ADVANCING MASS SPECTROMETRY-BASED PROTEOMIC ANALYSIS STRATEGIES FOR THE INVESTIGATION OF HUMAN HEALTH AND DISEASE

by

Justin McKetney

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

Doctor of Philosophy

(Biochemistry)

at the

UNIVERSITY OF WISCONSIN–MADISON

2021

Date of final oral examination: 05/14/2021

The dissertation is approved by the following members of the Final Oral Committee:
Joshua J. Coon, Professor, Biomolecular Chemistry
Melissa M. Harrison, Associate Professor, Biomolecular Chemistry
Anjon Audhya, Professor, Biomolecular Chemistry
Mark E. Burkard, Professor, Medicine
John M. Denu, Professor, Biomolecular Chemistry

To my parents, who made me all that I am

ACKNOWLEDGMENTS

Scientific research is by its nature a collective and collaborative effort, and the same goes for training scientists. When I came to Wisconsin, I knew basically nothing about mass spectrometry and was not interested in joining a group focused on technology. But there was something about the Coon group that changed my mind. I immediately felt comfortable in the group's atmosphere and was in awe of the scope of biological questions that could answered with mass spectrometry. I have often joked that one of my more valuable skills in graduate school was charm, because I needed constant help. I would ask the same question multiple times a day to different group members and it must have been annoying. In my five years as a member of the group (and approximately 100,000 questions) I have never seen the slightest irritation from these scientists, who certainly have better things to do, who have always made time to help me. This network of support has been invaluable to me and I don't think any part of this dissertation would have been possible without it. I want to thank Professor Coon for creating and cultivating that environment. Professor Coon has provided a level of research space, resources, ideas, guidance that is virtually unparalleled. He has helped me become a better writer and a defter navigator of the scientific process.

So much of what I know about mass spectrometry comes from the staff of the lab: Katie Overmyer, Michael Westphall, but especially Alex Hebert. I followed Alex around like a lost puppy dog for my first two years in the lab. He gave advice on LCs, mass spectrometers, data cleaning, and making figures. I cannot impress how much he improved all aspects

of my science, and despite his misanthropic nature, he was an amazing mentor. I cannot thank him enough for that guidance in the fledgling years of my training.

The Coon group also provided an amazing group of senior graduate students that guided, mentored, and exchanged ideas with me throughout my time in graduate school: Elyse Freiburg, Evgenia Shishkova, Erin Weisenhorn, Matt Rush, Paul Hutchins, and Nick Kwiecien. I spent most of our overlap in the lab simultaneously in awe of their scientific knowledge and acumen, and the fact that they would take time to answer my long list of questions. I want to specifically thank my two desk mates, Vanessa Linke and Dain Brademan, who provided conversations about mass spectrometry but also science more generally, and even topics outside of science. That commiseration, with all of the members of the lab, is what I have missed most in this strange last year. I also want to specifically thank Nick Riley and Edna Trujillo, two graduate students that provided two very different types of support. Nick Riley was one of the first to introduce me to mass spectrometry. His gregarious, inquisitive and driven nature helped me feel accepted in the Coon group, helped me understand mass spectrometry, and pushed me to be a better scientist. Edna Trujillo is perhaps the most positive, determined, upbeat and supportive peer that I have ever had. She will not be deterred. She will always provide feedback, support, an ear to vent to or do anything she can to help another lab member. As a scientist of color in Wisconsin, I cannot thank her enough for this spiritual and emotional support she gave throughout my time in the group.

Outside of the Coon group, I would like to thank my collaborators. We are a mass

spectrometry group at our core and so rely on so many other specialties to allow for our investigations into human tissues and diseases. I deeply appreciate the work of many of those who catalogue and store tissue or fluid samples to make any of this possible, and the other groups that help compose, draft and edit manuscripts to create polished final products.

I have been a part of many groups in my time at UW that have given me valuable support as a scientist and a person and have made me feel like a member of a community. Thank you to my IPIB cohort, specifically Adam Lewis, Eddie Rashan, and especially Zach Romero. Thank you to members of MBTG, and especially the director Christina Hull, who has provided mentorship far outside the requirement for her position. Thank you to all the members of SciMed and Abbey Thompson, for helping to make me feel comfortable when first coming to Wisconsin.

Thank you to my past mentors who provided a research foundation for me to even pursue graduate school, Edward Collins, Scott Bridgham, David Lipson, and especially Barry Stoddard. Without those opportunities I would not have known how I felt about science or had the resume to apply to graduate school. Barry mentored me in a very unique fashion while I was applying, interviewing, and deciding, and his advice was instrumental in coming to UW.

For the last four years one of my biggest supporters has been my future wife, Amanda. I am still, four years later, in shock that she chose to take this adventure to Wisconsin with me. Her love, compassion, infectious laugh, and steadfast belief in my success has been the

light to my darkest days. I cannot wait to spend the rest of my life with her.

Lastly, I would like to thank my parents. My entire life they have inspired me, supported me, and made me want to be better. Any aspect of my life that I have ever succeeded in should be credited to them. They have given me so much, supported me throughout my scientific career, and I hope every day that I can make them proud of the scientist, and the person that I have become.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

2D-LC	Two-dimensional liquid chromatography
AC	Alternating current
AD	Alzheimer's disease
AGC	Automatic gain control
AMY	Amygdala
ANOVA	Analysis of variance
API	Application programming interface
APP	Amyloid precorsor protein
a.u.	Arbitrary unit
AUC	Area under the curve
BEH	Ethylene bridged hybrid bead
CAA	Chloroacetamide
CAD	Collision-activated dissociation
CBM	Cerebellum
CHTC	Center for High-Throughput Computing
CID	Collision-induced dissociation
CNC	Caudate nucleus
COMPASS	Coon OMSSA Proteomic Analysis Software Suite
C#	C sharp, a programming language

CSMSL	C# Mass Spectrometry Library
CV	Compensating voltage
CV_{max}	Highest transmission compensating voltage for a peptide
Da	Dalton, the atomic mass unit
DAVID	Database for Annotation, Visualization and Integrated Discovery
DC	Direct current
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
DNA	Deoxyribonucleic acid
DV	Dispersion voltage
DTT	Dithiothreitol
ECD	Electron-capture dissociation
ECX	Entorhinal cortex
ER	Endoplasmic reticulum
ESI	Electrospray ionization
ETD	Electron-transfer dissociation
F_2	Accuracy score based on harmonic mean in which recall is favored over precision
FAIMS	High field asymmetric waveform ion mobility spectrometry
FASTA	A format for storing protein sequences
FDR	False discovery rate

FT	Fourier transform
FT-ICR	Fourier transform ion cyclotron resonance
GC	Gas chromatography
GnHCl	Guanidine hydrochloride
GO	Gene ontology
GUI	Graphical user interface
HCD	Higher-energy collisional dissociation
HESI	Heated electrospray ionization source
HPLC	High-performance liquid chromatography
HRAM	high resolution accurate mass
Hz	Hertz, inverse seconds
IDA	Intelligent data acquisition
IMS	Ion mobility spectrometry
IPL	Inferior parietal lobule
IRM	Ion routing multipole
IT	Ion trap
ITCL	Ion trap control language
LC	Liquid chromatography
m/z	Mass-to-charge Ratio
MALDI	Matrix-assisted laser desorption-ionization

MAPT	Microtubule-associated protein tau
MFG	Middle frontal gyrus
min	Minute
mRNA	Messenger ribonucleic acid
MS	Mass spectrometry
MS^1	Survey mass analysis
MS^n	Tandem mass spectrometry
MS/MS	Tandem mass spectrometry
NEO	Neocortex
NCE	Normalized collision energy
nLC	Nanoflow liquid chromatography
NN	Neural net
OMSSA	Open Mass Spectrometry Search Algorithm
p-value	Probability score
PCA	Principal component analysis
PCR	Polymerase chain reaction
PLS-DA	Partial least squares discriminant analysis
ppm	Part per million
PRM	Parallel reaction monitoring
psi	Pounds per square inch

