from the type I and type II training data, respectively.

2.4 CIUSuite Applications

Assessing the reproducibility of CIU for intact antibody analysis

Monoclonal antibodies (mAbs) are among the fastest growing class of therapeutics due to their high specificity and low incidence of side effects. Unlike most drugs, mAbs are complex macromolecules (~150 kDa), leading to a host of quality control and characterization challenges inherent in their development.[32] We have developed a CIU method capable of differentiating human IgG subtypes which mainly differ by inter-chain disulfide bonding numbers and patterns.[26] Our CIU method was validated by reproducibility tests enabled by CIUSuite, where large numbers of replicates were collected and analyzed.

In order to illustrate the function of CIUSuite_stats, we analyzed the CIU profiles for four IgG4 samples from four different suppliers (Sigma-Aldrich Co., St. Louis, Mo; MyBioSource Inc., San Diego, Ca; Abcam, Cambridge, UK; and Fitzgerald Industries International, Acton, Ma.), denoted as IgG4-1 (Lot

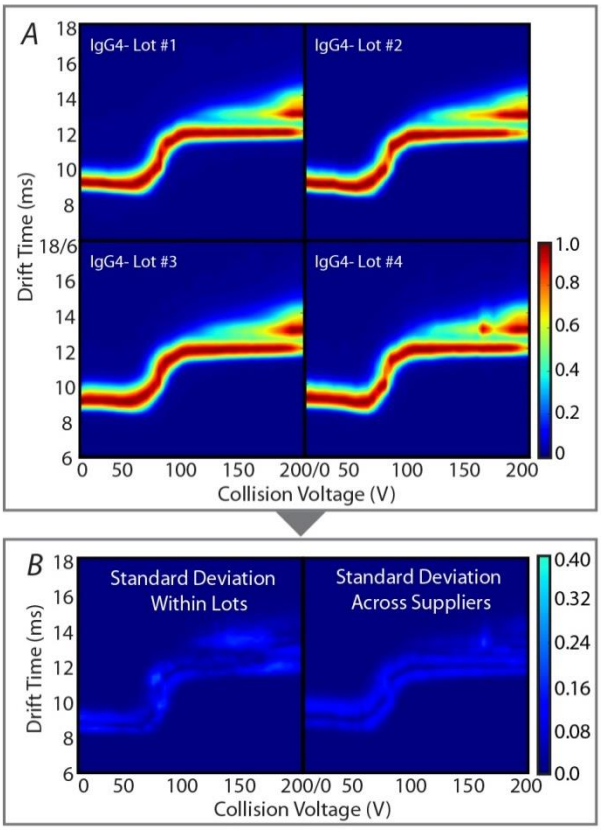


Figure 2-2. CIU of Biotherapeutics. (A) CIU fingerprints for Human IgG4 samples are plotted using CIUSuite_plot and display almost identical unfolding pathways. (B) Standard deviation plots are generated using CIUSuite_stats for CIU replicates of a single lot of IgG4 purchased from Sigma-Aldrich (Left) and CIU fingerprints of IgG4 samples from various suppliers (Right).

#1), IgG4-2 (Lot #2), IgG4-3 (Lot #3), and IgG4-4 (Lot #4) respectively.

For the CIU experiments, each sample was provided from the manufacturer in a 20 mM phosphate buffer that contained 150 mM NaCl, pH 7.4, with 0.05% sodium azide as a preservative except for the Sigma sample, which was provided in 20 mM Tris buffered saline, pH 8.0. All samples were purchased at a concentration of 1 mg/ml (~6.7 µM) and buffer exchanged into 100 mM ammonium acetate buffer using Micro Bio-Spin 30 columns (Bio-Rad, Hercules, CA) without further purification. ~7 µl aliquotes were introduced to a nano-electrospray ionization-quadrupole-ion mobility-time-of-flight mass spectrometer (Synapt G2, Milford, MA). Capillary voltages were set to 1.5-1.7kV, and the sampling cone was operated at 60V. The trap traveling wave ion guide was set to a pressure of 3.4×10^{-2} mbar of argon gas, and the traveling wave ion mobility separator was set to 3.5 mbar. The wave height and wave velocity was set to 40V and 600 m/s, respectively, to maintain ion mobility separation. The time-of-flight (ToF) was operated at a pressure of 1.7×10^{-6} mbar over a *m/z* range of 1,000-10,000. Collision energy was added to the ions in the trap traveling wave ion guide before the IM separator to unfold the antibodies. The 23+ charge state was isolated in the quadrupole and the collision voltage was ramped from 5-200V in 5V increments to unfold the ions, collecting IM-MS spectra at each voltage.

As shown in Figure 2-2 A, the 23+ charge state of intact IgG4 ions exhibit identical initial drift times prior to activation and highly similar unfolding pathways. For example, as the collision voltage is increased to the region of 70V – 100V, all IgG4 ions exhibit a gradual transition from an initial compact state, to an elongated unfolded state, resulting in an increase in recorded IM drift time. This unfolded species dominates the

CIU fingerprint for all species tested, until a larger unfolded protein ions are produced between 120V and 160V. At voltage values higher than 160V, two unfolded IgG4 forms are universally observed, at nearly equal intensities.

While small variations between replicate CIU fingerprints are apparent, a CIUSuite statistical analysis of the data reveals that these variations are insignificant, leading to the conclusion that the four IgG4 samples can be treated as 'identical'. The workflow for this analysis first involves the acquisition of multiple CIU data replicates for a single IgG4-1 lot. These CIU fingerprint replicates are then plotted and analyzed using CIUSuite_plot and CIUSuite_stats functions and their standard deviations comprehensively assessed (Figure 2-2B, left). Near-zero deviations are observed across the whole collision voltage range, consistent with previous observations.[26] With these "baseline" standard deviation values established, we can further evaluate CIU reproducibility by investigating variations between samples. To perform such an assessment, a standard deviation plot is generated using CIUSuite that compares CIU fingerprints acquired for IgG4-1, IgG4-2, IgG4-3, and IgG4-4 (Figure 2-2B, right). Standard deviations equal to or smaller than the baseline values are observed, further illustrating the excellent reproducibility of the CIU method. Taken together, the data shown in Figure 2-2 illustrate the capabilities of CIUsuite to evaluate CIU data for potential applications in biopharmaceutical characterization and quality control.

Quantifying Differences in the Unfolding Pathways of Homologous Albumins

We evaluated the use of the CIU fingerprint experiment for differentiation of homologous serum albumins for both analytical separations and comparative biology. Using albumins from bovine (BSA, 66,463Da) and human (HSA, 66,437Da) serum (both from Sigma Aldrich Co., St. Louis, MO, 10uM, buffer exchanged into 100mM ammonium acetate using a Biospin 6 column(Bio-Rad, Hercules, CA)), we isolated the 15+ charge state and subjected it to collisional activation over a voltage range of 26 to 188V, using instrument conditions similar to those published previously.[25] Bovine and human albumin are both three-domain proteins that share 76%

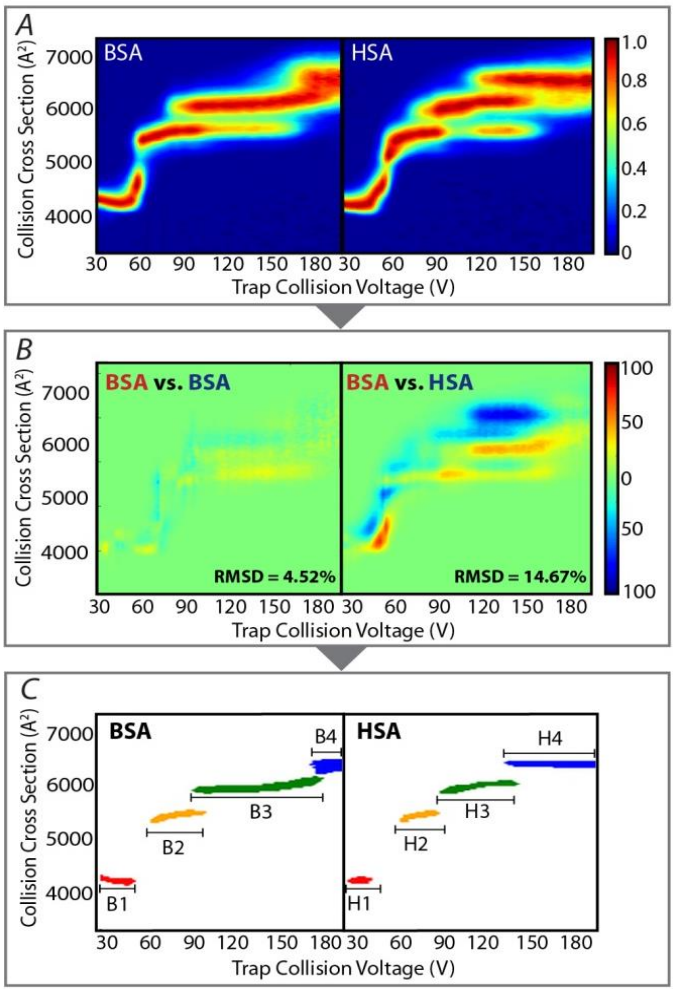


Figure 2-3. Analysis of homologous serum albumins reveals significant differences in their CIU fingerprints**. A)** Comparison of bovine (BSA) and human (HSA) serum albumin fingerprints using CIUSuite_plot reveals similar unfolding pathways, each having four CCS features present over the entire voltage range probed. **B)** A product of CIUSuite_compare, difference plots are shown where features that identify most strongly with one fingerprint are shown on a red intensity scale, and the other is shown in blue. The evaluation of technical BSA CIU replicates reveals a 5% RMSD, whereas a similar BSA-HAS comparison results in nearly 15%. **C)** Using feature detection and extraction from CIUSuite_detect, we quantified the cross section (A^2), centroid stability (V), and stability range (V) of each feature, finding that across homologous, the CCS of the unfolding intermediates vary negligibly, with most of the variability coming from the different CIU stabilities of the features observed.

sequence identity, and are indistinguishable in both m/z and ground state CCS in the context of a native mass spectrometry experiment. In Figure 2-3A, we first observe that both homologues produce four prominent CIU features, in agreement with our previous report noting the correlation between native protein domain structure and CIU.[25] Our previous data showed that at for low charge states, the number of CIU transitions observed should equal the number of domains, indicating uncoupling and unfolding of individual domains as a major source of CCS transitions in CIU.

Despite these similarities, we also observe qualitative differences in the fingerprints that require further investigation. To assess the reproducibility of the experiment and the significance of these differences, we used CIU_compare to assign RMSD values to BSA and HSA replicates, comparing these values with RMSDs obtained by

comparing BSA directly to HSA(Figure 2-3B). We find that in proteins such as BSA and HSA, the RMSDs observed for technical replicates are low, under 5% for all cases with satisfactory (>3) signal intensities. In contrast, the RMSD computed between BSA and HSA was found to be 14.7%, indicating significant differences in the unfolding process between homologues. The difference plot shown in Figure 2-3 allows us indicates a few major regions of the fingerprint where deviation occurs, specifically subtle changes in the region between 40V and 60V, and much more pronounced changes in the region from 120V to 150V.

To gain a more quantitative understanding of the changes in CIU between BSA and HSA, we extracted and characterized the most intense features from each fingerprint using CIU_detect (Figure 2-3C). The results of this analysis show that the

centroid drift time of each feature varies by only ~1% across homologues, whereas the centroid position of these features in the collision voltage dimension, and the stability of the feature, exhibit larger average variations, 9% and 20%, respectively. We interpret these results as support for our previous work that initially linked the native domain structure of proteins to their respective CIU pathways at low charge states.[25] Additionally, the differences in the stability of CIU features between homologous proteins may indicate potential for domain or interface-specific stability measurements to be used in biopharmaceutical or protein engineering applications.

2.5 Conclusions

Analysis of CIU data is an emerging challenge for those seeking to expand the information content of typical IM-MS experiments. Rapid and robust procedures for CIU fingerprint analysis are necessary for the continued development and application of such gas-phase unfolding experiments. Emerging applications, such as protein engineering and high throughput screening, which involve the rapid analysis of large numbers of samples, require streamlined quantitative analysis, provided by CIUSuite, in order to achieve realistic analysis capacities. Forthcoming challenges in this field may include the integration of CIU fingerprint data into databases, allowing for analysis of variability across instruments as well as the comparison of CIU data for quality control applications. Although the CIU analysis workflows contained herein overcome major bottle necks in experiment and analysis time, hurdles still exist in making CIU ready for diverse high throughput applications. One exciting area of exploration will surely be the integration of adaptive, data-dependent algorithms for optimization of signal and

analysis time in high throughput screens. Important features of these algorithms will be the ability to adaptively focus on collision voltages that are the most information-rich, and to rapidly tune instrumental conditions and acquisition times to ensure that adequate signal-to-noise is achieved for each measurement.

This software and the mathematical procedures contained within CIUSuite represent a framework for the continued study of gas-phase protein unfolding, and we anticipate that its application will lead to further discoveries regarding the basic biophysics of proteins in the absence of bulk solvent. In addition, as the study of protein unfolding analysis in the gas-phase is a relatively new area, the authors encourage modification and expansion of CIUSuite capabilities, so that the base approaches described here can be applied to data structures not yet conceived.

2.6 ACKNOWLEDGMENT

The authors would like to thank Dr. Kerby Shedden at the Center for Statistical Consultation & Research at the University of Michigan for critical discussions on the statistical analysis and Python code for CIUSuite. CIU method development in the Ruotolo group is supported by the National Science Foundation (CAREER, 1253384) and the University of Michigan department of Chemistry. JDE acknowledges support from the University of Michigan Rackham Graduate School in the form of a Research Award, and JNRG acknowledges support in the form of a Rackham Merit Fellowship and a Research Award.

2.7 SUPPORTING INFORMATION

Supporting Information can be found in Appendix I

2.8 REFERENCES

(1) Benesch, J. L. P.; Ruotolo, B. T. *Curr. Opin. Struct. Biol.* **2011**, *21*, 641.

(2) Hyung, S.-J.; Robinson, C. V.; Ruotolo, B. T. *Chem. Biol.* **2009**, *16*, 382.

(3) Uetrecht, C.; Rose, R. J.; van Duijn, E.; Lorenzen, K.; Heck, A. J. R. *Chem. Society Reviews* **2010**, *39*, 1633.

(4) Uetrecht, C.; Barbu, I. M.; Shoemaker, G. K.; van Duijn, E.; HeckAlbert, J. R. *Nat Chem* **2011**, *3*, 126.

(5) Sharon, M.; Mao, H.; Boeri Erba, E.; Stephens, E.; Zheng, N.; Robinson, C. V. *Structure* **2009**, *17*, 31.

(6) Politis, A.; Park, A. Y.; Hall, Z.; Ruotolo, B. T.; Robinson, C. V. *J. of Mol. Biology* **2013**, *425*, 4790.

(7) Ruotolo, B. T.; Benesch, J. L.; Sandercock, A. M.; Hyung, S.-J.; Robinson, C. V. *Nature protocols* **2008**, *3*, 1139.

(8) Niu, S.; Rabuck, J. N.; Ruotolo, B. T. *Curr. Opin. Chem. Biology* **2013**, *17*, 809.

(9) Xing, W.; Busino, L.; Hinds, T. R.; Marionni, S. T.; Saifee, N. H.; Bush, M. F.; Pagano, M.; Zheng, N. *Nature* **2013**, *496*, 64.

(10) Shepherd, D. A.; Marty, M. T.; Giles, K.; Baldwin, A. J.; Benesch, J. L. P. *Int. J. Mass Spectrom.* **2015**, *377*, 663.

(11) Pagel, K.; Natan, E.; Hall, Z.; Fersht, A. R.; Robinson, C. V. *Angewandte Chemie Int. Ed.* **2013**, *52*, 361.

(12) Soper, M. T.; DeToma, A. S.; Hyung, S.-J.; Lim, M. H.; Ruotolo, B. T. *Physical Chemistry Chem. Phys.* **2013**, *15*, 8952.

(13) Do, T. D.; Economou, N. J.; Chamas, A.; Buratto, S. K.; Shea, J.-E.; Bowers, M. T. *The J. of Physical Chemistry B* **2014**, *118*, 11220.

(14) Susa, A. C.; Wu, C.; Bernstein, S. L.; Dupuis, N. F.; Wang, H.; Raleigh, D. P.; Shea, J.-E.; Bowers, M. T. *J. Am. Chem. Soc.* **2014**, *136*, 12912.

(15) Pukala, T. L.; Ruotolo, B. T.; Zhou, M.; Politis, A.; Stefanescu, R.; Leary, J. A.; Robinson, C. V. *Structure* **2009**, *17*, 1235.

(16) Brocca, S.; Testa, L.; Sobott, F.; Šamalikova, M.; Natalello, A.; Papaleo, E.; Lotti, M.; De Gioia, L.; Doglia, Silvia M.; Alberghina, L.; Grandori, R. *Biophysical J.* **2011**, *100*, 2243.

(17) Niu, S.; Ruotolo, B. T. *Protein Science* **2015**.

(18) Politis, A.; Park, A. Y.; Hyung, S.-J.; Barsky, D.; Ruotolo, B. T.; Robinson, C. V. **2010**.

(19) Shelimov, K. B.; Clemmer, D. E.; Hudgins, R. R.; Jarrold, M. F. *J. Am. Chem. Soc.* **1997**, *119*, 2240.

(20) Han, L.; Hyung, S.-J.; Mayers, J. J.; Ruotolo, B. T. *J. Am. Chem. Soc.* **2011**, *133*, 11358.

(21) Han, L.; Hyung, S. J.; Ruotolo, B. T. *Angew. Chem.* **2012**, *124*, 5790.

(22) Rabuck, J. N.; Hyung, S.-J.; Ko, K. S.; Fox, C. C.; Soellner, M. B.; Ruotolo, B. T. *Anal. Chem.* **2013**, *85*, 6995.

(23) Hopper, J. T.; Oldham, N. J. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1851.

(24) Laganowsky, A.; Reading, E.; Allison, T. M.; Ulmschneider, M. B.; Degiacomi, M. T.; Baldwin, A. J.; Robinson, C. V. *Nature* **2014**, *510*, 172.

(25) Zhong, Y.; Han, L.; Ruotolo, B. T. *Angew. Chem.* **2014**, *126*, 9363.

(26) Marty, M. T.; Baldwin, A. J.; Marklund, E. G.; Hochberg, G. K.; Benesch, J. L.; Robinson, C. V. *Analytical chemistry* **2015**, *87*, 4370.

(27) Sivalingam, G. N.; Yan, J.; Sahota, H.; Thalassinos, K. *Int. J. Mass Spectrom.* **2013**, *345*, 54.

(28) Larriba, C.; Hogan, C. J. *J. of Comput. Phys.* **2013**, *251*, 344.

(29) Marklund, E. G.; Degiacomi, M. T.; Robinson, C. V.; Baldwin, A. J.; Benesch, J. L. *Structure* **2015**, *23*, 791.

(30) Bleiholder, C.; Wyttenbach, T.; Bowers, M. T. *Int. J. Mass Spectrom.* **2011**, *308*, 1.

(31) Van Rossum, G.; Drake Jr, F. L. *Python tutorial*; Centrum voor Wiskunde en Informatica, 1995.

(32) Beck, A.; Wagner-Rousset, E.; Ayoub, D.; Van Dorsselaer, A.; Sanglier-Cianférani, S. *Analytical chemistry* **2012**, *85*, 715.

Chapter 3: Chemical Probes and Engineered Constructs Reveal a Detailed Unfolding Mechanism for a Solvent-Free Multi-Domain Protein

Joseph D. Eschweiler, Rachel M. Martini, and Brandon T. Ruotolo
J. Am. Chem. Soc., 2017, 139 (1), pp 534–540
DOI: 10.1021/jacs.6b11678
References in this chapter are formatted according to J. Am. Chem. Soc. standards
Supplemental Information can be found in *Appendix I*

3.1 Abstract

Despite the growing application of gas-phase measurements in structural biology and drug discovery, the factors that govern protein stabilities and structures in a solvent-free environment are still poorly understood. Here, we examine the solvent-free unfolding pathway for a group of homologous serum albumins. Utilizing a combination of chemical probes and non-covalent reconstructions, we draw new specific conclusions regarding the unfolding of albumins in the gas-phase, as well as more-general inferences regarding the sensitivity of collision induced unfolding to changes in protein primary and tertiary structure. Our findings suggest that the general unfolding pathway of low charge state albumin ions is largely unaffected by changes in primary structure; however, the stabilities of intermediates along these pathways vary widely as sequences diverge. Additionally, we find that human albumin follows a domain associated unfolding pathway, and are able to assign each unfolded form observed in

our gas-phase dataset to the disruption of specific domains within the protein. The totality of our data informs the first detailed mechanism for multi-domain protein unfolding in the gas phase, and highlights key similarities and differences from the known the solution-phase pathway.

3.2 Introduction

A detailed understanding of protein structure is centrally-important in the post-genomic era, especially in the context of human disease.[1] Despite nearly sixty years of molecular-level observations, and an online repository of nearly 120,000 structural datasets, our ability to predict the three-dimensional fold of an amino acid sequence *ab initio* is mainly limited to small, single domain proteins.[2] In contrast, the successes of template-based methods of protein structure prediction, relying upon previously-captured structural data, can extend to much larger sequences.[3] Currently, such datasets are limited primarily to those gathered through X-ray diffraction, nuclear magnetic resonance (NMR) spectroscopy, or electron microscopy (EM).[4] While all highly enabling, high-resolution technologies in their own right, these techniques also bear significant limitations in terms of their throughput and their ability to access mixtures, thus rendering significant regions of the proteome absent from our current structural databases and refractory to rational drug design efforts.[5] Therefore, it is clear that in order to move forward our fundamental understanding of the forces that drive environment-dependent protein folding reactions, protein structure data from other experimental methods must be considered.

Beginning with the introduction of electrospray ionization (ESI)[6] and matrix-assisted laser desorption ionization (MALDI)[7,8] over twenty-five years ago, solvent-free biomolecular structure has been targeted in an effort to resolve some of the mysteries surrounding native protein folding. In the gas phase, a simplified state of biological matter can be accessed, free from its native environment and accessible to high resolution spectrometric techniques. Surprisingly, many aspects of native protein structure can be retained *in vacuo*, including protein complex binding stoichiometry and topology.[9] In addition, the locations of bound substrates[10,11] and overall protein folds[12,13] can exhibit a strong memory of their native forms when observed in the gas phase. Technologies including gas-phase hydrogen-deuterium exchange mass spectrometry,[14,15] action spectroscopy,[16] and ion mobility-mass spectrometry (IM-MS)[17] have revealed that gas-phase proteins are not 'inside-out' as originally surmised,[18] but are instead largely charge solvated, existing in multiple iso-energetic states, and can strongly resemble their native-like forms.[19]

A key observation from such gas-phase structural biology measurements is that solvent-free proteins can undergo unfolding following sufficient collisional heating, and that unfolding pathways can be monitored by IM-MS and mined for detailed structural information.[20] Generally, these experiments involve sequentially increasing the kinetic energy of ions as they enter a pressurized ion trap, and thereby collisionally heating them. Subsequent analysis of IM drift time distributions for ion populations post-activation generally reveal increases in ion collision cross sections as in a manner correlated with their increased internal energies. Although generalized correlations between gas-phase and native state protein stabilities are not yet available, gas-phase

protein unfolding has already demonstrated substantial promise as a fingerprinting technology in biomolecular analysis.[21] Early IM-MS measurements revealed that single domain protein ions containing disulfide bonds resist collision induced unfolding (CIU) more so than those that lack such bonding.[22] Subsequent CIU experiments targeted protein-protein and protein-ligand complexes, and highlighted the ability of gas-phase unfolding to detect minor differences in protein stability connected to changes in both local and global protein structure.[23-27] More recent experiments have discerned the ability of CIU to detect conformationally-selective ligand binding,[28] the cooperative stabilization upon ligand binding in multiprotein complexes,[29] the details of disulfide bond structure within intact antibodies,[30] and a domain-correlated mechanism of gas-phase unfolding overall.[31] Despite these insights, we still lack a clear, detailed picture of protein CIU. Information regarding the extent to which gas phase unfolding mimics such processes in solution, as well as a detailed view of domain-correlated unfolding events achieved in the absence of bulk solvent, could be transformative for both CIU as an analytical tool and our ability to predict protein structure.

In this report, we use a variety of homologous serum albumins to study the sensitivity of CIU to changes in primary structure. Additionally, we utilize domain-specific chemical probes and novel noncovalent constructs to assign CIU transitions to specific regions of Human Serum Albumin (HSA). Taken together, our results demonstrate, for the first time, a detailed mechanism of gas-phase protein unfolding that links individual increases in ion size to unfolding events within specific regions of a multi-domain protein. In addition, by comparing our gas-phase results with well-known mechanisms of HSA unfolding in solution[32] we are able to determine that elements of albumin CIU

strongly resembles albumin unfolding in solution, adding further evidence of solution-phase memory in gas-phase proteins and allowing us to point toward future applications of CIU in protein stability analyses.

3.3 Experimental Section

Sample Preparation. Wild type (WT) bovine, hominian, ovine, leporine, caprine, murine, and porcine serum albumin were purchased from Sigma Aldrich (St. Louis, MO) as lyophilized powders at purities greater than 97%. (Table S1) The lyophilized proteins were diluted to 100μM in 100mM Ammonium Acetate and stored at -80^{0}C. 8-Anilino-1-naphthalenesulfonic acid (ANS) ammonium salt hydrate (97%), Warfarin (WRF, analytical grade), indomethacin (IDM, 99%), L-thyroxine (98%), Bilirubin (98%), and Hemin (HMN, 98%) were also purchased from Sigma Aldrich, 10mM DMSO stocks were prepared prior to each experiment. A stock solution of 1mg/mL Diazepam (DZP) was generously provided by the Kennedy Group at the University of Michigan. Recombinant albumin domains 1, 2, and 3, as well as the domain 1-2 fusion protein, were purchased from Albumin Biosciences (Huntsville, AL) as lyophilized powders. Recombinant albumin domains were diluted to 90 or 180μM in 100mM ammonium acetate and stored at -80^{0}C.

IM-MS Data Collection. CIU fingerprints were obtained on a Synapt G2 IM-MS instrument (Waters Corp, Manchester, UK) as described previously.[31] Briefly, albumin samples were diluted to 10μM and loaded into homemade, gold coated borosilicate needles. The cone voltage was maintained at 1.5kV with the sampling and extraction cones set to 30V and 2V, respectively. The source pressure was set to 50mbar and the

source backing pressure was adjusted to 9 mbar. For all measurements, the quadrupole was set to isolate the 15^+ charge state between 4420 and 4450 m/z. The IM T-wave ion guide was operated at 4mBar with wave height and wave velocity values of 15V and 150m/s, respectively. Mass spectra and drift time distributions were obtained for the ions at multiple trap collision energies in steps of 2V from 20V to 188V. All collision cross-section (CCS) values, which relate IM drift times directly to ion size and shape, were calibrated using ions of known CCS as described previously and detailed in table S2.[33]

Chemical Probe CID Analysis. HSA was incubated with chemical probes at a ratio of 10uM protein to 100uM probe with a DMSO content of less than 5%. Experiments were performed as described above, although the quadrupole was adjusted to isolate either singly-bound and apo protein. Selected measurements were chosen for replication and their variabilities were found to be +/- 2 V, much lower than what is needed for our data analysis.

CIU/CID analysis of Reconstituted HSA. Noncovalent reconstitutions of HSA were prepared by mixing component domains and incubating on ice for 10 min. The domain 1-2 fusion protein was incubated with domains 1, 2 and 3 at concentrations of 45uM to provide noncovalent complexes. These complexes were then subjected to CIU and CID analysis without quadrupole selection . We similarly analyzed nonspecific trimers comprised of domains 1, 2 and 3, as well as the specific trimer of domains 1, 2 and 3.

IM-MS Data Analysis. CIU fingerprints, subsequent RMSD calculations, and feature analysis were carried out using CIUSuite.[34] CID analysis of both chemical probes and

reconstituted HSA complexes were carried out using Masslynx (Waters Corp. Manchester, UK)

3.4 Results and Discussion

Effects of Protein Primary Structure on CIU.

Previous studies from our group have indicated that the CIU behavior of proteins is sensitive to their domain structure.[31] To further understand the effects of intermolecular interactions governed by primary structure on the global CIU process, we undertook analysis of 7 homologous serum albumins. Each SA studied is a single polypeptide chain composed of three homologous domains (D1, D2, and D3), all of which share near-identical tertiary structure while differing significantly in primary structure from 70 to 90% sequence identity. Previous work has revealed that the CIU fingerprint of an ion is dependent on its charge-state.[31] Figure 3-1A shows the average unfolding pathway for 7 homologous albumins for the 14, 15, and 16^+ charge states. The standard deviation plots to the right characterize the variability between these structures caused by subtle changes to primary structure. As expected, the lowest charge state, 14^+, requires higher voltages to unfold and generally accesses fewer intermediate structures along the unfolding pathway. For the 15^+ and 16^+ ions, the average behavior of the 7 homologues is quite similar, giving rise to a plurality of intermediate conformer families during CIU. Interestingly, analysis of the standard deviation plots reveals significant differences in CIU response caused by small changes to primary structure. We identified the 15^+ ion not only as having the highest total deviation from the mean, but also as exhibiting significant deviations across the largest area of the unfolding fingerprint. We note that

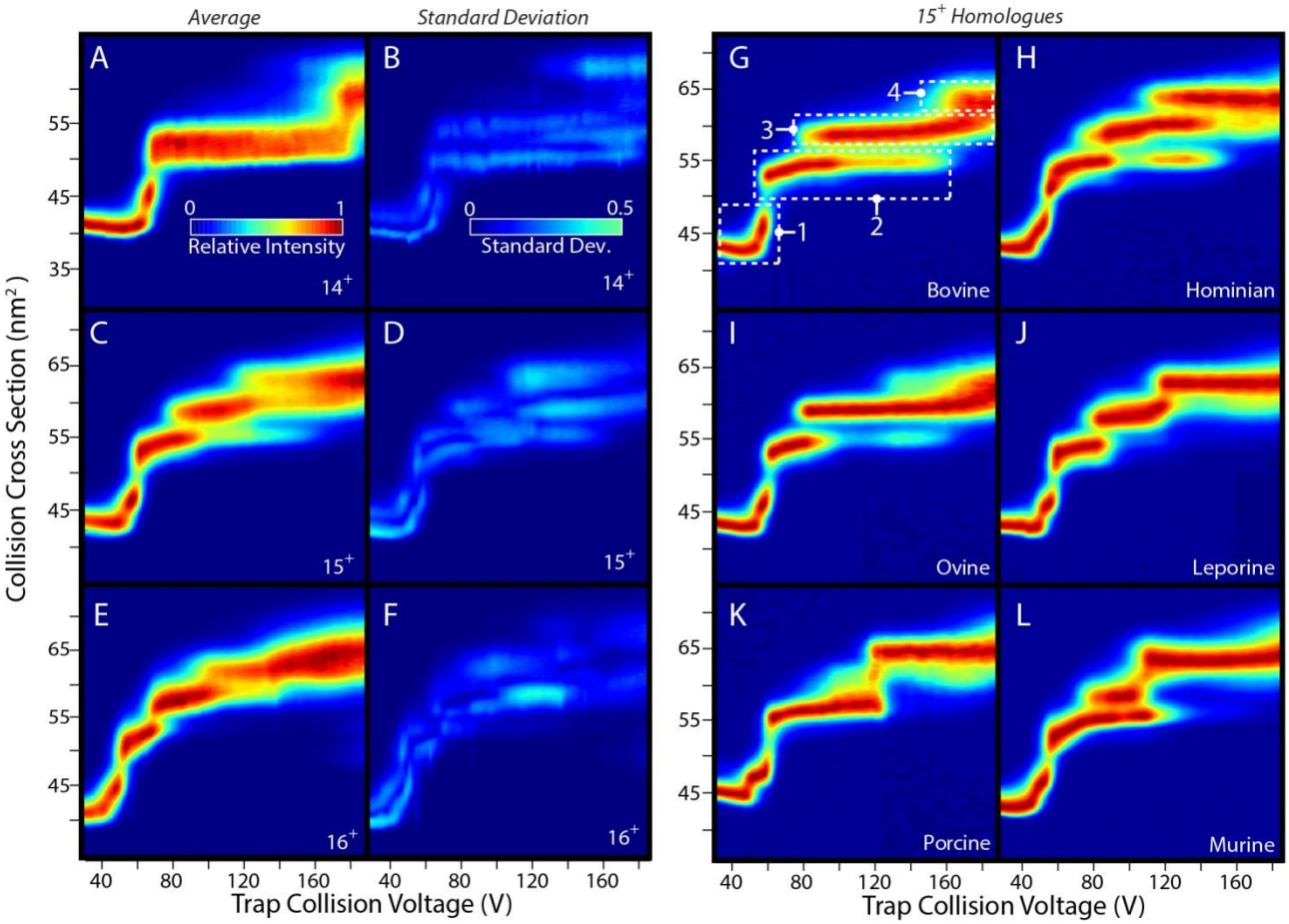


Figure 3-1. CIU screen of homologous serum albumins. Average and standard deviation CIU fingerprints of 7 homologous serum albumins at charge states 14^+ (A, B), 15^+ (C, D) and 16^+ (E, F). Examples of albumin CIU fingerprints from various species acquired for 15^+ ions (G-L as indicated on the figure). Four main conformer families (1-4, highlighted in G) are detected throughout.

standard deviation values between replicates of the same protein are generally at least 5 times lower than those reported here across homologues. Due to this potential richness of information, we choose to focus on this charge state for further analysis and discussion. The reader is directed to Figure *I*-1 for CIU fingerprints for individual 14^+ and 16^+ homologues.

Analysis of the resulting fingerprints for 15^+ serum albumins (Figures 3-1G-L, *I*-2) reveals clear similarities. For example, all albumins appear to undergo the same structural transitions upon collisional activation, resulting in a total of N+1 conformer families, where N is the number of domains in the native structure (labelled 1-4 in Figure

3-1G). This behavior is predicted from our previous work, which describes a method for predicting optimal charge-states for CIU analysis.[31] It should be noted that transitions and conformer families are assigned based on their stabilities, signal intensities, and resolution in CCS/energy space (See *Appendix I* for more details on CIU feature assignments).

Further analysis using the feature extraction and characterization functions of CIUSuite revealed additional information regarding the gas-phase unfolding pathways of homologues (Tables *I*-3-6, Figures *I*-3 and *I*-4). We characterized each CIU conformer family in terms of its centroid drift time, the range of voltages over which the conformer family is stable, and the centroid collision voltage value for the conformer family in order to determine the main factors that drive the absolute deviations observed in CIU fingerprints. Our results indicate that nearly all the albumin homologues tested access virtually identical unfolded conformer families, as defined by their centroid collision voltage and drift time values, when subjected to CIU. Critically, however, these same conformer families differ substantially in terms of their stability values. Based on this data, we draw two major conclusions: 1) The stabilities of CIU features are sensitive to small changes in protein primary structure, and 2) The number of CIU features observed, centroid IM drift times, and activation voltage values are conserved across different protein homologues and are instead linked to native protein domain structure.

These results, therefore, indicate potential future applications for CIU in the context structure predictions for large proteins of unknown folds based on CIU data, as well as high-throughput local stability measurements of domains within larger protein or multiprotein constructs. Furthermore, we anticipate future CIU-based separations of iso-

mass, iso-CCS proteoforms for the purposes of protein identification and quantitation (Figure *I*-5) based on the principles outlined in Figure 3-1.

Domain-Specific Chemical Probes for the Structural Interpretation of CIU Fingerprints.