PSM	Peptide-spectrum match
ptau	Phosphorylated tau protein
PTM	Post-translational modification
q-OT-QLT	Quadrupole Orbitrap linear ion trap hybrid mass spectrometer
q-value	Corrected p-value for FDR calculations
q-value	Unit-less dimension for ion motion derived from Mathieu equation
QLT	Quadrupole linear ion trap
R	Statistical scripting language
RF	Radio frequency
RFC	Random forest classifier
RNA	Ribonucleic acid
ROC	Receiver operator characteristic0
ROC-AUC	Receiver operator characteristic area under the curve
RP	Reverse phase
RSD	Relative standard deviation
RT	Retention time
s	Second
STG	Superior temporal gyrus
S/N	Signal-to-noise ratio
SNP	Single Nucleotide Polymorphism

SRM	Selected reaction monitoring
SCX	Strong-cation exchange
TCEP	Tris(2-carboxyethyl)phosphine
TFA	Trifluoroacetic acid
Th	Thomson, the unit of the mass-to-charge ratio
THA	Thalamus
tRNA	Transfer ribonucleic acid
TIC	Total-ion chromatogram
UHPLC	Ultra high performance liquid chromatography
UPR	Unfolded protein response
V	Volts
VCX	Visual cortex
WADRC	Wisconsin Alzheimer's Disease Research Center
WISL	Wisconsin Initiative for Science Literacy
XIC	Extracted-ion chromatogram
z	Charge

ABSTRACT

This dissertation describes advances in mass spectrometry-based analysis of the human proteome and applies the most current technologies to investigations of age, neurodegeneration and stress. Although direct protein measurements in human tissues provide valuable insight into the development and progression of disease, the complex and dynamic nature of human proteomes provides several challenges. **Chapter 1** details some of these hurdles along with the basic concepts of bottom-up proteomics in terms of sample preparation and instrument operation. **Chapter 2** describes regional protein signatures for nine neuroanatomically distinct regions of the aged human brain. These region-specific proteins are then compared to proteins associated with Alzheimer's disease (AD). An efficient and scalable method for proteomic analysis of AD in cerebrospinal fluid is demonstrated in **Chapter 3** using an age- and sex-matched sample cohort. In **Chapter 4**, a proteomic analysis is performed on saliva of soldiers before and after a simulated combat training exercise in order to quantify the proteomic effects of stress. To expand the method development toolkit for proteomics of human tissues and biofluids, a machine learning model is developed for predicting peptides' transmissive compensating voltage when using high field asymmetric waveform ion mobility spectrometry (FAIMS)(**Chapter 5**). Conclusions and future directions for these projects, including expanded analyses and continued technological development are discussed in **Chapter 6**. A chapter conveying aspects of this dissertation to a general public audience is included after the conclusion (**Chapter 7**).

Chapter 1

BACKGROUND AND INTRODUCTION

Proteins as the Building Blocks of Life, the Impetus

The human proteome: a monumental challenge Almost two decades ago, researchers sequenced the full human genome[1–3]. Unbeknownst to them at the outset of this impressive international endeavor, the completion of this project would initiate a new research focus[4]: a race to identify the expression of the 20,000 protein-coding genes encoded within the human genome. Even with the efforts of 25 labs across 21 countries, experimental observation of 90% of these genes has only recently been achieved[5]. When taking into account the expansion in protein diversity that occurs throughout transcription, translation and post-translationally, the sequencing of genomes seems almost trivial by comparison, with estimates of the true size of the human proteome ranging from 20,000 to several million entities[6–10]. During transcription, splicing greatly expands the eukaryotic proteome. More than 94% of human transcripts contain multiple exons, with 92-97% of multi-exon mRNA transcripts suspected to be alternatively spliced[11]. Although some alternative transcripts are produced at negligible amounts, approximately 86% of genes are thought to generate isoforms at 15% of total transcript levels or greater[12]. Proteins can also be modified post-translationally by the addition of chemical motifs or cleavage by proteases. More than 300 protein chemical modifications have been identified, with at least 230,000 sites of chemical post-translational modification identified in the human proteome[13]. The number of protein-modifying enzymes has expanded throughout vertebrate evolution[14] with the human proteome containing more than 90 kinases alone[15]. Proteolytic processing also plays

a role in the diversity of proteins and peptides found in human tissues[16], with regulatory roles beyond simple degradation[17].

Why study proteins? Proteins are core to the function and continued life of all human cells, allowing response to stimuli and effectively connecting genotype to phenotype[18]. Proteins perform and regulate processes key to eukaryotes, such as DNA replication[19], transcription[20] and translation[21,22], in some cases initiating their own expression by genomic insertion[23]. Enzymes in the mitochondria generate the fuel necessary to power the cell[24]. Differences in protein splicing[25–27], abundance[28], and localization[26] play roles in the diversification of mammalian cell types and cell fate. This cellular diversity along with protein-regulated cell growth and division allows the development of organs and intra-organ structures such as the cerebral cortex[25,26,29,30]. Proteins allow mammals to experience the world with signal transduction roles in all sensory pathways, including vision[31], hearing[32] and touch[33,34]. This multitude of function means that understanding normal biological processes of protein expression, degradation, sequestration, and modification provides valuable insights into the alterations observed in disease, allowing improved detection, diagnosis, and treatments.

Proteomics technologies Despite the integral role that proteins play in human biology, much of the investigation into gene expression in the late 20th century focused on nucleic acids. Part of this nucleotide centricity stemmed from the capacity to amplify nucleotides

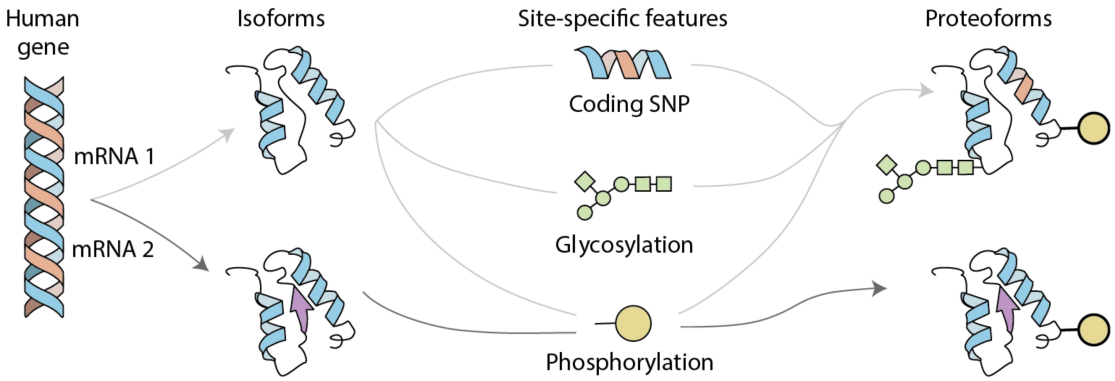


Figure 1.1: Human protein diversity expands through a variety of molecular events. Depicted is a single human gene and two of its isoforms, which differ by the coding for several different amino acids of a protein primary sequence (at left); isoforms commonly arise from alternative splicing of RNA and from use of different promoters or translational start sites. Isoform variation combines with site-specific changes to generate human proteoforms (at right); three examples of site-specific changes include single-nucleotide polymorphisms (SNPs) and co- or post-translational modifications like N-glycosylation or phosphorylation, respectively. Reprinted by permission from Macmillan Publishers Ltd.: NATURE, Aebersold, et al. Nat. Chem. Bio. 2018, 14, 206-214. Copyright 2018.

using polymerase chain reaction (PCR) allowing for extremely sensitive assays. No comparable expansion technology exists for polypeptides. PCR was just one of many developments that occurred in the DNA and RNA technology space during the 1970s and 80s. Nucleotide sequencing also allowed for reading of individual DNA bases. And, although Frederick Sanger had determined the sequence of bovine insulin by 1955[35,36], protein sequencing technologies such as Edman degradation were limited in scope, amenable protein size, and analysis speed[37]. Given these limitations, linking of the genome to proteins was frequently inferred using transcriptomics and estimations of genome accessibility despite discrepancies between these metrics and protein levels[38,39]. Interestingly, the 1995 manuscript in which Mark Wilkins originally coined the term "proteomics" relied on a combination of both mass spectrometry and several of these early proteomic technologies: Edman degradation and electrophoresis; all in pursuit of what was at the time an impressive characterization of 19 proteins[40,41].