Our data linking the conformational families accessed by CIU and native protein domain structure motivated us to develop a mechanistic understanding of serum albumin unfolding in the gas phase. For this series of experiments we chose to focus on the human variant of serum albumin (HSA), as it is arguably the most well studied and is supported by a large amount of crystallographic data associated with ligand binding.[35] Our approach involves correlating ligand dissociation energies with CIU features, where the binding location of the ligand is known from robust experimental data. First, we chose hemin as a marker for domain 1(D1), as many datasets, including an x-ray structure, indicate a highly-specific hemin binding site in this region for the HSA sequence.[35,36] Results from 15^{+} ions (Figure 3-2) indicated that the hemin binding pocket on D1 is preserved through at least the first two CIU structural transitions we observe, and the ligand is finally dissociated from conformer family 3 as it begins to transition to family 4. Surprisingly, the third albumin domain, D3, positioned on the opposite end of the polypeptide chain from D1, shows similar results, indicating the preservation of the D3 drug binding site for both ANS[31] and Diazepam[32,35] through the first two albumin conformer families observed. In contrast, ligands such as Indomethacin and Warfarin[1] that bind to a site on D2, are rapidly dissociated during the initial HSA CIU events observed at low collision voltages, despite similar overall binding affinities to the D1 and D3 binders described above. Of the ligands presented in Figure

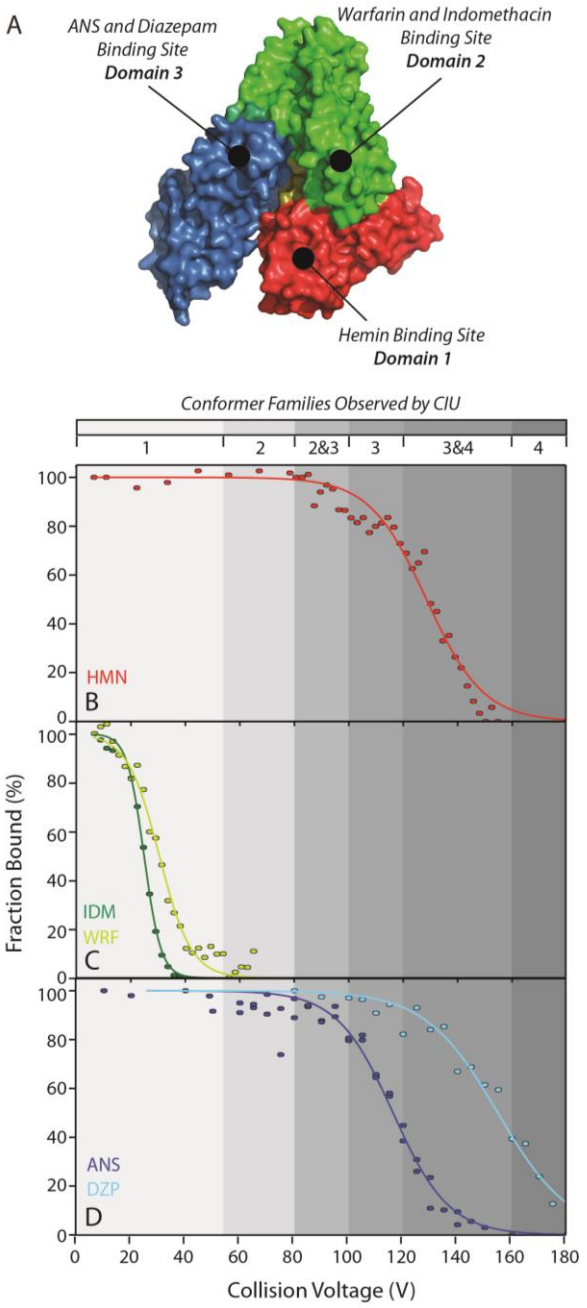


Figure 3-2. HSA Domain-Specific Chemical Probes of CIU. A) Surface representation of HSA based on PDB ID: 4K2C with ligand binding sites indicated. CID breakdown curves for HSA-ligand complexes, fitted to sigmoid functions. Datasets are shown for binders associated with domain 1 (B), 2 (C), and 3 (D). See text for abbreviation definitions. Overlaid on all plots is a color scale indicating the voltage ranges where different CIU conformer families are observed (see legend, top).

3-2, we found no significant differences in the CIU fingerprints of ligand bound and apo species, with the exception of diazepam complexes, for which we observe large CID thresholds and stabilization of conformer family 3 (Figure *I*-6). To validate these results, we examined ligand dissociation behavior of hemin, warfarin, and diazepam from the 14^+ and 16^+ HSA ions. (Figure *I*-7) These datasets reinforced our hypotheses, as nearly identical behavior to the 15^+ ions described above were observed for 16^+ ions, and only minor deviations were detected in the 14^+ case. Additionally, we examined the behavior of two larger molecules, Iodipamide[35] and Thyroxine[37] (Figure *I*-8) that are known to bind in two and four locations within the HSA structure respectively. Data acquired for these ligands also support the hypothesis that the D2 binding site is affected early in collisional activation, however structural interpretation of the

CID data from these ligands proved to be difficult as it was clear that both displayed evidence of significant cooperative stabilization. (Figure *I*-9).

In addition to the above-noted correlations between CID and CIU data, we interpret our HSA chemical probe data in the context of previous experiments correlating protein-ligand interactions with CID energies.[38-41] These early studies employing native ESI and CID demonstrated a strong correlation between the polar surface binding area of a ligand and the corresponding threshold collision energy required for ligand CID, although no similar correlation could be found for nonpolar binders.(Figure *I*-10) For each ligand in our study, we calculated the total polar surface area[42] as well as the total number of polar contacts for each albumin-ligand complex from available X-ray datasets.[43] We observe no correlation between our observed CID threshold energies for HSA-ligand complexes and any description of the native contacts formed between ligands and their respective protein binding pockets (Figure *I*-10). This result, taken in context with the CIU/CID correlations observed in Figure 3-2, strongly indicates that that the collisional ejection of a ligand from a multi-domain protein system, such as HSA, is most strongly correlated with the structural cohesion of its resident domain, rather than the number of local contacts developed within a protein-ligand binding site.

Noncovalent Albumin Constructs Further Reveal the CIU Mechanism of HSA.

To build on our understanding of the unfolding of HSA we designed a series of CIU/CID experiments utilizing HSA constructs built as noncovalent complexes comprised of individual HSA domains (Figures *I*-11- *I*-19). First, we incubated covalently attached albumin domains 1 and 2 (D12) with D3 to generate a noncovalent dimer that mimics full length HSA, (D12-D3). These results were compared with nonspecific dimers

where D3 was replaced with either D1 or D2, creating D12-D1 and D12-D2 respectively. Surprisingly, we were able to generate stable, noncovalent albumin mimics that had ground state drift time values nearly identical to their covalently-bound, native counterparts. CID analysis of these complexes (Figure 3-3 A) revealed that the two mismatched dimers showed nearly identical CID behavior, both possessing increased stability relative to D12-D3. Specifically, D12-D1 and D12-D2 begin to dissociate at around 80V, similar to D12-D3, however neither achieves the expected sigmoidal trend between intact dimer intensity and collision voltage, and instead both proceed to dissociate along an apparently frustrated, near-linear trend-line. A comparison of the CIU fingerprints for these mis-matched dimers (Figure *I*-20) reveals that D12-D1 and D12-D2 are unable to access CIU conformer family 3, which we find is necessary for the efficient dissociation of D3

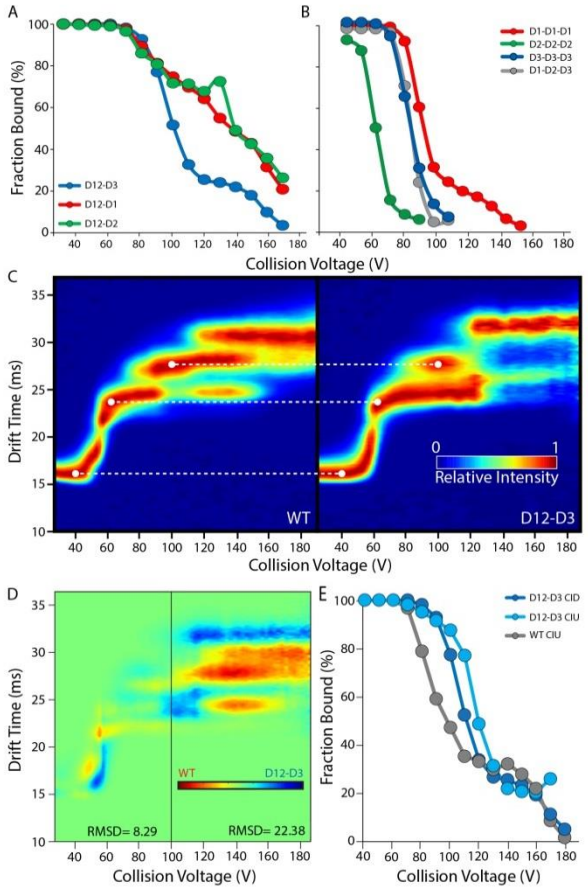


Figure 3-3. CIU/CID analysis of 15^{+} noncovalent, reconstituted albumins. A) CID breakdown curves representing the dissociation of a noncovalently-bound D3 domain from the covalent D12 fusion protein. B) CID breakdown curves representing the dissociation of a noncovalent subunit from a noncovalent homo-trimer of albumin domains. C) CIU comparison of WT HSA with the noncovalent D12-D3 construct. Dashed lines indicate strong correlation between the first three CIU features observed. D) CIU Difference plot between WT and D12-D3 HSA. RMSD values are calculated for before and after the transition to non-sigmoidal CID behavior (black line). E) Correlation between the structural transition from CIU conformer 2 to 3 and the CID behavior of D12-D3

from the D12-D3 dimer. Next, we used individual HSA domains to construct both all the possible homotrimers and a heterotrimer of native-like composition for CIU and CID analysis. CID data shown in Figure 3-3B illustrates that homotrimers constructed entirely from D2 are least stable, and those comprised of D1 are most stable, with D3 and D1-D2-D3 trimers presenting intermediate, and nearly identical, stabilities (Figure 3-3E).

Figure 3-3C compares the CIU fingerprint for native-like HSA (left) with the D12-D3 construct (right) discussed above. We find a striking correlation in the positions of CIU features and transitions for these two constructs at collision voltages lower than the CID threshold for D3 ejection from the D12-D3 dimer. In order to quantify the similarity of the CIU data shown in figure 3-3C, we computed two RMSD values the datasets: one for all CIU data collected at collision voltages lesser than 100 V (at which CID has depleted D12-D3 by 50%), and another for all CIU data collected above that value (Figure 3-3D). Aside from some additional stability imparted in the noncovalent complex, a difference analysis shows that these fingerprints are nearly identical at lower collision voltages, as evidenced by a relatively low CIU RMSD value of 8.29. In contrast, the RMSD value computed for CIU data acquired above 100 V is 22.28, strongly indicating that D12-D3 can neither efficiently access conformer family 3, nor any of conformer family 4. Instead, D12-D3 appears to access a new final unfolded state for CIU/CID above 100 V. Surprisingly, CIU fingerprints of both D1 and D3 homotrimers, as well as for the D1-D2-D3 heterotrimer, showed similar levels of correlation with native HSA CIU prior to their respective CID thresholds (Figure *I*-21). In Figure 3-3E, we compare the transition from conformer family 2 to 3 in both native and D12-D3 albumins with the CID breakdown

curve that tracks D3 ejection from the D12-D3 dimer. The same non-sigmoidal trends are observed in all three datasets, indicating a mechanistic connection between the appearance of CIU conformer family 3 in both WT HSA and D12-D3, as well as the ejection of D3 from D12-D3. We observe an average charge state for dissociated D3 of ~8^+, which amounts to 53% of the parent ion charge, thus strongly indicating D3 is unfolded prior to ejection from the D12-D3 complex.[44]

Mechanism of Gas-Phase Albumin Unfolding.

Taken together, our data allows us to generate the first detailed unfolding model for a solvent-free multi-domain protein. Our chemical probe studies indicate that ligands bound to D2 generally dissociate iso-energetically with the transition from conformer family 1 to 2, leading us to assign CIU conformer family 2 to a D2 unfolding event. Supporting this assignment is our CID/CIU data for D2 non-covalent trimer ions, indicating that D2 forms the least stable

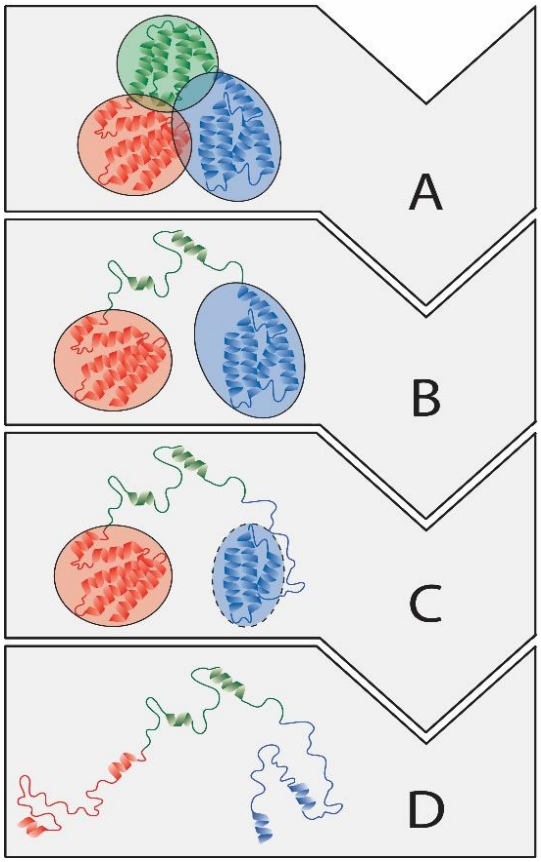


Figure 3-4. A Modular Unfolding Mechanism for 15^+ Human Serum Albumin. Proposed structural transitions that agree with experimental evidence are depicted in a cartoon. Domains are indicated in Red for D1, Green for D2, and Blue for D3. (A) Albumin compacts upon entry to the gas-phase, and all three domains are in a native like conformation (represented by circles). (B) As ion energy is increased, D2 undergoes unfolding, leaving the other two domains in a relatively native-like state. (C) Partial unfolding of D3 (indicated by the dashed ellipse) is achieved only at higher ion energies. (D) At the highest ion energies access in our experiment, all native-like protein structure is lost, including D1.

homotrimer out of all those studied here. Next, our CID data for D12-D3 inform our assignment of CIU conformer family 3 as related to unfolding of D3. Interestingly, our ligand binding studies indicate that D3-bound ligands can survive activation past this transition despite relying upon a relatively low number of polar contacts to remain within the D3 binding site. Taking this into account, we assign this transition to the partial unfolding of D3 in a manner that leaves its diazepam binding site intact. Finally based on both the large stabilities of D1-based constructs and D1/D3 chemical probe data, we assign CIU conformer family 4 to a coupled unfolding event involving both D1 and the remainder of D3.

By combining WT HSA CIU data with CID data from D12-D3, we can also infer a role for charge migration in the unfolding of large multi-domain proteins. The dissociation products of D12-D3 are those expected from multi-protein CID: highly charged, unfolded D3, and charge stripped, compact D12.[44] Although such results have been observed in CID data for many multi-protein complexes, the finding takes on new meaning in the context of understanding the CIU of a single protein chain. Considering the remarkable similarities between the measured unfolding data for HSA and D12-D3, we argue that the CIU of WT HSA must involve asymmetric charge migration during the first unfolding steps, and that the charge is likely redistributed evenly across newly-revealed protein surfaces as unfolding continues.

3.5 Conclusions

This work presents the most thorough investigation to date of the gas-phase unfolding of a multidomain protein. While the scope of our study is limited to albumins, a

careful analysis of our data indicates trends that may be generalizable across many classes of proteins and protein complexes. Our results indicate the sensitivity of CIU experiments to subtle changes in primary structure, where the tertiary structure remains essentially unchanged. Stated more specifically, we show that the conformational intermediates accessed during unfolding are dictated entirely by the tertiary structure of the protein, whereas the stability of those intermediates is determined by the underlying primary structure.

In addition to the insights above, our dataset reveals previously unknown correlations between gas-phase protein dissociation and unfolding. For example, our data show a clear correlation between the CIU of individual domains within a protein and the threshold voltage associated with CID-based ligand ejection. Interestingly, we did not find any correlation between our CID data and ligand-protein polar contacts/surface areas, in contrast to previous literature for smaller, single-domain protein systems. Additionally, a comparison of our CIU and CID data from noncovalent models of HSA strongly indicates that surface charges are re-distributed during the CIU of multi-domain proteins, similar to the mechanism proposed to describe multi-protein CID.[44]

A comparison of the mechanism shown in figure 3-4 with previously reported solution-phase measurements [32,35-37] indicates key similarities between solvent-free and solvated albumin unfolding. Guanidine hydrochloride-based denaturation of albumin bound to many of the probes used in our studies has identified the unfolding of D3 as a relatively early participant in the overall albumin unfolding process, and D1 has the most stable albumin domain. Additionally, these studies describe a modular, domain-centric unfolding pathway for albumin in solution, in agreement with our gas-phase studies. In

contrast, solution-phase studies do no capture large conformational shifts, like those that we attribute to D2 unfolding. While it is not surprising that gas-phase CIU does not precisely mimic protein denaturation in solution, the level of correlation that we observe projects a tantalizing future for CIU measurements in understanding the fundamental forces that drive native protein stability and predicting domain organization.

Although the experimental data presented in this report provides our most detailed insights into gas-phase protein unfolding to date, the resolution of our model is certainly limited. In order to improve our understanding of both CIU and protein structure in general, it is clear that improvements in both gas-phase molecular dynamics simulations and experimental IM-MS techniques will be required. For example, tandem IM technologies,[45] capable of both assessing the direct connectivity between CIU conformer families and revealing the fine structure within such families, will undoubtedly prove useful in achieving CIU models of greater detail than shown here. In addition, the synergy between CIU datasets and recent advancements that combine charge migration algorithms with atomistic molecular dynamics simulations of large proteins[46] is clear. Future efforts in both theory and experiment will likely build on the insights discussed here, along with accumulated CCS data from these studies (Tables S7 and S8), in order to move our understanding of solvent-free protein folding forward, acquire structural data for refractory regions of the proteome, and access canonically challenging targets for the discovery of next-generation therapeutics.

3.6 ACKNOWLEDGMENTS

The authors would like to thank the National Science Foundation (CAREER, 1253384) and the University of Michigan department of Chemistry for their support. JDE also acknowledges support from the University of Michigan Rackham Graduate School in the form of a Research Award. Additionally, we thank Prof. Robert Kennedy for his help with preparing diazepam treated albumin samples for CIU analysis.

3.7 SUPPORTING INFORMATION

Supporting Information can be found in *Appendix I*

3.8 REFERENCES

(1) Metallo, S. J. Curr. Opin. Chem. Biol. 2010, 14, 481.
(2) Dill, K. A.; MacCallum, J. L. Science 2012, 338, 1042.
(3) Yang, Z. BMC Bioinform. 2008, 9, 40.
(4) Berman, H. M.; Kleywegt, G. J.; Nakamura, H.; Markley, J. L. Structure 2012, 20, 391.
(5) Ward, A. B.; Sali, A.; Wilson, I. A. Science 2013, 339, 913.
(6) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Science 1989, 246, 64.
(7) Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T. Rapid Commun. Mass Spectrom. 1988, 2, 151.
(8) Karas, M.; Hillenkamp, F. Anal. Chem. 1988, 60, 2299.
(9) Benesch, J. L. P.; Ruotolo, B. T.; Simmons, D. A.; Robinson, C. V. Chem. Rev. 2007, 107, 3544.
(10) Xie, Y. M.; Zhang, J.; Yin, S.; Loo, J.A. J. Am. Chem. Soc. 2006, 128, 14432.
(11) Liu, L.; Bagal, D.; Kitova, E. N.; Schnier, P. D.; Klassen, J. S. J. Am. Chem. Soc. 2009, 131, 15980.
(12) Wyttenbach, T.; Bowers, M. T. J. Phys. Chem. B 2011, 115, 12266.
(13) Pierson, N. A.; Chen, L.; Valentine, S. J.; Russell, D. H.; Clemmer, D. E. J. Am. Chem. Soc. 2011, 133, 13810.
(14) Wood, T. D.; Chorush, R. A.; Wampler, F. M.; Little, D. P.; O'Conner, P. B.; McLafferty, F. W. Proc. Natl. Acad. Sci. U. S. A. 1995, 92, 2451.
(15) Valentine, S. J.; Clemmer, D. E. J. Am. Chem. Soc. 1997, 119, 3558.
(16) Oomens, J,; Polfer, N.; Moore, D. T.; van der Meer, L.; Marshall, A. G.; Eyler, J. R.; Meijer, G.; von Helden, G. Phys. Chem. Chem. Phys. 2005, 7, 1345.
(17) Wyttenbach, T.; von Helden, G.; Bowers M. T. J. Am. Chem. Soc. 1996, 118, 8355.
(18) Wolynes, P. G. Proc. Natl. Acad. Sci. U. S. A. 1995, 92, 2426.
(19) Ruotolo, B.T.; Robinson, C. V. Curr. Opin. Chem. Biol. 2006, 10, 402.
(20) Niu, S.; Rabuck, J. N.; Ruotolo, B. T. Curr. Opin. Chem. Biol. 2013, 17, 809.
(21) Mayer, P. M.; Martineau, E. Phys. Chem. Chem. Phys. 2011, 13, 5178-5186.
(22) Shelimov, K. B.; Clemmer, D. E.; Hudgins, R. R.; Jarrold, M. F. J. Am. Chem. Soc. 1997, 119, 2240.
(23) Ruotolo, B. T.; Hyung, S.-J.; Robinson, P. M.; Giles, K.; Bateman, R. H.; Robinson, C. V. Angew. Chemie Int. Ed. 2007, 46, 8001.
(24) Hopper, J. T. S.; Oldham, N. J. J. Am. Soc. Mass Spectrom. 2009, 20, 1851.
(25) Hyung, S.-J.; Robinson, C. V.; Ruotolo, B. T. Chem. Biol. 2009, 16, 382.
(26) Han, L.; Hyung, S.-J.; Mayers, J. J. S.; Ruotolo, B. T. J. Am. Chem. Soc. 2011, 133, 11358.

(27) Zhou, M.; Dagan, S.; Wysocki, V. H. Analyst 2013, 138, 1353.
(28) Rabuck, J. N.; Hyung, S.-J.; Ko, K. S.; Fox, C. C.; Soellner, M. B.; Ruotolo, B. T. Anal. Chem. 2013, 85, 6995.
(29) Niu, S.; Ruotolo, B. T. Protein Sci. 2015, 24, 1272.
(30) Tian, Y.; Han, L.; Buckner, A.C.; Ruotolo, B. T. Anal. Chem. 2015, 87, 11509.
(31) Zhong, Y.; Han, L.; Ruotolo, B. T. Angew. Chemie Int. Ed. 2014, 126, 9363.
(32) Ahmad, B.; Ahmed, M. Z.; Haq, S. K.; Khan, R. H. Biochim. Biophys. Acta, Proteins Proteomics 2005, 1750, 93.
(33) Ruotolo, B. T.; Benesch, J. L.; Sandercock, A. M.; Hyung, S.-J.; Robinson, C. V. Nat. Protoc. 2008, 3, 1139.
(34) Eschweiler, J. D.; Rabuck-Gibbons, J. N.; Tian, Y.; Ruotolo, B. T. Anal. Chem. 2015, 87, 11516.
(35) Ghuman, J.; Zunszain, P. A.; Petitpas, I.; Bhattacharya, A. A.; Otagiri, M.; Curry, S. J. Mol. Biol. 2005, 353, 38.
(36) Zunszain, P. A.; Ghuman, J.; Komatsu, T.; Tsuchida, E.; Curry, S. BMC Struct. Biol. 2003, 3, 6.
(37) Petitpas, I.; Petersen, C. E.; Ha, C.-E.; Bhattacharya, A. A.; Zunszain, P. A.; Ghuman, J.; Bhagavan, N. V.; Curry, S. Proc. Natl. Acad. Sci. U. S. A. 2003, 100, 6440.
(38) Wu, Q.; Gao, J.; Joseph-McCarthy, D.; Sigal, G. B.; Bruce, J. E.; Whitesides, G. M.; Smith, R. D. J. Am. Chem. Soc. 1997, 119, 1157.
(39) Rogniaux, H.; Van Dorsselaer, A.; Barth, P.; Biellmann, J. F.; Barbanton, J.; van Zandt, M.; Chevrier, B.; Howard, E.; Mitschler, A.; Potier, N.; Urzhumtseva, L.; Moras, D.; Podjarny, A. J. Am. Soc. Mass Spectrom. 1999, 10, 635.
(40) El-Kabbani, O.; Rogniaux, H.; Barth, P.; Chung, R. P. T.; Fletcher, E. V.; Van Dorsselaer, A.; Podjarny, A. Proteins: Struct., Funct., Bioinf. 2000, 41, 407.
(41) Stojko, J.; Fieulaine, S.; Petiot-Becard, S.; Van Dorsselaer, A.; Meinnel, T.; Giglione, C.; Cianferani, S. Analyst 2015, 140, 7234.
(42) Ertl, P.; Rohde, B.; Selzer, P. J. Med. Chem. 2000, 43, 3714.
(43) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. Protein Eng. 1995, 8, 127.
(44) Hyung, S.-J.; Ruotolo, B. T. Proteomics 2012, 12, 1547.
(45) Koeniger, S. L.; Merenbloom, S. I.; Valentine, S. J.; Jarrold, M. F.; Udsth, H. R.; Smith, R. D.; Clemmer, D. E. Anal. Chem. 2006, 78, 4161.
(46) McAllister, R. G.; Metwally, H.; Sun, Y.; Konermann, L. J. Am. Chem. Soc. 2015, 137, 12667.

Chapter 4: Coming to Grips with Ambiguity: Ion Mobility-Mass Spectrometry for Protein Quaternary Structure Assignment

Joseph D. Eschweiler, Aaron T. Frank, and Brandon T. Ruotolo

Supplemental Information can be found in *Appendix II*

4.1 Abstract

Multiprotein complexes are central to our understanding of cellular biology, as they play critical roles in nearly every biological process. Despite many impressive advances in structural characterization techniques, large and highly-dynamic protein complexes are too often refractory to analysis by conventional, high-resolution approaches. To fill this gap, ion mobility-mass spectrometry (IM-MS) methods have emerged as a promising approach for characterizing challenging structural targets due in large part to the ability of these methods to characterize the composition, connectivity, and topology of large, labile complexes. In this Critical Insight, we present a series of bioinformatics studies aimed at assessing the information content of IM-MS datasets for building models of multiprotein structure. Our computational data highlights the limits of current coarse-graining approaches, and compelled us to develop an improved workflow for multiprotein topology modeling, which we benchmark against a subset of the multiprotein complexes within the PDB. This improved workflow has

allowed us to ascertain both the minimal experimental restraint sets required for generation of high-confidence multiprotein topologies, and quantify the ambiguity in models where insufficient IM-MS information is available. We conclude by projecting the future of IM-MS in the context of protein quaternary structure assignment, where we predict that a more complete knowledge of the ultimate information content and ambiguity within such models will undoubtedly lead to applications for a broader array of challenging structural targets.

4.2 Introduction

Structural characterization of the multicomponent complexes that form the functional units of the "interactome", specifically protein complexes, represents a major challenge for structural biology.[1,2] Due to their large size, low copy numbers, and intrinsic heterogeneity and lability, important targets are too often refractory to analysis by traditional techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, or electron microscopy, despite impressive advances in these fields.[3,4] Alternative approaches for characterizing difficult multicomponent structures may result in low-resolution or sparse datasets, such as those generated from small-angle scattering[5] or covalent labeling/crosslinking methodologies.[6] Circumventing the limitations of a single technique, integration of datasets from multiple experiments has been shown to be a potent approach for characterizing multiprotein complexes,[7] as often times these datasets provide complementary information. This family of methods, commonly referred to as integrative structural biology, have progressed rapidly due largely to advances in computational techniques that have made it possible to encode

different types of experimental datasets as spatial restraints in a single modeling workflow.[8]

Generally, an integrative modeling workflow is an iterative process described by four major steps: 1) the gathering of experimental data, 2) the translation of such data into spatial restraints, 3) the generation of an ensemble of putative structures that satisfy the experimentally-defined restraints, and 4) the characterization of the ensembles generated, where ambiguities are identified and used to refine structural hypotheses. This process may then be iterated as necessary in order to resolve ambiguities to the extent allowed by the experimental restraints utilized.[8] MS-based methods such as chemical crosslinking,[9-11] native-MS,[12,13] and ion mobility-MS[14] have garnered much attention as valuable experiments within such integrative structural biology frameworks. Of these MS-based technologies, ion-mobility-mass spectrometry (IM-MS) is uniquely positioned for interrogating multiprotein structure.[15] Unlike solution-phase measurements which may report on the average of an ensemble of proteoforms, conformers, or oligomerization states, IM-MS datasets can be used to discern the relative proportions of these species within mixtures, and interrogate their composition, connectivity, and collision cross sections individually.[16] Due to its unique capabilities in protein structure analysis, IM-MS is often deployed to determine coarse-grained (CG) protein topology models for assemblies that have resisted previous characterization attempts, often in combination with other forms of biophysical data.[17,18]

Figure 4-1A illustrates the potential information content often derived from native MS datasets. While direct analysis of the masses of intact complexes can often provide

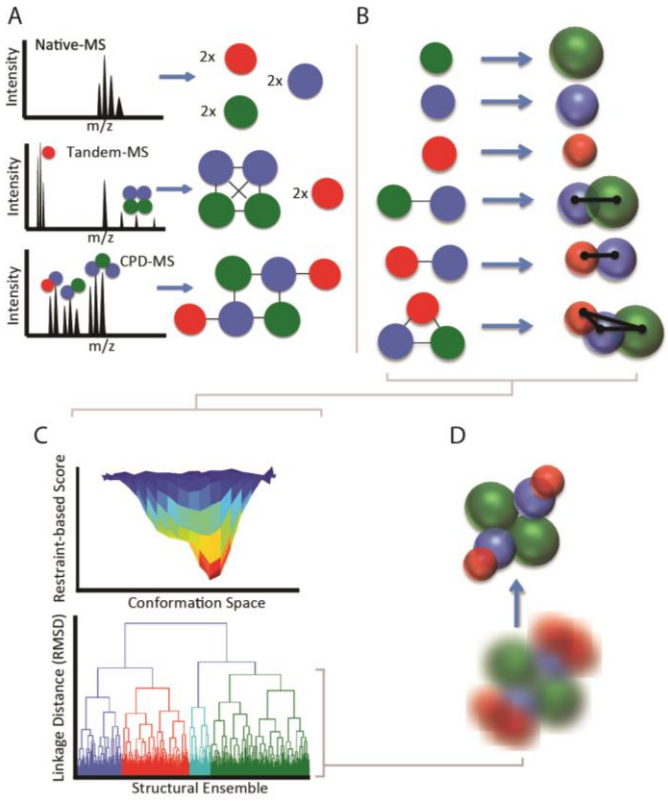


Figure 4- 1 A general workflow for IM-MS-based modeling. A) native-MS, tandem-MS, and solution-phase disruption-MS yield increasing amounts of composition and connectivity information for a multiprotein complex. This information can be encoded with varying levels of ambiguity based on the information available. B) IM-MS data can be included to build a 3D topology mode. Individual subunits or domains can be encoded as spheres with radius derived from their measured CCS, while exact distances between subunits can be derived from CCS measurements of dimeric and trimeric species. C) Optimization of the experimentally-defined scoring function using a Monte Carlo method provides unbiased sampling of potential structures for high-stoichiometry complexes. These structures form an ensemble which is subjected to clustering analysis to mine for predominant structural families D) Structural families detected by clustering can be characterized in aggregate using kernel density functions, mean structures and standard deviations, or individual structures can be identified as representative of the family.

unambiguous information about the protein composition and stoichiometry,[19] it is also useful to interrogate solution or gas-phase disassembly products to further elucidate connectivity and structural modularity. To this end, methods for solution[20,21] and gas-phase[22-24] disruption of multiprotein complexes are actively being developed to increase the number of observable sub-complexes, and therefore the overall information content of the experiment. In addition to the composition and connectivity information garnered by MS, IM-MS (Figure 4-1B) provides 3D structural information on both monomeric and oligomeric protein ions in the form of collision cross sections (CCSs).[25,26] Since multiple methods are available for the accurate calculation of CCS values from *in silico* models,[27-29] it is possible to assign putative structures to the signals observed in the IM-MS experiment.

Despite being used to restrain rigorous dynamics experiments for peptides [30] and small proteins [31] for decades, our ability to extract structural information from CCS measurements of large proteins and multiprotein complexes is still evolving. A recent comprehensive analysis of the PDB revealed that the general amount of CCS variance exhibited by proteins increases for high mass and stoichiometry protein complexes, indicating increased information content from CCS measurements in this regime.[28] These observations corroborate earlier experimental results showing that the oligomerization patterns of homomeric protein complexes can be discerned in many cases based on CCS trends as a function of complex stoichiometry.[32]

Methods for extracting topological information for large, heteromeric protein complexes are, however, less developed. Early procedures for optimizing pairwise and trimeric subunit interactions were based on a linear search for conformations, using spherical subunit representations that satisfied experimental CCS restraints.[33] Although the spherical representation of protein subunits possesses obvious limitations when modeling highly aspherical subunits such as multidomain proteins, spheres still represent the primary component in IM-MS based modeling due to their trivial geometric relationship to the CCS parameter, their ease of implementation in computational workflows, and their facile relationship to protein-protein interaction geometries. Subsequently described IM-MS workflows aimed at the generation of protein quaternary structure models (Figure 4-1C) utilized a Monte Carlo approach for sampling orientations of spheres that satisfied excluded volume, symmetry, connectivity, and CCS restraints in order to yield an ensemble of structures that can be interrogated via

hierarchical clustering methodologies.[33-35] Such IM-MS derived models have been favorably compared to structures produced using more mature workflows, indicating a promising level of cooperativity between CCS measurements and other biophysical parameters commonly used in protein complex model generation.[34] This general approach has been used to elucidate the topological features of the DNA replisome,[33,36] ribosomal initiation factor complexes,[37] and ATPases,[38] all providing critical structural insights as well as methodological enhancements. More recently, surface induced dissociation (SID) coupled to IM-MS and covalent labeling has been applied to build a complete model of the toyocamycin nitrile hydratase complex[39] by leveraging the sub-complexes produced both through controlled disruption in solution and SID.

Despite these promising examples, many questions remain about the ability to unambiguously assign protein topology using IM-MS datasets (Figure 4-1D). Most of these questions surround the potential errors introduced when high levels of coarse graining is applied, the interpretation of structural ensembles generated from IM-MS modeling approaches, and the confidence levels associated with IM-MS structures in a general sense.[40] Additionally, questions remain regarding the extent of structural rearrangement apparent in some proteins and complexes in the gas-phase; a topic that has been investigated in detail elsewhere.[35,41] In this Critical Insight, we seek to critically evaluate the information content of IM-MS for protein quaternary structure assignment in cases where we can assume a strong memory of solution-phase structure. Based on many of the challenges described above, we develop a new generalized algorithm for translating IM-MS datasets into structural models and benchmark our new method against many known topologies present in the PDB. We continue by quantifying, for the

first time, the ambiguity present in under-restrained models, and suggest approaches mitigating such effects. We conclude by projecting the future of IM-MS derived models of protein quaternary structure.

4.3 Assessing Coarse-Graining Errors in Multiprotein Models Generated from IM-MS Data

In workflows that utilize IM-MS data to restrain models of protein quaternary structure, it is typically assumed that the protein components of the assembly can be accurately represented by spheres defined by either their measured or estimated CCS. Although many reports have demonstrated a strong correlation between experimental CCS measurements and CCS values extracted from solution-phase protein models, the strength of this correlation can depend on the domain structure and globularity of the protein analyte in question.[42,43] Moreover, the magnitude and nature of the errors incorporated into IM-MS multiprotein models through the coarse-graining process are currently unknown. In order to investigate such coarse-graining errors, we extracted a non-redundant set of 191 high-resolution protein complex structures from the 3D complex set database,[44] and developed a method for the rapid generation of CG structures based on these entries where the extent of coarser graining can be treated as a variable. The first step in our protocol involves extracting coordinates and center-of-mass values for each subunit within the protein complex. Next, the CCS values are calculated for each subunit using the projection approximation function within the IMPACT library.[28] To generate the initial CG model at subunit resolution, we placed

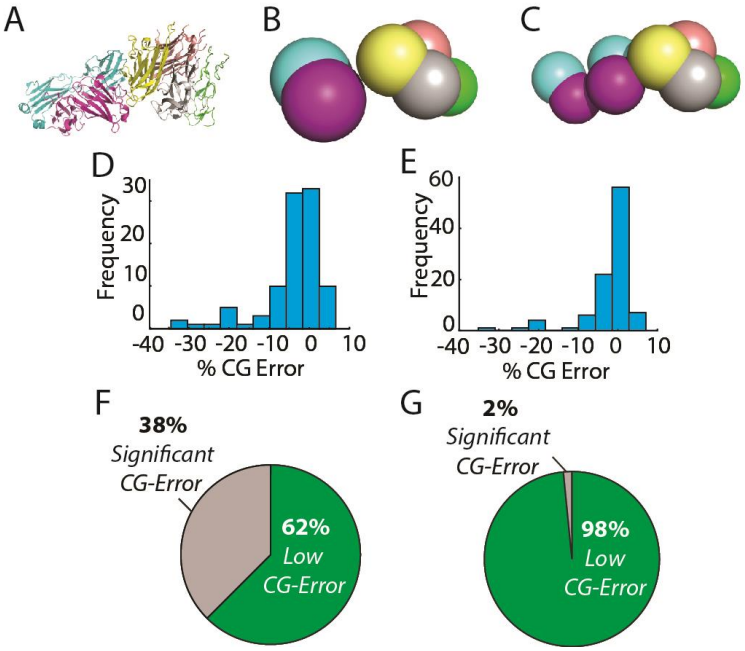


Figure 4- 2 Coarse-graining Error for domain and subunit-level representations. A) An example high resolution reference structure PDB ID 4MXW with subunits color coded. B.) A coarse-grained model of 4MXW at the subunit level. C) A coarse-grained model of 4MXW at the domain-level. D) 191 non-redundant protein topologies were coarse grained at the subunit-level. The coarse-graining error distribution for this level of coarse-graining is shown. E) Subunit-level coarse graining introduced significant CCS errors for 28% of the complexes in our set. F) The coarse-graining error distribution for the same set of protein topologies coarse-grained at the domain level. G) When coarse-grained at the domain level, only 2% of topologies had significant coarse-graining errors introduced.

spheres having radii corresponding to the projected area of the subunits at the center of mass for each subunit in the complex. To evaluate the model, the projected area of the high-resolution structure was compared to that calculated from the CG model.

Our results suggest that a significant number of the protein complexes currently available within the PDB contain subunits that are not accurately represented when subunit-level coarse-graining is applied. As shown in Figure 4-2A-C, subunit-level coarse-graining very often results in large deviations in CCS compared to the reference. We define CG error as the total percent of atoms found within the high-resolution structure that fit within an average of CG representations determined by our workflow (see *Appendix II* for Details). We used a 5% deviation in the CCS values obtained for CG models when compared to reference CCSs for the corresponding all-atom reference structure to define a ‘significant’ error threshold in our analysis, as such defects reflect, in our view, both the maximum error

that can be introduced into a model before losing significant topology information, as well as the maximum error value carried by experimental restraint information recovered for large protein complexes by IM-MS.[45] Specifically, over 28% of the protein complexes studied here contained significant errors (greater than 5%) when this level of coarse graining was applied. We also note that the error distribution associated with this level of coarse graining is highly asymmetric, containing many structures having CG errors greater than 10%.

A more detailed analysis of the structures within the survey reveals that proteins with multiple domains are most susceptible to high CG errors, especially those proteins having domains connected by long linker regions. Interestingly, however, we found no correlation between the CCS/mass ratio of individual subunits and their propensity to introduce error into the model, indicating that the overall packing density of the protein does not play a major role in the CG errors on display in Figures 4-2D and 4-2E. Based on this data we hypothesized that coarse-graining at the domain level should eliminate the majority of the errors we observed from our subunit-resolution CG modeling experiments. To investigate this, we implemented a k-means clustering method[46] in SciPy[47] to heuristically detect protein domain structure over a range of thresholds associated with protein and domain mass (See *Appendix II* for Details). The results associated with these higher-resolution CG structures are shown in Figures 4-2F and 2G, and reveal a strong relationship between the resolution of the CG structures and the propensity for CG error we record during our analysis. Figure 4-2G, for example, shows that the fraction of protein complexes with significant errors drops to ~2% when domain-level CG is applied to the same pool of structures analyzed in Figure 4-2E.

4.4 Benchmarking the Information Content of IM-MS Datasets for Modeling Known Protein Complexes

To generate ensembles of putative structures based on IM-MS-derived data, we developed a program for interpretation and optimization of diverse MS and/or IMS-derived restraint sets. This program, referred to as IMMS_modeler, was built using connectivity and distance restraints from the Integrated Modeling Platform (IMP)[8], some of which were implemented previously.[34] Novel aspects of our approach include: 1) the use of a restraint file for facile input of new data, 2) the ability to use new functional forms within the scoring function, and 3) a new Monte Carlo algorithm that enables a significantly broader sampling restraint space. By default, IMMS_modeler generates ensembles of 1000 structures that satisfy all of the declared restraints. We found this amount of structures to be a representative sample of structural space for most complexes, and have based the following experiments on these ensembles. All CCS calculations were performed offline using the projected area function in the IMPACT library (See *Appendix II* for Details).

In order to thoroughly evaluate our method as a general approach for modeling multiprotein complexes, we set out to benchmark IM-MS modeler against known protein complex topologies with varying levels of restraint information. In these experiments, we generated CG models at the resolution of individual protein subunits for a small subset of complex topologies used in the previous experiment. For simplicity, we focused this stage of our analysis only on those protein complexes that did not show significant CG error, as described in the above section (Figure 4-2). Despite these limitations in the

scope of our benchmarking, the geometric principles described here are transferrable to models created at higher levels of CG resolution.

On the Positive Predictive Power of IM-MS datasets

In order to characterize the information content associated with CCS measurements of intact protein complexes and sub-complexes when used to define inter-protein distances and geometry in the context of a search of potential quaternary structures (which we define as 'internal restraints'), we simulated IM-MS datasets for at least five non-redundant complex topologies for protein trimers, tetramers, pentamers, and hexamers (Figure 4-3). Although some of the complexes used to generate the analysis shown in Figure 4-3 contain symmetric elements, no symmetry restraints were implemented to avoid bias. All restraint sets contained detailed information regarding the connectivity of the complex, as well as the CCS of the intact assembly. In addition to this information, restraint sets contained varying numbers of the 'internal restraints' described above, which correspond to the pair-wise distance restraints that are commonly obtained from native IM-MS datasets.[21,32] We note that although 3D systems

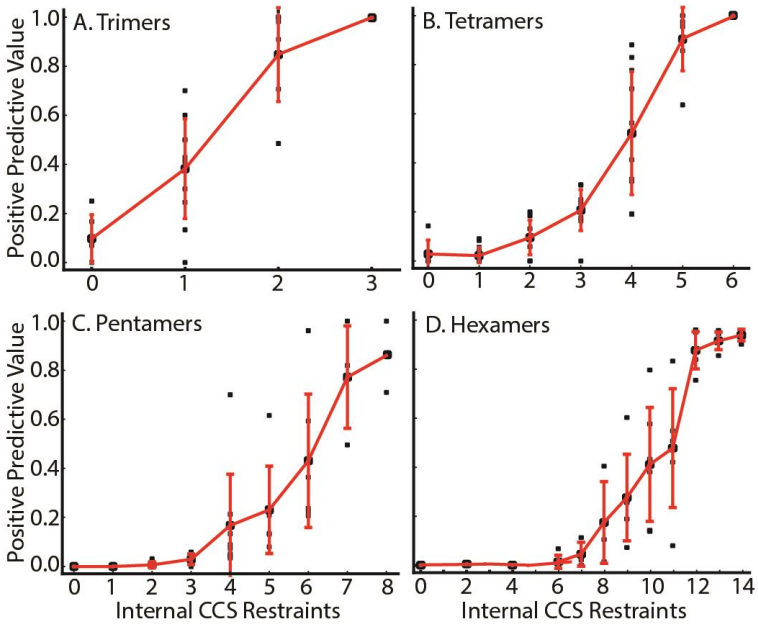


Figure 4- 3 Positive Predictive Values of the IM-MS restraint sets plotted as a function of the number of internal CCS-derived restraints. At least 5 non-redundant topologies from the PDB were considered for each number of subunits, A) Trimers B) Tetramers C) Pentamers and D) Hexamers. Each restraint set was manually curated to ensure the data reflected data that could be reasonably generated through existing IM-MS technologies.

are generally restrained by a minimum of 3N-6 restraints (where N is number of bodies), our restraint sets attempt to simulate restraints from real IM-MS datasets with built in errors, often causing producing predictive values less than predicted by the precise distance geometry. For purposes of this analysis, the structures generated using our method were defined as true positives (native-like topologies) if they had an RMSD values of less than 5Å relative to the reference structure; and we defined a positive predictive value (PPV) as the fraction of true positive structures within the ensemble of structures sampled using a given restraint set.

As expected, our results reveal a positive relationship between the number of internal CCS restraints available for a complex and the positive predictive value for a given modeling effort. For trimeric protein complexes (Figure 4-3A), the ensemble is enriched for true positives with the addition of internal distance restraints between subunits. Here, due to the trivial relationship between the CCS and the angle of subunits within the complex, the model should be fully restrained by the global CCS plus any two IM-derived distance restraints.[33] Notably, there is one outlier structure that seemingly refutes this general conclusion; however, our analysis also suggests that the CCS restraint becomes less sensitive when large disparities exist in the CCS of each component, allowing us to rationalize all of the results shown. (Figure *II*-2) Higher stoichiometry complexes (Figure 4-3B-D), exhibit similarly strong increases in PPV in a manner correlated with the number of internal restraints included. We note that the number of restraints necessary to reach a PPV > 0.8, where 80% of the structures identified in the ensemble are within 5Å of the 'true' structure, increases rapidly as the number of subunits increases, further motivating the need to develop new methods and

technologies for the comprehensive generation of native-like sub-complexes for IM-MS analysis.[20,23,24]

4.5 Characterizing Ambiguity in the Structural Ensembles Defined by IM-MS

Although the PPV is a valuable metric for comparing the information content of multiple restraint sets, interpretation of PPV values for individual datasets can be challenging. This is due to the fact that members of a structural ensemble generated by the IMMS-modeling approach described here are not randomly distributed, and in many cases can be clustered into distinct sub distributions, or structural families. Pairwise relationships between structures within an ensemble can be described by a pairwise RMSD matrix, which can in turn be interrogated using hierarchical clustering to determine groups of highly related structures, and the relationships between those groups. Alternatively, other similarity measures can be implemented to describe structural relationships between models, including the ultrafast similarity score,[48] or distance matrix RMSD,[49] which each may have their own advantages depending on the geometries present in the ensemble. For the computational data described in this Critical Insight, a detailed

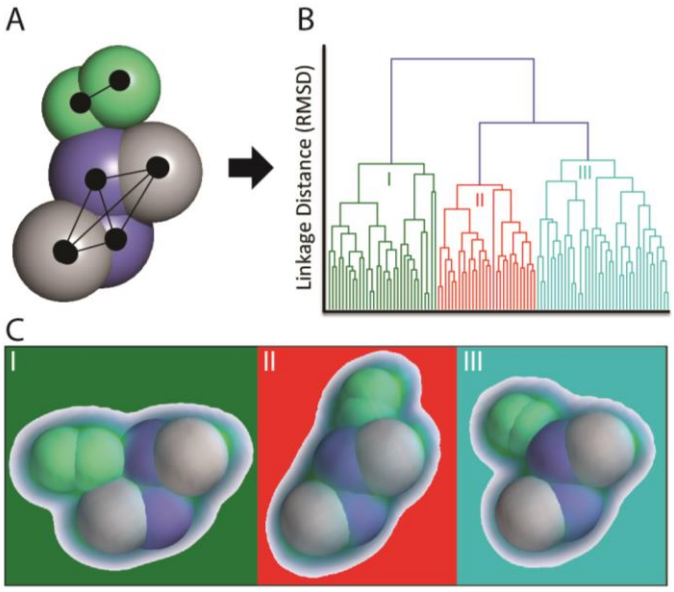


Figure 4- 4 Parsing Structural Ensembles Generated with Ambiguous Restraint Sets. A) A restraint set was generated for 2AFH, a nitrogenase heterotetramer (purple and grey) bound to the dimeric nucleotide switch protein (green). The binding location and pose for the nucleotide switch protein is not defined in the restraint set and a CCS-filtered structural ensemble contains many putative structures. B) Hierarchical clustering of the ensemble reveals three distinct structural families within the ensemble, greatly simplifying the analysis. C) Plotting the kernel density function of each structural family reveals high resolution within all families.

analysis of the structural ensemble produced from an IM-MS restrained search of protein topology space regularly reveals useful information, in addition to what is provided by the PPV value analysis shown in Figure 4-3 alone. In the sections below, we discuss the interpretation of hierarchical clustering da tasets in the context of such IM-MS restrained models, focusing on our recent efforts to define and quantify the ambiguity and resolution within the IM-MS data.

A hierarchical clustering dendrogram (as shown in Figure 4-4) illustrates the relationship between all structures within an ensemble. The number of clusters depends on the 'cut point' chosen during dendrogram analysis, a value that is typically a user defined parameter. For example, our algorithm automatically defaults to a dendrogram cut point that generates clusters at linkages that exhibit greater than 70% of the maximum RMSD in the entire matrix analyzed. Our ensemble analysis workflow evaluates the in-cluster RMSD as it compares to the average RMSD of the ensemble, as well as the cross-cluster RMSD, revealing distinct structural families that define the identified clusters (Figure4- 4). It is worth noting that the application of IM-MS restraints often leads to the type of model ambiguity shown in Figure4- 4 for large hetero-protein targets.[34] Indeed, such ambiguity may, in some cases, represent the native ensemble of protein complex structures associated with function.[38,50] Commonly, however, such uncertainty is due to incomplete structural information and can be resolved either through the application of additional restraints [18,51] (see below for examples).

As mentioned above, the in-cluster RMSD can be a valuable metric for quantitatively expressing the ambiguity within a cluster. However, when evaluating biomolecular structures, qualitative and visual expression of ambiguity is often more

facile to interpret. In order to fill this gap for IM-MS derived models, we developed a new method for visualizing the ambiguity within a structural family using kernel density functions.[47,52] In this method, the coordinates within a structural family or ensemble are aligned, and each subunit coordinate is uniformly populated with protein density as a sphere corresponding to its collision cross section. Next, the Gaussian kernel function is estimated for this volume of coordinates, and then visualized. For the workflow described here, we utilize the Mayavi Library[53] in Python to visualize the kernel densities. As illustrated in Figure 4-4C, this kernel density function approach allows for the visualization of structural ambiguity present within an ensemble; information that is likely vital for the detailed interpretation of structural ensembles defined by sparse sets of restraints.

4.6 Leveraging Symmetry and Modularity to Resolve Ambiguity within IM-MS Model Ensembles

To further evaluate IM-MS based quaternary structure assignments in a general sense, as well as the newly-developed methods described here, we chose two case studies that illustrate real-world examples of challenging modeling targets. As shown in Figure 4-3, the number of restraints needed to accurately recapitulate the topology of a multiprotein complex with greater than 5 subunits increases exponentially, creating challenges for integrative modeling of these complexes. However, in the data shown below, we demonstrate that by leveraging modularity and symmetry within high-stoichiometry complexes, it is possible to circumvent these limitations.

As an example of a symmetry restraint applied in order to resolve ambiguity within an IM-MS restrained ensemble of protein quaternary structures, we built models

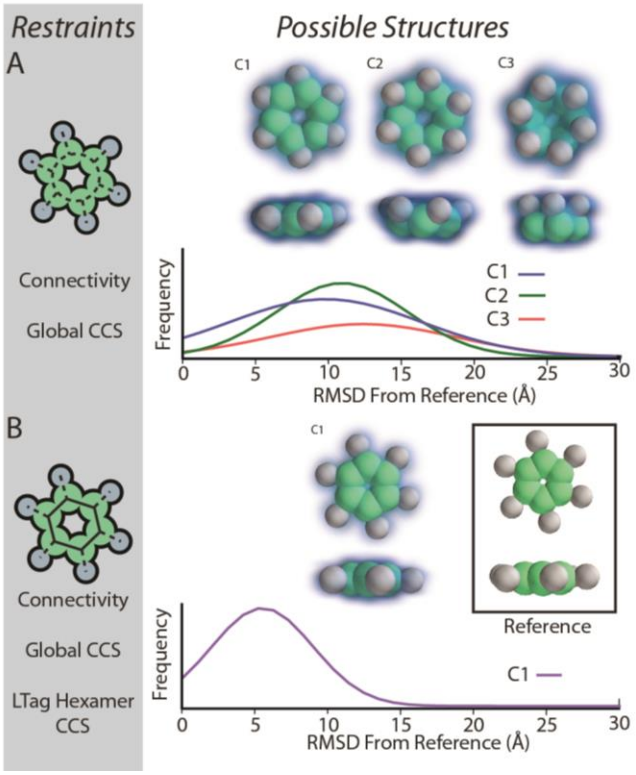


Figure 4- 5 Modeling the topology of hexameric LTag bound to p53 using the symmetry restraint. Two restraint sets (left panels, A and B) were used to generate structural ensembles that were evaluated using hierarchical clustering, kernel density functions, and RMSD distributions. (Right panels).

of the Large T-antigen (LTag) complex bound to p53. LTag is a hexameric ring structure that binds p53 monomers in a stoichiometric and symmetric fashion around the ring.[54] Assuming a comprehensive protein-protein connectivity dataset from Native MS, we searched for a minimal IM-MS restraint set to recapitulate the known topology of LTag-p53 with C6 symmetry. Our first attempt utilized only connectivity and global CCS information to generate a structural ensemble. (Figure 4-5A) For this ensemble, we observe three structural families, with relatively little resolution between them. Each family is represented by a very broad distribution of RMSD values relative to the reference structure, indicating that both the accuracy and effective resolution of the structural models created in this search are low. The kernel density function estimated for each structural family also illustrates the poor resolution generated from this restraint set.

In order to resolve the above ambiguity, we add restraints associated with the CCS of the LTag hexamer and the overall C6 symmetry of the complex, a likely result given the interface structure known for this assembly.[54] The resulting IM-MS restrained ensemble is homogenous and gives rise exclusively to highly accurate models (Figure

4-5B). This monomodal ensemble of structures is characterized by a significantly narrower distribution of RMSD values when compared with the distributions observed in Figure 4-5A, and is centered at an RMSD of 6Å relative to the reference. Such RMSD values are typically achieved by our modeling workflow for structures where additional symmetry restraints can be coupled to the distances mined from IM-MS data.

For our second example, we sought to apply our method to a large, asymmetric protein complex that has been interrogated using MS methods previously.[55] The Actin-Related Protein 2/3 (ARP2/3) complex structure was recently solved by X-ray crystallography (PDB ID 1K8K).[56] In addition, a previous native mass spectrometry study identified two modules within the heptameric complex, the trimeric Actin Localization Module (ALM) and the tetrameric Nucleating Module (NM). Extrapolating from the data shown in Figure 4-3, we predict that the heptameric ARP2/3 requires between 16 and 19 internal CCS restraints to reach a PPV value of 80%. When modeling the ALM and NM individually, we find that even minimal simulated IM-MS restraint sets lead to highly accurate models. We generated high-confidence models for the trimeric ALM using 2 IM-derived distance restraints and a global CCS restraint. In parallel, the correct structure was readily found for the NM using 3 IM-derived distance restraints plus the global CCS restraint. These results agree well with data shown in Figure 4-3 for trimeric and tetrameric protein complexes.

Next, we attempted to find the minimal IM-MS restraint sets necessary for localization of ALM binding to NM, leading to a precise assignment of ALM-NM topology. We started by attempting to model this complex without providing any information about points of connectivity between ALM and NM, and filtered the resulting

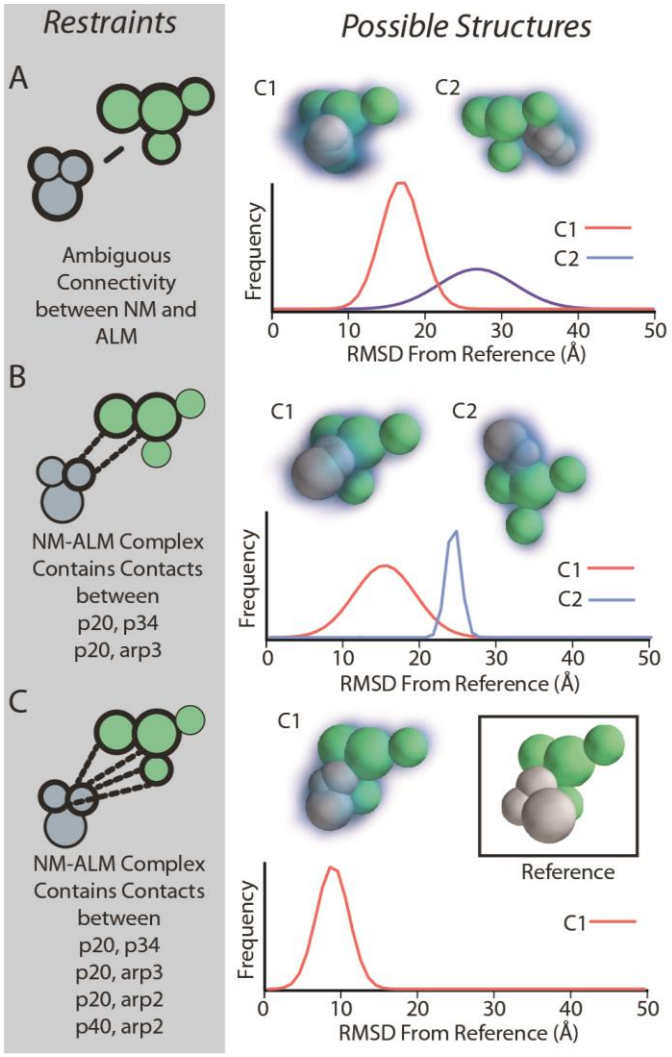


Figure 4- 6 Docking modules within the ARP2/3 complex using connectivity restraints. After encoding the structures of the nucleating module (NM) and the actin localization module (ALM), we tested the global CCS in conjunction with various sets of connectivity restraints (left panels, A, B, and C) for their ability to restraint the docking location and pose of NM on ALM. Structural ensembles were evaluated by hierarchical clustering and the structural families, kernel density functions, and RMSD distributions from the reference are provided.

ensemble based on global CCS alone. (Figure 4-6A) The resulting ensemble features two structural families, a larger population family with an RMSD distribution centered on 15 Å from the reference structure, and a less populated cluster with a very broad RMSD distribution centered on 28 Å. Interestingly, although the resolution within both families is poor, the major family appears to correctly localize the general ALM binding site on the NM surface. To reduce the ambiguity in the models, we then added two restraints that enforced connectivity between the p20 subunit of the ALM and the p34 and arp3 subunits of the NM (Figure 4-6B). This new connectivity information, along with the global CCS restraint gives rise to a new ensemble of potential structures. The new restraint set acts to eliminate the majority of the incorrect structures found in Figure 4-6A; however, it gives rise to a new, more highly-resolved distribution of structures centered on 25 Å from the reference. Interestingly, we note that the major structural family identified for this

restraint set remains essentially unchanged from the one identified in Figure 4-6A, where the correct localization of ALM is determined, but having a broad RMSD distribution centered on ~15 Å from the reference structure

Finally, we applied a new restraint set with 4 total connectivity restraints linking p20 from the ALM with p34, arp2, and arp3 from the NM; and linking p40 from the ALM with arp2 from the NM. (Figure 4-6C) These restraints represent the full complement of protein connectivity information accessible through MS methods.[57] When combined with sufficient connectivity information, we find that the global CCS restraint can define not only the location of ALM on the surface of NM, but also the relative orientation of the two sub-complexes. We observe a single, well-resolved family of structures centered around and RMSD value of 9 Å relative to the reference structure. Furthermore, when structures within this family are averaged, the resulting mean structure has an RMSD of only 2 Å from the reference, indicating that in this example, the mean structure is in much closer agreement with the reference than any individual structure in the ensemble. Combining the connectivity restraints used here with the distance and internal CCS restraints used to build models for each module, we recapitulated the correct topology using only 11 internal restraints, one third fewer internal restraints than that the number of restraints one would predict based on PPV alone (extrapolated from Figure 4-3).

4.7 Conclusions and Future Directions

In this report we explored several questions related to the generation of CG multi-protein topology models restrained using IM-MS data. We outlined a workflow based on integrative modeling principles that allows for facile translation of IM-MS data into

ensembles of putative structures for hypothesis refinement or integration with high resolution docking tools. We explored the limits of coarse-grained modeling, and demonstrated that many protein topologies found in the PDB are not amenable to coarse-graining at the subunit-level, mostly due to their intricate domain architectures. However, when sufficient data is available, domain-level coarse-graining is high fidelity, resulting in significant errors in only 2% of cases.

We benchmarked our CG modeling workflow against protein topologies extracted from the PDB, exploring the ambiguity in IM-MS derived structural ensembles as a function of the information content contained in restraint sets. Our results indicated a predictable relationship between the PPV of an ensemble, and the number of internal IM-MS restraints used to generate it. Although the estimated PPV may be used as a benchmark to predict the ambiguity within a CG modeling ensemble, in many cases it underestimates the total possible information content of the IM-MS experiment, as such an analysis does not account for the structural relationships between members of an ensemble. We found that applying hierarchical clustering yields, in many cases, highly resolved conformational families that can inform future experiments, or be reported as likely structures based on available data. Additionally, we undertook two case studies that showed that highly symmetric or modular complexes can be modeled with high fidelity using smaller numbers of internal restraints than those predicted by a PPV analysis.

Although the computational results presented in this Critical Insight are encouraging, there are still many challenges ahead in fully harnessing the information content available in IM-MS datasets. Our CG error analysis (Figure 4-2) clearly

motivates the development of domain-level IM-MS models of protein quaternary structure, and a move away from CG at the intact subunit level. The development of IM-MS tools for the generation of such information on protein tertiary structure, such as collision induced unfolding (CIU),[58,59] as well as efforts to integrate IM-MS data with other sources of experimental data sensitive to local protein structure[51,60,61] and computational domain assignment algorithms[62] will, therefore, become increasingly important in future protein topology modeling efforts. Similarly, our analysis of ambiguity in IM-MS models of protein quaternary structure strongly points to the need for improved methodologies capable of detecting protein complex connectivity and symmetry. As such, the development of technologies that produce a comprehensive population of protein sub-complexes, either in the gas-phase or in solution, will prove highly valuable.[20,23,24] Finally, the ability of our IMMS-Modeler algorithm to assess, for the first time, the ambiguity present within IM-MS restrained models of protein complex structure will likely lead to a greater ability to integrate such datasets with other forms of structural restraints, derived both from MS and other forms of data. Future iterations of IMMS-Modeler will incorporate the ability to build models based on custom shapes, interface directly with domain-prediction software, and utilize next-generation scoring functions that enable multi-factorial assessments of model fitness. Although not discussed in detail here, it is also clear that increases in CCS precision will drive concomitant increases in the PPV of IM-MS restraints.[63-65] On the other hand, our data demonstrate that much can be accomplished using current IM-MS capabilities and that the proper application of restraints can be used to build high-confidence models of

multi-protein complexes with both full knowledge of their precisions and informed estimates of their accuracies.

4.8 Supplemental Information

All of the software used in this work for modeling and analysis will be made freely available at: https://sites.lsa.umich.edu/ruotolo/software/IMMS_Modeler

Supplemental Information can be found in *Appendix II*.

4.9 Acknowledgements

Protein topology modeling efforts in the Ruotolo lab are supported through the National Institute of General Medical Sciences, National Institutes of Health (R01 GM095832). Additionally, we gratefully acknowledge the support of Erik Marklund (Uppsala), Matteo Degiacomi (Oxford), and Justin Benesch (Oxford), who helped us to integrate IMPACT CCS calculations into our computational workflows.

4.10 References

(1) Robinson, C. V.; Sali, A.; Baumeister, W. *Nature* **2007**, *450*, 973.
(2) Marsh, J. A.; Teichmann, S. A. In *Annu. Rev. Biochemistry* 2015; Vol. 84, p 551.
(3) Hansen, M. R.; Graf, R.; Spiess, H. W. *Accounts Chem. Res.* **2013**, *46*, 1996.
(4) Skiniotis, G.; Southworth, D. R. *Microscopy* **2016**, *65*, 9.
(5) Mertens, H. D. T.; Svergun, D. I. *J. of Struct. Biol.* **2010**, *172*, 128.
(6) Gingras, A.-C.; Gstaiger, M.; Raught, B.; Aebersold, R. *Nat Rev Mol Cell Biol* **2007**, *8*, 645.
(7) Alber, F.; Dokudovskaya, S.; Veenhoff, L. M.; Zhang, W.; Kipper, J.; Devos, D.; Suprapto, A.; Karni-Schmidt, O.; Williams, R.; Chait, B. T.; Rout, M. P.; Sali, A. *Nature* **2007**, *450*, 683.
(8) Russel, D.; Lasker, K.; Webb, B.; Velázquez-Muriel, J.; Tjioe, E.; Schneidman-Duhovny, D.; Peterson, B.; Sali, A. *PLoS Biology* **2012**, *10*.
(9) Shi, Y.; Fernandez-Martinez, J.; Tjioe, E.; Pellarin, R.; Kim, S. J.; Williams, R.; Schneidman-Duhovny, D.; Sali, A.; Rout, M. P.; Chait, B. T. *Mol. Cell. Proteomics* **2014**, *13*, 2927.

(10) Stengel, F.; Aebersold, R.; Robinson, C. V. *Mol. Cell. Proteomics* **2012**, *11*.
(11) Aebersold, R.; Mann, M. *Nature* **2016**, *537*, 347.
(12) Hyung, S. J.; Ruotolo, B. T. *Proteomics* **2012**, *12*, 1547.
(13) Smits, A. H.; Vermeulen, M. *Trends in Biotechnology*, *34*, 825.
(14) Zhong, Y.; Hyung, S.-J.; Ruotolo, B. T. *Expert Review of Proteomics* **2012**, *9*, 47.
(15) Ruotolo, B. T.; Benesch, J. L.; Sandercock, A. M.; Hyung, S.-J.; Robinson, C. V. *Nature protocols* **2008**, *3*, 1139.
(16) Chorev, D. S.; Ben-Nissan, G.; Sharon, M. *Proteomics* **2015**, *15*, 2777.
(17) Marcoux, J.; Cianferani, S. *Methods* **2015**, *89*, 4.
(18) Politis, A.; Borysik, A. J. *Proteomics* **2015**, *15*, 2792.
(19) Snijder, J.; Heck, A. J. R. *Annu. Rev. Analytical Chemistry* **2014**, *7*, 43.
(20) Zhong, Y.; Feng, J.; Ruotolo, B. T. *Anal. Chem.* **2013**, *85*, 11360.
(21) Marsh, J. A.; Hernández, H.; Hall, Z.; Ahnert, S. E.; Perica, T.; Robinson, C. V.; Teichmann, S. A. *Cell* **2013**, *153*.
(22) Benesch, J. L. P. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 341.
(23) Zhou, M. W.; Wysocki, V. H. *Accounts Chem. Res.* **2014**, *47*, 1010.
(24) Samulak, B. M.; Niu, S.; Andrews, P. C.; Ruotolo, B. T. *Anal. Chem.* **2016**, *88*, 5290.
(25) Uetrecht, C.; Barbu, I. M.; Shoemaker, G. K.; van Duijn, E.; Heck, A. J. *Nature chemistry* **2011**, *3*, 126.
(26) Bush, M. F.; Hall, Z.; Giles, K.; Hoyes, J.; Robinson, C. V.; Ruotolo, B. T. *Anal. Chem.* **2010**, *82*, 9557.
(27) Bleiholder, C.; Contreras, S.; Bowers, M. T. *Int. J. Mass Spectrom.* **2013**, *354*, 275.
(28) Marklund, E. G.; Degiacomi, M. T.; Robinson, C. V.; Baldwin, A. J.; Benesch, J. L. *Structure* **2015**, *23*, 791.
(29) Larriba, C.; Hogan Jr, C. J. *The J. of Physical Chemistry A* **2013**, *117*, 3887.
(30) Silveira, J. A.; Fort, K. L.; Kim, D.; Servage, K. A.; Pierson, N. A.; Clemmer, D. E.; Russell, D. H. *J. Am. Chem. Soc.* **2013**, *135*, 19147.
(31) Shi, H. L.; Pierson, N. A.; Valentine, S. J.; Clemmer, D. E. *J. Phys. Chem. B* **2012**, *116*, 3344.
(32) Pukala, T. L.; Ruotolo, B. T.; Zhou, M.; Politis, A.; Stefanescu, R.; Leary, J. A.; Robinson, C. V. *Structure* **2009**, *17*, 1235.
(33) Politis, A.; Park, A.; Hyung, S.-J.; Barsky, D.; Ruotolo, B. T.; Robinson, C. V. *PLoS ONE* **2010**, *5*.
(34) Hall, Z.; Politis, A.; Robinson, C. V. *Structure* **2012**, *20*, 1596.
(35) Ruotolo, B. T.; Giles, K.; Campuzano, I.; Sandercock, A. M.; Bateman, R. H.; Robinson, C. V. *Science* **2005**, *310*, 1658.
(36) Politis, A.; Park, A.; Hall, Z.; Ruotolo, B. T.; Robinson, C. V. *J. of Mol. Biology* **2013**, *425*.
(37) Politis, A.; Schmidt, C.; Tjioe, E.; Sandercock, A. M.; Lasker, K.; Gordiyenko, Y.; Russel, D.; Sali, A.; Robinson, C. V. *Chem. Biol.* **2014**.
(38) Zhou, M.; Politis, A.; Davies, R. B.; Liko, I.; Wu, K.-J.; Stewart, A. G.; Stock, D.; Robinson, C. V. *Nature Chemistry* **2014**, *6*, 208.
(39) Song, Y.; Nelp, M. T.; Bandarian, V.; Wysocki, V. H. *ACS Central Science* **2015**, *1*, 477.
(40) Schneidman-Duhovny, D.; Pellarin, R.; Sali, A. *Curr. Opin. Struct. Biol.* **2014**, *28*, 96.
(41) Han, L.; Hyung, S.-J.; Mayers, J. J.; Ruotolo, B. T. *J. Am. Chem. Soc.* **2011**, *133*, 11358.
(42) Pagel, K.; Natan, E.; Hall, Z.; Fersht, A. R.; Robinson, C. V. *Angew. Chem. Int. Ed.* **2013**, *52*, 361.
(43) Pacholarz, K. J.; Porrini, M.; Garlish, R. A.; Burnley, R. J.; Taylor, R. J.; Henry, A. J.; Barran, P. E. *Angew. Chem. Int. Ed.* **2014**, *53*, 7765.
(44) Levy, E. D.; Pereira-Leal, J. B.; Chothia, C.; Teichmann, S. A. *PLoS Comput Biol* **2006**, *2*, e155.
(45) Ruotolo, B. T.; Benesch, J. L. P.; Sandercock, A. M.; Hyung, S.-J.; Robinson, C. V. *Nature Protocols* **2008**, *3*, 1139.
(46) Arthur, D.; Vassilvitskii, S. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*; Society for Industrial and Applied Mathematics: New Orleans, Louisiana, 2007, p 1027.
(47) Jones, E.; Oliphant, T.; Peterson, P. 2001.

(48) Ballester, P. J.; Richards, W. G. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science* **2007**, *463*, 1307.
(49) Kloczkowski, A.; Jernigan, R. L.; Wu, Z.; Song, G.; Yang, L.; Kolinski, A.; Pokarowski, P. *J. of structural and functional genomics* **2009**, *10*, 67.
(50) Lanucara, F.; Holman, S. W.; Gray, C. J.; Eyers, C. E. *Nat Chem* **2014**, *6*, 281.
(51) Politis, A.; Stengel, F.; Hall, Z.; Hernandez, H.; Leitner, A.; Walzthoeni, T.; Robinson, C. V.; Aebersold, R. *Nat Meth* **2014**, *11*, 403.
(52) Weiss, R. *J. of the American Statistical Association* **1994**, *89*, 359+.
(53) Ramachandran, P.; Varoquaux, G. *Computing in Science & Engineering* **2011**, *13*, 40.
(54) Lilyestrom, W.; Klein, M. G.; Zhang, R.; Joachimiak, A.; Chen, X. S. *Genes & Development* **2006**, *20*, 2373.
(55) Chorev, D. S.; Moscovitz, O.; Geiger, B.; Sharon, M. *Nature Communications* **2014**, *5*, 3758.
(56) Robinson, R. C.; Turbedsky, K.; Kaiser, D. A.; Marchand, J.-B.; Higgs, H. N.; Choe, S.; Pollard, T. D. *Science* **2001**, *294*, 1679.
(57) Sinz, A.; Arlt, C.; Chorev, D.; Sharon, M. *Protein Science : A Publication of the Protein Society* **2015**, *24*, 1193.
(58) Zhong, Y.; Han, L.; Ruotolo, B. T. *Angew. Chem.* **2014**, *126*, 9363.
(59) Eschweiler, J. D.; Martini, R. M.; Ruotolo, B. T. *J. Am. Chem. Soc.* **2017**, *139*, 534.
(60) Hambly, D. M.; Gross, M. L. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 2057.
(61) Schmidt, C.; Macpherson, J. A.; Lau, A. M.; Tan, K. W.; Fraternali, F.; Politis, A. *Anal. Chem.* **2017**, *89*, 1459.
(62) Ansari, E. S.; Eslahchi, C.; Pezeshk, H.; Sadeghi, M. *Proteins: Structure, Function, and Bioinformatics* **2014**, *82*, 1937.
(63) Benigni, P.; Marin, R.; Molano-Arevalo, J. C.; Garabedian, A.; Wolff, J. J.; Ridgeway, M. E.; Park, M. A.; Fernandez-Lima, F. *Int. J. for Ion Mobility Spectrometry* **2016**, *19*, 95.
(64) Glaskin, R. S.; Ewing, M. A.; Clemmer, D. E. *Anal. Chem.* **2013**, *85*, 7003.
(65) Hamid, A. M.; Garimella, S. V. B.; Ibrahim, Y. M.; Deng, L.; Zheng, X.; Webb, I. K.; Anderson, G. A.; Prost, S. A.; Norheim, R. V.; Tolmachev, A. V.; Baker, E. S.; Smith, R. D. *Anal. Chem.* **2016**, *88*, 8949.