Although molecular analysis by mass spectrometry (MS) has been in use since the early 20th century, its application to proteomics was limited by generation of gas-phase ions. The development of two techniques for ionization of large biomolecules revolutionized the field in 1990: electrospray ionization (ESI)[42] and matrix assisted laser desorption ionization (MALDI)[43,44]. The substantial impact of these techniques is underscored by the fact that their inventors – Koichi Tanaka (MALDI) and John Fenn (ESI) – were co-awarded the 2002 Nobel Prize in Chemistry[45] for their work. These techniques have allowed mass spectrometry to emerge as a premiere technology for proteomics since the turn of the century, with

the capacity for global proteome analysis as well as targeted peptide sequence investigation. This is evidenced by the key role mass spectrometry has played in characterizing the proteome of a variety of eukaryotic organisms[46–50], including humans, with 90% of proteins in the human proteome project supported by MS data[5]. More recent advances have enabled mass spectrometry to serve as a powerful tool in characterizing the proteomic heterogeneity of both healthy and diseased human populations, tissues and cells. However, as the field edges closer to complete coverage of the human proteome, it has become clear that the scope of its diversity and variability requires continued advancements in analysis speed, sensitivity, precision, and adaptability to novel sequences. This dissertation will present technologies and techniques to advance those objectives as well as demonstrate applications to human health that capitalize on the most current technologies.

Mass Spectrometry and Associated Technologies

Some basic familiarity with the operation of mass spectrometry technology is essential to the understanding of the work within this dissertation. Mass spectrometry relies on the predictable behavior of molecular ions in the gas phase to determine their ratio of mass to charge. The acceleration of molecular ions in an electric field is driven by only two characteristics, the force applied to them as dictated by their charge differential, and their mass. By accelerating and decelerating these ions using alternating electric fields we can determine the ratio of these two properties, the mass-to-charge, also known as the m/z or Thompson, the unit namesake of British physicist J. J. Thompson. The work described

in this dissertation relies on the Thermo Scientific Quadrupole-Orbitrap-Linear Ion Trap Hybrid Mass Spectrometers, abbreviated q-OT-QLT for the three mass analyzers contained within. Although two different models were used, the Orbitrap Fusion Lumos and the Orbitrap Eclipse, their configuration and operation are essentially identical (**Figure 1.2**). Chapter 3 and Chapter 5 also rely on the utilization of high field asymmetric waveform ion mobility spectrometry or FAIMS, which is both conceptually and physically described later.

The quadrupole mass filter The quadrupole mass filter functions by selectively transmitting ions based on their stable trajectory within an oscillating electric field. Physically, the quadrupole consists of four metal rods equally spaced around a central z-axis along the length of the rods (**Figure 1.3A**)[51]. The rods are electrically organized into pairs along the x- and y-axis, allowing the generation of differential voltage across both these axes (**Figure 1.3B**). A radio frequency (RF) waveform electric field is applied across each rod pair, constraining the ions as they pass along the z-axis of the device (**Figure 1.3A**). A supplemental direct current (DC) offset, termed the resolving DC, is applied with equal amplitude and opposite polarity for each rod pair, creating a positive and negative axis (**Figure 1.3B**). If no resolving DC was applied, ions would be constrained to the center path and no filtering would occur. However, the two different DC offsets destabilize different subpopulations of ions based on their m/z by causing them to drift into the poles and be neutralized. The positive DC offset pair imparts this destabilizing effect on the ions above a certain m/z, creating a low-pass filter. The negative DC offset pair in contrast, destabilizes the ions

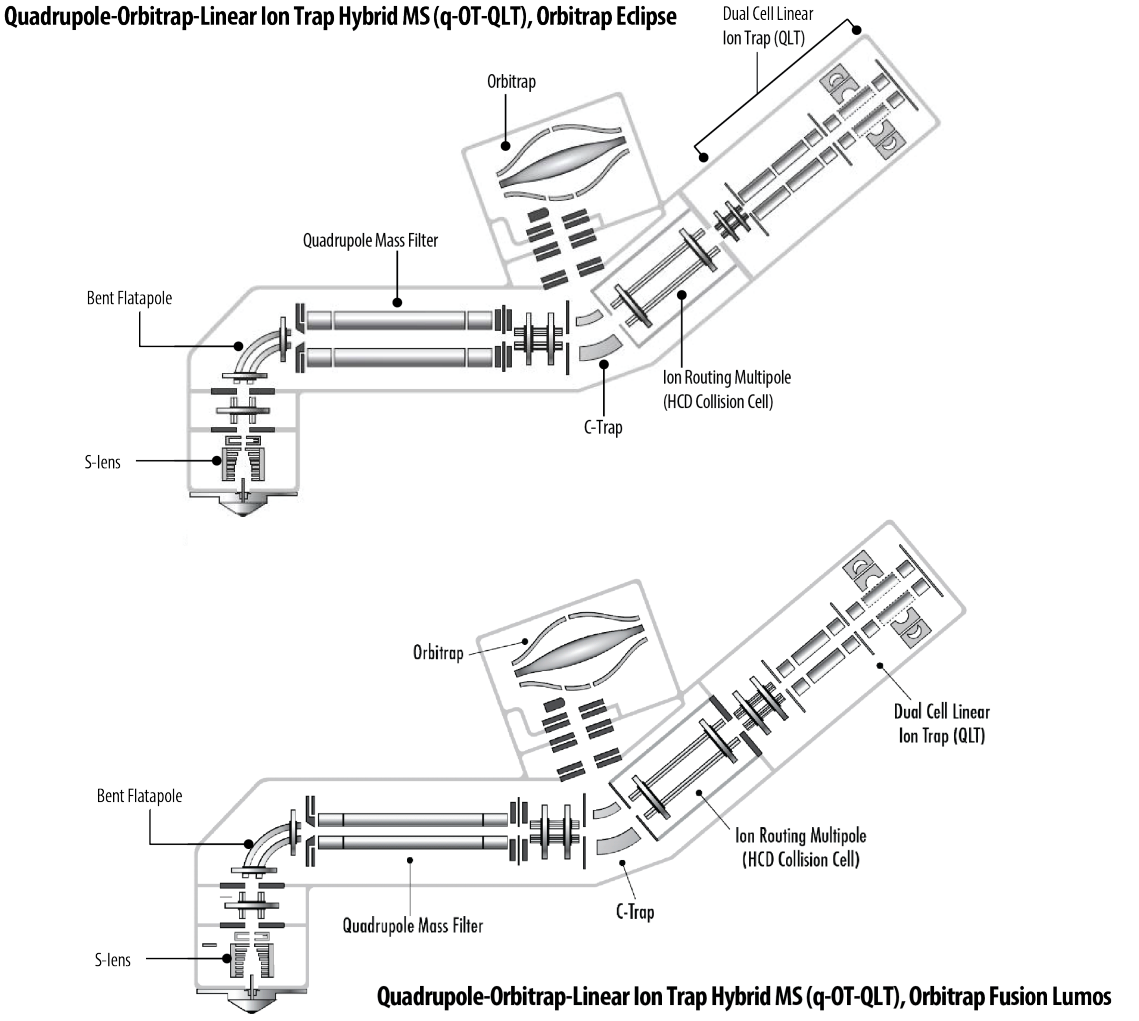


Figure 1.2: Instrument platform used for research. Both the Orbitrap Eclipse and the Orbitrap Fusion Lumos are quadrupole-Orbitrap-quadrupole linear ion trap (q-OT-QLT) hybrid MS systems.

below a certain mass-to-charge, providing a high-pass filter[52]. Together these two rod pairs allow for the selection and transmission of a specific m/z range known as the band pass region (**Figure 1.3C**). The width of this band pass region is controlled by the ratio of the amplitude of the RF waveform and the DC offset, allowing effective isolation at widths as low as 0.5 m/z[51] (**Figure 1.3D**). As this ratio decreases, the width of the isolation window increases along with the number of ions with stable trajectories through the quadrupole mass filter, leading to greater transmission (**Figure 1.3D**). Although addition of detection devices would allow for the quadrupole to be used in an analysis capacity, the work here utilizes it exclusively as a mass filter. This circular arrangement of rod pairs in quantities greater than four, such as hexapoles and octupoles, collectively called multipoles, are used in several other regions of the instrument for the transfer of ions between segments. The stability range for these multipoles is greater than that of the quadrupole making them favorable for "all pass" applications[53–55].