Chapter 5: A Structural Model of the Urease Activation Complex Derived from IM-MSIon Mobility-Mass Spectrometry and Integrative Modeling

Joseph D. Eschweiler, Mark. A. Farrugia, Robert P. Hausinger, Brandon T. Ruotolo

Supporting information can be found in *Appendix III*

5.1 Abstract

The activation of K. aerogenes urease via an 18-subunit enzyme-accessory protein complex has been well-studied biochemically, but thus far this complex has remained refractory to direct structural characterization. Using ion mobility-mass spectrometry,we characterized several protein complexes between the core urease enzyme and its accessory proteins, including the 610 kDa (ureABC)3(ureDFG)3 complex. Using our recently-developed computational modeling workflow, we generated ensembles of putative (ureABC)3(ureDFG)3 consistent with experimental restraints and characterized the structural ambiguity present in these models. By integrating structural information from previous studies, we increased the resolution of the ion mobility-mass spectrometry derived models substantially, and we observe a discrete population of structures consistent with all of the available data for this complex.

5.2 Introduction

Protein-protein interactions are critical to nearly all complex cellular processes, making structural characterization of such interactions imperative to our understanding of biology.[1,2] These interactions are diverse, however, ranging from discrete protein dimers and other small oligomers [3] to large, labile interaction networks comprised of dozens of protein chains.[4] One such multiprotein system, the urease activation complex, features a wide range of protein subunit sizes, interaction strengths, and stable subcomplexes within a putative 21-subunit network that has been the focus of diverse structural biology efforts.[5-7] Despite the presence of numerous structural datasets for this system, few direct measurements of its higher-order complexes have been made, and thus relatively little is known about the structure of the urease activation complex or its mode of action in vivo.

Ureases are an important class of bacterial enzymes responsible for the hydrolysis of urea to ammonia and carbamate.[5] Significant attention has been paid to this class of enzymes due to their impacts on human health,[8] and agriculture.[9] *K. aerogenes* urease, the subject of this study, is composed of three protein chains, UreA, UreB and UreC which form a trimer of trimers, $(UreABC)_3$ with molecular weight around 250 kDa.[7] Although X-ray crystallography has elucidated the details of the quaternary and tertiary structure for this urease, including details about its dinuclear Ni^{2+} active site featuring a carbamylated lysine,[10] much less is known about the GTP and CO_2-dependent assembly of the urease active site by the urease accessory proteins UreD, UreE, UreF, and UreG which are co-expressed in *K. aerogenes*.[7] Biochemical studies of these accessory proteins have provided insight into their specific roles in urease

activation. Briefly, UreD is a relatively insoluble protein in aqueous environments that been found to bind directly to urease, but is not competent for urease activation without the other accessory proteins.[11] Recent studies have provided experimental evidence for a Ni^{2+} channel through UreD, indicating a unique ability to provide Ni^{2+} to urease specifically.[12] UreF, a similarly insoluble protein in aqueous solvent, has been found to act as a GTPase modulator to the GTPase UreG,[13] a dimeric protein known to bind Ni^{2+} ions[14,15] and UreE is a known nickel chaperone with known interactions with ureG.[16] Early hypotheses for urease activation proposed sequential binding of ureD, ureF, and ureG to the three urease active sites to form an octadecameric pre-activation complex, $(ureABC)_3(ureDFG)_3$, that could accept Ni^{2+} ions from ureE before performing a GTP-dependent Ni insertion event.[6,7] More recently, our group and others observed a soluble, stable complex of $(ureDFG)_2$ that accepts Ni^{2+} from ureE prior to formation of the $(ureABC)_3(ureDFG)_3$ complex.[17,18]

Despite its importance, the $(ureABC)_3(ureDFG)_3$ pre-activation complex has eluded detailed structural characterization. One factor precluding such a structural characterization for this complex is its lability, which results in multiple coexisting subcomplexes that make interpretation of any dataset that doesn't include a high-resolution separation step extremely difficult. Importantly, several subcomplexes have been identified by native mass spectrometry (MS)[17] and chemical crosslinking[17,19] including $(UreABC)_3(ureDFG)$ and $(UreABC)_3(ureDFG)_2$, $(ureABC)_3(UreD)_3$, and $(ureABC)_3(ureDFF)$. Detailed analysis of crosslinked peptides by tandem MS revealed putative interaction sites of urease with UreD,[19] which have also been supported by SAXs datasets for samples containing $(ureABC)_3(ureD)_3$.[20] These datasets were

recently integrated with molecular docking to provide the most comprehensive picture of the $(ureABC)_3(ureDFG)_3$ structure to date,[21] however without direct observation of the complex it is difficult to assign a confidence level to the model produced.

In this study, we use ion mobility-mass spectrometry (IM-MS) to characterize complex samples relating to the $(ureDFG)_2$ complex as well as complexes formed between UreDFG and urease. IM-MS is a tandem methodology that separates proteins and protein complex ions produced using nano-electrospray ionization (nESI) under native conditions first by size by IM and then by m/z using MS. IM-based size separations can be calibrated to produce orientiationally-averaged collision cross section (CCS) values that can be used, along with connectivity information recovered from native MS to restrain modeling efforts. We utilize a previously-reported maltose-binding-protein:ureD fusion protein to increase the solubility of the system, while still allowing for formation of key protein complexes that are competent activators of urease.[11] Furthermore, we report the first observations of the fully assembled $(ureABC)_3(ureDFG)_3$ complex, and using IM-derived CCS information[22] we develop a method for coarse-grained modeling[23,24] of $(ureABC)_3(ureDFG)_3$ and its subcomplexes that allows us to characterize the conformational space of $(ureABC)_3(DFG)_3$ consistent with our data.

5.3 METHODS

Sample preparation. $(UreDFG)_2$ $(UreAC)_3$ and $(UreABC)_3$ were expressed and purified as reported previously.[17,11,25] $(UreABC)_3(ureDFG)_x$ samples were prepared by incubating $(ureABC)_3$ with $(ureDFG)_2$ for 30m before flash freezing and storage at -80C.

Ion mobility-mass spectrometry. Samples were buffer exchanged into 200mM ammonium acetate using Micro Bio-Spin P-30 columns (Bio-Rad, Hercules, CA) at an initial concentration of ~1uM. The final concentrations of the samples were unknown, however likely range from 100 to 900 nM based on expected losses during exchange. IM-MS Experiments were performed on a Synapt G2 ion mobility-mass spectrometry platform (Waters Corp., Milford, MA) with a nano-electrospray ionization source equipped. Briefly, the capillary voltage was set to 1.5 kV, with sampling and extraction cone voltages set to 0 V to preserve noncovalent interactions. The trap and transfer collision energies were both set to 4V. Optimal IMS parameters were as previously published,[26] IMS gas pressure approximately ~4 mBar with a wave height and wave velocity set to 15 V and 150 m/s, respectively. Data was processed using Masslynx and Driftscope (Waters Corp., Milford, MA). Mass assignments were calculated using the maximum entropy method as implemented in ESIprot.[27]

CCS Calibration. CCS values were determined as previously described.[22,28] cytochrome C, avidin, alcohol dehydrogenase, and glutamate dehydrogenase were used as calibrants for the wide range of CCS values observed. Typical calibration curves produced correlation coefficients of greater than 0.99.

Coarse-grained Modeling. Our general method for coarse-grained modeling is similar to previously described protocols[23,29] and more specific details on our method are provided in an accompanying manuscript (Chapter 4). Briefly, individual protein subunits are represented as spheres with radii corresponding to their experimental or calculated CCS. Higher order complexes are restrained by specific geometric constraints or more ambiguous connectivity restraints as defined by the experimental

data acquired. An ensemble of putative models is generated by repeated optimization of a scoring function built from the above restraints. Typical ensembles ranged from 1000 to 25,000 structures but were drastically reduced in size by filtering these model pools using experimental and biophysical restraints not included in the original scoring function.[23] CCS for each model were determined using the projection approximation function in IMPACT,[30] and CCS values with an uncertainty of +/- 3% were used as an experimental filter for the ensemble. After CCS filtering, models that agree with biophysical and experimental data are analyzed using hierarchical clustering to determine the predominant structural families present in the ensemble. Ensembles or subsets thereof are then represented by kernel-density functions,[31] with the median structure shown for reference. These same groups were also characterized by the average RMSD from the mean, which provides an estimate of the relative resolutions between ensembles.

5.4 RESULTS

IM-MS of $(MBP\text{-}UreDFG)_2$ and $(UreABC)_3(MBP\text{-}ureDFG)_3$ containing samples

IM-MS analysis of samples containing the fusion protein MBP-UreD, UreF, and UreG revealed the predicted $(MBP\text{-}UreDFG)_2$ complex plus a number of subcomplexes consistent with our previous study.[17] Analysis of the CCSs of these subcomplexes revealed values consistent with those predicted from a homologous urease accessory complex, $(hypHFG)_2$, for which a crystal structure exists,[18] indicating a degree of topological agreement between these two complexes. After evaluation of the complexes derived from $(MBP\text{-}UreDFG)_2$, we incubated this complex with $(UreABC)_3$ before subsequent IM-MS analysis. (Figure 5-1) The resulting dataset featured many of the

subcomplexes found in the $(MBP\text{-}UreDFG)_2$ sample, plus a host of complexes derived from interactions between MBP-UreDFG and $(UreABC)_3$. Notably, in this dataset we a observe a new complex, the linear $(MPB\text{-}UreDF)_2$, which was not observed in the absence of urease, which has been observed in the *H. pylori* urease activation pathway.[18] In contrast to our previous study, these new datasets also contain signals from the fully assembled $(UreABC)_3(MBP\text{-}DFG)_3$ complex, as well as the subcomplexes $(UreABC)_3(MBP\text{-}DFG)_2$ and $(UreABC)_3(MBP\text{-}DFG)$. These results provide further evidence for modular addition of MBP-UreDFG to urease to form the pre-activation

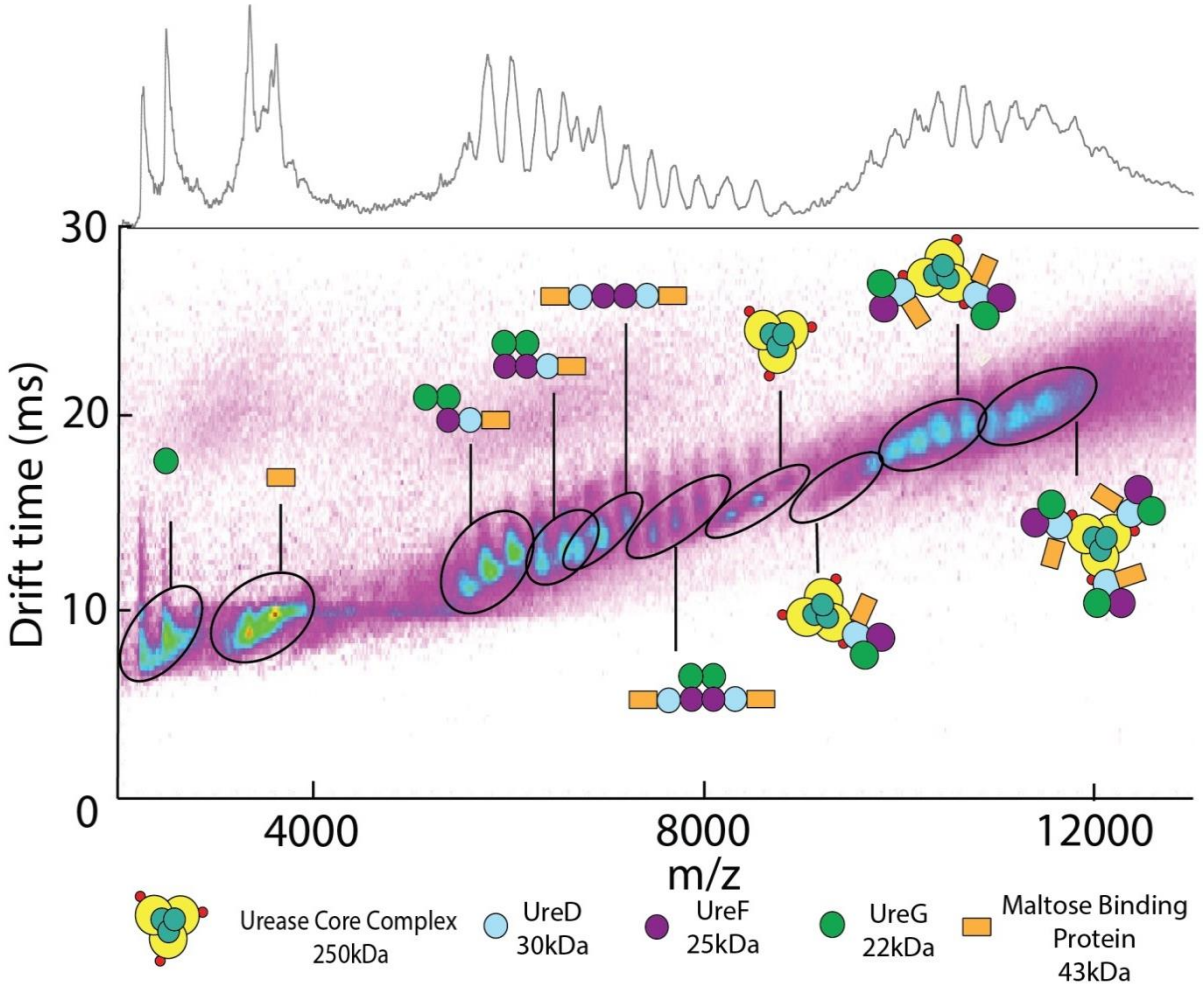


Figure 5-1 IM-MS analysis of $(UreABC)_3(MBP\text{-}DFG)_3$ and its subcomplexes. IM-MS analysis of $(UreABC)_3(MBP\text{-}DFG)_3$ is presented as a plot of drift time vs m/z, with the standard m/z dimension projected on top of the figure. This data reveals the masses and collision cross sections of many subunits, and subcomplexes that comprise the 610 kDa $(UreABC)_3(MBP\text{-}DFG)_3$ octadecamer.

complex, and provide the first direct observation of the long-hypothesized $(UreABC)_3(MBP\text{-}DFG)_3$ complex.

IM-MS-based Modeling of $(UreABC)_3(MBP\text{-}DFG)_3$

After calculating CCS values for $(UreABC)_3(MBP\text{-}DFG)_3$ and its subcomplexes, as well as incorporating other datasets that provided CCS data for the urease core and other subcomplexes, specifically $(ureA)_3$, $(UreAC)_3$, $(UreABC)_3$, and $(ureABC)_3$-(MBP-ureD), (Figure 5-2A and figures *III-1 – III-5*) we hypothesized that the data would be sufficient to restrain a coarse-grained model of the $(UreABC)_3(MBP\text{-}DFG)_3$ complex. We started by translating each protein within the complex into a sphere with radius corresponding to its measured or calculated CCS. Specifically, CCS values for UreA, UreC, UreF, and UreD were derived from the trajectory method approximation in IMPACT, while values for ureB, UreG, and MBP were derived from calibrated experimental drift times. In terms of the urease core complex $(ureABC)_3$, our experimental CCS values were in close agreement with trajectory method approximations from IMPACT, with an error of -1.4%. We leveraged this data, as well as experimental data for $(ureA)_3$ and $(UreAC)_3$ to build a coarse-grained model of the urease core that matched our experimental dataset with errors <1%.(Figure 5-2 B and 5-2 C) Importantly, we found it was necessary to model ureC as two spheres representing each domain within the protein chain, as it was impossible to accurately recapitulate the shape and CCS of the complex otherwise.

Once an accurate model of the urease core was developed using gas-phase restraints, we developed a restraint-based scoring function for a Monte-Carlo search for a representative sample of structures for the complete pre-activation complex that agree

with our experimental IM-MS data. The restraints in our initial scoring function included a rigidly restrained model of MBP-UreD, however all other restraints were defined simply by connectivity, which allows subunits to adopt a range of distances and orientations as long as they remain in contact. Some of these connectivity restraints were derived from previous studies which demonstrated connectivity between UreC:UreD, UreB:UreD, UreD:UreF, and UreF:UreG.[19,13] Although detailed structural information is available for the UreD:UreF and UreF:UreG found in homologous complexes,[18] we chose to only assign connectivity rather than rigid distance restraints to avoid biasing the model toward interactions that may not be present in the $(UreABC)_3(MBP\text{-}DFG)_3$ structure. We generated 25,000 possible conformations for the complex of MBP, UreD, UreF, and UreG with respect to the urease core structure, with the same C3

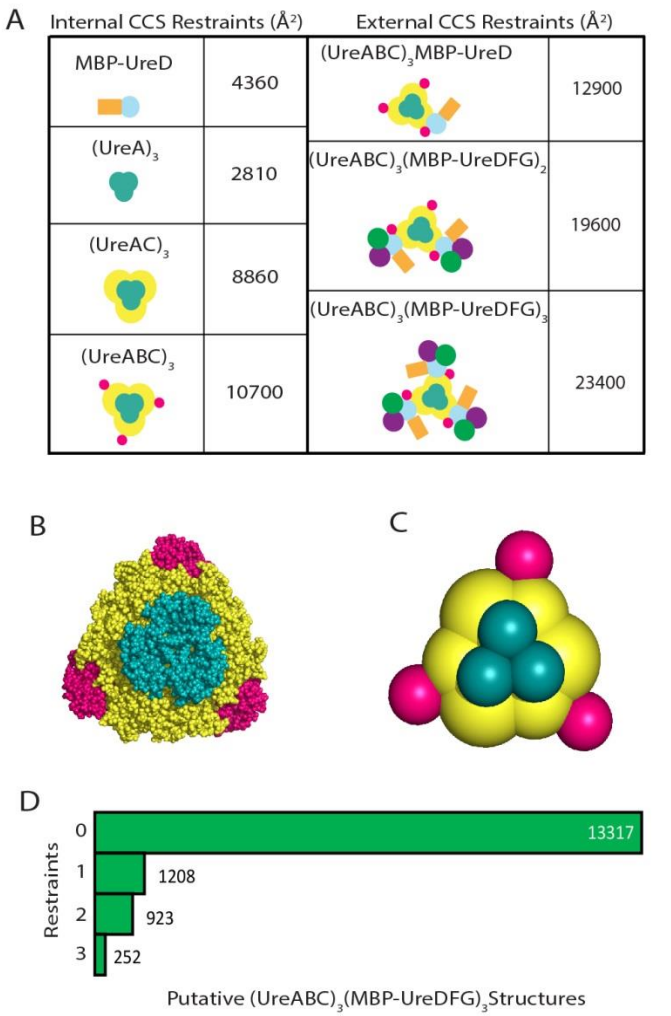


Figure 5-2 IM-MS restraints for building molecular models of $(UreABC)_3(MBP\text{-}DFG)_3$. A) left panel: CCS restraints used to build the urease core scaffold $(UreABC)_3$, and to restrain the interaction between maltose binding protein (MBP) and ureD. Right panel: CCS restraints for target complexes used to filter large ensembles of structures generated using a Monte Carlo search in IM-MS_modeler. Comparison of the X-ray structure of $(UreABC)_3$ (B) with a coarse-grained model generated with IM-MS data (C) that is used as a scaffold to for modeling of $(UreABC)_3(MBP\text{-}DFG)_3$. D) Results of filtering an ensemble of 25,000 putative structures of $(UreABC)_3(MBP\text{-}DFG)_3$ based on biophysical and experimental data. "0" restraints filters only by general biophysical parameters related to the interaction geometries of proteins, "1", "2", and "3" restraints incorporate filters for the experimental CCS values +/- 3% for $(UreABC)_3(MBP\text{-}ureD)$, $(UreABC)_3(MBP\text{-}DFG)_2$, and $(UreABC)_3(MBP\text{-}ureDFG)_3$, respectively.

symmetry enforced as found in the core.[12] In the next step, we filtered models based on biophysical restraints that govern maximum and minimum levels of inter-digitation observed for protein-protein interfaces, which manifests as an overlap between the spheres representing single-domain protein subunits in our models.[24] This filter brought the ensemble down to 13,317 models, which were subjected to the next round of filtering based on agreement with experimental CCS data. In this step, we used IMPACT to calculate CCS values for each model, as well as subcomplexes within a given model. Models were passed into the filtered ensemble if they agreed with the experimental CCS value for $(UreABC)_3(MBP\text{-}DFG)_3$, $(UreABC)_3(MBP\text{-}DFG)_2$, and $(UreABC)_3(MBP\text{-}D)$ within +/- 3%. In Figure 5-2 D, each CCS restraint decreases the size of the ensemble, indicating that each adds unique information about the structure of the pre-activation complex. The structures in the resulting ensemble, matching all biophysical and experimental CCS restraints, were then subjected to structural analysis by hierarchical clustering. The resulting dendrogram (Figure 5-3 A) reveals strong clustering of structures within the final ensemble into 3 families, denoted as cluster 1, cluster 2, and cluster 3. In Figure 5-3 B, each cluster is visualized by plotting the median structure of each cluster plus the kernel density function that represents probability density of structures around the median. Visual analysis of these clusters reveals key ambiguities present in our data. First, the position of the MBP (brown) cannot be resolved, which is not surprising because it can adopt many orientations around UreD (blue) within the model. Next, the position of ureD itself is changed substantially between clusters, anchoring the ureDFG assembly to the right of UreB (red) in cluster 1, forward of ureB in cluster 2, and to the left of ureB in cluster 3.

Our models suggest this positioning plays a role in the possible configurations for ureF (purple) and ureG (green) within the ensemble, where the clusters with ureD in plane with ureB and UreC (clusters 0 and 2) adopt similar configurations for UreF and UreG, and are in contrast to the out-of-plane conformations found in cluster1.

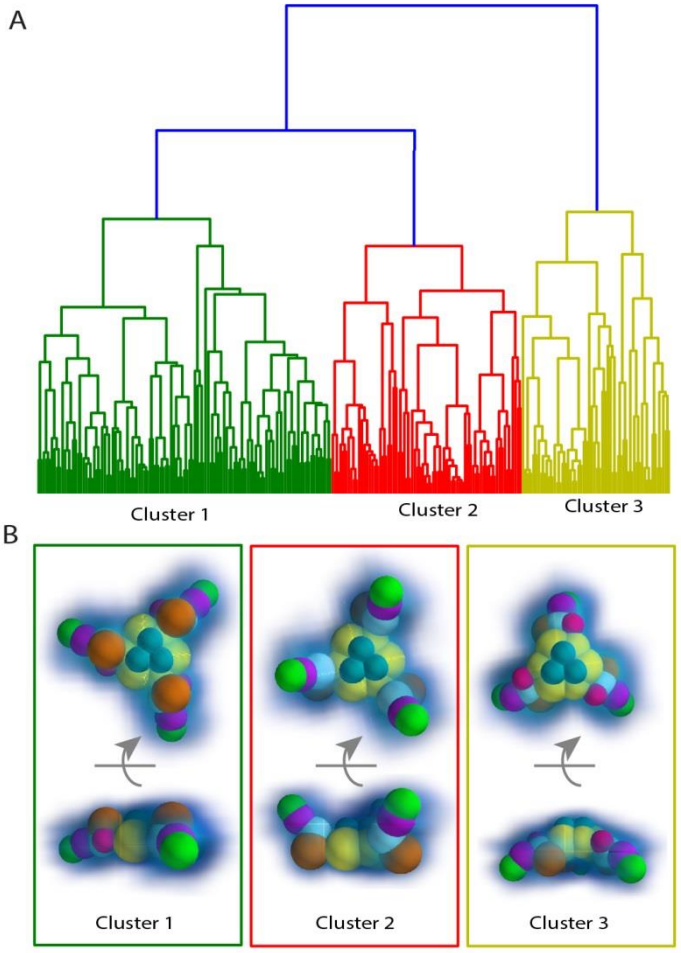


Figure 5-3 Hierarchical clustering reveals ambiguity in under-restrained models. A) An ensemble of 252 structures that agreed with all experimental restraints was subjected to hierarchical clustering analysis to identify structural families within the group. In this example, the ensemble clustered strongly into 3 structural families. B) Each structural family is represented by the median structure and the kernel density function estimated from the structural ensemble. Using this technique, we qualitatively assess the resolution and structure of each family.

Although significant ambiguity was present in these models, we were encouraged by the strong clustering of the ensemble into distinct groups having structural differences that were easily assessed qualitatively. Since the major sources of ambiguity for non-MBP urease accessory proteins hinged on the positioning of ureD relative to ureB and ureC, we sought to incorporate data from other sources into our model improve the confidence in our structure assignment for the pre-activation urease complex.

Integration of Additional Structural Data for Improved Structural Resolution

To resolve structural ambiguity within our IM-MS-derived models, we looked to previously reported structural data to incorporate into our model. Ligabue-Braun and colleagues have previously reported a model for the $(UreABC)_3(MBP\text{-}DFG)_3$ complex based on molecular docking that broadly agrees with existing crosslinking[19] and SAXS[20] datasets relating to the $(ureABC)_3(ureD)_3$ complex.[21] Specifically, the positioning of the UreD subunit in the LigaBue-Braun $(UreABCD)_3$ model is broadly consistent with chemical crosslinks between ureC K401 and the UreD N-terminus, UreB K76 and UreD N-terminus, and the deactivation of UreC 515 crosslinking upon binding of UreD.[19] The simulated SAXS profile for the this $(ureABC)_3(ureD)_3$ conformation is also in agreement with experimental data. In light of this broad agreement with other experimental sources, we used the $(UreABCD)_3$ structure put forth by Ligabue-Braun to restrain the position of UreD in our coarse-grained model. Although Ligabue-Braun and colleagues in provided a model for the fully assembled $(UreABC)_3(MBP\text{-}DFG)_3$ complex, we chose not include any restraints related to higher order complexes into our model, as we could find little experimental support for the positioning of ureF and ureG within the $(UreABC)_3(MBP\text{-}DFG)_3$ complex. When we implement these new restraints for the $(ureABC)_3(ureD)_3$ subcomplex, the resulting structures are largely in agreement with the ureB:ureC:ureD configuration in cluster 2 from our initial modeling effort (Figure 5-3B), where MBP-UreD is found proximal to the active site, on the left side of UreB and oriented toward the back of the urease complex. (Figure 5-4 A-B)

Using this new model as a scaffold, we repeated our Monte-Carlo search for 50,000 possible structures that represent the conformational space available to UreF and UreG within this model. Filtering this ensemble by the previously discussed biophysical restraints resulted in 16,308 candidate structures, which were then filtered by agreement with experimental CCS to yield a population of 443 structures. Remarkably, upon clustering of these structures, only weak clustering was observed, indicating two closely-related structural families. (Figure 5-4 C; note that spheres corresponding to MBP have been removed for clarity) Indeed, the standard deviation of this pre-clustering ensemble was consistent with the post-clustering standard deviation from the previous modeling output, indicating that restraining UreD resulted in a significantly narrowed ensemble of models for the pre-activation complex. (Figure 5-4B). Visualization of these

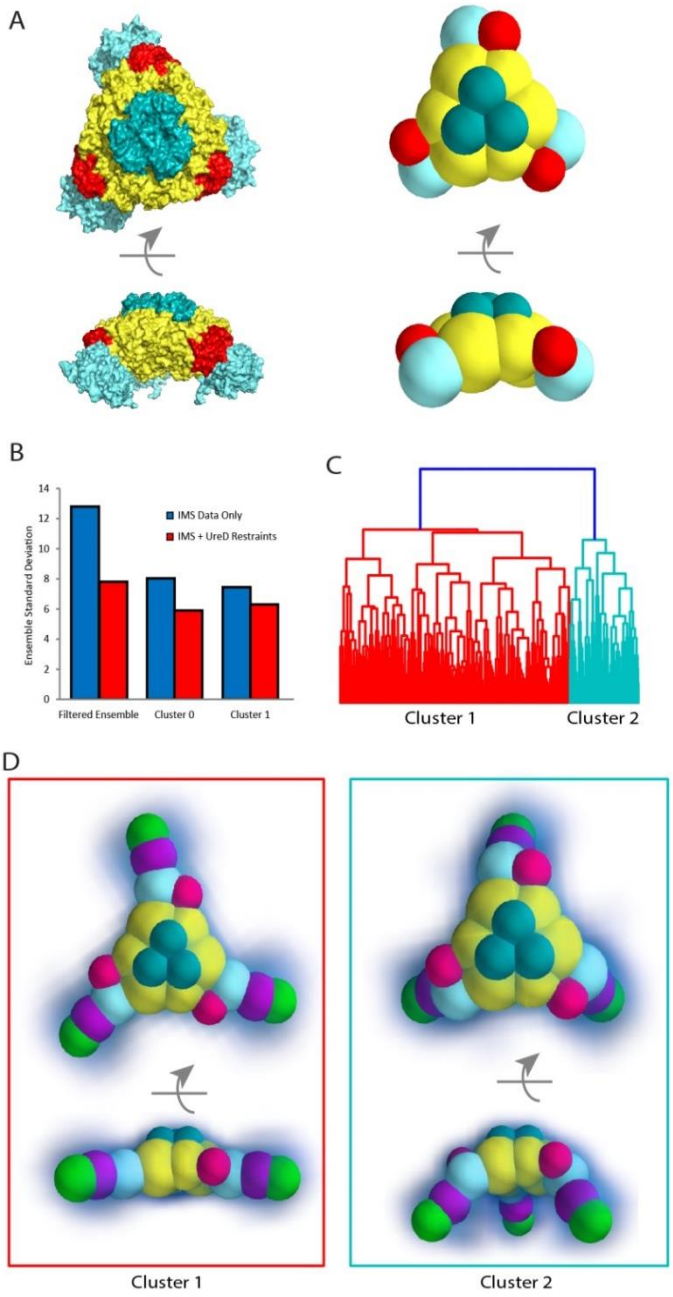


Figure 5-4 Resolving ambiguity by integrating new data. A) A previously published all-atom model of $(UreABC)_3(ureD)_3$ was used to restrain the position of ureD in our coarse-grained scaffold structure. B) Additional restraints significantly increase the resolution of the model as measured by the standard deviation within the ensemble. Blue bars indicate IM-MS data only, red bars indicate IM-MS data plus restraints on ureD from a previous model. The standard deviations of the entire ensemble and each cluster are reported. C) Hierarchal clustering of the new ensemble reveals weak clustering into two structural families, indicating a more homogenous population of structures. D) Median structures and kernel density estimates for the two structural families identified.

structural families reveal that the experimental data define a discrete structural space that lies between a largely planar and extended structure, and a slightly more compact structure featuring ureDFG modules directed toward the back of $(ureABC)_3$ (Figure 5-4D).

Comparison with previous models. Although our coarse-grained model incorporated some elements of the model put forward by Ligabue-Braun,[21] a detailed analysis of our IM-MS-derived structural models reveals significant differences in the expected conformations for the urease $(UreABC)_3(MBP\text{-}DFG)_3$ complex. Figure 5-5A shows the RMSD distributions for two structural ensembles derived from the coarse-grained version of the Ligabue-Braun model. The first ensemble, represented by a blue histogram with a black Gaussian fit, incorporates no experimental CCS data, and represents the entirety of conformational space that can be adopted by UreF and UreG on a scaffold of $(UreABCD)_3$. The same ensemble after filtering by our experimental

CCS restraints is represented in green with an orange Gaussian fit (The frequency axis is scaled in order to enable unbiased comparison, but the actual frequencies are approximately 10-fold lower for the filtered ensemble). This result reveals that an ensemble filtered by our experimental CCS data is only minimally enriched for models akin to the Ligabue-Braun model, indicating only weak agreement between our experimental data and the model. Since direct comparison of experimental CCS with that calculated from the Ligabue-Braun model is difficult due to the presence of the MBP tag in our experimental data, we also compared the CCS values for our entire ensemble of experimentally-restrained models with CCS values calculated from the Ligabue-Braun model. In figure 5-5 B, we compare CCS values for our experimentally restrained models and theoretical CCS values computed from the all-atom structures proposed by Ligabue-Braun. CCS values for our experimentally

restrained models are shown as blue dots, with error bars representing 2 standard deviations within the ensemble. CCS values for the Ligabue-Braun model were calculated using a linearly scaled projection approximation method (PA*1.15), as well as the trajectory method estimation (TJM) within IMPACT in red and green circles, respectively. Interestingly, although the TJM values agree very well with our experimental measurements for smaller complexes like $(ureABC)_3$ and $(ureABC)_3(ureD)$, we also note increasing deviation between these values as additional subunits are added. In contrast, the scaled projection approximation values for the high-resolution models put forth by Ligabue-Braun are consistently 12% to 20% lower than the CCS values associated with models generated from IM-MS. To understand these deviations from predicted model CCS

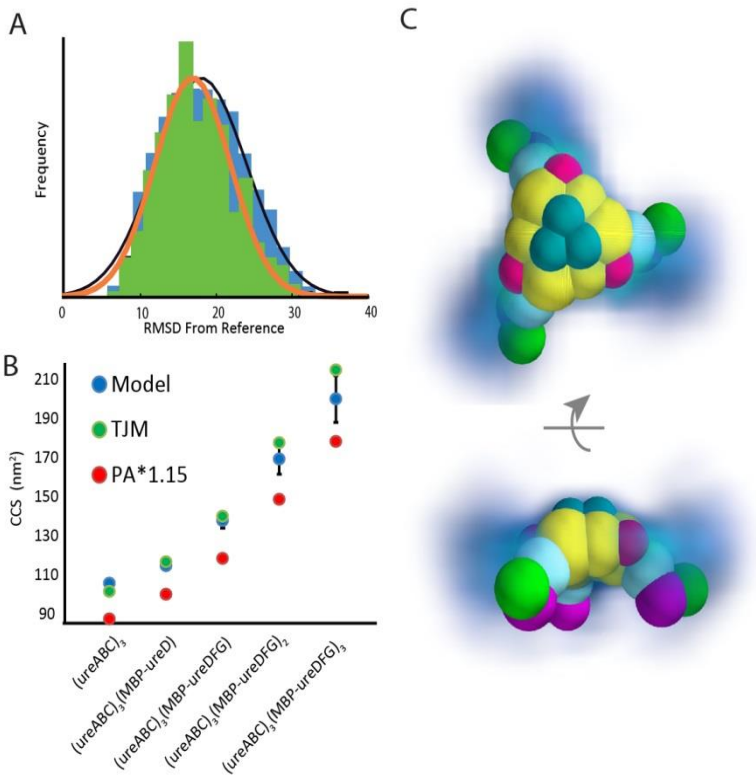


Figure 5-5 Comparing IM-MS-derived models with structures from molecular docking. A) RMSD distributions of IM-MS derived models from a reference model generated using molecular docking and integrative modeling. The blue distribution (black Gaussian fit) represents an ensemble of structures generated from the $(UreABC)_3(ureD)_3$ scaffold with no filtering by experimental CCS restraints. The green ensemble (orange fit) represents the same ensemble filtered by experimental CCS (frequency axis normalized). B) Experimental and predicted CCS values for several urease complexes. Blue dots indicate IM-MS-derived models where the error bars represent two standard deviations, green and red dots represent calculated CCS values for the reference structure by the trajectory method approximation and scaled projection approximation, respectively. C) Qualitative comparison of the kernel density function of an IM-MS-derived ensemble with a coarse-grained representation of the reference structure.

values, we built a coarse-grained model based on Ligabue-Braun's structure and superimposed it into the density cloud calculated from our IM-MS restrained ensemble. (Figure 5-5 C) This representation reveals that the ensemble restrained by our experimental CCS generally adopts a more extended conformation than the previously-reported model. Although a portion of the experimentally-derived models do agree well with the Ligabue-Braun model, specifically those in cluster 2 shown within figure 5-4, it is clear that our IM-MS experiments sample a somewhat different ensemble of protein quaternary structures preferentially.

5.5 Conclusions

In this study, we characterized the 610 kDa, 18-subunit urease pre-activation complex using IM-MS. To our knowledge, this complex, among the largest heterocomplexes to be characterized by IM-MS, has not been directly observed by any other method. We used CCS values derived from ion mobility drift times of the fully assembled complex, as well as several subcomplexes to build coarse-grained models revealing possible gas-phase structures of the complex. Our IM-MS data alone was not sufficient for unambiguous structural assignment, but when combined with data from chemical crosslinking, SAXS, and molecular modeling, we were able to define a narrow population of possible structures falling within our experimental restraints.

Our model shares major structural features with other models proposed by computational docking, however it differs in the angle of ureDFG modules relative to the urease core structure. By estimating kernel density functions for ensembles of experimentally-restrained structures, we visualized the discrepancies between our experimental data and the previously reported model. These discrepancies may be due to gas-phase rearrangements of

the (UreABC)3(MBP-DFG)3 complex, they may also be representative of how the complex may alter its structure under different experimental conditions, or related to the scarcity of experimental data restraining previous models. We note that due to the size and putative structures of the proteins involved, a scenario that rationalized the above-described differences based solely on a gas-phase rearrangement is unlikely, and that the relative flexibility of the urease pre-activation complex has been discussed in detail previously.

In summary, the model of the urease pre-activation complex presented in this report represents the most restrained structure of the assembly to date, representing a consensus of datasets acquired through IM-MS, chemical cross-linking, and SAXS experiments reported from multiple laboratories. Clearly, urease activation includes additional steps and protein binding events in order to load the enzyme with its required dinuclear Ni2+ core, but given the information content presented in this report we expect that our model will drive new discussions surrounding the role of this activation complex in the context of current urease activation mechanisms. Furthermore, the lability and size of this complex represents a frontier for the IM-MS technique in terms of its capabilities to build structural models of such large protein hetero-oligomers, and points to a bright future for the tool in similar structural biology efforts.

5.6 Supporting Information

Supporting Information can be found in *Appendix III*

5.7 References

1. Robinson, C.V., Sali, A., Baumeister, W.: The molecular sociology of the cell. Nature **450**(7172), 973-982 (2007). doi:10.1038/nature06523
2. Marsh, J.A., Teichmann, S.A.: Structure, dynamics, assembly, and evolution of protein complexes. In: Annu. Rev. Biochemistry, vol. 84. pp. 551-575. (2015)
3. Venkatakrishnan, A.J., Levy, Emmanuel D., Teichmann, Sarah A.: Homomeric protein complexes: evolution and assembly. BioChem. Society Transactions **38**(4), 879-882 (2010). doi:10.1042/bst0380879
4. Perkins, J.R., Diboun, I., Dessailly, B.H., Lees, J.G., Orengo, C.: Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. Structure **18**(10), 1233-1243 (2010). doi: 10.1016/j.str.2010.08.007

5. Mobley, H.L., Island, M.D., Hausinger, R.P.: Molecular biology of microbial ureases. Microbiological Reviews **59**(3), 451-480 (1995).
6. Carter, E.L., Flugga, N., Boer, J.L., Mulrooney, S.B., Hausinger, R.P.: Interplay of metal ions and urease. Metallomics **1**(3), 207-221 (2009). doi:10.1039/B903311D
7. Farrugia, M.A., Macomber, L., Hausinger, R.P.: Biosynthesis of the Urease Metallocenter. J. of Biological Chemistry **288**(19), 13178-13185 (2013). doi:10.1074/jbc.R112.446526
8. Collins, C.M., Dorazio, S.E.F.: BACTERIAL UREASES - STRUCTURE, REGULATION OF EXPRESSION AND ROLE IN PATHOGENESIS. Mol. Microbiol. **9**(5), 907-913 (1993). doi:10.1111/j.1365-2958.1993.tb01220.x
9. Bremner, J.M.: Recent research on problems in the use of urea as a nitrogen fertilizer. Fertilizer research **42**(1), 321-329 (1995). doi:10.1007/bf00750524
10. Pearson, M.A., Michel, L.O., Hausinger, R.P., Karplus, P.A.: Structures of Cys319 Variants and Acetohydroxamate-Inhibited Klebsiella aerogenes Urease. Biochemistry **36**(26), 8164-8172 (1997). doi:10.1021/bi970514j
11. Carter, E.L., Hausinger, R.P.: Characterization of the Klebsiella aerogenes Urease Accessory Protein UreD in Fusion with the Maltose Binding Protein. J. of Bacteriology **192**(9), 2294-2304 (2010). doi:10.1128/jb.01426-09
12. Farrugia, M.A., Wang, B., Feig, M., Hausinger, R.P.: Mutational and Computational Evidence That a Nickel-Transfer Tunnel in UreD Is Used for Activation of Klebsiella aerogenes Urease. Biochemistry **54**(41), 6392-6401 (2015). doi:10.1021/acs.biochem.5b00942
13. Boer, J.L., Hausinger, R.P.: Klebsiella aerogenes UreF: Identification of the UreG Binding Site and Role in Enhancing the Fidelity of Urease Activation. Biochemistry **51**(11), 2298-2308 (2012). doi:10.1021/bi3000897
14. Boer, J.L., Quiroz-Valenzuela, S., Anderson, K.L., Hausinger, R.P.: Mutagenesis of Klebsiella aerogenes UreG To Probe Nickel Binding and Interactions with Other Urease-Related Proteins. Biochemistry **49**(28), 5859-5869 (2010). doi:10.1021/bi1004987
15. Moncrief, M.B., Hausinger, R.P.: Characterization of UreG, identification of a UreD-UreF-UreG complex, and evidence suggesting that a nucleotide-binding site in UreG is required for in vivo metallocenter assembly of Klebsiella aerogenes urease. J. of Bacteriology **179**(13), 4081-4086 (1997). doi:10.1128/jb.179.13.4081-4086.1997
16. Merloni, A., Dobrovolska, O., Zambelli, B., Agostini, F., Bazzani, M., Musiani, F., Ciurli, S.: Molecular landscape of the interaction between the urease accessory proteins UreE and UreG. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics **1844**(9), 1662-1674 (2014). doi: 10.1016/j.bbapap.2014.06.016
17. Farrugia, M.A., Han, L., Zhong, Y., Boer, J.L., Ruotolo, B.T., Hausinger, R.P.: Analysis of a Soluble (UreD:UreF:UreG)(2) Accessory Protein Complex and its Interactions with Klebsiella aerogenes Urease by Mass Spectrometry. J. Am. Soc. Mass Spectrom. **24**(9), 1328-1337 (2013). doi:10.1007/s13361-013-0677-y
18. Fong, Y.H., Wong, H.C., Yuen, M.H., Lau, P.H., Chen, Y.W., Wong, K.-B.: Structure of UreG/UreF/UreH Complex Reveals How Urease Accessory Proteins Facilitate Maturation of Helicobacter pylori Urease. PLoS Biology **11**(10), e1001678 (2013). doi:10.1371/journal.pbio.1001678
19. Chang, Z., Kuchar, J., Hausinger, R.P.: Chemical Cross-linking and Mass Spectrometric Identification of Sites of Interaction for UreD, UreF, and Urease. Journal of Biological Chemistry **279**(15), 15305-15313 (2004). doi:10.1074/jbc.M312979200
20. Quiroz-Valenzuela, S., Sukuru, S.C.K., Hausinger, R.P., Kuhn, L.A., Heller, W.T.: The structure of urease activation complexes examined by flexibility analysis, mutagenesis, and small-angle X-ray scattering. Archives of Biochemistry and BioPhys. **480**(1), 51-57 (2008). doi:10.1016/j.abb.2008.09.004

21. Ligabue-Braun, R., Real-Guerra, R., Carlini, C.R., Verli, H.: Evidence-based docking of the urease activation complex. J. of BioMol. Structure and Dynamics **31**(8), 854-861 (2013). doi:10.1080/07391102.2012.713782
22. Ruotolo, B.T., Benesch, J.L., Sandercock, A.M., Hyung, S.-J., Robinson, C.V.: Ion mobility–mass spectrometry analysis of large protein complexes. Nature protocols **3**(7), 1139-1152 (2008).
23. Politis, A., Park, A., Hyung, S.-J., Barsky, D., Ruotolo, B.T., Robinson, C.V.: Integrating Ion Mobility Mass Spectrometry with Molecular Modelling to Determine the Architecture of Multiprotein Complexes. PLoS ONE **5**(8) (2010). doi:10.1371/journal.pone.0012080
24. Hall, Z., Politis, A., Robinson, C.V.: Structural modeling of heteromeric protein complexes from disassembly pathways and ion mobility-mass spectrometry. Structure **20**(9), 1596-1609 (2012). doi:10.1016/j.str.2012.07.001
25. Carter, E.L., Boer, J.L., Farrugia, M.A., Flugga, N., Towns, C.L., Hausinger, R.P.: Function of UreB in Klebsiella aerogenes Urease. Biochemistry **50**(43), 9296-9308 (2011). doi:10.1021/bi2011064
26. 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 **136**(17), 3534-3541 (2011). doi:10.1039/C0AN00987C
27. Winkler, R.: ESIprot: a universal tool for charge state determination and molecular weight calculation of proteins from electrospray ionization mass spectrometry data. Rapid Comm. Mass Spectrom. **24**(3), 285-294 (2010). doi:10.1002/rcm.4384
28. Bush, M.F., Hall, Z., Giles, K., Hoyes, J., Robinson, C.V., Ruotolo, B.T.: Collision Cross Sections of Proteins and Their Complexes: A Calibration Framework and Database for Gas-Phase Struct. Biol.. Anal. Chem. **82**(22), 9557-9565 (2010). doi:10.1021/ac1022953
29. Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B., Sali, A.: Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. PLoS Biology **10**(1) (2012). doi:10.1371/journal.pbio.1001244
30. Marklund, E.G., Degiacomi, M.T., Robinson, C.V., Baldwin, A.J., Benesch, J.L.: Collision cross sections for structural proteomics. Structure **23**(4), 791-799 (2015). doi:10.1016/j.str.2015.02.010
31. Weiss, R.: Multivariate Density Estimation: theory, practice, and visualization. J. of the American Statistical Association **89**, 359+ (1994).

Chapter 6: Applications of IM-MS for Studying the Self-Assembly of both Natural and Engineered Protein Complexes

6.1 Abstract

The ability of proteins to assemble into higher-order structures is fundamental to many biological processes.[1] Additionally, this phenomenon presents intriguing opportunities for engineering macromolecular structures or machines from rationally designed proteins.[2,3] The study of protein interfaces has shown that these interfaces are incredibly diverse, and feature hydrophobic and polar interactions that are highly tuned for modulating the strength and specificity of the interaction.[4] In some cases, unstructured interaction domains facilitate highly regulated protein-protein interactions based on allostery or solution conditions.[5] As discussed in previous chapters, methods for accurately measuring the stoichiometric and conformational equilibrium of protein complexes are limited, and IM-MS methods provide can provide unique insight into equilibrium and structure simultaneously.[6-8] In the first sections of this chapter, work is presented where the equilibrium stoichiometry of several engineered protein complexes is assessed. In the last section, I discuss the use of the IM-MS modeling techniques developed in the previous chapters to elucidate gas-phase structures of ApoE tetramers

6.2 IM-MS Evaluates de novo -designed coiled coils as off-the-shelf components for protein assembly

Protein engineering seeks to build amino acid sequences either from scratch or as derivatives of existing sequences to access new structures and functions.[9] Control of protein-protein interactions is a longstanding goal within this field, as it may allow for modulation of enzymatic or regulatory activities of proteins, or directed assembly into macromolecular protein structures.[2] Although the composition of protein interfaces throughout nature is well understood, rational design of existing interfaces, and especially *de novo* design of interfaces remains extremely difficult.[10] One emerging technique for driving oligomerization of proteins leverages α-helical coiled-coil domains inserted at the ends of polypeptide chains.[11] The self-assembly of coiled-coil domains is one of the most well-studied protein-protein interaction types, where the sequence of the repeating canonical heptad of amino acids in the coil governs the interactions between the coils. Indeed, many sequence variants of coiled coils have been shown by crystallography to self-assemble into various oligomer numbers,[12] and software tools are also available for prediction of self-assembly based on variations to canonical sequences.[13] Despite the wealth of knowledge in this area, the oligomerization of coiled coils has not been examined in detail under the conditions commonly used for protein engineering.[14]

In this study, we use ion mobility mass spectrometry to evaluate protein complexes engineered to self-assemble using coiled coils. In the first experiments, we used green fluorescent protein as a monomeric model system to provide proof of concept for the use of coiled coils to form GFP oligomers. IM-MS is capable of assigning both the oligomerization states and the integrity of protein folds for samples

like these, as has been reported previously.[15] Next, we leveraged the symmetry of protein interactions within native esterase trimers to direct self-assembly into protein cages using coiled-coils. The resulting complexes are macromolecules between 400 and 800 kDa, consisting of at least 12 subunits. Some of the work presented here is previously published[1614]

6.1.1 Methods

All protein constructs were designed by the Marsh group (University of Michigan, Department of Chemistry) and expressed and purified using standard protocols [14] including SEC as the last step before IM-MS sample preparation. For green fluorescence protein constructs and tetrahedron constructs, samples were prepared for mass spectrometry concentrated to 40 µL. The minimum concentration of protein required for analysis was 1 µM. Samples were then loaded into gold plated needles prepared in house as previously described.[15] Nano-electrospray-ion-mobility-TOF mass spectrometry was performed using a Synapt G2 Traveling-Wave instrument (Waters Corp, Manchester, U.K.). Ions were generated by applying a voltage of 1.5 kV between the needle and the instrument source, with further voltage drops aiding in acceleration and desolvation as ions passed through the skimmer region of the instrument. The quadrupole region was set to RF-only mode for collection of complete mass spectra, and in some cases was tuned to isolate selected peaks for MS/MS analysis. A range of collision energies was tested for enhanced transmission and desolvation of the ions, and in some cases dissociation of the ion into its component subunits. The base values for collision energies were 20–50 V; however, energies up to 150 V were utilized for dissociation experiments. The IMS region of the instrument was operated at 4 mBar of

nitrogen, with wave heights and wave velocities of 15 V and 150 m s−1, respectively. The instrument time of flight mass analyzer was operated in sensitivity mode, and mass spectra were collected from 1000 to 15 000 m/z. Data analysis was performed using the manufacturer-provided Masslynx software.

In order to analyze octahedral structures, after SEC, samples were concentrated to ~5 mg/mL and then buffer-exchanged into 200 mM ammonium acetate, pH 7.0, using a Bio-spin P30 column (Bio-Rad, Inc.); 2–3 µL of the sample was loaded into glass capillary (approximate o.d. of 1.5–1.8 mm and wall thickness of 0.2 mm) before mounting to the source of an Exactive Plus EMR mass spectrometer (Thermo Fisher Scientific, housed in the Ohio State University Campus Chemical Instrument Center). An electrospray voltage of 1.2 kV was applied to the sample using a platinum wire inserted into the capillary, the source temperature was set to 175 °C, in-source CID was minimized to 1 V or 2 V, HCD was 20 V, the resolution was set to 17,500, and other instrument parameters were set as described previously. Data processing was preformed using MMass.[17]

6.2 Results and Discussion

6.2.1 Evaluation of GFP-coiled-coil oligomerization states To evaluate the propensity of coiled coils to drive the oligomerization of monomeric GFP, we analyzed the native MS spectra of several constructs coupling specific coiled coil sequences to the c-terminus of GFP (Figure 6.2.1 A) using the Q-ToF analyzer platform. Accurate mass measurements were obtained for the WT GFP monomer at 29,053 +/- 5 Da, and no oligomers were observed, indicating that effects from artifactual concentration-dependant ESI oligomers were neglible.[18] (Figure 6.2.1 B) When coiled-coils were

added to GFP at the C-terminus, however, clear signs of oligomerization were observed. New signals with masses around 65,000 and 98,000 Da are clearly observable in these spectra, indicating an

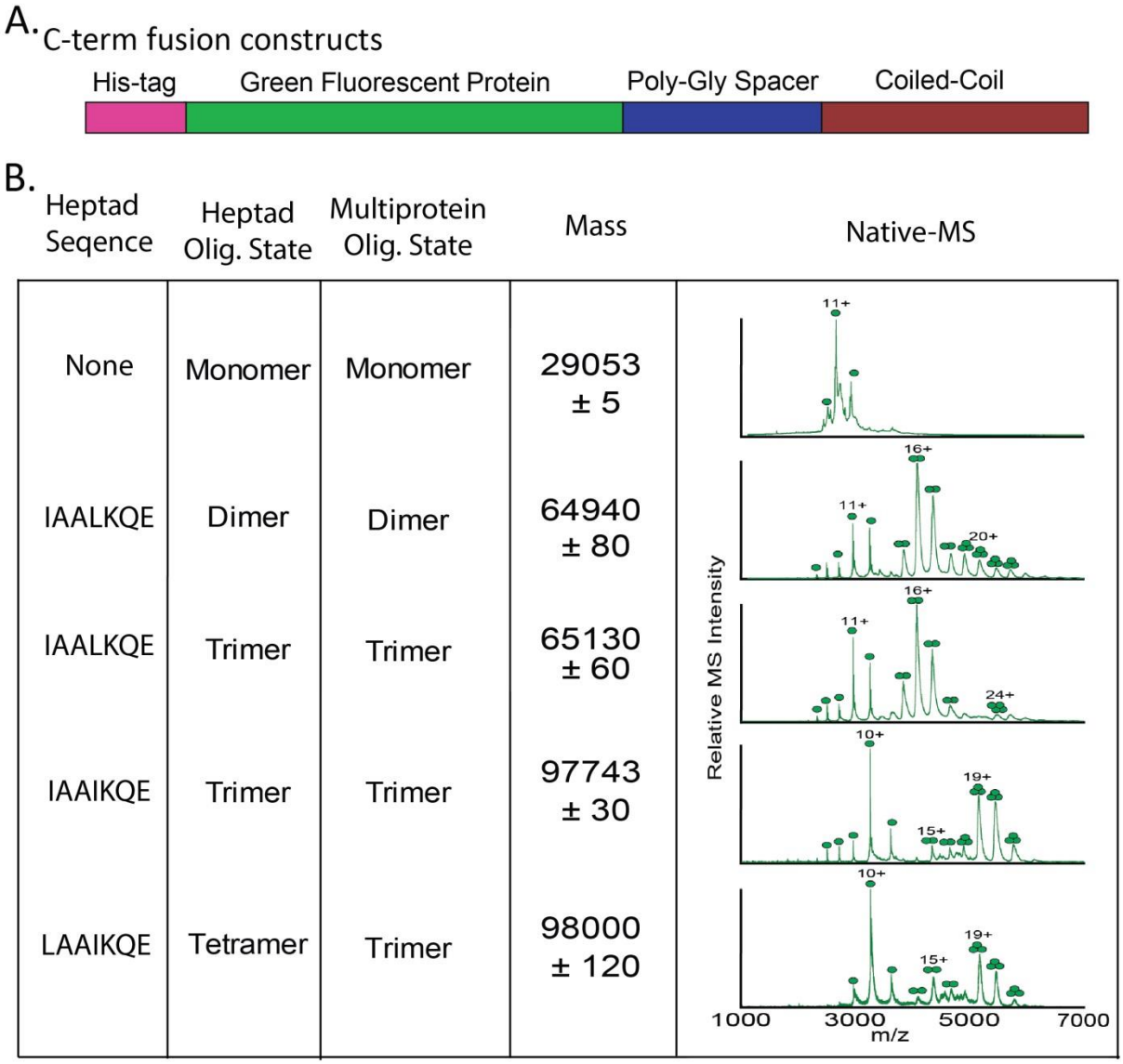


Figure 6- 1 Evaluation of Coiled Coils for Oligomerization of GFP monomers. A) Schematic of the amino acid sequences for the C-terminal fusion proteins screened in this study. B) Characterization of the Oligomerization of Fusion proteins containing Heptad Repeat Coiled Coils. The heptad oligermization state is the predicted oligomerization for coiled coils in isolation, whereas the multiprotein oligomerization state is the observed state from our experiments. The mass is column represents the mass of the most abundant oligomer in the sample.

equilibrium of monomer, dimer, and trimer populations exist when coiled coils are implemented. Generally, the oligomerization states observed GFP constructs correspond to those predicted from X-ray and computational analysis of the coiled coil sequences. Conversely, analysis of the LAAIKQE sequence by native MS revealed tetramers for the coiled coils alone but signals only up to trimer were for GFP constructs incoreprating this sequence. Overall, these results indicate that although most coiled coils operate as expected when assembling large protein complexes, steric effects may inhibit the formation of some higher order structures (e.g. tetramers).

6.2.2 Evaluation of the effects of coil position on GFP structure

Preservation of protein tertiary structure following sequence manipulation is a key factor in the success of protein engineering efforts, as the disruption of major structural components can lead to misfolding and uncontrolled aggregation. To further understand the effects of coiled coil domains on the native structure of GFP, we analyzed the transposition of the IAAIKQE GFP construct shown in the previous section to successfully form trimers. This transposition constructs were structured with the IAAIKQE heptad repeat domain now on the N-terminus of the GFP, as opposed to C-terminal coiled coils from Figure 6.2.1 A in the previous section. Surprisingly, we found that that placement of the coiled coil on the N-terminus greatly disrupts the GFP structure. Figure 6.2.2 A shows native state ion mobility and MS-only data for the N-terminal construct II4, which correlates directly to the transposition of IAAIKQE. Here, we observe that the transposition of the oligomerization domain to the N-terminus removes the assembly potential for these proteins, resulting in signal for only

monomeric subunits. We then characterized the effects of strengthening the interaction of the IAAIKQE n-terminal oligomerization domain by adding another heptad repeat. These studies showed that the addition of extra repeating units to these helices increased their ability to form larger oligmers substantially.[12] In figure 6.2.2 B we show

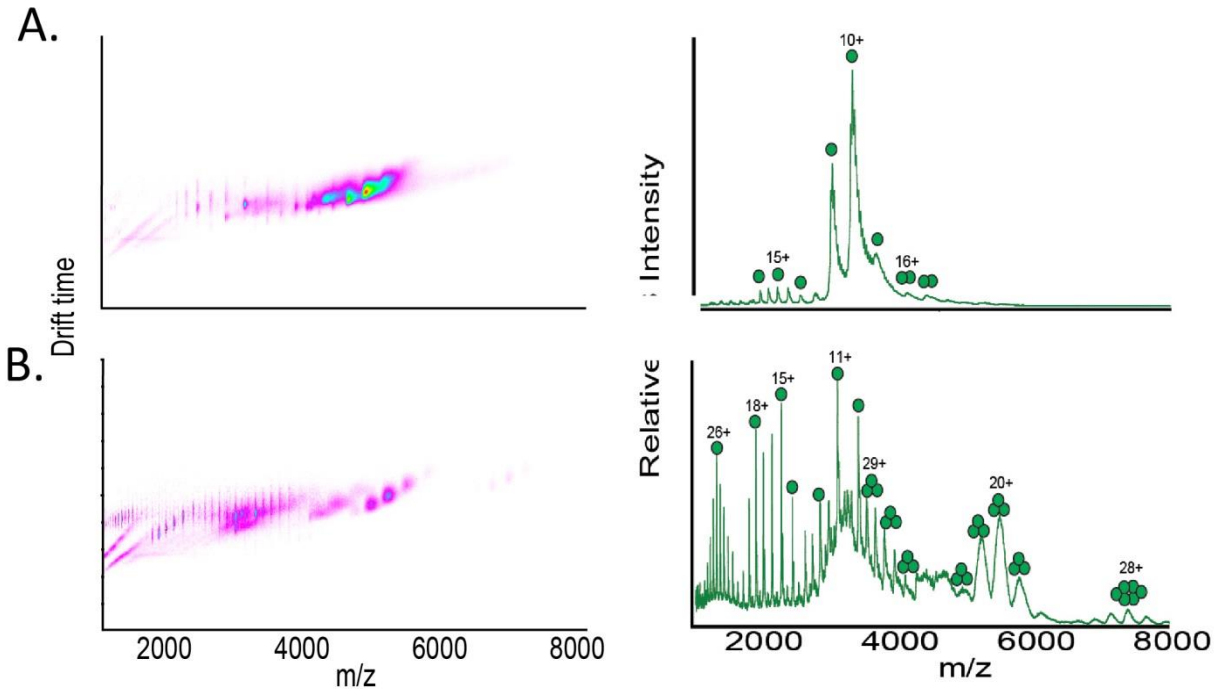


Figure 6- 2 Evaluation of the effects of coiled-coil fusion on protein structure. A) Transposition of the IAAIKQE 4-heptad repeat from the previous section to the N-terminus of GFP was evaluated, IM-MS data and standard MS data is provided to show oligomer distributions and relative signal intensities. B) A strengthened oligomerization domain featuring a 5-heptad repeat is evaluated.

that in this N-terminal construct the integrity of the GFP monomer structure appears to be affected, as evidenced by large amounts of highly charged, low m/z signals that indicate the increased surface area of the GFP monomers and are symptomatic of protein

unfolding.[19,20] Additionally, the IM-MS spectra for these constructs revealed increased drift times for these highly charged signals, providing further evidence of protein unfolding. Despite the predominance of signals associated with unfolded proteins in these spectra, we do observe evidence of assembled oligomers, including the targeted trimers. We take this result with caution, however, as we also observe hexameric complexes that indicate the presence of nonspecific aggregation in this sample.

6.2.3 Symmetry-directed Assembly of Protein Cages using Coiled Coils

After evaluating the fidelity of coiled coils for engineering the self-assembly of monomeric proteins, we set out to develop protein cages by leveraging the natural symmetry of a trimeric esterase. In this esterase, the c-terminus of each monomer is oriented toward the apex of the trigonal structure formed by the protein complex, making it a strong candidate for potential self-assembly into a protein cage with the addition of c-terminal coiled coils. We introduced the IAAIKQE coiled coil into the c-terminus with an 8-glycine linker and to introduce a C3 oligomerization element at the apex of each esterase trimer. The predicted geometry for the resulting self-assembled structure would be a tetrahedron, as shown in figure 6.2.3 A. Indeed, comparison of native MS spectra for a control fusion protein with non-oligomerizing coils (figure 6.2.3 B) and those with the IAAIKQE, coiled coils reveals high propensity for oligomerization into a monodisperse population of complexes with N = 12 subunits. Within the resulting native MS data, we observe a smaller population of esterase trimers with N = 3 subunits that did not oligomerize. We attribute this behavior to the solution-phase equilibrium for these interactions, however due to differences in ionization and transmission efficiency

in native MS analysis between N =3 and N =12, we predict the efficiency of the self-assembly process is higher than what is indicated in our data.[21,22]

Next, we undertook a similar study utilizing tetramer-forming LAAIKQE coiled coils to drive C4 symmetry around the c-termini of the esterase trimer, which we predict would result in an octahedral protein cage. Due to the size of this complex, we were not able to obtain charge-state resolution using the Synapt G2 IM-MS platform, making determination of m/z ratios for the signals observed impossible. The Orbitrap Exactive EMR platform is in some cases able to obtain better resolution for large complexes, due largely to its heated-capillary inlet that allows for more efficient desolvation of large ions. Using this platform, we observed signals corresponding to two Gaussian-like distributions in m/z, one

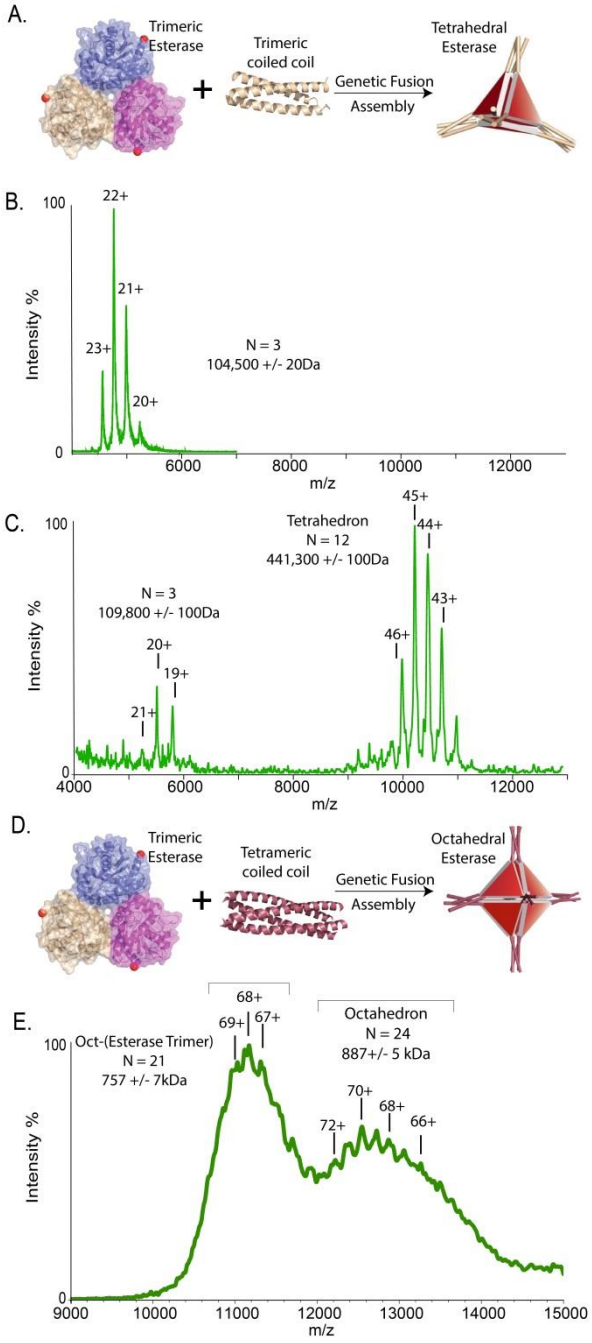


Figure 6- 3 Symmetry-directed Assembly of Protein Cages. A) Schematic of the symmetry directed approach. B) Control native MS data for the esterase trimer with “dummy” coiled coil fusions. C) Native MS data for the trimeric esterase fusion with XXX coiled coils. C) Native MS data for the trimeric esterase fusion with LAAIKQE coiled coils.

centered around 11,000 m/z and the other around 13,000 m/z. By increasing the in-source CID energy from 1V to 2V, we increased the observed intensity for the signals observed at 11,000 m/z substantially, indicating some amount of disruption for the intact assembly. However these settings also allowed for charge-state resolution of both distributions at around 90% of the max intensity. Deconvolution of the m/z peaks revealed the masses of the two distributions 757 ± 7 kDa and 887 ± 5kDa, corresponding to N = 21 and N = 24 subunits, respectively. With the help of cryo-electron microscopy, we verified that the N= 24 subunit structure did in fact form an octahedral geometry. SEC and ultracentrifugation data for these samples revealed a homogenous population, with masses corresponding to N =24, so we attribute our observation of N = 21 species to dissociation events occurring during the buffer exchange process for native MS, or perhaps an non-canonical gas-phase dissociation event, as the relative populations of these species were highly dependent on the in-source CID energies (data not shown).

6.2.4 Conclusions

In this section, we demonstrated the utility of native IM-MS for aiding in the study of protein self-assembly, specifically in the context of protein engineering. IM-MS provides sequence analysis to confirm the identity of an engineered construct, a snapshot of the equilibrium distribution of stoichiometries the sample, and information regarding the integrity of protein tertiary structures in a single experiment. Using this technology, dozens of engineered GFP protein variants were screened in a matter of a few hours, providing robust information to guide the next round of engineering. In the

future, higher throughput analyses may be possible using automated screening workflows.

6.3.0 Elucidating the Structure of Gas-Phase ApoE Tetramers

Apoplipoprtoein E (apoE) is the primary cholesterol transporter in the brain, and is implicated in both Alzheimer's and cardiovascular disease.[23] Genetic sequencing has correlated the presence of ApoE isoforms with these diseases, although at the chemical level, apoE isoforms only differ by amino acid substitutions at two sites.[24] The ε4 allele produces an apoE proteoform with arginines at sites 112 and 158, and has the strongest risk for Alzheimer's disease associated with it compared with ε3 and ε2 which replace arginine with cysteine at one or both sites, respectively. Lipid-free apoE is known to assemble into a homotetramer at physiological concentrations in the low micromolar range, and these associations are believed to be important in apoE pathogenesis due to the potential sequestration of lipid binding sites within the protein complex.[25] High resolution structural information on near-native apoE is limited to an NMR structure of a monomeric mutant in which 5 C-terminal residues were deleted to inhibit oligomerization.[26] In this structure, the C-terminal domain is revealed to be a helix that wraps around the largely globular N-terminal domain. Other studies, however, have provided evidence for more extended apoE structures, where the C-terminal domain is proposed to generate a new conformation that is independent of the N-terminus.[27-29]

In this study, we integrated IM-MS with molecular modeling, ECD and CIU to provide the first structural characterization of the ApoE tetramer. Here, we relied on IM-MS datasets collected by collaborators at Washington University in St. Louis to build a

coarse-grained model of apoE. As discussed in chapter 4, coarse-graining can introduce errors into a molecular model, and care must be taken to ensure that all structural elements are represented properly. In this case, we encountered a system where coarse-graining errors made it impossible to find models consistent with all of the input data, guiding us toward a higher-resolution depictation of the protein complex and a more informative result.

6.3.1 Materials and Methods

Coarse-grained modeling was performed using IM-MS_Modeller, a Monte-carlo annealing algorithm built within the Integrative Modeling Platform library in Python described in Chapter 4.Experimental data was input in the form of spherical radii corresponding to measured collision cross sections, and in some cases distance restraints between multiple spheres that optimized agreement with experimental cross sections. Ensembles of 10,000 models were generated according to scoring functions with and without C4 symmetry constraints, and models with tetrahedral symmetry were generated by applying the appropriate transformation to the C4 models. For each ensemble, we calculated CCS values using the IMPACT projection approximation and considered only models within +/- 3% of the experimental cross section. Models agreeing with all CCS datasets were then clustered into an average-linkage hierarchy using a pairwise RMSD matrix, and the probability density function of each cluster was estimated using gaussian kernels.

6.3.2 Results and Discussion

A typical IM-MS modeling workflow involves representation of protein subunits as spheres with radii corresponding to experimental CCS.[7] Additionally, common restraints on the interactions between subunits include rigidly defined distances, or connectivity restraints that are more ambiguous in terms of their absolute geometry. Moreover, symmetry may be incorporated in some models to increase sampling efficiency where symmetric elements are hypothesized. In the case of apoE, native IM-MS was performed for all variants and a population of species was observed, with tetramers as the dominant oligomer, accounting for at least 60% of the total signal. Monomers were observed at appreciable levels around 20-30% total intensity, with small amounts of dimer and trimer accounting for the remaining signals. Interestingly, we found no significant differences between the apoE isoforms in terms of oligomerization potential. The similarities between isoforms extended into the ion-mobility dimension, with each isoform giving rise to CCS values within 2% of the mean value of 2420 $Å^2$ and 7540 $Å^2$ for the monomer and tetramer, respectively.