The ion trap mass analyzer Although data-dependent acquisition typically involves the Orbitrap as the first mass analyzer, the ion trap operates in a manner more similar to the quadrupole, and so will be addressed here. The hybrid instruments rely on a quadrupole linear ion trap (QLT) or 2D trap as opposed to the 3D or Paul trap due to a greater capacity for storage and sensitivity[52,56]. Much like the quadrupole, the ion trap consists of four long segmented electrodes, with parabolic cross sections. These electrodes are split into three sections: front, center and rear (**Figure 1.4A**). While the quadrupole provides a

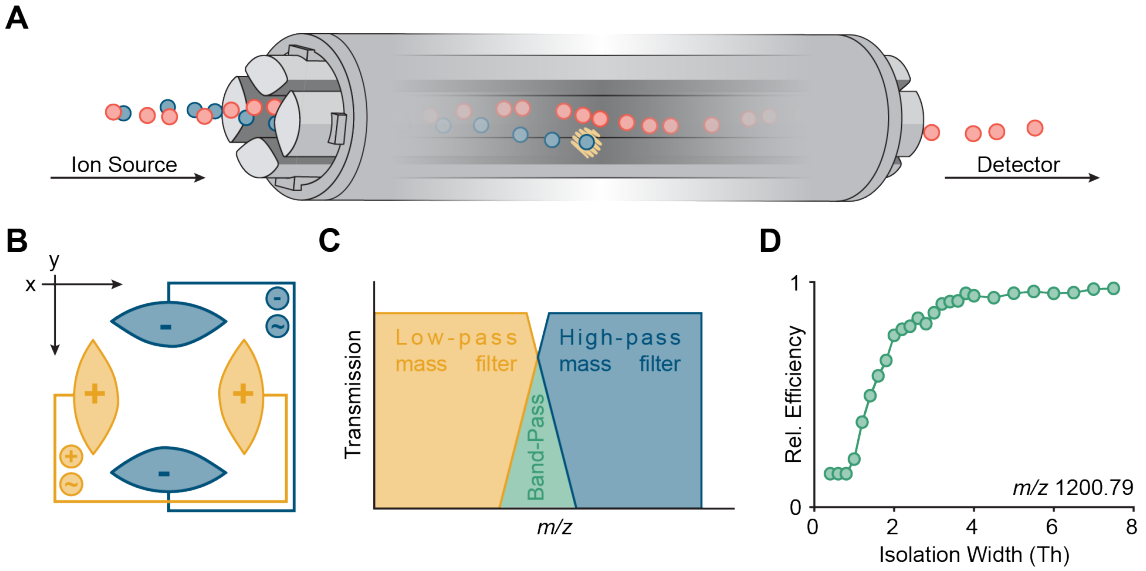


Figure 1.3: Quadrupole Mass Filter. (A) Cutaway illustration of a quadrupole mass filter for selective transmission of ions towards a detector.(B) Layout of opposing pairs of hyperbolic quadrupole rods connected electronically for generation of quadrupole RF. Here, the positive and negative symbols denote the sign of the rod pair DC offset, not its permanent polarity.(C) Combination of quadrupole low-pass and high-pass mass filters to isolate a specific m/z range of cations. Note: for isolation of anions, the DC offset/mass filter relationship is reversed.(D) Relative transmission efficiency for m/z 1200.79 ions collected across varying quadrupole isolation widths. Reprinted with permission from Paul D. Hutchins: Hutchins, 2019. Copyright 2019.

filtering component, the ion trap stores ions, destabilizing ions from a cloud rather than a beam. Ions are constrained in z-direction by a DC voltage potential well created by raising the voltage potential of the front and rear sections (**Figure 1.4B**). The ions in the trap are constrained in the x and y direction by an electric field with a RF waveform, termed the main rf, similar to that in the quadrupole. This RF waveform imparts kinetic energy causing secular motion, with the ions traveling in a roughly corkscrew path within the trap (**Figure 1.4B**). This kinetic energy and motion are proportional to the m/z value of the ions and the amplitude of the RF field. This energy is defined by the q-value as derived from the Mathieu equation[57]. Ions with smaller m/z have greater kinetic energy, and in turn larger q-values due to their greater force (as derived from charge) per unit mass. Due to these larger q-values, small m/z ions are destabilized by smaller amplitude fields. When ions are destabilized in the trap they are ejected through small slits in the side of the trap, where they collide with an electron detector, generating an electrical signal (**Figure 1.4A**). As the amplitude of the main RF is increased, ions of increasing m/z are ejected from the trap, striking the detector and generating sequential signal peaks across the designated range. These signal peaks can then be assigned to a particular point in m/z space based on the voltage amplitude required for their ejection from the trap. The resolution of these m/z identifications can be controlled by the speed with which the RF amplitude is increased[56]. In practice, the stability threshold of the ions is lowered by application of a supplemental RF to the center region of the trap. Once the increasing amplitude of the main RF has increased ions q-value to be in resonance with this supplemental RF, the ions experience a

rapid increase in energy causing their destabilization and ejection from the trap.

The Orbitrap mass analyzer The Orbitrap has become one of the premier mass analyzers for proteomics due to its capacity for high resolution accurate mass (HRAM) analysis, allowing discrimination of minor chemical modifications[58]. Prior to the Orbitrap's release in 2005[59], HRAM data could only be collected using Fourier transform ion cyclotron resonance (FT-ICR) instruments which measure ion frequency in uniform magnetic field[60,61]. These instruments required large cryogenically cooled magnets, leading to challenges in practicality and acquisition speed. In contrast to the ion trap, the Orbitrap measures m/z as a function of frequency rather than stability. Its physical design consists of a roughly cylindrical, axially symmetrical, inner electrode that widens towards the center of the z-axis, surrounded by a barrel-shaped outer electrode, split into two halves, with a slit to allow the entrance of ions (**Figure 1.5A**). Within the q-OT-QLT instruments, ions are collected in a secondary quadrupole linear ion trap called the ion routing multiple (IRM) (**Figure 1.2**). Once the appropriate ion count has been reached, ions move to another trapping device called the C-trap, due to the arced shape of its electrodes, which accelerates ions into the Orbitrap (**Figure 1.2**). Initially, a RF waveform constrains ions in the C-trap while also functioning to focus ions towards the center of the trap. Once all ions have been transferred into the C-trap the RF voltage of one electrode is rapidly lowered causing ions to accelerate through a series of ion guides into the Orbitrap (**Figure 1.2**). Upon ion entrance into the Orbitrap an initial voltage ramp at the center electrode causes the ions to squeeze towards the center

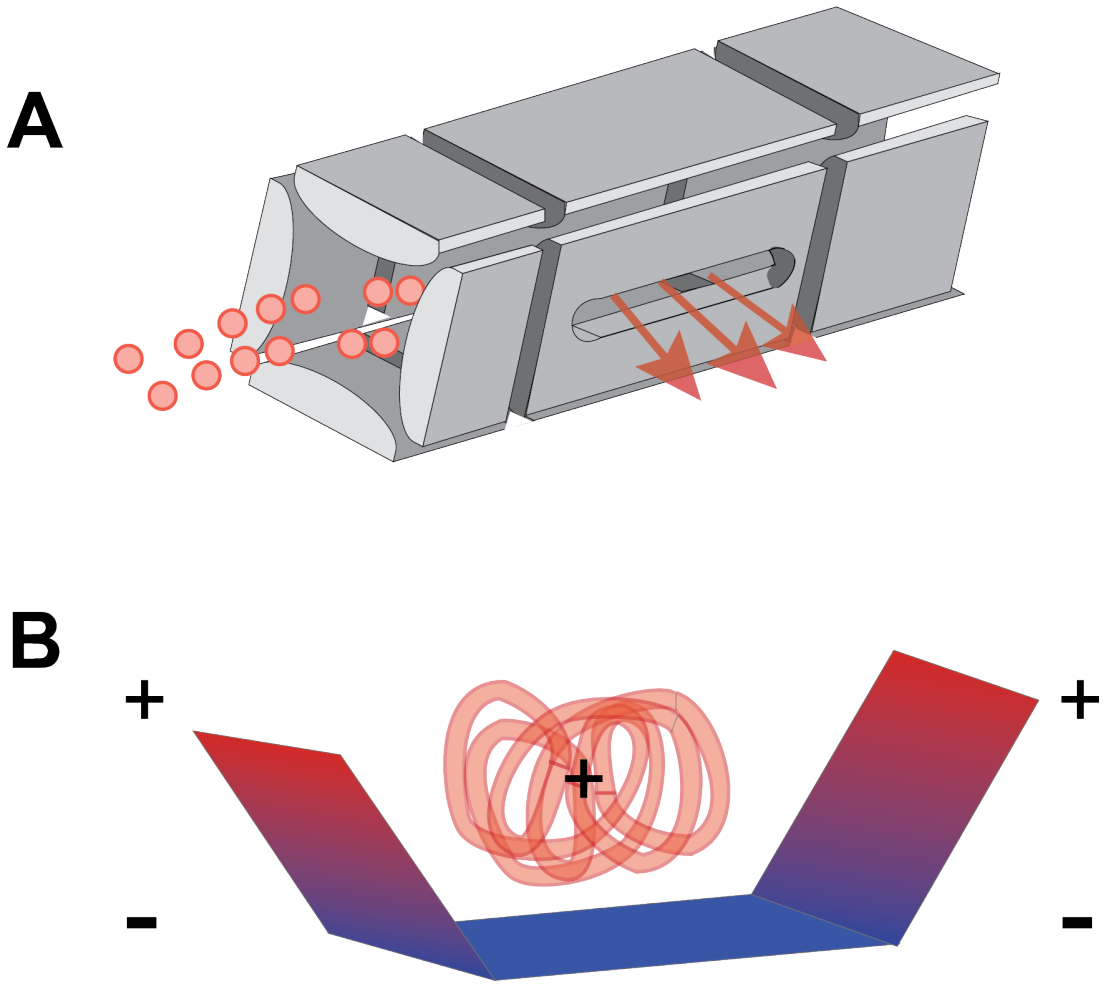


Figure 1.4: The two-dimensional ion trap. (A)The four quadrupolar rods of the 2D ion trap are split into three segments: a front section, a center section and a rear section. Ions enter along the z-axis. As ions are destabilized, they are ejected along the x-axis through the ejection slit, where they collide with a detector to generate signal. (B) As the main RF constrains ions in the x- and y-direction a DC trapping voltage creates "potential well" by raising the potential of the front and rear segments. These various constraining forces cause the secular motion of ions in the ion cloud.

and begin to spin radially around it, after which no additional voltage change is applied, and the trap operates electrostatically[62–64]. A precisely controlled quadro-logarithmic field generated from the shape of the electrodes causes ions to move axially along the central electrode as they rotate around it in a stable orbit (**Figure 1.5A**). The m/z of ions dictates their movement within this complex electrostatic field, causing ions to quickly form disc-like packets with specific frequencies of axial movement and radii of orbits (**Figure 1.5B**). Although several modes of detection were proposed at the inception of the Orbitrap[65], the path of these ions is detected using a current image taken by measuring and amplifying the current differential between the two halves of the outer electrode. This complex signal generated from the aggregate ion paths within the trap is converted into the frequency domain by a modified Fourier transform which is used to derive m/z and intensity[66,67] (**Figure 1.5C**). Increasing the duration of the scan, or the transient, allows for a greater number of ion oscillations and in turn greater m/z resolution. As the ions oscillate, their initial energy dissipates and ions begin to de-phase due to collisions with gas molecules, dampening the image current amplitude and providing some physical limitation to the resolution of the spectra collected.

Peptide dissociation strategies Although the Orbitrap has capacity for high resolution mass analysis, dissociation is required for greater peptide sequence elucidation in proteomics. Dissociation occurs when an unfragmented peptide ion, or precursor, absorbs enough energy to result in fragmentation of one or more it's molecular bonds[68] leading to

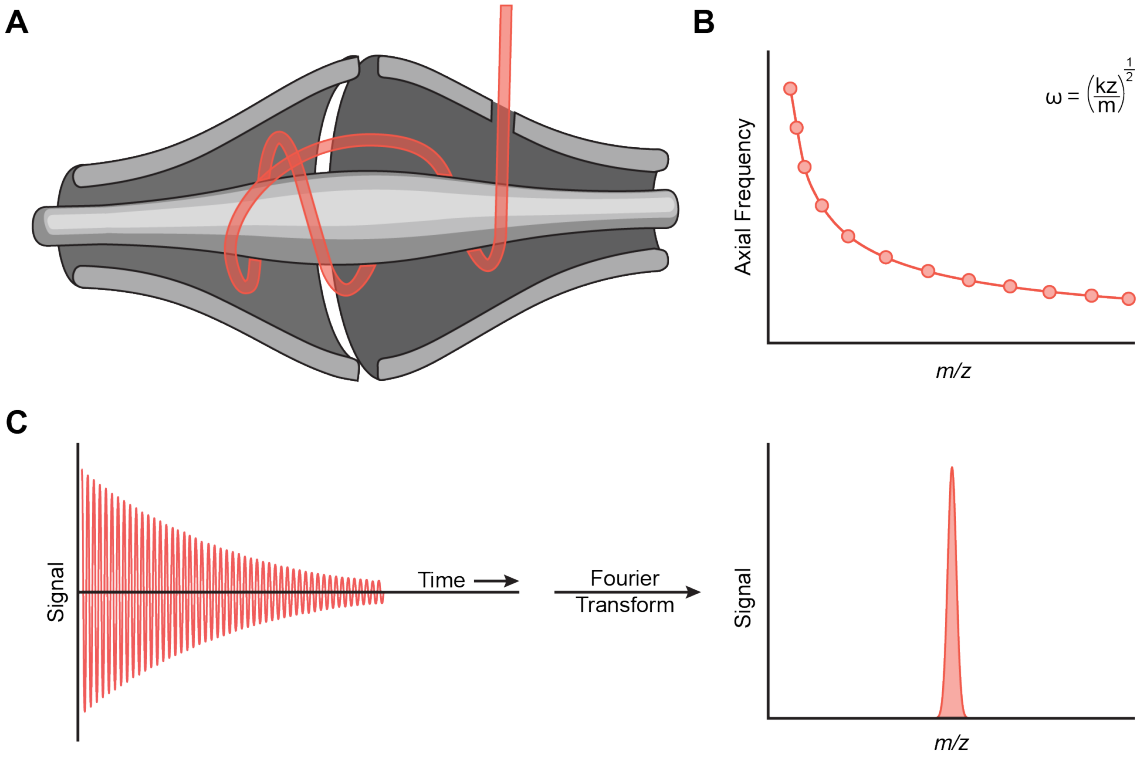


Figure 1.5: The Orbitrap mass analyzer. (A) Cutaway illustration of Orbitrap mass analyzer and depiction of complex ion movement around the center electrode. (B) Mathematical relationship between axial ion frequency and m/z in the Orbitrap which permits mass analysis. (C) Representative raw sinusoidal transient signal from single ion and conversion to m/z using a Fourier Transform. Reprinted with permission from Paul D. Hutchins: Hutchins, 2019. Copyright 2019.

the formation of fragment or product ions which can then be analyzed by MS/MS. These fragment ions are named based on the specific bond that is broken (**Figure 1.6**). This energy can be imparted by application of photons[69] or collisions with gas molecules[70], as well as chemical dissociation through electron transfer using a carrier gas[71]. The work within this dissertation relies on dissociation by collision with neutral gas molecules in a process called higher energy collision dissociation (HCD) or beam-type collision-induced dissociation. When performing HCD within the hybrid mass spectrometers, peptide precursors are accelerated through the C-trap and collide with nitrogen gas in the ion routing multipole. It is possible for this collisional energy to break several different bonds, including those in the amino acid side chains. However, HCD typically produces primarily b- and y-type fragments due to the thermally labile nature of the peptide bond.