A simple model of the apoE tetramer was build using spheres corresponding to monomeric CCS that employed only basic connectivity restraints, requiring that each subunit was connected to at least two other subunits. The rationale for this connectivity restraint was that a linear connectivity was unlikely given that oligomerization did not extend indefinitely through pentamers, hexamers, and other larger stoichiometry assemblies. After generating 1,000 putative structures using this set of restraints, we filtered these structures based on the tetrameric CCS as measured by the IMPACT projection approximation. Surprisingly, only 5 structures out of the ensemble of 1,000

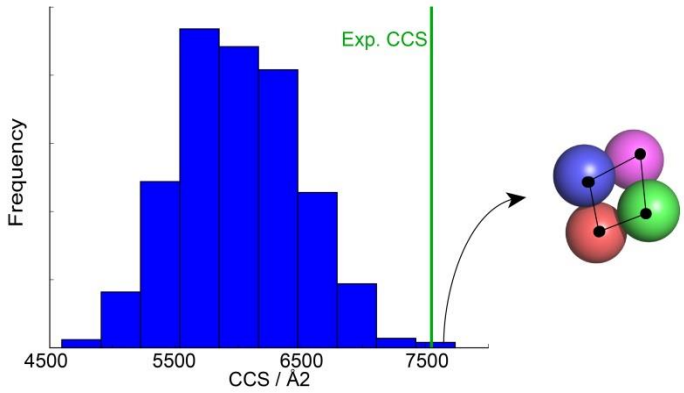
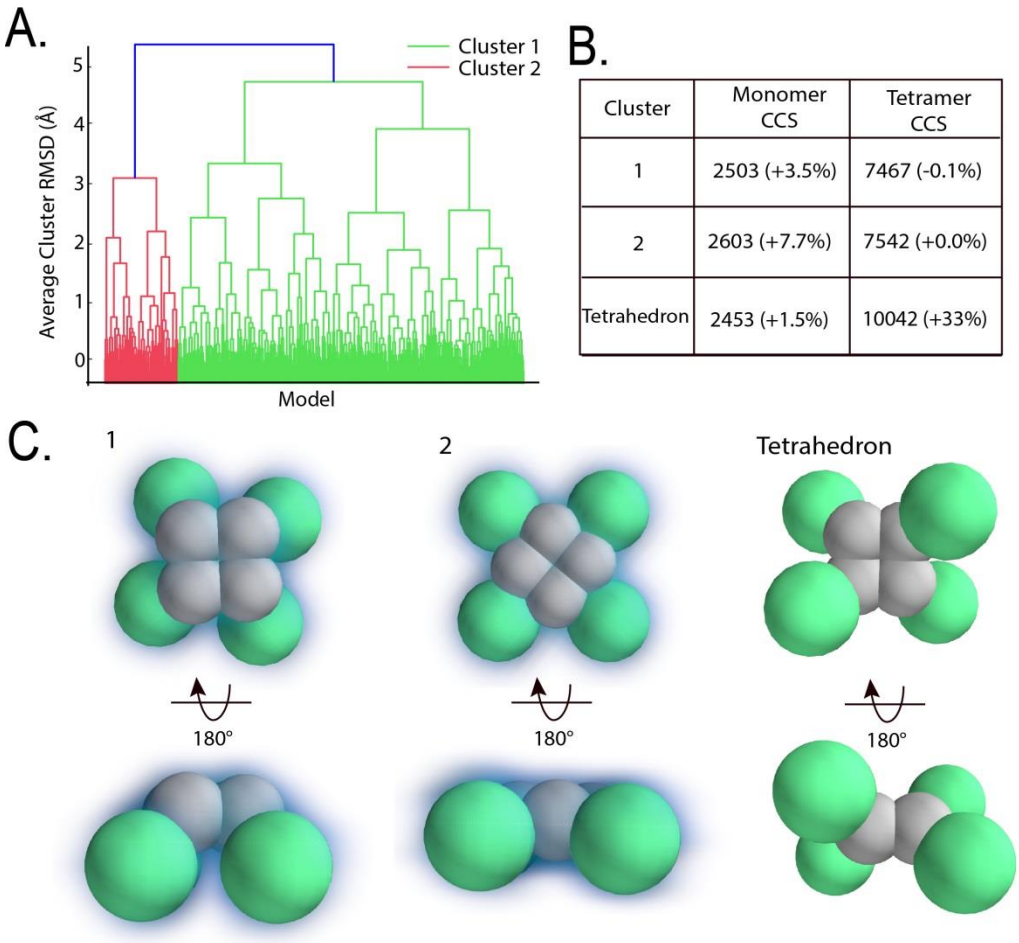


Figure 6- 4 CCS Distribution of apoE tetramers modeled at subunit resolution. The CCS distribution reveals that models generated from sbunit-level representations of the apoE monomer are significantly more compact than our experimental measurements suggest. An example of a high scoring model reveals that the inter-subunit distances within the most extended models in this ensemble do not agree with biophysical restraints on interprotein distances.

were within 3% of our experimental CCS. Further analysis of the ensemble of structures reveals that distribution of CCS values falls well below the experimental dataset, and only the most extreme CCS values approach our experimental result.(Figure 6.4) Analysis of the distances between each subunit in the group of structures agreeing with our experimental CCS revealed that although the connectivity restraint had been satisfied, connecting subunits only had a spherical overlap value of about 5%, well under the threshold for biologically relevant interactions, which is generally a minimum of 15%. From this result, we hypothesized that the assumption of sphericity, which holds for many globular, single domain proteins, is not appropriate for the dynamic two-domain structure of apoE.[29] That tetrameric apoE cannot be calibrated as a globular protein based on size-exclusion chromatography has also been reported earlier.[29] Therefore, we developed a two-domain model for the monomer subunit, which agrees well with our experimental CCSs, based on the proposed structures for WT.[27,29] We used a computational model[27] that has two distinct domains, in contrast to the NMR structure where the domains are intermeshed. Residue 183 was used to "cut" the domains to make a two-body structure. We then assembled

four copies of this two-domain monomer via connectivity restraints applied to the C-terminal domain for the generation of tetramer models. Modeling conducted with and without C4 symmetry enforced on the complex yielded similar structural ensembles that agreed with the recorded experimental CCS values. For simplicity, we present the results that possess C4 symmetry, and consider tetrahedral symmetry separately as a control. Through analyzing the dendrogram that represents hierarchical clustering of the



Cluster	Monomer CCS	Tetramer CCS
1	2503 (+3.5%)	7467 (-0.1%)
2	2603 (+7.7%)	7542 (+0.0%)
Tetrahedron	2453 (+1.5%)	10042 (+33%)

Figure 6- 5 Figure 6- 6 Modeling the apoE tetramer at domain resolution from IM-MS data. A.) Hierarchical clustering of models satisfying all experimental and biophysical restraints represented as a dendrogram with two structural families identified. B.) CCS values for median structures in each cluster agree with experimental results (relative error in parenthesis), with the exception of tetrahedral structures. C.) Median structures for each cluster (plus an example tetrahedral structure) are plotted within their kernel density functions to visualize the most probable structures and the ambiguity within each cluster.

ensemble of C4-symmetric structures that exhibit good agreement with all experimental data, two distinct families of structures are revealed, with each having a median structure which falls within 1% of the experimental CCS for the tetramer (Figure 6.5A and B). Visualizing the kernel density function and median structures of each family reveals the presence of planar and non-planar families. The most common family, the non-planar structure pictured as cluster 1 within Figure 6.5 C represents over 80% of the possible solutions as compared to the planar structure which represents the other 20%. To avoid biasing our modeling approach toward C4-symmetric structures, we transformed a selection of the C4 models into tetrahedral structural to measure the impact of such a step on the eventual model CCS values. For these tetrahedral structures, we observe that the resultant CCS increases to values on average 30% larger than our experimental result, strongly indicating that apoE does not produce tetrahedral symmetry.

6.3.3 Conclusions

Combined with evidence from other targeted structural biology techniques, IM-MS has provided the first structural model for the apoE tetramer. ECD datasets (data not shown) also complement our IM-MS-based models, where data indicates that the C-terminal regions of apoE undergo significant amounts of electron capture and dissociation, indicative of their relatively large exposed surface area. The IM-MS and ECD data discussed above are at odds with a previously-hypothesized tetrahedral geometry for the apoE tetramer, and point more strongly to parallel association along a C4 axis. Our analysis highlights the importance of properly representing protein domain structure within IM-MS derived models . Our inability to generate physical models that

matched our experimental data at the subunit-level of coarse graining led us to develop a two-domain model for the apoE monomer that was consistent with both the literature and our experimental results. Importantly, IM-MS based modeling of apoE tetramers did not lead to anunambiguous structural determination. In our analysis, two putative structures emerged as well-defined structural families within an ensemble of possible structures. We hope that reporting the ambiguity present in our data results in future hypothesis-driven structural studies using other experimental methods capable of properly annotating such uncertainty.

6.4 Acknowledgements

Work from section 6-2 was performed in collaboration with several graduate students from the Neil Marsh group at the University of Michigan. Ajitha Christie-David preformed the molecular biology and protein purification to generate all of the GFP-related constructs studied in this chapter. Somayesadat Badieyan and Aaron Sciore were responsible for designing the tetrahedral and octahedral protein cages, respectively. Section 6.3 was performed in collaboration with the Michael Gross group at the University of Washington in St. Louis. Hanliu Wang was responsible for providing the ion mobility datasets for modeling, and provided valuable insights into the structure of apoE along the way.

6.5 References

(1) Marsh, J. A.; Teichmann, S. A. In *Annu. Rev. Biochemistry* 2015; Vol. 84, p 551.
(2) Papapostolou, D.; Howorka, S. *Mol. BioSystems* **2009**, *5*, 723.
(3) Lai, Y.-T.; King, N. P.; Yeates, T. O. *Trends in Cell Biology* **2012**, *22*, 653.
(4) Shoemaker, B.; Panchenko, A. *PLoS Comput Biol* **2007**, *3*, e42.
(5) Jubb, H.; Blundell, T. L.; Ascher, D. B. *Prog. Biophys. Mol. Biol.* **2015**, *119*, 2.
(6) Pukala, T. L.; Ruotolo, B. T.; Zhou, M.; Politis, A.; Stefanescu, R.; Leary, J. A.; Robinson, C. V. *Structure* **2009**, *17*, 1235.
(7) Politis, A.; Park, A.; Hyung, S.-J.; Barsky, D.; Ruotolo, B. T.; Robinson, C. V. *PLoS ONE* **2010**, *5*.
(8) Zhong, Y.; Hyung, S.-J.; Ruotolo, B. T. *Expert Review of Proteomics* **2012**, *9*, 47.

(9) Channon, K.; Bromley, E. H. C.; Woolfson, D. N. *Curr. Opin. Struct. Biol.* **2008**, *18*, 491.
(10) Norn, C. H.; Andre, I. *Curr. Opin. Struct. Biol.* **2016**, *39*, 39.
(11) Lupas, A. N.; Gruber, M. In *Advances in Protein Chemistry*; Academic Press: 2005; Vol. Volume 70, p 37.
(12) Fletcher, J. M.; Boyle, A. L.; Bruning, M.; Bartlett, G. J.; Vincent, T. L.; Zaccai, N. R.; Armstrong, C. T.; Bromley, E. H. C.; Booth, P. J.; Brady, R. L.; Thomson, A. R.; Woolfson, D. N. *ACS Synthetic Biology* **2012**, *1*, 240.
(13) Wood, C. W.; Bruning, M.; Ibarra, A. Á.; Bartlett, G. J.; Thomson, A. R.; Sessions, R. B.; Brady, R. L.; Woolfson, D. N. *Bioinformatics* **2014**, *30*, 3029.
(14) Cristie-David, A. S.; Sciore, A.; Badieyan, S.; Escheweiler, J. D.; Koldewey, P.; Bardwell, J. C. A.; Ruotolo, B. T.; Marsh, E. N. G. *Mol. Systems Design & Engineering* **2017**.
(15) Ruotolo, B. T.; Benesch, J. L. P.; Sandercock, A. M.; Hyung, S.-J.; Robinson, C. V. *Nature Protocols* **2008**, *3*, 1139.
(16) Sciore, A.; Su, M.; Koldewey, P.; Eschweiler, J. D.; Diffley, K. A.; Linhares, B. M.; Ruotolo, B. T.; Bardwell, J. C. A.; Skiniotis, G.; Marsh, E. N. G. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 8681.
(17) Strohalm, M.; Kavan, D.; Novák, P.; Volný, M.; Havlíček, V. *Anal. Chem.* **2010**, *82*, 4648.
(18) Han, L.; Ruotolo, B. T. *Anal. Chem.* **2015**, *87*, 6808.
(19) Fernandez de la Mora, J. *Analytica Chimica Acta* **2000**, *406*, 93.
(20) Chowdhury, S. K.; Katta, V.; Chait, B. T. *J. Am. Chem. Soc.* **1990**, *112*, 9012.
(21) Liu, J.; Konermann, L. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 408.
(22) Gabelica, V.; Rosu, F.; De Pauw, E. *Anal. Chem.* **2009**, *81*, 6708.
(23) Puglielli, L.; Tanzi, R. E.; Kovacs, D. M. *Nat Neurosci* **2003**, *6*, 345.
(24) Gau, B.; Garai, K.; Frieden, C.; Gross, M. L. *Biochemistry* **2011**, *50*, 8117.
(25) Perugini, M. A.; Schuck, P.; Howlett, G. J. *J. of Biological Chemistry* **2000**.
(26) Garai, K.; Mustafi, S. M.; Baban, B.; Frieden, C. *Protein Science* **2010**, *19*, 66.
(27) Hsieh, Y.-H.; Chou, C.-Y. *J. of Biomedical Science* **2011**, *18*, 4.
(28) Chen, J.; Li, Q.; Wang, J. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 14813.
(29) Hatters, D. M.; Peters-Libeu, C. A.; Weisgraber, K. H. *Trends in BioChem. Sciences* **2006**, *31*, 445.

Chapter 7: Conclusions and Future Directions

This dissertation has focused on new contributions to gas-phase structural biology and drug discovery by way of two major techniques: Multiprotein complex modeling via IM-MS datasets and the study of proteins and protein-ligand complexes by CIU. In both of these areas, significant progress has been made in designing experimental workflows as well as informatics tools for extracting maximal information from complex systems and processes.

7.1 Findings and Future Directions for Integrative Modeling of Multiprotein Complexes

Our studies of multiprotein complexes have focused on the utility of IM-MS datasets in restraining the connectivity and structure of heterogeneous protein complexes. Although previous work in this area has provided proof-of-concept for such studies,[1-3] we were unable to confidently model some reference complexes based on simulated IM-MS data using current approaches, indicating that more development had to take place before IM-MS data could be widely used for protein topology assignment. The studies in this dissertation highlight the power of IM-MS to restrain models of multiprotein complexes; however, we are also approaching a clearer definition of the limitations of the technique. Although we envisage IM-MS being used in conjunction with other datasets as part of a much broader integrative modeling workflow,[4-6] the details of IM-MS-only datasets must first be worked out before

understanding how this data can complement other types of information. In Chapter 4, we undertook a series of bioinformatics experiments focused on the ability of IM-derived CCS restraints to accurately restrain the topology of multiprotein complexes. We found that one limitation is the use of heavily coarse-grained models at the subunit resolution leads to the introduction of CCS errors in model structures. On the other hand, we find that only modest increases in resolution, where subunits are modeled at the domain level, can nearly eliminate coarse-graining errors in the context of protein complexes. In Chapter 6, we outlined a method for generating domain-level coarse grained models from partial high-resolution structures and bioinformatics. When this information is not available, however, CIU may be a powerful tool for assignment of domain structure, and even the sizes of such domains. Previous work in the lab, and work from chapter 2 and 3 in this dissertation has found high correlation between CIU and protein domain structure,[7] and as we continue to gain understanding of the mechanism of CIU, we believe this data could be directly integrated into domain-level models of protein subunits.[8]

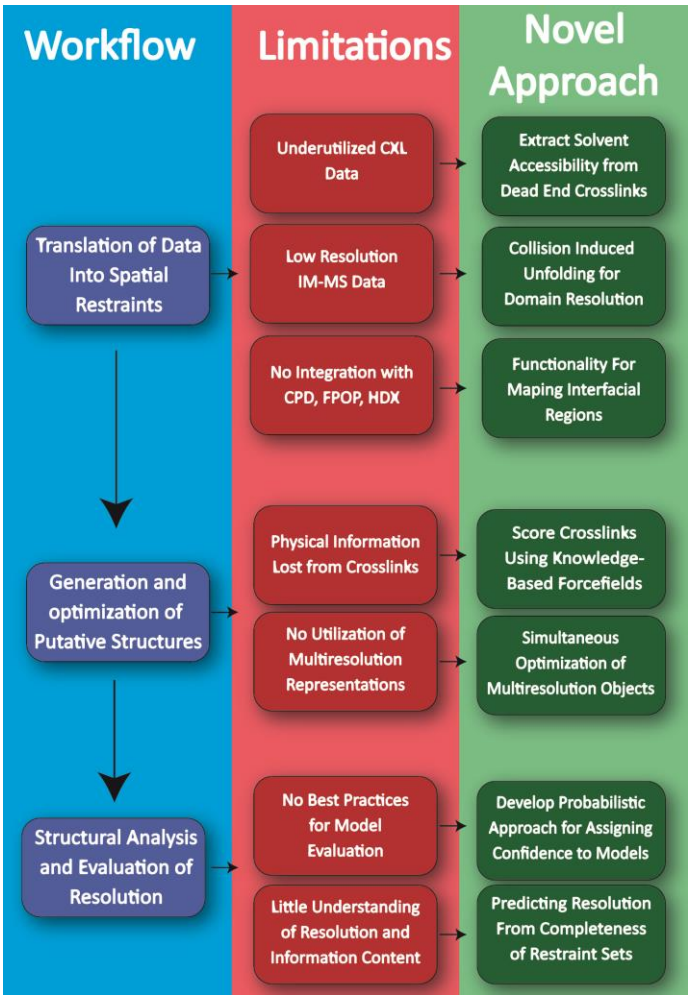


Figure 7- 1 Challenges in Integrative Modeling of Protein Complexes From MS datasets.

In order to sample structures that fit experimental datasets with more confidence, we developed a novel Monte Carlo algorithm[9] that efficiently samples the entirety of structural spaced defined by the input restraints. Using this tool, we defined the information content of CCS measurements in terms of their effects on the positive predictive power when modeling known complexes. Moreover, the effects of symmetry and modularity were explored, pointing toward new directions for the integrative modeling field. We envisage a platform for facile integration of MS-based datasets to understand protein complex structure. In Figure 7-1, we outline current workflows, challenges, and potential solutions for such a platform. The studies in this dissertation have focused on predicting the resolution and accuracy of models based on the completeness of the input restraints, as well as utilizing a probabilistic approach for assigning confidence to models. In the immediate future, multiresolution objects must be integrated into this workflow in order to harness the power of covalent labeling experiments such as HDX,[10] FPOP,[11] and CXL.[12] Once these multiresolution objects are implemented and restrained by experimental data, the structural landscape to explore will become much more complex, and so we also see the need for more rigorous investigation of optimization algorithms incorporating replica exchange[13,14] or dynamic weighting of restraints.[15]

7.2 Conclusions and Future Directions for CIU as a Structural and Drug Discovery Tool

Work in this dissertation focused on understanding the CIU process and importantly, developing robust data analysis tools to help bring CIU measurements toward the mainstream of structural biology tools.[16] In chapter 2, we demonstrate our

software package CIUSuite that allows users from many IM-MS groups worldwide to batch process large amounts of data, and draw quantitative comparisons between the unfolding processes of different analytes. Although the principles of CIU used in the context of drug screening have been demonstrated,[17-19] widespread adaptation the technology is hindered by a lack of understanding of the physical process of CIU, and specifically how it relates to solution-phase structure.

In Chapter 3, we investigated the mechanism of unfolding for a series of multidomain serum albumins.[8] In our studies we found that the stability of unfolded intermediates of protein ions is sensitive to subtle changes in primary structure, but the overall CIU pathway is dependent on the tertiary structure and specifically the domain organization of the protein. Additionally, we found that the ejection of small molecules during the CIU process was highly correlated with specific structural transitions, which allowed us to develop a model of albumin unfolding based on CID of small molecules of known binding locations. These studies not only shed light on the unfolding processes of multidomain proteins, but indicated new applications of CIU and CID for studying localized changes in stability or ligand binding.

To fully understand CIU processes at the mechanistic level, more detailed case studies of proteins of different tertiary structures must be conducted. Importantly, it is extremely difficult for these studies to provide information at the atomic level, and therefore it is necessary to develop molecular dynamics tools to that capture complex ion heating and charge migration processes in the gas-phase.[20] Although the effects of charge state on the CIU of proteins is understood broadly,[21-24] it is still difficult to predict

how the presence or absence of charge may affect the information content of the experiment.

Future directions for the analytical side of this experiment include development of new technologies for increasing the throughput and information content of the experiment. A major limitation for CIU is the requirement that single charge states be isolated and unfolded to avoid convolution of signals from charge stripped products. Although this workflow increases the confidence in the observed unfolding fingerprints, it either throws out useful information from other charge states, or vastly increases the

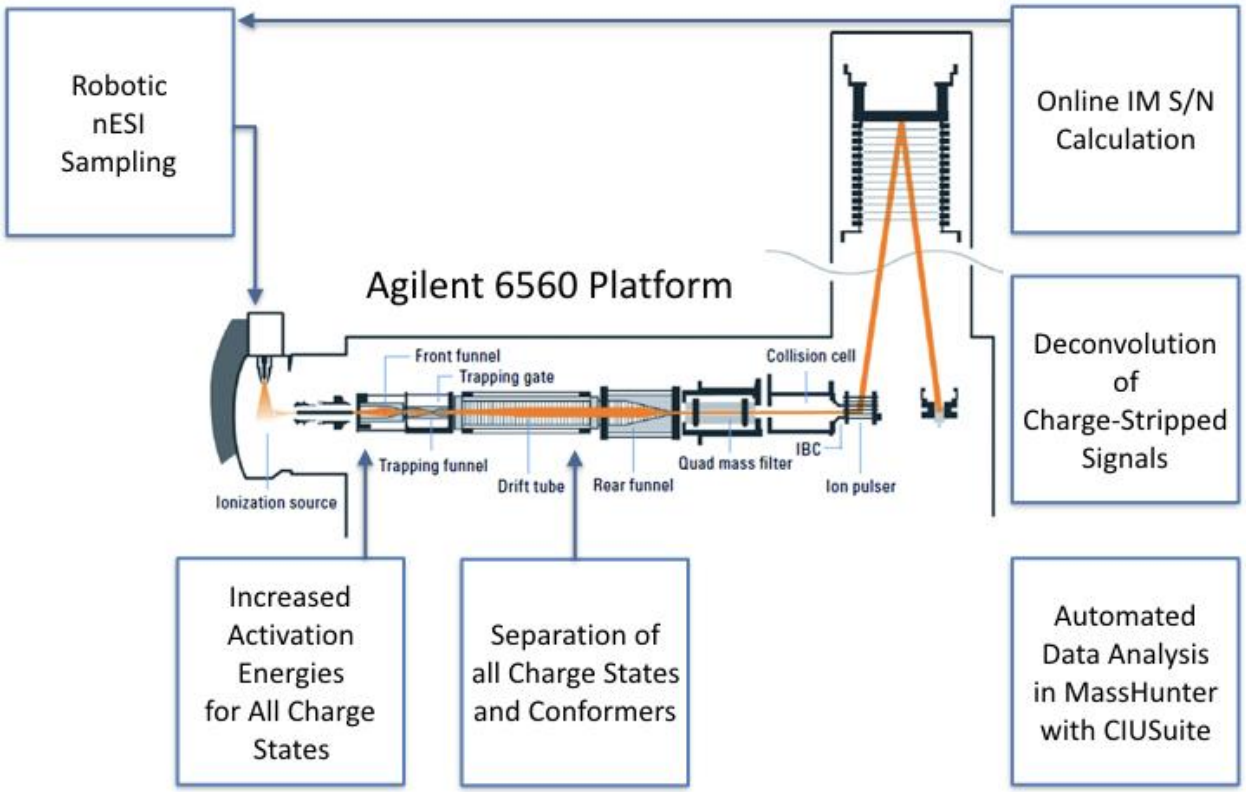


Figure 7- 2 Adaptation of the Agilent 6560 IM-MS platform for high throughput, charge multiplexed CIU experiments. We envision integration of instrument control software with data analysis tools, allowing for maximum acquisition efficiency. Further improvements in ion activation and IMS resolution will also be required for maximizing the potential for high throughput CIU.

time of the experiment if CIU for more charge states is required. To obviate these problems, an algorithm for deconvolution of signals from charge-stripped ions should be

developed that allows CIU fingerprints for all observable charge states to be collected simultaneously. Moreover, automated sample handling and online signal-to-noise detection are key improvements to the CIU workflow that may lead to truly high throughput applications.

Recent work has focused on developing CIU workflows for the Agilent 6560 IM-MS platform.[25] This platform has the potential for high-throughput CIU experiments, although many parts of the process require further development. Figure 7-2 provides a schematic of the platform, annotated with several necessary advancements that are required for reaching maximum throughput and CIU information content. The first consideration is integrating automated sample handling and instrument control with online data processing, allowing for great reductions in acquisition time. This improvement will allow the signal-to-noise ratio of the acquired data to dictate the actions of the instrument: when sufficient signal-to-noise is achieved, the instrument will automatically move to a new collision energy or to the next sample; if the signal drops to minimal levels, the instrument can automatically reload the sample. Other challenges for this instrument will include engineering of the capillary region of the instrument to maximize the collision energies achieved, as they are currently not capable of fully activating large ions. Additionally, optimal conditions need to be found for separating unfolded ions across multiple charge states simultaneously.

All of these engineering advances will be coupled with new features within CIUSuite. These features will include novel data mining techniques that may expand into the areas of Gaussian modeling and even machine learning. In the near term, however, the field should focus on identifying regions of maximal heterogeneity within

CIU fingerprints to decrease the time spent acquiring signal at less informative collision energies. With a combination of the tools mentioned in this section, we feel that the time requirements for CIU experiments can be decreased from hours to seconds per sample.

7.3 References

(1) Pukala, T. L.; Ruotolo, B. T.; Zhou, M.; Politis, A.; Stefanescu, R.; Leary, J. A.; Robinson, C. V. *Structure* **2009**, *17*, 1235.

(2) Politis, A.; Park, A.; Hyung, S.-J.; Barsky, D.; Ruotolo, B. T.; Robinson, C. V. *PLoS ONE* **2010**, *5*.

(3) Hall, Z.; Politis, A.; Robinson, C. V. *Structure* **2012**, *20*, 1596.

(4) Russel, D.; Lasker, K.; Webb, B.; Velázquez-Muriel, J.; Tjioe, E.; Schneidman-Duhovny, D.; Peterson, B.; Sali, A. *PLoS Biol.***2012**, *10*.

(5) Politis, A.; Borysik, A. J. *Proteomics* **2015**, *15*, 2792.

(6) Marcoux, J.; Cianferani, S. *Methods* **2015**, *89*, 4.

(7) Zhong, Y.; Han, L.; Ruotolo, B. T. *Angew. Chem.* **2014**, *126*, 9363.

(8) Eschweiler, J. D.; Martini, R. M.; Ruotolo, B. T. *J. Am. Chem. Soc.* **2017**, *139*, 534.

(9) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem.Phys.***1953**, *21*, 1087.

(10) Zhang, Z.; Smith, D. L. *Protein Science* **1993**, *2*, 522.

(11) Hambly, D. M.; Gross, M. L. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 2057.

(12) Shallan, M.; Radau, B.; Salnikow, J.; Vater, J. *Biochimica Et Biophysica Acta* **1991**, *1057*, 64.

(13) Swendsen, R. H.; Wang, J.-S. *Physical Review Letters* **1986**, *57*, 2607.

(14) Sugita, Y.; Okamoto, Y. *Chem. Phys. Letters* **1999**, *314*, 141.

(15) Wong, W. H.; Liang, F. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94*, 14220.

(16) Eschweiler, J. D.; Rabuck-Gibbons, J. N.; Tian, Y.; Ruotolo, B. T. *Anal. Chem.* **2015**, *87*, 11516.

(17) Hyung, S.-J.; Robinson, C. V.; Ruotolo, B. T. *Chem. Biol.* **2009**, *16*, 382.

(18) Niu, S.; Ruotolo, B. T. *Protein Science* **2015**, *24*, 1272.

(19) Rabuck, J. N.; Hyung, S.-J.; Ko, K. S.; Fox, C. C.; Soellner, M. B.; Ruotolo, B. T. *Anal. Chem.* **2013**, *85*, 6995.

(20) Popa, V.; Trecroce, D. A.; McAllister, R. G.; Konermann, L. *The J. of Physical Chemistry B* **2016**, *120*, 5114.

(21) Shelimov, K. B.; Jarrold, M. F. *J. Am. Chem. Soc.* **1997**, *119*, 2987.

(22) Valentine, S. J.; Anderson, J. G.; Ellington, A. D.; Clemmer, D. E. *The J. of Physical Chemistry B* **1997**, *101*, 3891.

(23) Badman, E. R.; Myung, S.; Clemmer, D. E. *J Am Soc Mass Spectrom* **2005**, *16*, 1493.

(24) Hopper, J. T.; Oldham, N. J. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1851.

(25) Kurulugama, R. T.; Darland, E.; Kuhlmann, F.; Stafford, G.; Fjeldsted, J. *Analyst* **2015**, *140*, 6834.

Appendix I: Supporting information for Chapter 3

Table I-1. Comparison of Albumins used in this study

Species	Sequence Mass	Experimental Mass(Da)	Experimental CCS(A^2)
Bovine	66432	66453	4060*
Human	66472	66460	4060
Sheep	66327	66384	4050
Goat	66313	66361	3940
Pig	66797	67045	4170
Rabbit	66015	66150	4020
Rat	65916	66024	3940

*Literature CCS 4100 A^2

Table I-2. Protein Calibrants used for CCS Calculation[1]

Calibrant Protein	Mass	Charge	m/z	Lit CCS
avidin	64000	16	4001.01	3640
avidin	64000	17	3765.71	3640
Concanavalin A	103000	20	5151.01	5550
Concanavalin A	103000	21	4905.77	5550
Concanavalin A	103000	22	4682.83	5480
Concanavalin A	103000	23	4479.27	5450
Alcohol Dehydrogenase	143000	25	5721.01	6830
Alcohol Dehydrogenase	143000	26	5501.01	6720

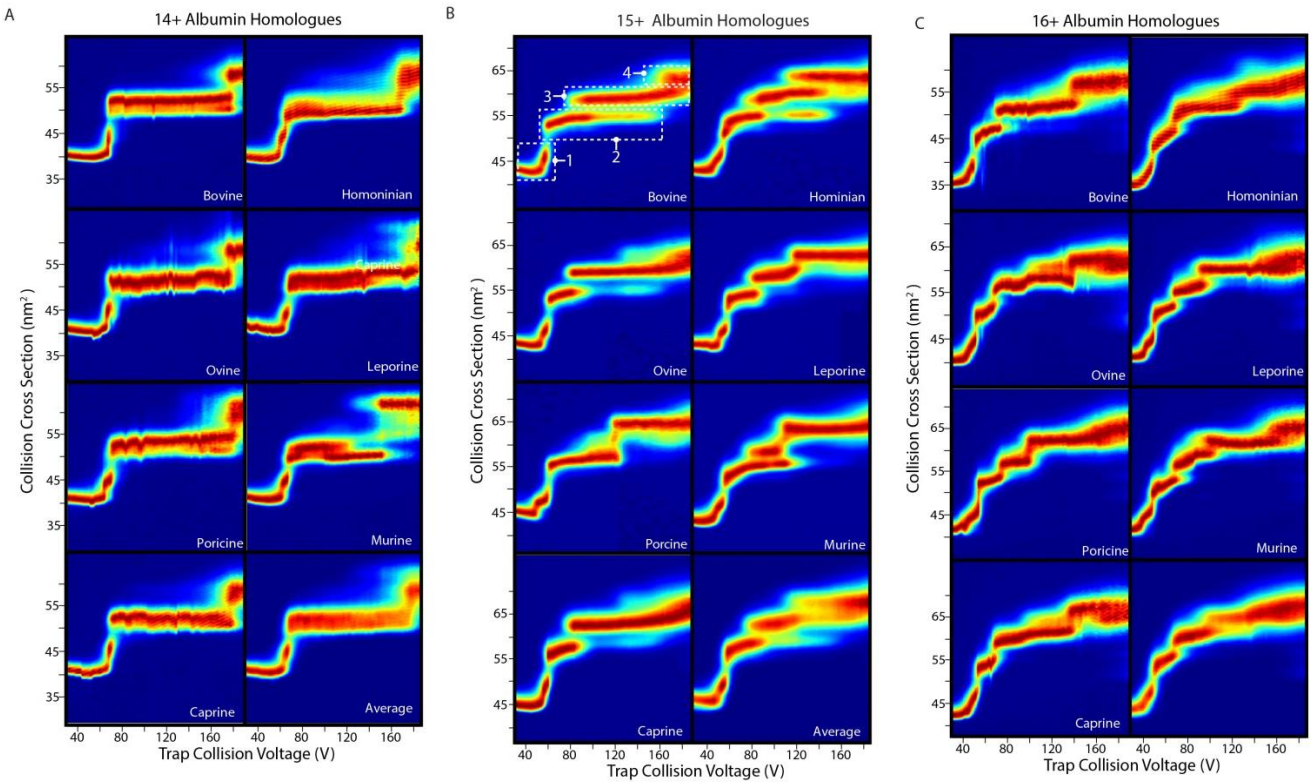
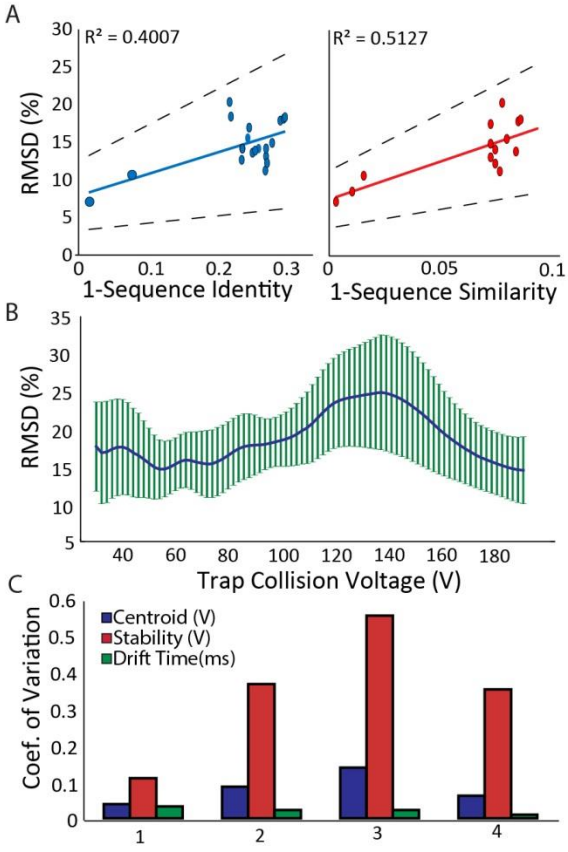


Figure I-1. CIU Fingerprints of All Homologues for 14+, 15+ and 16+ charge states

Supplementary Discussion: CIU Feature Assignments

For our analysis of the CCS and stability of folded and unfolded features observed in our CIU fingerprints, we used qualitative and quantitative (CIUSuite) metrics for deciding what was counted and not counted as a feature. Feature 1 is L shaped, and we have chosen not to treat the tail as a separate feature because it is not always differentiated from the base of feature 1, it does not remain stable for more than 10V in any of our datasets, and presents as a contiguous feature in drift time/energy space. While this feature represents more than one conformational family, we have constructed our analysis paradigm so that all elements of this L-shaped signal are counted as a single, unresolved feature.

In our Human Albumin dataset, we also note a reappearance of signal in the drift time range of feature 2 around 120V. This signal is artificially amplified due to the normalization procedures we use in CIUSuite to project our fingerprint plots. As collision energy is increased, the signal intensity for all ions decreases due to CID pathways. Thus, we attribute this signal as an fragmented structural population related to feature 2. We note that the in other homologues, e.g. BSA, this anomalous feature occupies a similar range.



A. Pairwise Comparison of CIU Fingerprints. We calculated pairwise RMSD values for all CIU fingerprints, and plotted these values against differences in albumin primary structure, as measured by sequence identity and similarity.

B. Regions of the Fingerprint Enriched for Sequence-Dependent Variation. Comparing all pairwise RMSD values revealed that the area from 120 to 140V showed higher incidence of significant deviations in the fingerprint.

C. Differences in Stability for Individual Conformers Drive Sequence-Dependent Variation in CIU. After quantitative measurement of drift times, centroid voltages, and stabilities of each feature, we found that feature-specific stability drove the vast majority of deviations in CIU. Most ablumins studied underwent near-identical unfolding pathways, achieving the same set of intermediate conformers, as defined by centroid voltage (V) and Drift Time (ms), as evidenced by the coefficient of variation for these two metrics.