High field asymmetric waveform ion mobility (FAIMS) FAIMS relies on the separation of molecules based on their differential mobility in a high and low amplitude electric field at atmospheric or near atmospheric pressure[71–73]. Although a variety of configurations of FAIMS exist[74–76] the FAIMS Pro interface utilized here consists of a cylindrical inner electrode and a cylindrically concave outer electrode between which gas phase pass before entering the mass spectrometer (**Figure 1.7A**)[77,78]. While passing through the device, ions are carried by flow of carrier gas towards the mass spectrometer and subjected to an asymmetric waveform electric field which alternates between high and low fields of opposite polarities (**Figure 1.7B**). The two fields are applied with different durations, proportional

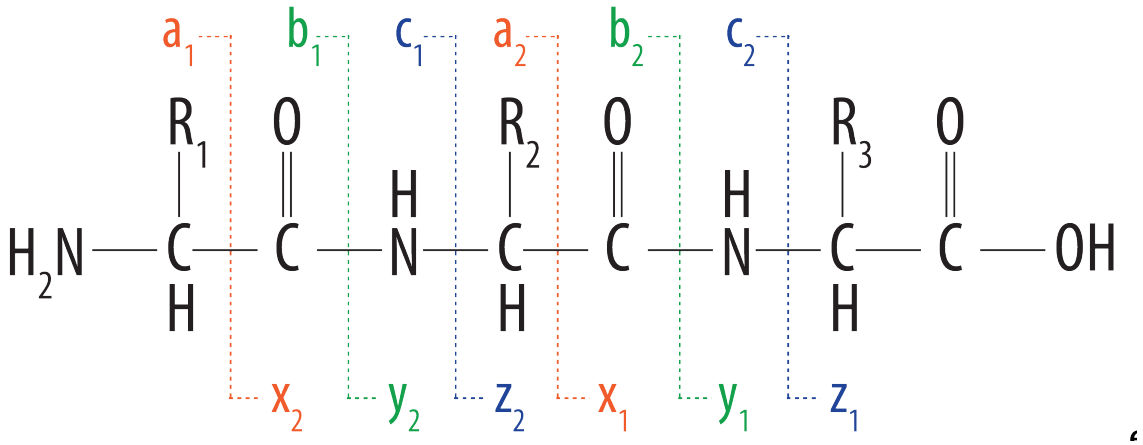


Figure 1.6: Possible peptide fragment ions. Different bond breakage leads to formation of different complementary fragment ion pairs with a, b and c complementing with x, y and z, respectively. Higher energy collisional dissociation utilized here promote primarily formation of b and y ions.

to their difference in amplitude, leading to equal area under the curve between the positive and negative polarities (**Figure 1.7B**). In a vacuum, this alternating field would lead to neutral displacement of the ions keeping them on a stable trajectory, similar to operating the quadrupole without a DC offset. However, due to the presence of gas molecules within the device, the mobility of peptide ions is altered based on the amplitude of the field, causing differential displacement between the high and low voltages, resulting in ion drift (**Figure 1.7C**). Without additional supplemental voltage, this drift would cause the collision of the peptide ions with one of the electrodes (**Figure 1.7C**). To prevent this, a stabilizing voltage is applied known as compensating voltage (CV). This voltage allows for a specific subpopulation of ions to be transmitted to the instrument based on their differential mobility in high and low fields, as the compensating voltage must neutralize the ion's specific displacement. This allows separation and filtration of gas-phase ions based on aspects of structure that affect mobility. Although FAIMS can be applied for a variety of charged analytes, the efficacious range for peptides largely falls between -10 V and -120 V.

Bottom-up Proteomics using Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS)

Bottom-up proteomics describes the process of proteomic characterization by measuring peptide ions within the mass spectrometer before assembling them into proteins by matching to a protein database. Although mass spectrometry technology has the capacity for sensitive measurements of abundance at high-resolution accurate masses, these capacities must be complemented by appropriate sample preparation strategies, chromatographic

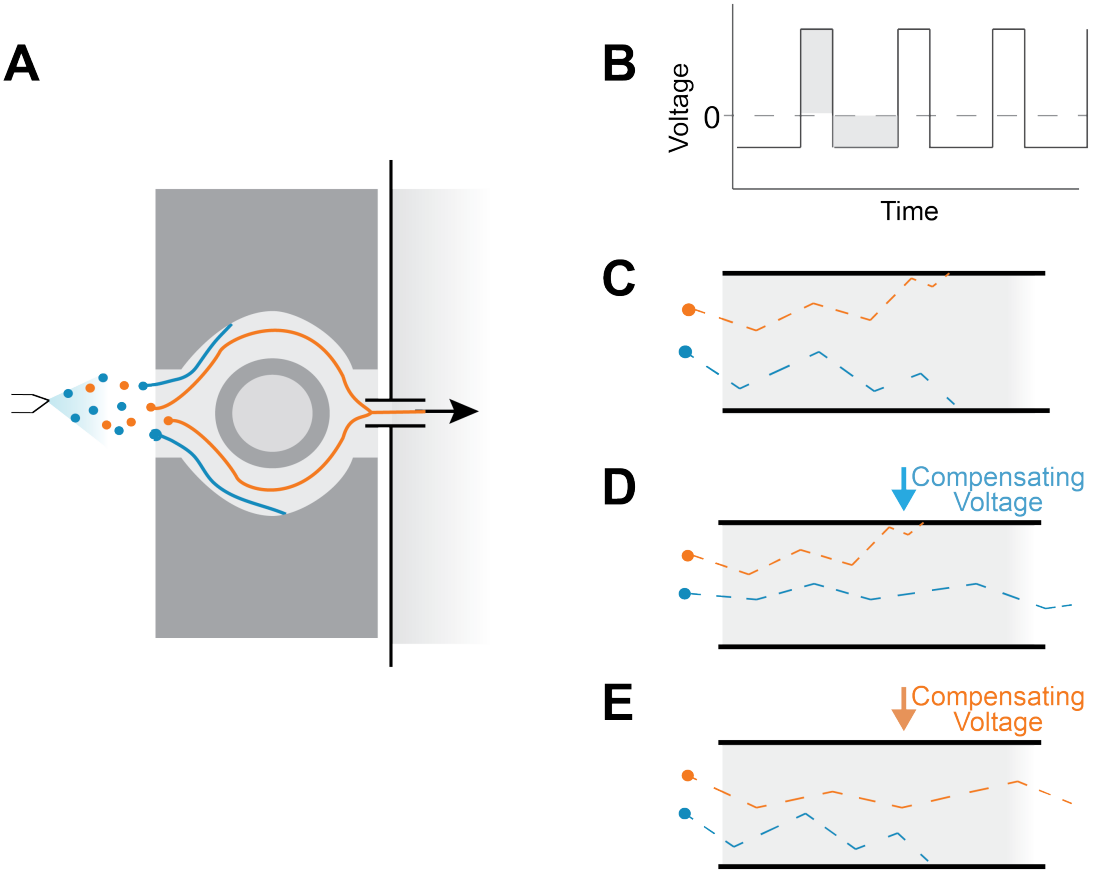


Figure 1.7: High field asymmetric waveform ion mobility spectrometry (FAIMS). (A) Schematic showing ion path between the inner and outer electrode prior to entrance into the instrument (B) Depiction of the asymmetric waveform of the destabilizing electric field. The two grey rectangles have equal area. (C) If no compensating voltage was applied peptide ions would collide with electrodes and be neutralized. (D, E) Different compensating voltage settings stabilize different subpopulations of ions allowing their transmission to the instrument.

separations, data acquisition, and data processing in order to reach their full potential.

Sample preparation Proteins must first be processed into amenable peptides to allow for their analysis by mass spectrometry. Cell or tissue structures are lysed to release intracellular proteins and those contained within organelles (**Figure 1.8**). Protein structure needs to be denatured, and disulfide bonds broken to allow access for proteases. This requires the application of chemical denaturants such as urea or guanidine as well as reducing agents such as chloroacetamide. The alkylation of cysteine side chains prevents the reformation of disulfide bonds. Once proteins have been linearized, they are digested into peptides at known motifs by proteases (**Figure 1.8**). The most widely used protease is trypsin, which cleaves the peptide backbone at the C-terminus of lysine and arginine residues. The popularity of trypsin in bottom-up proteomics stems from the generation of peptides with proton-accepting residues (lysine or arginine), increasing the average charge of peptides at low pH. Although specific reagents are named here, this general process of peptide preparation from mammalian cells and tissues exists in a state of constant development in the field, in a quest for increased efficiency, speed and proteomic coverage[79].