Figure I-2. Quantitative analysis of CIU Differences across Homologues

Table I-3. Pairwise RMSD matrix for Albumin Homologues Generated by CIUSuite_compare

	Bovine	Bovidian	Hominian	Porcine	Leporine	Murine	Ovine
Bovine	0	-	-	-	-	-	-
Bovidian	11.7	0	-	-	-	-	-
Hominian	14.1	16.9	0	-	-	-	-
Porcine	17.5	20.3	12.6	0	-	-	-
Leporine	14.9	13.1	13.6	13.9	0	-	-
Murine	17.9	18.1	14.1	14.1	11.2	0	-
Ovine	8.5	7.2	15.5	18.4	12.3	18.3	0

Table I-4. Pairwise Sequence Identities for Albumin Homologues

	Bovine	Hominian	Ovine	Porcine	Murine	Leporine	Bovidian
Bovine	1	-	-	-	-	-	-
Hominian	0.758	1	-	-	-	-	-
Ovine	0.922	0.75	1	-	-	-	-
Porcine	0.792	0.759	0.775	1	-	-	-
Murine	0.702	0.734	0.695	0.723	1	-	-
Leporine	0.714	0.743	0.722	0.739	0.724	1	-
Bovidian	0.921	0.748	0.985	0.777	0.697	0.723	1

Table I-5. Pairwise Sequence Similarities for Albumin Homologues

	Bovine	Hominian	Ovine	Porcine	Murine	Leporine	Bovidian
Bovine	1	-	-	-	-	-	-
Hominian	92.8	1	-	-	-	-	-
Ovine	99	92.3	1	-	-	-	-
Porcine	93	91.1	92.8	1	-	-	-
Murine	91.8	93.5	91.4	91.6	1	-	-
Leporine	93	94.5	92.8	91.9	92.6	1	-
Bovidian	98.5	92.5	99.7	92.5	91.7	93	1

Table I-6. Raw numerical output from CIUSuite_Detect

	Feature1 Centroid	Feature1 Stability	Feature1 dt
Bovine	38	24	15.85
Goat	37	22	15.196
Human	38	24	16.83
Pig	35	18	16.38
Rabbit	37	22	15.56
Rat	34	18	16.29
Sheep	37	22	15.743
Avg.	36.57142857	21.42857143	15.97842857
Std.	1.399708424	2.32115383	0.513073054
	0.038273277	0.108320512	0.032110358
	Feature2 Centroid	**Feature2 Stability**	**Featuer2 dt**
Bovine	76	36	23.3
Goat	69	22	22.566
Human	72	32	22.93

Pig	89	54	24.297
Rabbit	72	20	22.839
Rat	72	32	23.387
Sheep	70	20	23.114
Avg.	74.28571429	30.85714286	23.20471429
Std.	6.340668863	11.20495517	0.515585025
	0.085355158	0.363123547	0.022218978
	Feature3 Centroid	**Feature3 Stability**	**Featuer3 dt**
Bovine	132	84	27.2
Goat	134	108	27.6
Human	113	50	27.2
Rabbit	102	36	26.3
Rat	94	12	26.2
Sheep	134	108	27.8
Pig			
Avg.	118.1666667	66.33333333	27.05
Std.	16.1494754	36.33944903	0.604841577
	0.136666929	0.54783089	0.022360132
	Feature4 Centroid	**Feature4 Stability**	**Featuer4 dt**
Bovine	179	18	30
Human	159	58	30
Rabbit	155	66	29.2
Rat	152	72	29.7
Pig	155	66	29.8
Sheep			
Goat			
Avg.	160	56	29.74
Std.	9.757048734	19.51409747	0.293938769
	0.060981555	0.348466026	0.009883617

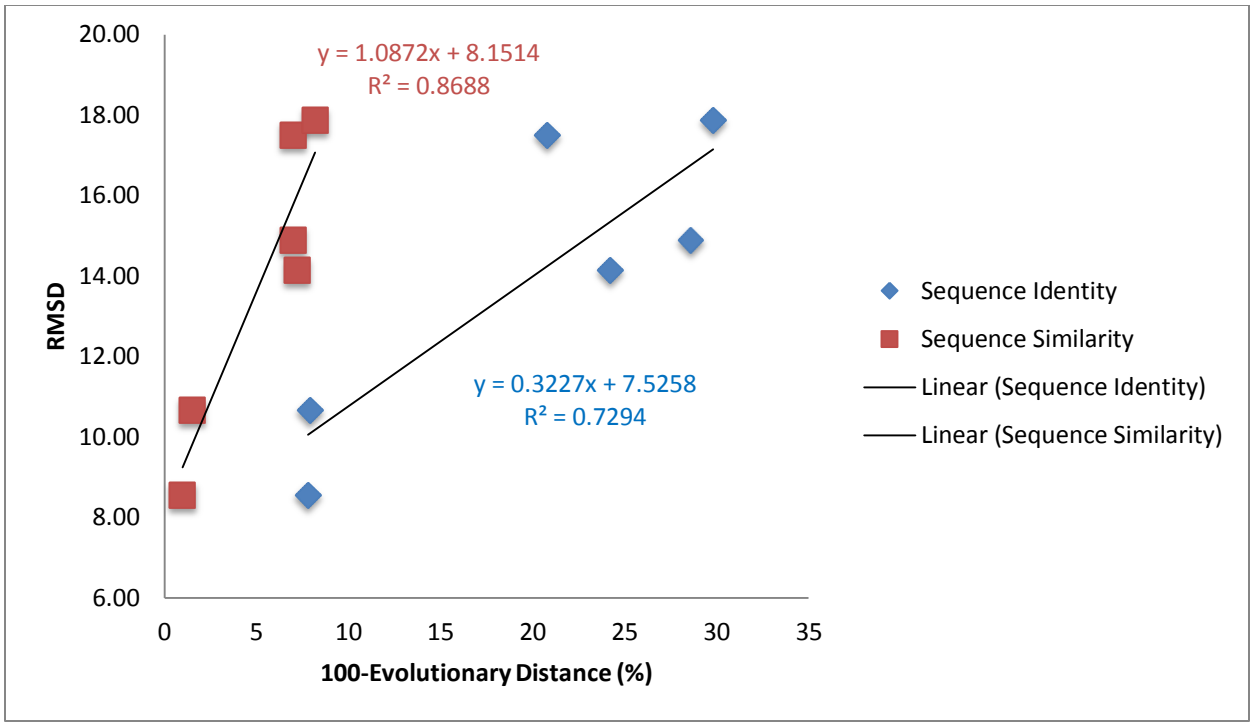


Figure I-3. Correlations between a BSA-based evolutionary distance and the CIU RMSD for all albumin homologues.

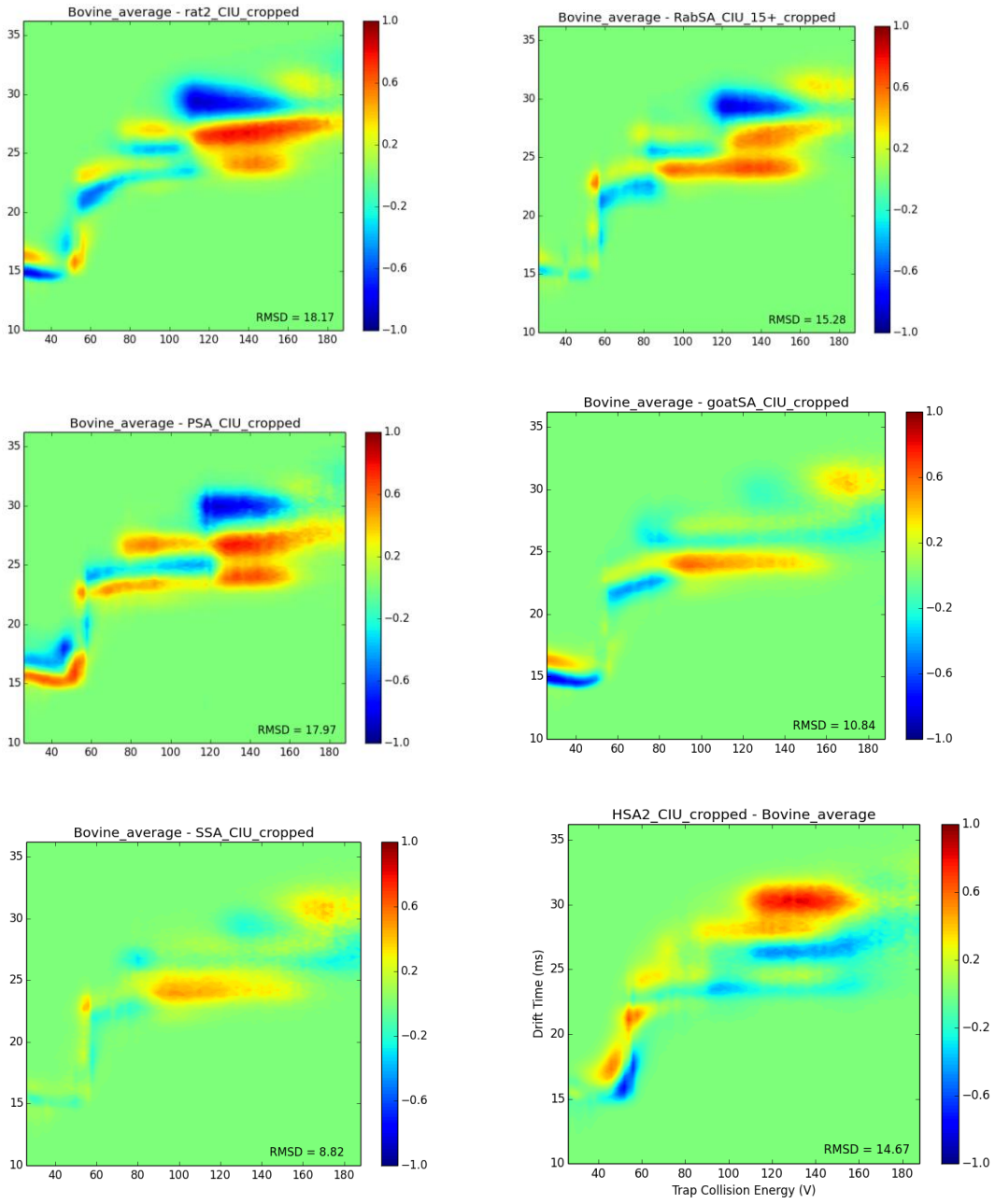


Figure I-4. Raw output from CIUSuite_compare, comparing Bovine and Human Albumins to other species

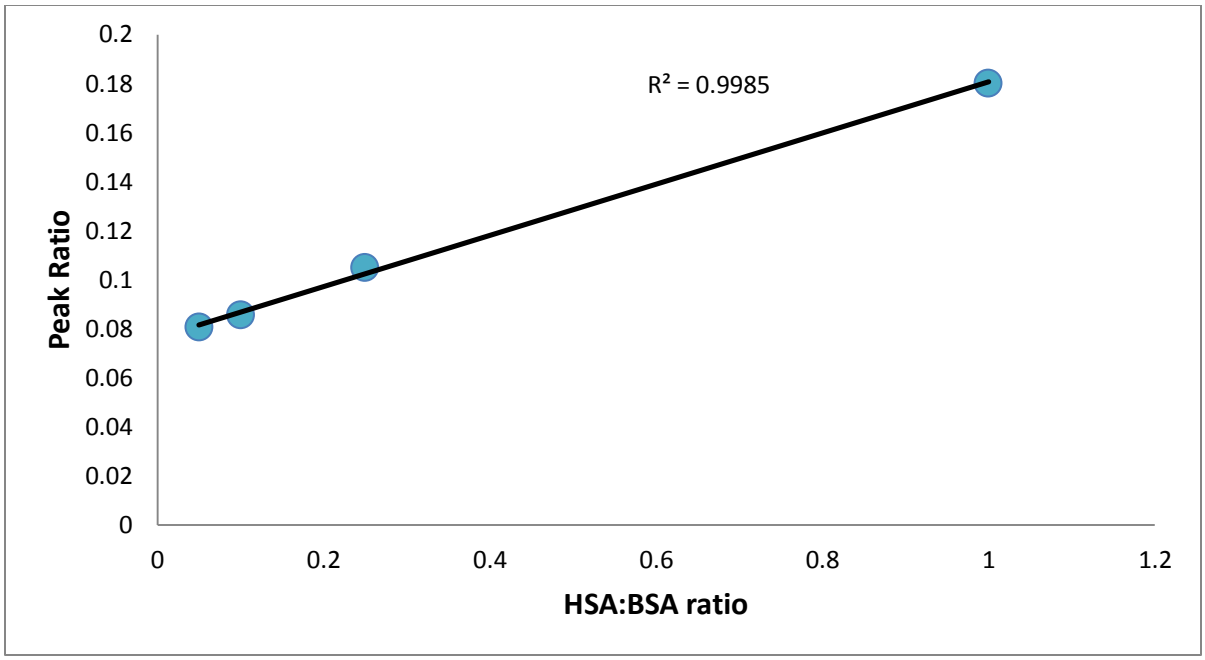


Figure I-5. Quantitation of BSA:HSA Ratio Using Peak Ratio of Unfolded States Present at 120V

Section 2: Supporting Information for Albumin-Ligand complexes and dissociation curves

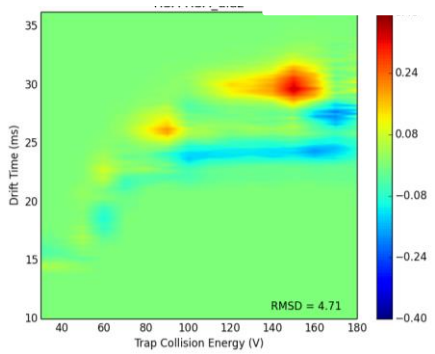


Figure I-6. A CIU difference plot from CIUSuite, showing that HSA bound to diazepam (blue) stabilizes late transitions relative to apo HSA (red).

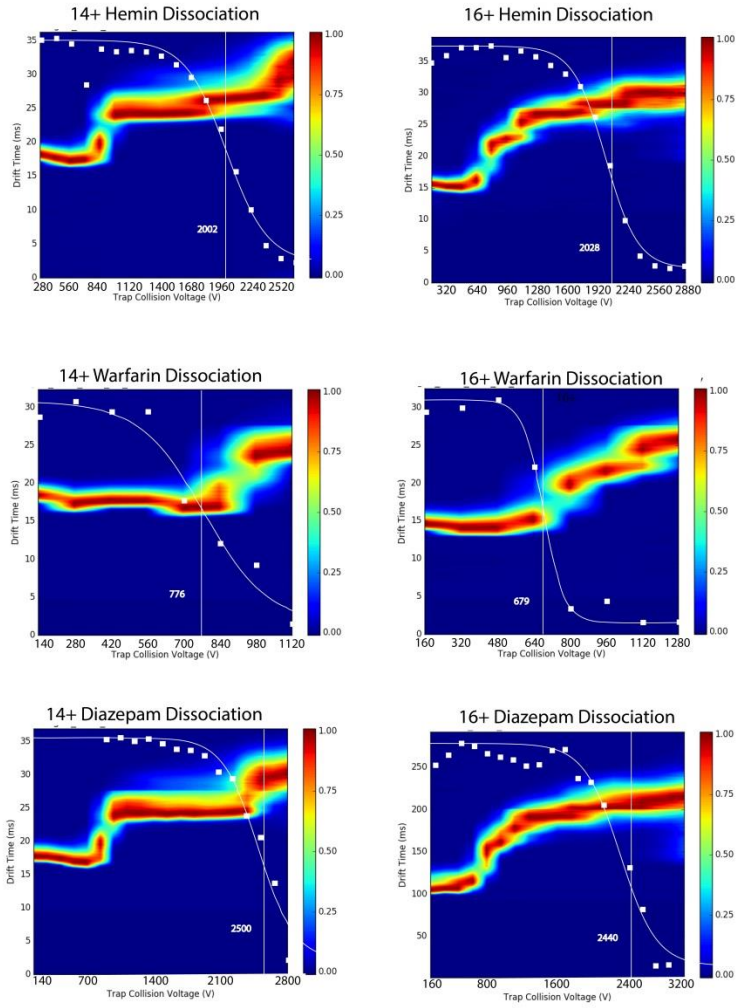


Figure I-7. CIU and CID Datasets for HAS bound to selected Ligands at 14^{+} and 16^{+} charge state

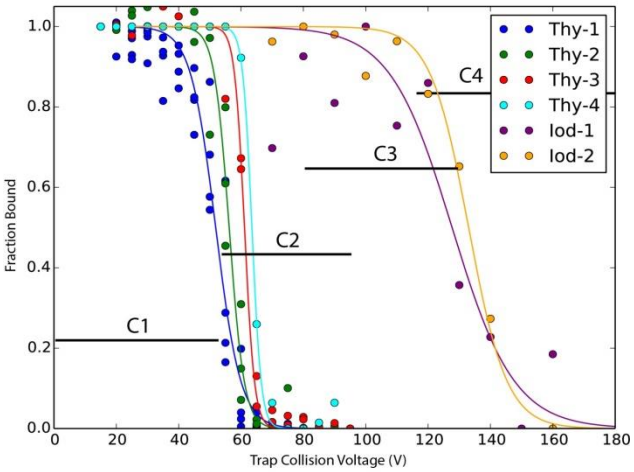


Figure I-8. A graph of fraction bound versus collision voltage in the ion trap prior to IM separation, reveals that Iodipamide and Thyroxine form multiply bound, cooperatively stabilized Albumin Complexes.

Supporting Text for Figure S8. CID curves for all observed species of Thyroxine and Iodipamide. Crystal structures for Thyroxine bound to HSA reveal two binding sites embedded in D2, and two binding sites on the perifiery of D3.[2] Our data suggests all of these ligands dissociate in rapid succession upon collisional activation, before, during, and after the first structural transition from C1 to C2. Considering the disparate positions of the ligands and relatively low CID transition energy, it is difficult to interpret this data in terms of individual dissociation events for each binding location. (PDB ID 1HK1) Likewise, Iodipamide is a large ligand that has multiple binding locations within HSA.[3] Crystallographic analysis reveals iodipamide binding at the D2 drug site, and also shows some electron density for the ligand in the cleft formed between all three domains. We utilized our individual domain constructs to further localize this binding, finding that Iodipamide binds to both D2 as well as D3. Due to the size of this ligand, we assume the electron density from the crystal structure is a tail of the D3-Bound ligand protruding into the interdomain cleft. Despite the ambiguity in assigning binding locations, our CID analysis shows that Iodipamide behaves similarly to thyroxine, releasing ligands within a narrow range of collision voltages, and thus providing evidence for cooperative effects during dissociation. Further CIU analysis also reveals conformational stabilization of the protein (Figure S9, see below).

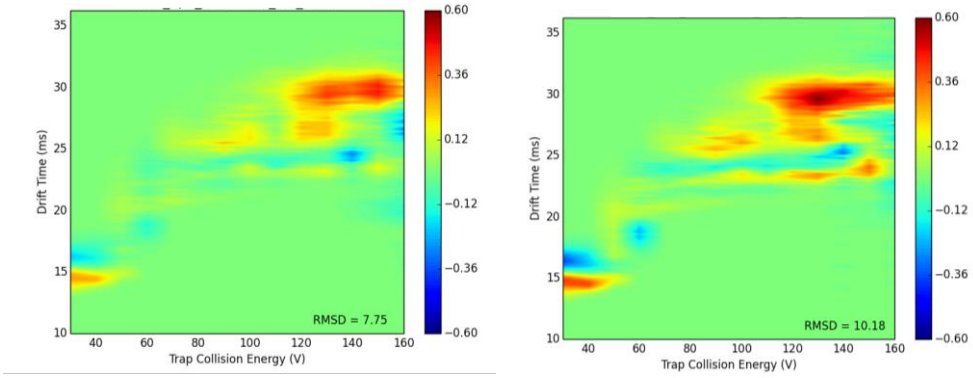


Figure I-9. Iodipamide binding causes significant conformational shifts in HSA, exaggerated at larger stoichiometries. CIU Difference plots between apo HSA and one (left plot) and two (right plot) iodipamide bound species reveal signficant increases in drift time for the bound forms, as well as large degree of stabilization for the final CIU transition observed.

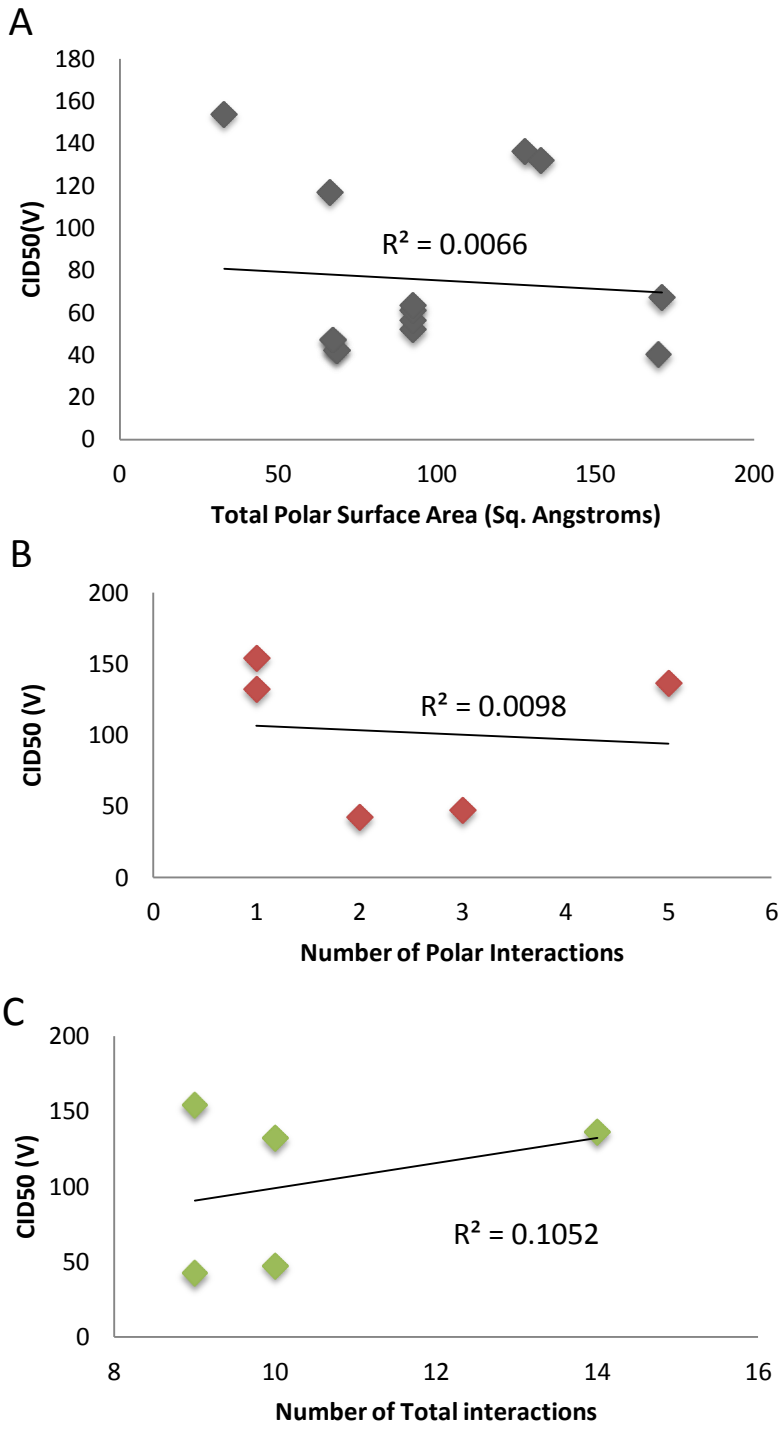


Figure I-10. We observe a lack of any correlation between the CID energy required to eject ligands from HSA and the total polar surface area (A), the total number of polar interactiosn in the binding pocket (B), or the number of total interactions in the binding pocket (C).[4] Correation coefficients (R^2 values) for a linear fit are shown on each graph.

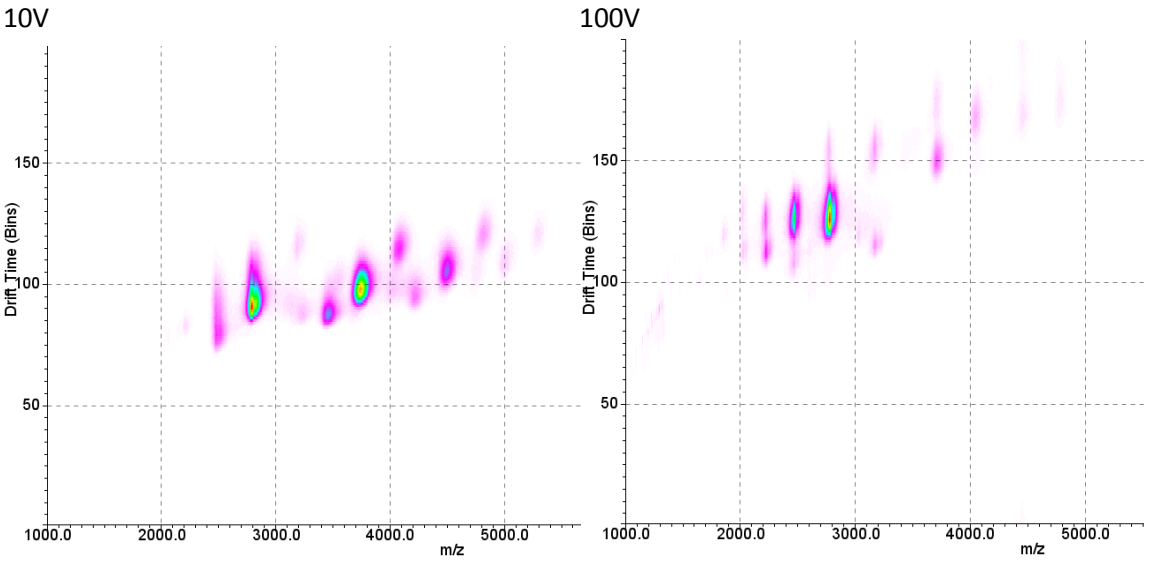


Figure I-11. IM-MS spectra for HSA domain 1 constructs at 10V (left) and 100V (right) collisional activation in the ion trap prior to IM separation, having a measured mass of 22,167. In both spectra, signals for monomers, dimers and trimers of the domain are observed. More CIU is observed at 100V, as expected.

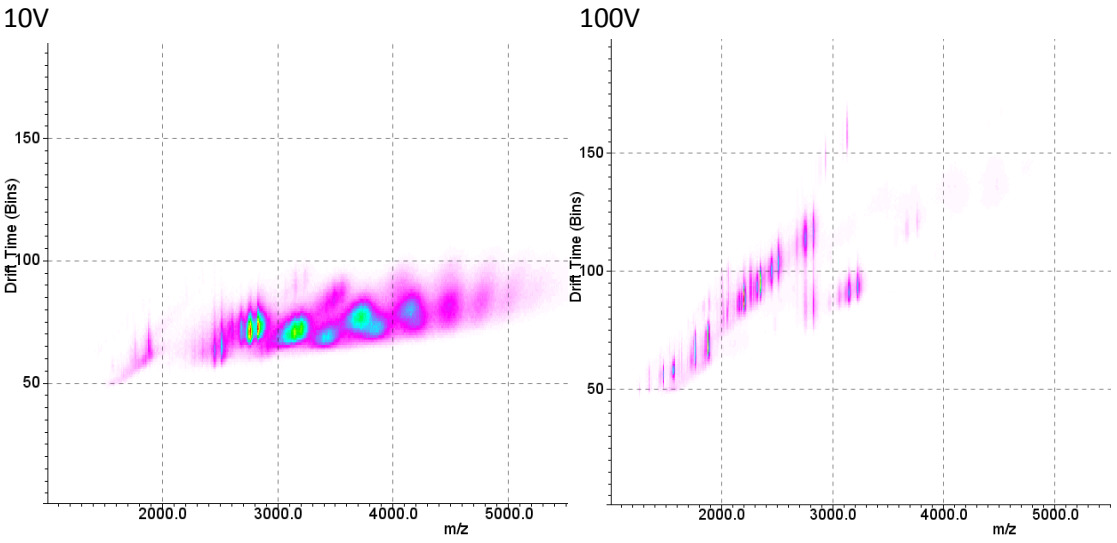


Figure I-12. IM-MS spectra for HSA domain 2 constructs at 10V (left) and 100V (right) collisional activation in the ion trap prior to IM separation, having a measured masses of 21,968 and 22,501. In the 10V spectra, signals for monomers, dimers and trimers of the domain are observed. In the 100V spectrum, significant fragmentation of the covalent domain 2 backbone is observed.

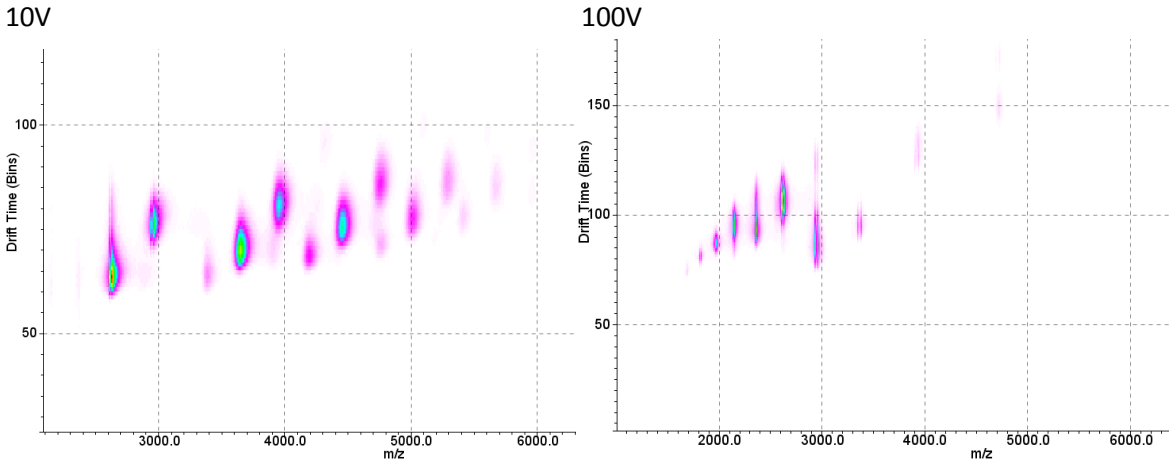


Figure I-13. IM-MS spectra for HSA domain 3 constructs at 10V (left) and 100V (right) collisional activation in the ion trap prior to IM separation, having measured masses of 23,365 and 23569. In the 10V spectrum, signals for monomers, dimers, trimers, tetramers and pentamers of the domain are observed. Signal associated with domain 3 monomer dominates the 100V spectrum.

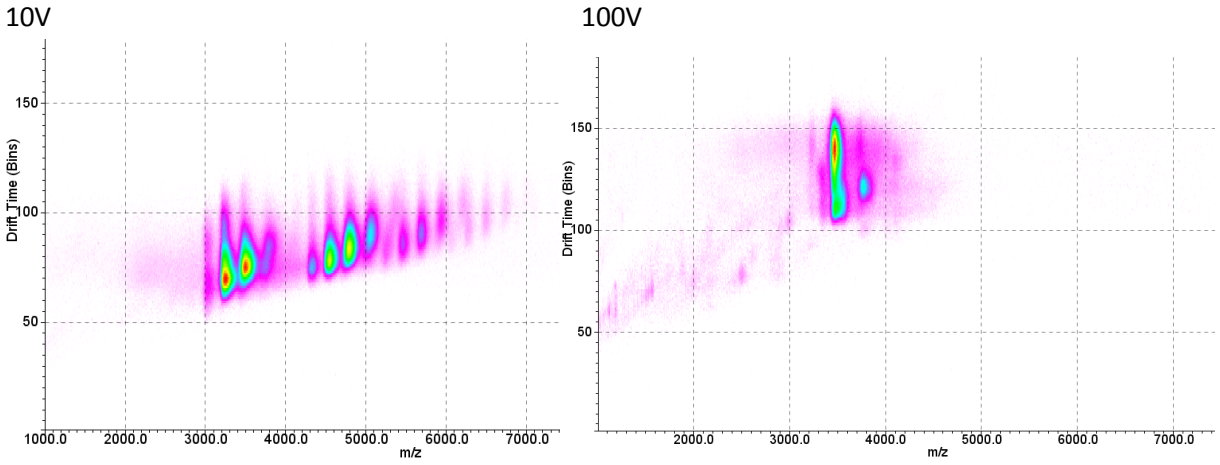


Figure I-14. IM-MS spectra for HSA domains 1 and 2, covalently linked, at 10V (left) and 100V (right) collisional activation in the ion trap prior to IM separation, having a measured mass of 45,266. In the 10V spectrum, signals for monomers, dimers and trimers of the construct are observed. Signal associated with domain 1-2 monomer dominates the 100V spectrum.

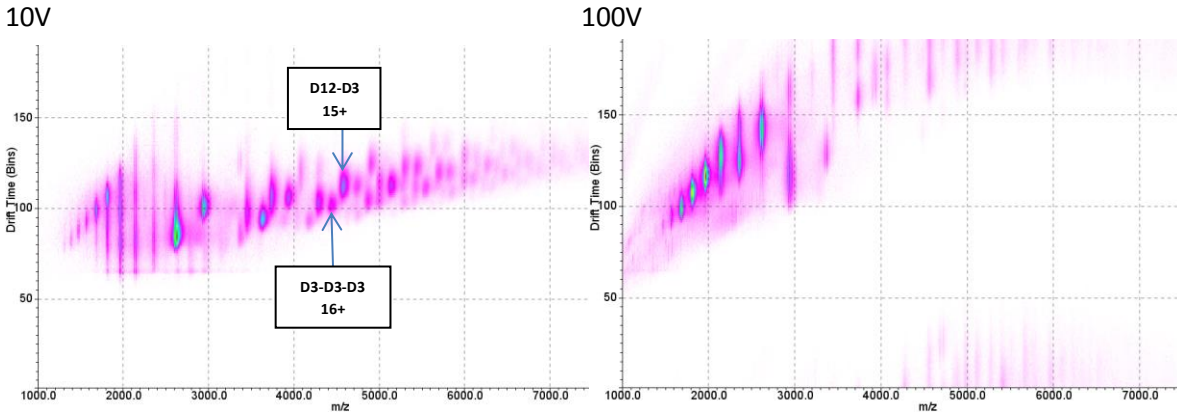


Figure I-15. IM-MS spectra for HSA domain 1-2 construct, mixed with domain 3, at 10V (left) and 100V (right) collisional activation in the ion trap prior to IM separation. The D12-D3 dimer has an expected mass of 68,835 Da, whereas the domain 3 trimer has an expected mass of 70707 Da. Based on the identifications made in the 10V spectrum, we observe both a species at 68409 and one at 70793, which we assign to D12-D3 and D3 trimer respectively. Monomer signal for both D3 and D12 dominates the 100V spectrum, and some traveling wave IM roll-over is observed at low drift time and high m/z values.

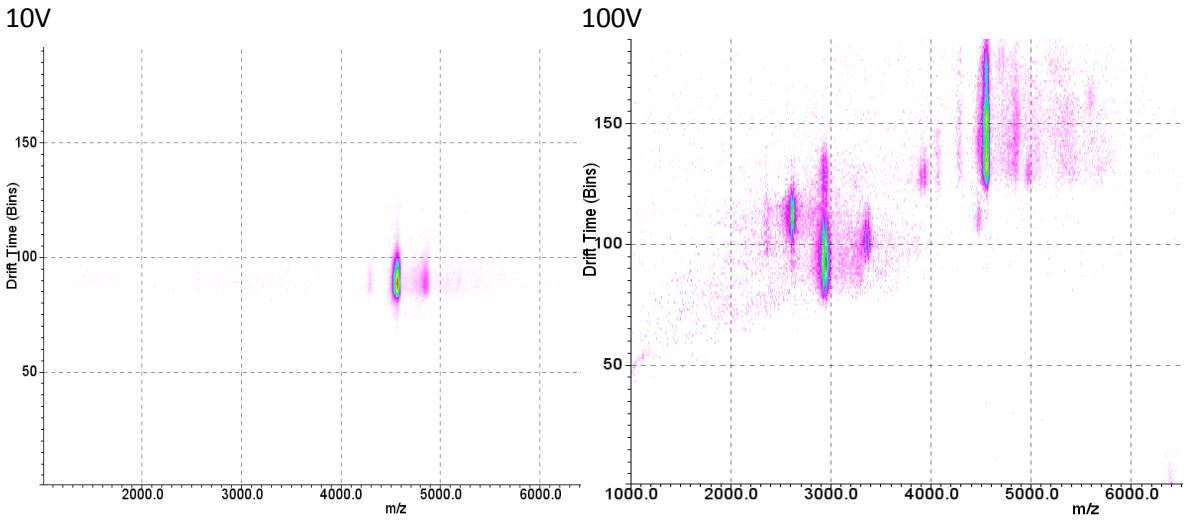


Figure I-16. MS/MS data, acquired using a quadrupole selection window centered on m/z 4560, acquired at a voltage of 10V (left) and 100V (right), for the signal putatively assigned as the D12-D3 dimer from Supplementary Figure 25. Two CID products are observed, corresponding to 23,359 (D3, at low m/z) and 45266 (D12, charge stripped at higher m/z), confirming our D12-D3 dimer assignment.

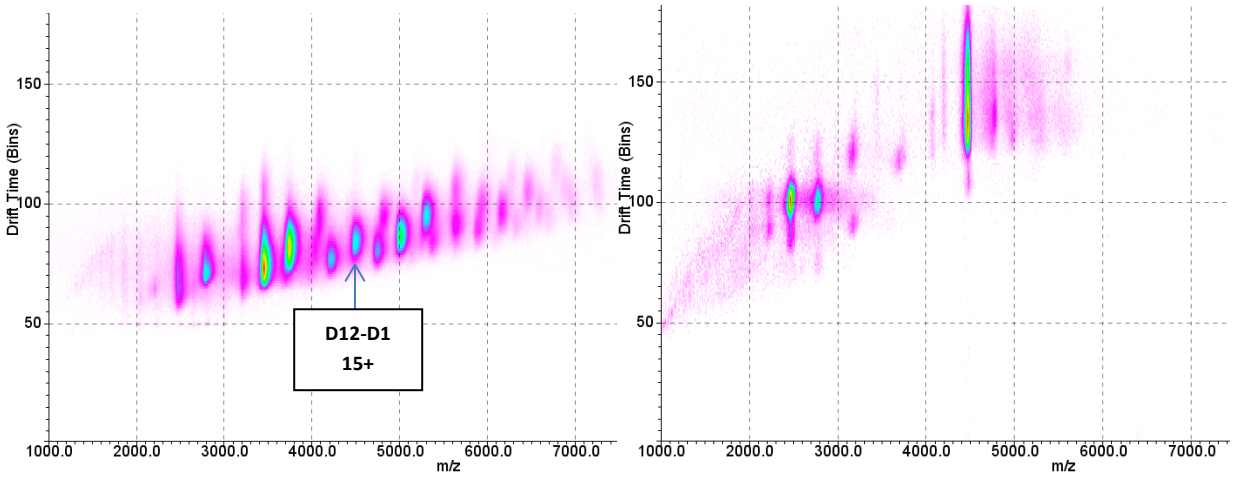


Figure I-17. IM-MS and MS/MS data for samples where D12 and D1 were mixed in solution. The left spectrum was acquired using 10V activation voltage in the ion trap, and we observe strong evidence of D12-D1 dimer complexes (as indicated), having a measured mass of 67,288 Da. D1 and D12 multimers are also observed. The right hand spectrum was acquired using a quadrupole selection window centered on the m/z value highlighted in the left hand spectrum. Two CID products are observed, corresponding to D1 (at low m/z) and D12 (at higher m/z), confirming our D12-D1 dimer assignment. See figures above for individual monomer masses.