Liquid chromatography and ESI Once proteins have been denatured and digested, they are separated using liquid chromatography before being ionized and entering the instrument in the gas phase. A number of liquid chromatography systems exist with a variety of parameters including volumes, flow rates, and separation characteristics. Peptides can be

separated based on charge[80,81], size[82] and various chemical tags[83], all of which have been integrated with mass spectrometry. However, the work included in Chapters 2-5 of this dissertation relies on separation of peptides based on hydrophobicity, using reverse-phase liquid chromatography[84]. The typical material limitations of human samples favor the utilization of relatively small volumes (<10 microliters) and low flow rates (<500 nanoliters per minute), commonly termed nanoflow. In this system, peptide solutions flow through fused silica columns over beads with extended hydrocarbon tails. Peptides are loaded onto this column in a highly aqueous, acidic solution to promote interaction between hydrophobic elements and provide protons for charged droplet formation. An increasingly organic gradient elutes peptides sequentially based on their hydrophobicity.

When peptides in solution elute from the chromatographic column, they exit a narrow opening at the tip of the column forming a charged droplet (**Figure 1.8**). The column rests within an electric field with a voltage differential between the tip and the mass spectrometer inlet, causing electrospray ionization[42]. Although several models exist to explain the process of electrospray ionization and the conversion from droplets to gas-phase peptide ions[85], the result is a molecular beam of ions entering the mass spectrometer.

Data-dependent acquisition Peptide ions enter the instrument through a heated transfer capillary, helping to remove any remaining solvent. Given the large differential in pressure between the ionization source (1 bar) and the first vacuum-controlled region of the mass spectrometer(2 mbar), entering ions undergo a rapid jet expansion[85]. This expansion is

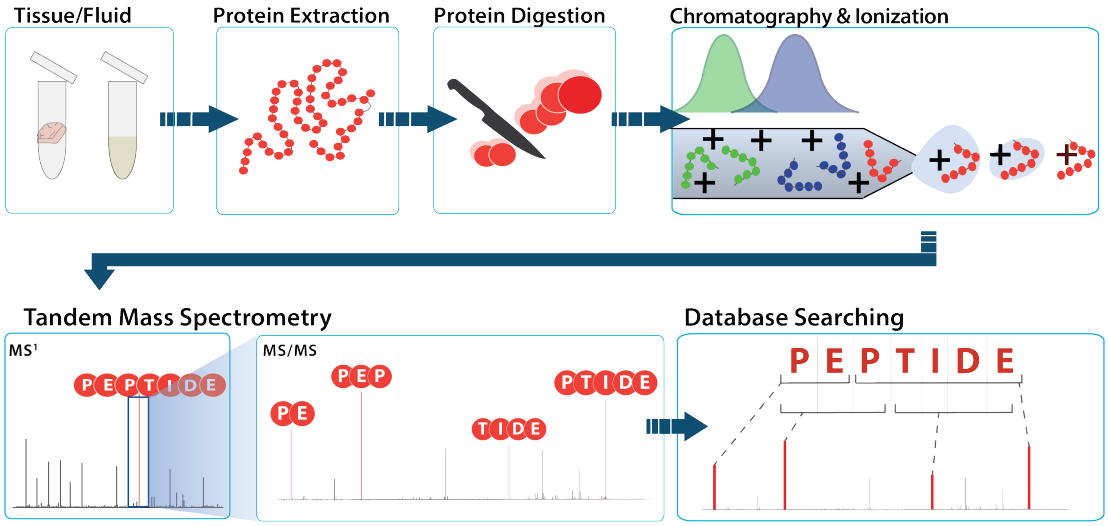


Figure 1.8: Sample preparation and analysis workflow. Proteins are first extracted from tissue or fluid before being denatured to allow for digestion by proteases. The resulting peptide mixture is then separated using liquid chromatography before ionization into the gas-phase to allow for mass spectrometry analysis. Precursor ions identified in survey scans are isolated and fragmented for scanning by MS/MS. These MS/MS scans are matched to a database of peptides digested and fragmented *in silico*.

counteracted by the focusing and constraining feature of the entrant ion funnel of the S-lens (**Figure 1.2**). In the S-lens an RF waveform is applied to a series of stacked electrodes along a narrowing path. The applied RF waveform of adjacent electrode plates are 180 degrees out of phase with one another providing a focusing force, with the narrowing path forming a roughly conical shape (**Figure 1.2**)[86]. After traveling through several ion optics, ions are transmitted through the bent flatpole, a set of electrodes turning at a 90-degree angle with an opening slit allowing for non-ionic molecules to be ejected before reaching the mass analyzers (**Figure 1.2**).

In data-dependent acquisition, a survey scan (MS^1), is used to identify precursor ions which are then isolated, fragmented and scanned in a secondary MS/MS (**Figure 1.9**). This process of isolation, fragmentation and scanning of fragments is commonly termed "sequencing". The multiple mass analyzers of the q-OT-QLT hybrid instruments (**Figure 1.2**) allow these different scans to occur in parallel, with precursor ions identified in $MS^1{}_n$ isolated and fragmented for MS/MS sequencing while $MS^1{}_{n+1}$ is scanning (**Figure 1.9**). Survey scans in the Orbitrap determine accurate mass and charge state of all eluting peptides at a given point in chromatographic retention time. Charge state is determined by the distribution of naturally occurring isotopes of each peptide species. For survey scans, the quadrupole mass filter is highly permissive, allowing a wide m/z range of ions to be transmitted and collected in the IRM. Once the correct number of ions has been collected as set by the user, this ion cloud is moved to the C-trap and enters into the Orbitrap for analysis. At the outset of acquisition, an initial survey scan must occur, from which precursor ions

are identified for MS/MS. In MS/MS, ions travel through the front lenses and enter the quadrupole in the same fashion, but rather than allowing all ions through, a narrow m/z window is isolated, centered around the target precursor's m/z (**Figure 1.9**). This isolated ion packet is transferred through the C-trap and accelerated into the IRM, where the ions collide with nitrogen gas, causing fragmentation (**Figure 1.2**). Once the ion quota has been met, the fragment ion packet is transferred to the terminal QLT where ions are sequentially ejected and detected as described above. Once a precursor has been sequenced via MS/MS its mass is placed on an exclusion list to avoid redundant sequencing events. The ion trap is typically utilized for MS/MS scans due to greater speed and sensitivity as compared to the Orbitrap. Precursor ions are typically preferential selected for sequencing based on intensity, moving from highest to lowest intensity (**Figure 1.9**). This dependence of sequencing events (MS/MS) on identification of precursor ions in survey scans leads to the data-dependent nomenclature.

Database searching This liquid chromatography tandem mass spectrometry (LC-MS/MS) acquisition strategy can produce more than 100,000 MS/MS spectra in only a single hour of analysis time. These spectra must then be matched against a database constructed from the full organismal proteome. Starting with all known organism proteins, proteins are digested and fragmented into full sets of appropriate fragments (here b- and y- ions) *in silico*. In addition, a set of known false peptides or decoy peptides is added to this peptide database as a metric of erroneous matching. These decoys are peptides known to be absent from