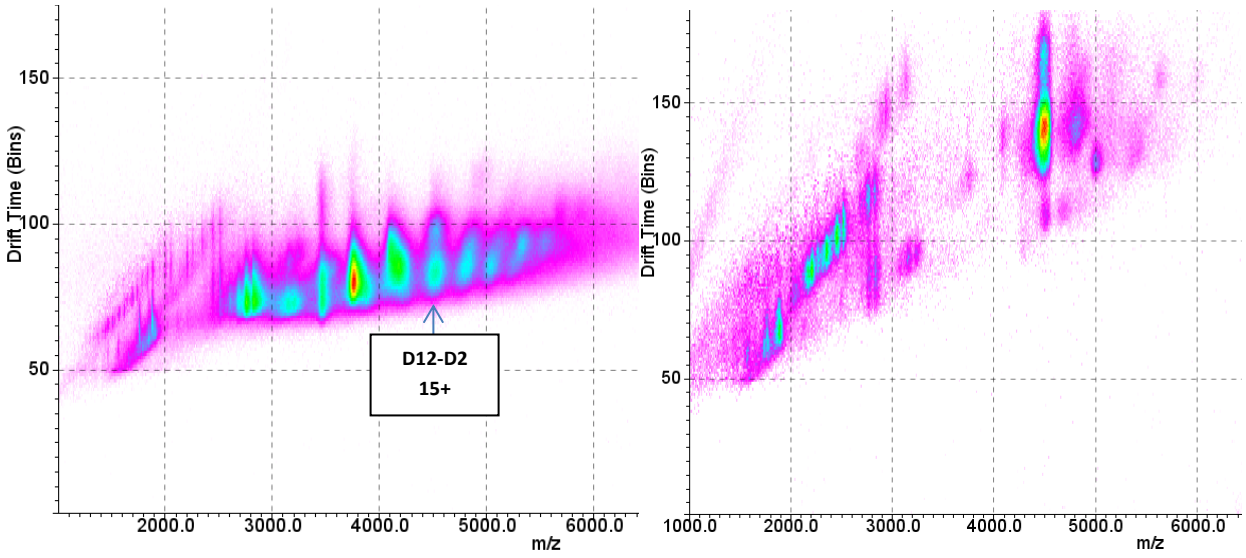


Figure I-18. IM-MS and MS/MS data for samples where D12 and D2 were mixed in solution. The left spectrum was acquired using 10V activation voltage in the ion trap, and we observe strong evidence of D12-D1 dimer complexes (as indicated), having a measured mass of 67,288 Da. D1 and D12 multimers are also observed. The right hand spectrum was acquired using a quadrupole selection window centered on the m/z value highlighted in the left hand spectrum. Two CID products are observed, corresponding to D1 (at low m/z) and D12 (at higher m/z), confirming our D12-D1 dimer assignment. See figures above for individual monomer masses.

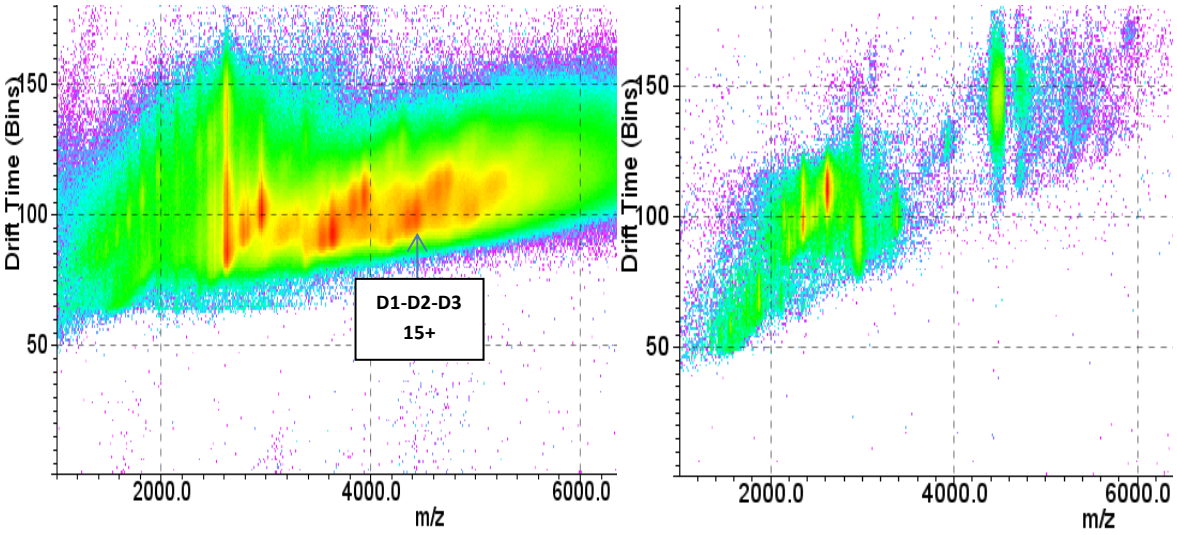


Figure I-19. IM-MS and MS/MS data for samples where D1, D2 and D3 were mixed in solution. The left spectrum was acquired using 10V activation voltage in the ion trap, and we observe strong evidence of D1-D2-D3 trimer complexes (as indicated), having a measured mass of 67,767 Da. Other D1, D2 and D3 multimers are also observed. The right hand spectrum was acquired using a quadrupole selection window centered on the m/z value highlighted in the left hand spectrum. Two CID products are observed, corresponding to D3 (at low m/z) and the noncovalent D12 dimer (at higher m/z), confirming our D1-D2-D3 dimer assignment. In addition to the MS/MS results, we can identify the signal at m/z of 4450 is 15+ because it shows near-identical CIU behavior to all other 15+ ions studied at <50 V. (shown below). See figure captions above for individual monomer masses.

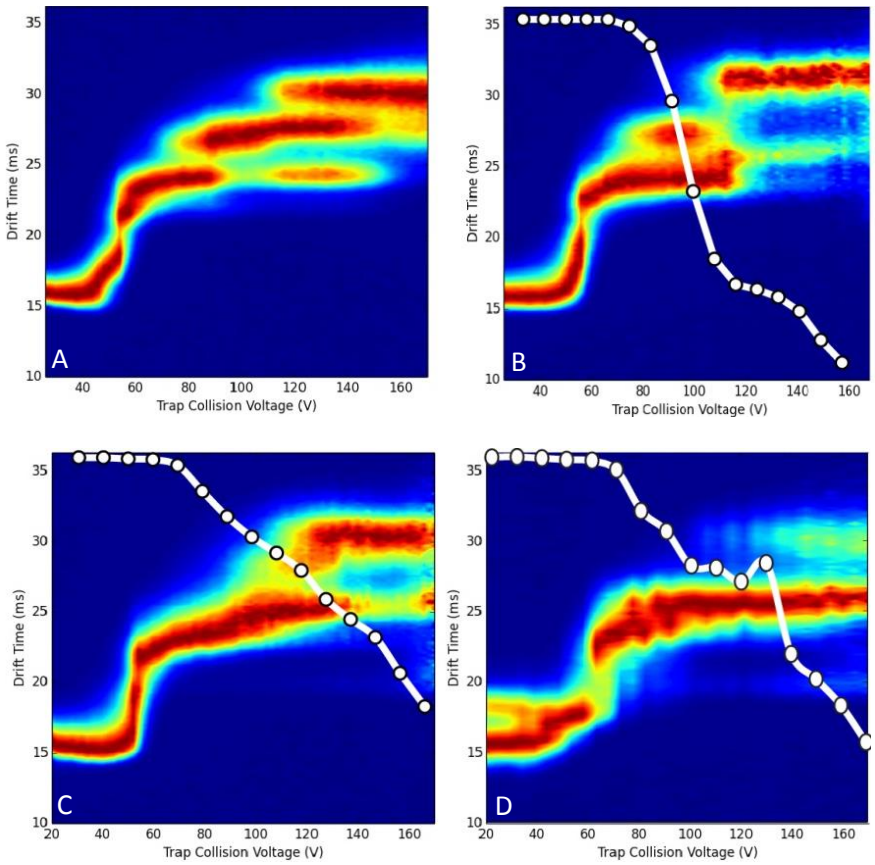


Figure I-20. CIU data for WT HSA (A), compared with CIU data for D12-D3 (B), D12-D1 (C) and D12-D2 (D), overlaid with CID breakdown curves (white) for the ejection of the smaller bound monomer within each dimer. A detailed comparison of native HSA with noncovalent reconstructions thereof highlights interesting mechanistic changes to the unfolding pathway that effect observed CID trends. D12-D3 appears to follow a near-sigmoidal curve coinciding with the appearance of confrmer family 3 (see main paper text). When this conformer family can no longer be efficiently accessed, the CID breakdown curve clearly changes. Likewise, D12-D1 and D12-D2 are unable to access this conformational family, and thus can only proceed through a frustrated CID mechanism that accesses alternative conformational states.

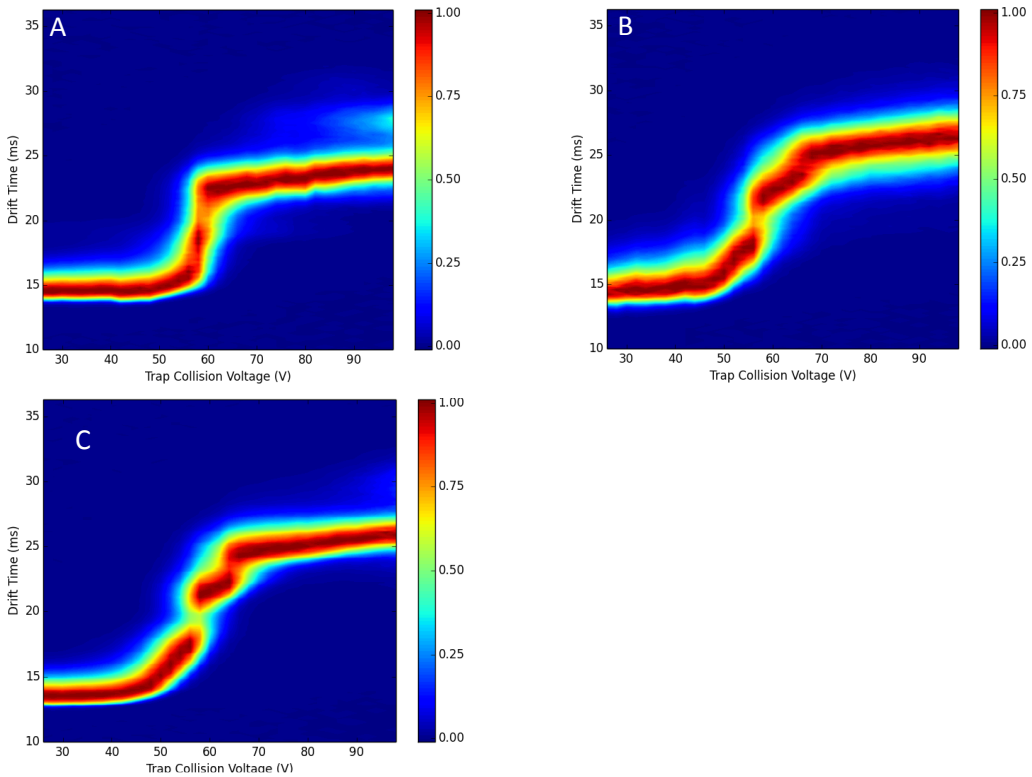


Figure I-21. CIU fingerprint data for reconstituted noncovalent HSA constructs for (A) the D1 trimer, (B) the D1-D2-D3 trimer, and (C) the D3 trimer. A complete CIU fingerprint for the D2 trimer could not be acquired due to its low stability.

Table I-7. Experimental and Calculated Cross Section Values for Albumin Domains and Multidomain Constructs

	calculated	exp	unfolded	2nd Unfolded	%ERROR	UNFOLDED-FOLDED	% CHANGE
D1 7+	2081	1484	1916	2354	-40.2	432	29
D1 8+	2081	1838	2341		-13.2	503	27
D1 9+	2081	1816	2636		-14.5	820	45
D1 10+	2081	2071	2595	2905	-0.48	524	25
D2 6+	2119	1361	1496		-55.6	135	10
D2 6+	2119	1361	2002		-55.6	641	47
D2 7+	2119	1626	1916		-30.3	290	17
D2 8+	2119	1634	2379		-29.6	745	45
D2 9+	2119	1874	2710		-13.0	836	44
D2 10+	2119	1903	2811		-11.3	908	47
D3 8+	2068	1840	2096		-11.0	228	12
D3 9+	2068	1900	2814		-8.1	914	
D12 13+	3410	3150	3901		-8.2	751	23
D12 13+	3410	3150	4136		-8.2	986	31
D12 15+	3410	3265	5036		-4.4	1771	54
D12 16+	3410	3367	5538		-1.2	2171	64
WT 15+	4388	4136	6132		-6.0	1996	48

Table I-8. Measured CCS values for Unfolded Albumin Conformations at 15^{+}

	Calculated	Observed	ΔCCS	%Change
Ground State	4388*	4136	-	-
Unfld 1		5153	1017	24.6
Unfld 2		5740	587	11.4
Unfld 3		6107	367	6

*From 4K2C using PA*1.15

References:

(1) Bush, M. F.; Hall, Z.; Giles, K.; Hoyes, J.; Robinson, C. V.; Ruotolo, B. T. *Analytical Chemistry* **2010**, *82*, 9557.

(2) Petitpas, I.; Petersen, C. E.; Ha, C.-E.; Bhattacharya, A. A.; Zunszain, P. A.; Ghuman, J.; Bhagavan, N. V.; Curry, S. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 6440.

(3) Ghuman, J.; Zunszain, P. A.; Petitpas, I.; Bhattacharya, A. A.; Otagiri, M.; Curry, S. *J. of Mol. Biology* **2005**, *353*, 38.

(4) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. *Protein Engineering* **1995**, *8*, 127.

Appendix II: Supporting information for Chapter 4: Coming to Grips with Ambiguity: Ion Mobility-Mass Spectrometry for Protein Quaternary Structure Assignment

Generating coarse-grained structures at variable resolution.

To estimate the contributions of multidomain and other non-globular protein subunits to coarse-graining errors, we developed a fast, online method for adding resolution to our coarse grained models. For any protein subunit greater than 500 residues, we invoked k-means clustering of its coordinates to determine if it contained clusters of greater than 100 residues that were resolved in space. If multiple clusters did exist within a single subunit, we developed a coarse-grained model using our standard workflow, albeit assigning a sphere to each cluster rather than each subunit. By applying this method against many known multiprotein topologies and varying the thresholds for cluster assignment, we assessed the coarse-graining error as a function of the residues per sphere in the coarse-grained model.

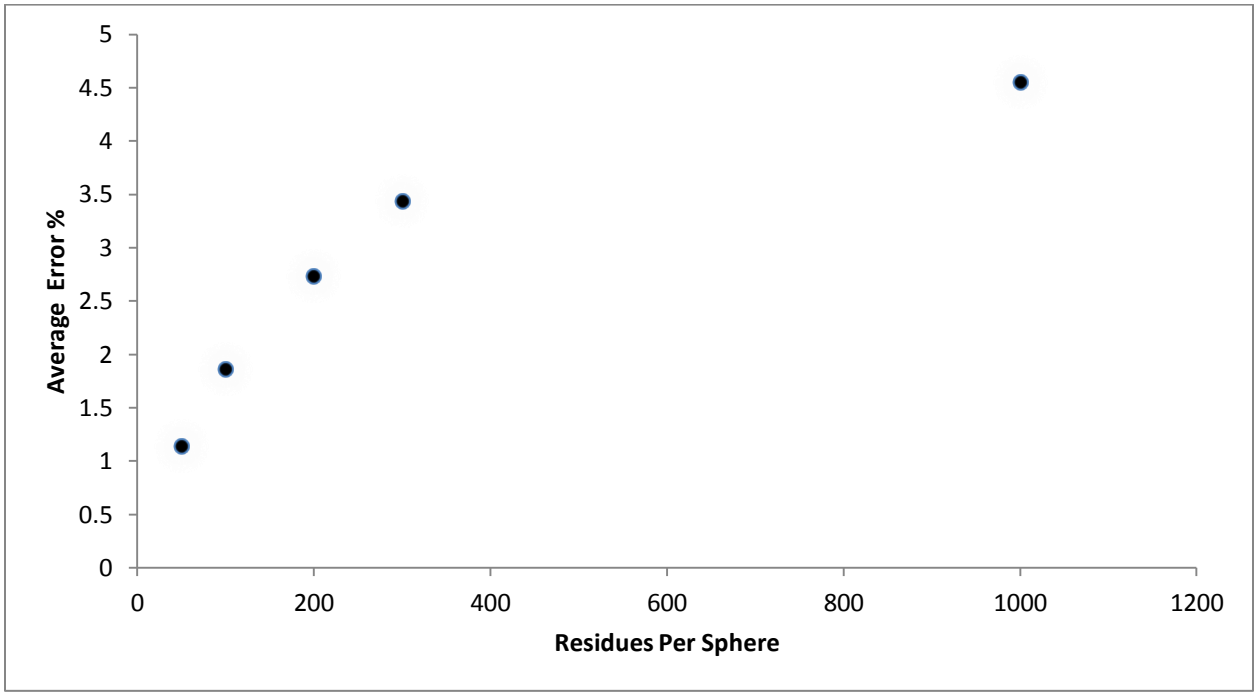


Figure II-1. *Coarse-graining CCS error as a function of Subunit Residues/Sphere*

Generation of Coarse-Grained Multiprotein Models based on IM-MS datasets.

Restraint file. Although previous methods have been published for generating structural models from IM-MS datasets, these have generally relied on custom scripts that are difficult to implement for those not well-versed in computational modeling or coding. Building on this work and integrating many aspects already developed, we sought to build a general method for translation of IM-MS restraints into structural models that does not require any changes to the source code, but instead relies on simple input files that are read by the program and translated into an ensemble of putative structures. Input files allow for facile input of CCS information for subunits, subcomplexes, and complexes; connectivity information at varying levels of ambiguity; explicit protein-protein distance information, and symmetry information. These four types of information represent all levels of information derived from IM-MS experiments, and can also represent data from other experiments, especially crosslinking-MS. For more information, we refer the reader to our website where annotated input files as well as the MS modeler source code and data analysis scripts are available:

https://sites.lsa.umich.edu/ruotolo/software/IMMS_Modeler

Interpretation of Restraints. IMMS-Modeler integrates information from the input file into a model system with a scoring function that is later optimized many times to generate an ensemble of putative structures. The model consists of multiple spheres representing each protein (or domain) with radii derived from the collision cross section. Each sphere is initially positioned randomly within a bounding box. The size of the bounding box scales with the expected CCS of the target complex in order to reduce biases related to the initial coordinates of the subunits and to reduce the need for unnecessary optimization of coordinates placed too far apart in space.

After generation of the geometric system, a scoring function is built by combining the connectivity, distance, and symmetry restraints into a single optimizable function. Connectivity restraints are expressed as a tree structure that allows for varying levels of ambiguity: explicit subunit-subunit contacts are scored by a square function that allows the two spheres to interpenetrate to varying degrees within biophysical thresholds for interacting proteins, previously established at 15% to 45%.[37] Interacting subunits that interpenetrate by greater or less than these values will not be accepted. For more ambiguous connectivity inputs, only a single connectivity restraint between a pair of subunits must be satisfied. For example, if subcomplex [A,B] has connectivity to subcomplex [C,D], any pairwise connectivity within the set ([A,C],[A,D],[B,C],[B,D]) would satisfy the restraint.

Explicit subunit-subunit distances can also be implemented into the scoring function. For instance, the CCS of a dimeric protein complex, where the CCS of both monomers are known, restrains the distance between those subunits explicitly within the model. The same is true for trimeric proteins where CCS information is known for all subcomplexes. When this information is known, it is often useful to input it directly into the model to increase sampling efficiency. Additionally, high resolution models may be available for some subcomplexes, in which case all the pairwise distances between subunits can be easily extracted and used to define the subunit structure explicitly within the model. Since pairwise

distances generated from IM-MS restraints have some error associated with them, they are implemented as a harmonic restraint for which the force constant can be scaled appropriately.

In order to increase sampling efficiency when symmetry is expected within the protein complex, we also implemented symmetry constraints on the system. These constraints enforce a given symmetry operation on a group of particles which forces them to maintain symmetry throughout the optimization cycle. When the exact symmetry is unknown, we have had success at comparing ensembles generated in parallel with different putative symmetries.

Unbiased Sampling with a Monte Carlo / Monte Carlo Search

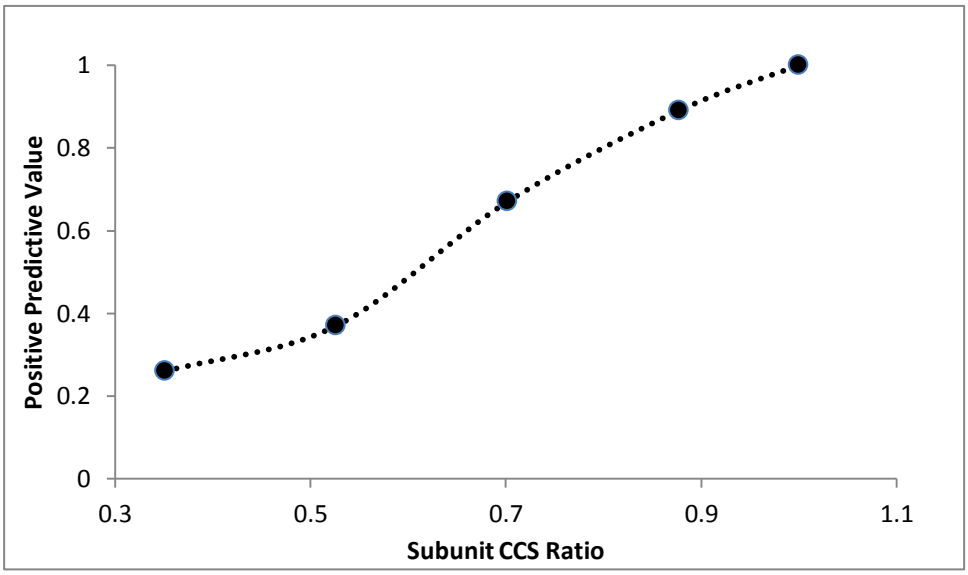
Our program optimizes the scoring function of the system while attempting to sample as much structural space as possible. First, starting from the random initial coordinates of the system, an annealing-type Monte Carlo (MC) search is utilized to locate a local minimum within the energy landscape. The annealing type MC proposes moves for each subunit randomly, and accepts or rejects those moves based on their effect on the scoring function relative to the temperature of the system. After the annealing-type MC has found a local minimum within energy landscape, that structure is recorded before the temperature of the system is then raised in order to allow for less biased exploration of nearby structures. In this way, each local minima within the system is not characterized by a single structure, but an ensemble of structures. This process can then be repeated, starting with new, randomized initial coordinates, to characterize many local minima within the system. We have found that 100-1000 optimization cycles, which each output an ensemble of 10-100 structures within a family, is generally sufficient to characterize all possible geometries that satisfy common IM-MS-based restraint sets.

Filtering and Characterizing the Ensemble

After an ensemble of putative structures is generated, it is necessary to filter the dataset based on biophysical parameters as well as experimental cross section results. Filtering the dataset to ensure no non-physical interactions exist is important because we have chosen not to penalize subunits not restrained by the connectivity restraint for interpenetrating beyond physical norms. The reason for this is that it allows our sampling algorithm more freedom in exploring new areas of conformational space, albeit with some sacrifice of sampling efficiency. Similarly, we have not included the experimental CCS of the target complex into the optimization function. Until recently, calculation of the CCS for a candidate model on the timescale of a MC optimization would have been intractable, even for the relatively fast projection approximation method. The development of IMPACT, however, has changed this paradigm and brought projection approximation to molecular dynamics timescales. Despite these advances, we still have some concern about the potential for a CCS restraint within the optimization function to create high barriers to exploring conformational space, and hence we have chosen to filter the ensemble after optimization, at the expense of some sampling efficiency. Our filtering procedures remove models that feature unphysical interactions, and those that have absolute deviations from the experimental CCS of greater than 3%. Once a filtered ensemble is obtained that satisfies both the optimization function parameters and post-optimization parameters, characterization of the ensemble is of paramount

importance for generating new hypotheses about the complex structure and planning further experiments to resolve ambiguities. As previously proposed, we implemented average-linkage hierarchical clustering as a tool for determining the structural families present within an ensemble. This algorithm utilizes pairwise structural distances, in the form of RMSD values, to generate clusters of similar complex structures; and provides quantitative relationships between clusters.

Figure II-2. CCS restraints increase in selectivity when subunits are similar in size. In our analysis of the PPVs for trimeric protein complexes from the 3D complex set, we found one outlier in the dataset (PDB ID 2INC) showing extremely low PPV even when restrained by a set of 2 internal CCS restraints and a global CCS restraint. We observed that in this protein complex, one subunit was approximately 75% smaller than the other proteins in the system. In order to determine the impact of this size disparity on the observed PPV, we artificially inflated the CCS of the small subunit until it was in parity with the others. Indeed, as the CCS of the small subunit was increased, the corresponding restraint set yielded a higher PPV.



Appendix III: Supporting information for Chapter 5: Structural Models of the Urease Activation Complex Derived from Ion Mobility-Mass Spectrometry and Integrative Modeling

Joseph D. Eschweiler, Mark. A. Farrugia, Robert P. Hausinger, Brandon T. Ruotolo

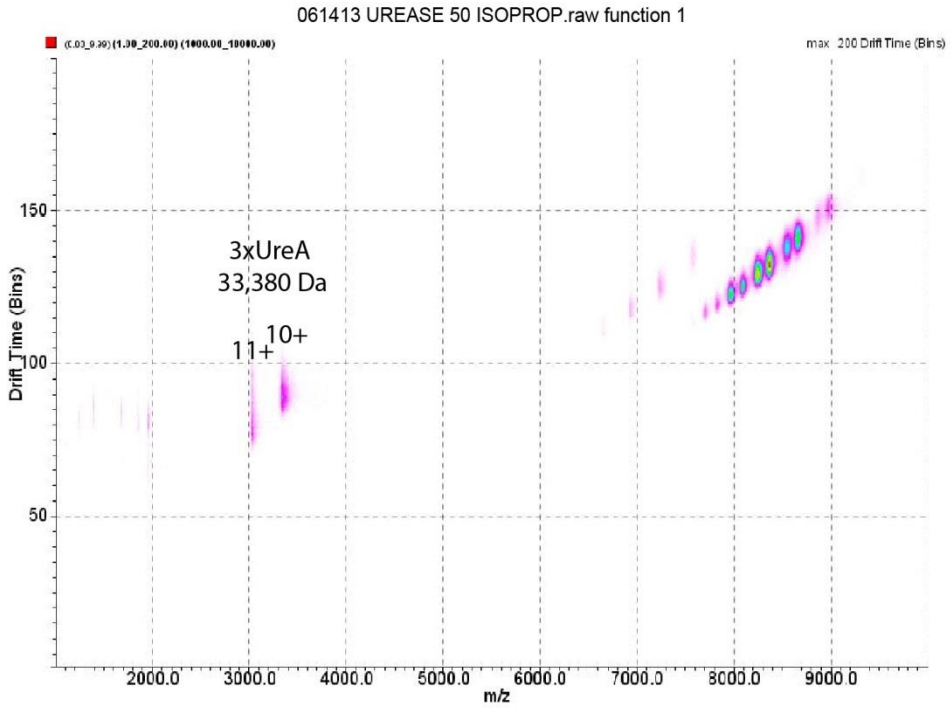


Figure III-1. Solution-phase disruption reveals the $(UreA)_3$ subcomplex of urease. We sprayed the urease core complex from a solution of 50% isopropanol and 50% 200mM ammonium acetate. The IM-MS spectrum under these conditions reveals the presence of a modular ureA trimer at 10+ and 11+. Also observed in the spectra are various losses of ureB from the urease core complex. Expected mass of $(UreA)_3$: 33254 Da

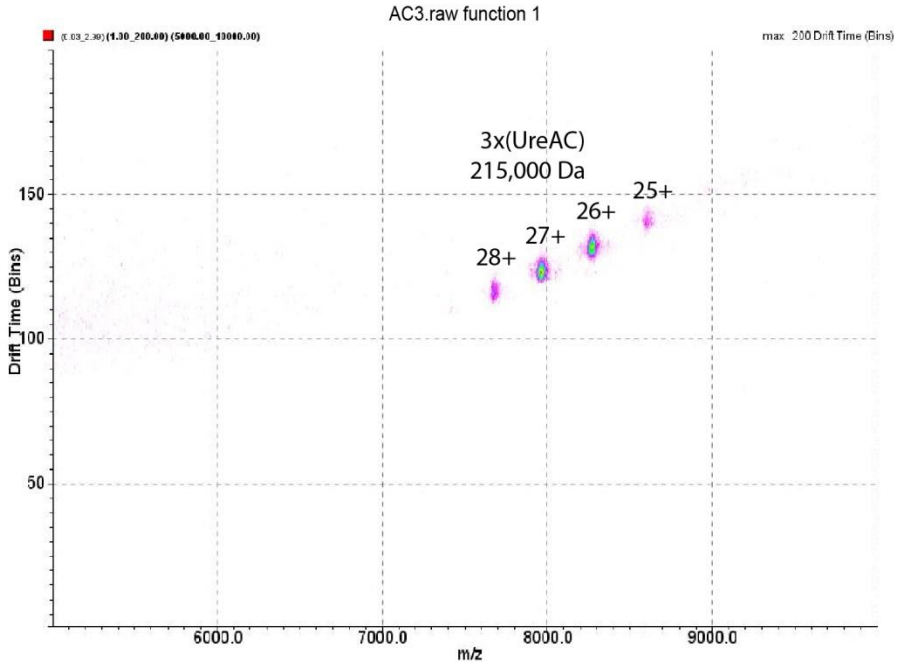


Figure III-2. Expression of urease with a ureB gene knockout yields a homogenous population of $(UreAC)_3$ hexamers at 215,000 Da. Expected mass : 215:024 Da

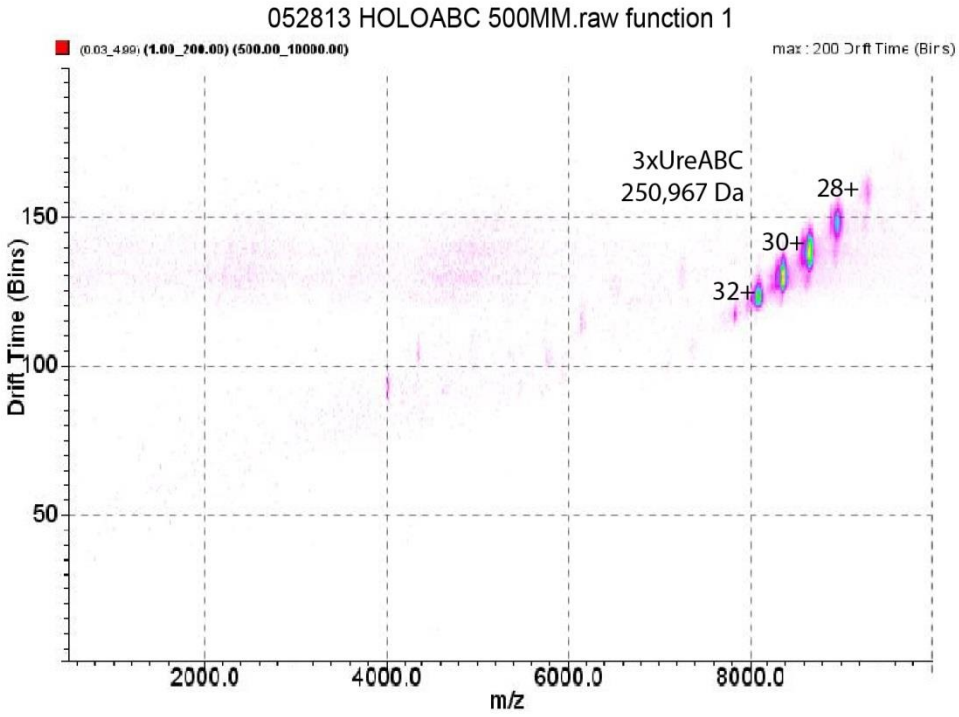


Figure III-3. IM-MS of the urease holoenzyme (ureABC)3 reveals predominant signals with a mass in close agreement with the expected mass of 250,308 Da.

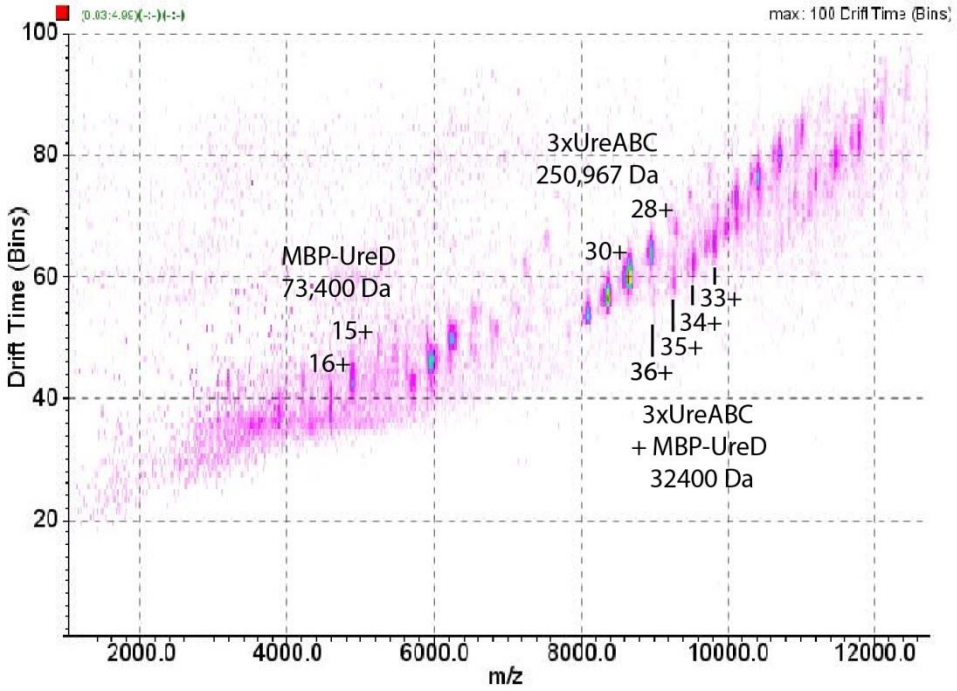


Figure III-4. IM-MS of a sample containing all urease accessory components including MBP-ureD. Key complexes are annotated. Although higher order complexes were identified, it was impossible to assign configurations to these different stoichiometries, and thus they were not included in the model.

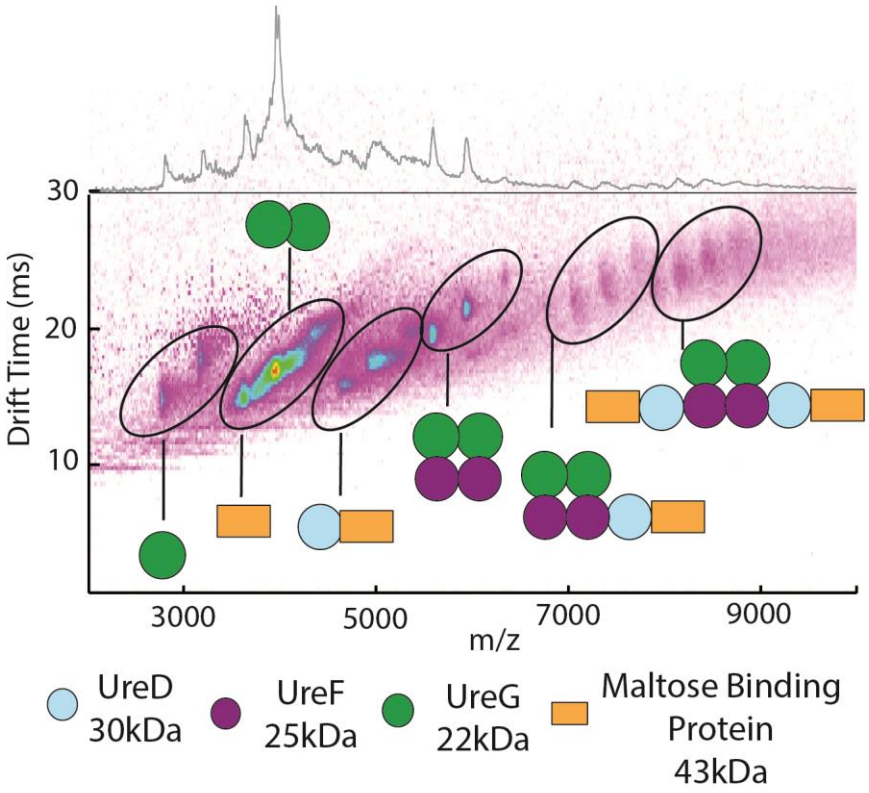


Figure III-5. Fully annotated IM-MS spectrum of the $(ureDFG)_2$ complex and its subcomplexes.