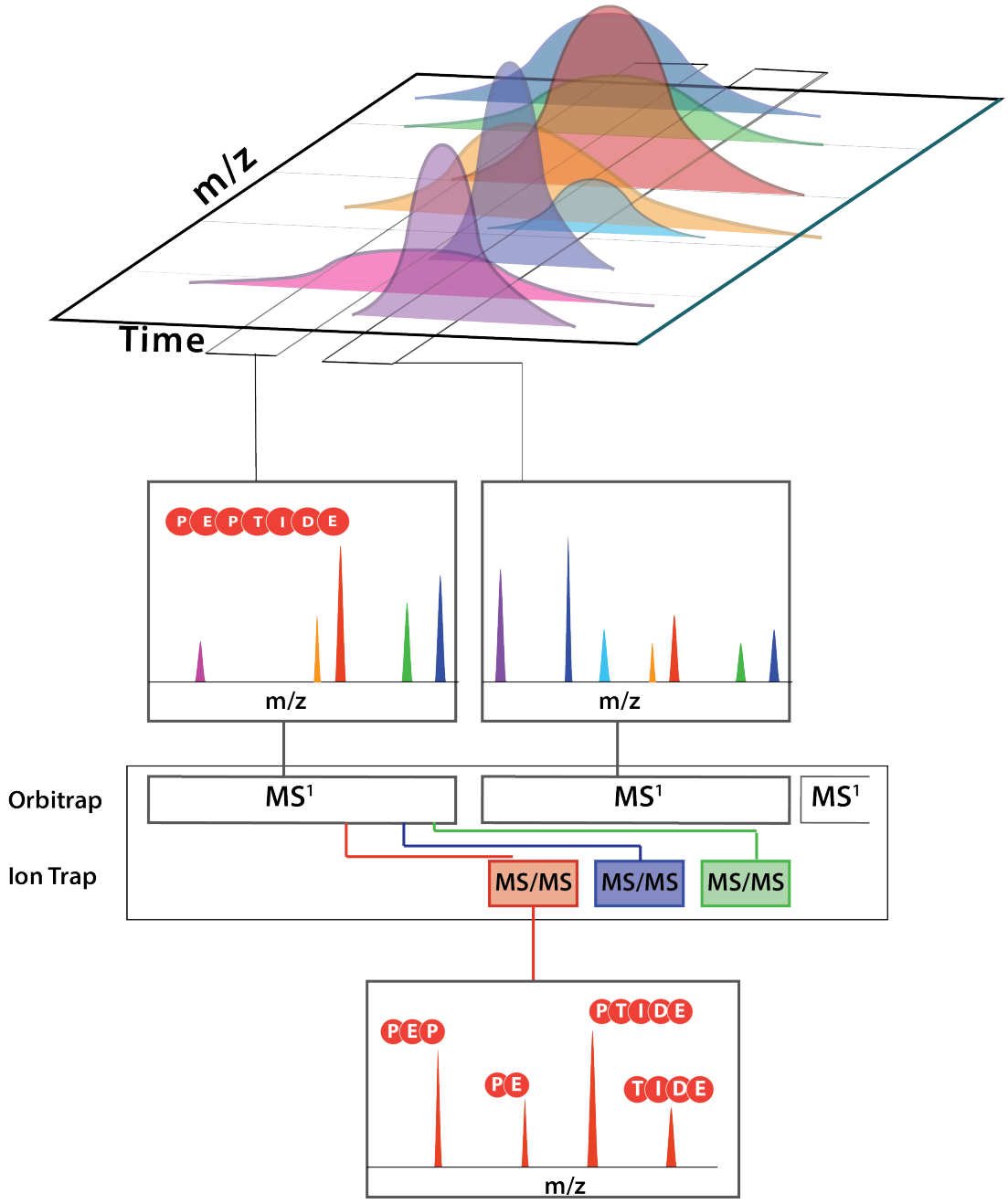


Figure 1.9: Simultaneous analysis of precursors and fragments in data dependent acquisition. As peptides elute from the chromatographic column, they are identified using a survey scan with a wide m/z window, typically collected in the Orbitrap. Identified peptides are then isolated, fragmented and scanned in the ion trap sequential while the next survey scan is completed in the Orbitrap. Parallelization allows for increased scan frequency and efficiency.

the sample and are often generated by shuffling or reversing the amino acid order of the true peptides. This strategy is termed the target-decoy approach and allows for users to measure and control the false positive rate, or the false discovery rate (FDR) of peptide spectral matches[87].

When data is searched, a narrow mass tolerance window is drawn around the accurate mass of each identified precursor. Database peptides falling within that window are considered for matching. When matching, the experimental MS/MS and the *in silico* generated fragment spectra are compared and scored based on similarity. Once all MS/MS have been matched and scored against the database, they are ordered by score, with some precursors matched incorrectly to decoy peptides. Assuming that erroneous matches to target peptides and decoy peptides are of equal likelihood, the percent of precursors incorrectly matched to decoy peptides provides an estimate of the FDR. Score cutoffs can then be selected based on the lowest similarity score in which results fall within the acceptable FDR. Typically operating within an FDR of 1% is common in proteomics[87].

This list of FDR-filtered peptides must then be reconstructed into the biologically relevant proteins. The protein inference process often operates parsimoniously, with the simplest combination of proteins explaining all identified peptides. In the eukaryotic proteome the existence of redundant peptide sequences and those derived from protein isoforms often leads to ambiguous peptide-protein assignment. These ambiguous assignments force the assembly of protein groups, in which all contained proteins could generate the included peptides. Once spectra, peptides, and proteins have been matched, relative

abundance can be determined. In label-free approaches, as applied here, relative quantitation is determined based on area-under-the-curve of each precursor across survey scans[87,88]. Although mass spectrometry instruments scan with relatively high frequency, differences in sample or conditions can lead to peptide sequencing in some experiments and not others. To compensate for this, precursors can be matched without MS/MS sequencing, if identified by sequencing in another experiment at the same m/z and chromatographic retention time. This strategy is called "match between runs" in the data searching software MaxQuant[89,90] and is based on the principles of accurate mass and time tags[91].

Analyzing Human Proteomes in Disease

In many ways, measuring protein expression in human disease represents a paramount challenge for proteomics. In addition to the challenging size of the human proteome, target abundance changes in tissues and biofluids can be relatively small in comparison to technical or biological variability[92–97]. Nucleotide polymorphisms increase population-level genomic diversity and can affect transcription, splicing, translation and stability of proteins in an individualized manner[98–100]. Researchers have gone to great lengths to replicate the mosaic nature of the human genome in other mammals, in hopes of controlling for these types of variables when studying disease[101]. The complexity of human biology leads to heterogeneous cell populations within many tissues and organs, further increasing variation. These proteins are not static, moving rapidly between organelles and cellular locations, as well as between intracellular and extracellular environments. In addition to physical

movement, proteins can transition between insoluble and soluble populations, and exist in an equilibrium between their translation and degradation. Virtually all forms of stimuli alter protein abundances, with age, sex, fasting, circadian rhythm, psychological stress, and the physical environment all having observed effects on the mammalian proteome[102–113]. Yet, it is this dynamism that makes proteins so crucial in understanding human health. Proteins, and in turn these transition events, are critical to the detection, development and treatment of disease.

The study of the age-associated neurodegenerative condition, Alzheimer's disease (AD), as discussed in Chapters 2 & 3, encapsulates many of these challenges. The microtubule associated protein tau (MAPT) and the amyloid precursor protein (APP), which eventually forms the pathogenic amyloid beta peptide, both play key roles in the disease. Despite identification of these key proteins more than 30 years ago[114,115] and the large health burden of AD[116], the mechanism by which aggregate formation leads to neuronal cell death remains unclear[117,118]. This is due in part to the difficulty in probing the human brain at a molecular level in living individuals, forcing the use of post-mortem tissues. In addition, highly diverse cell populations inhabit the different structures of the brain, with many of the major cells of the brain including neurons, microglia, and astroglia, exhibiting some regional expression patterns[119–121]. The interconnected function of these structures can make replication of this system difficult in model organisms[122,123] and cell culture, even with advances in organoids[124] and three-dimensional culturing[124,125]. Although cerebrospinal fluid (CSF) provides a more accessible view into the central nervous system,

with greater ease in sampling from living participants, it also comes with challenges. Like many biofluids[126,127], CSF contains a small number of high-abundance proteins leading to a large dynamic range that can impede proteomic depth and the detection of smaller magnitude, biologically relevant changes[128–130]. The variability in AD also extends beyond the CNS tissues, with substantial phenotypic heterogeneity in terms of both symptoms[131,132] and disease progression[133]. This variation in symptoms has proved a confounding factor in a number of recent studies[133,134] and indicates the need for further investigation into the associated proteomic variability.

These complexities are found across human tissues and disease but can be overcome by MS-based experiments with altered study design, preparation, and acquisition, as exemplified by mass spectrometry's use in identifying several recent therapeutic targets[79,135]. Nearly all pathological proteomic investigations center around two objectives: (1) more detailed characterization of the molecular mechanism of disease; and (2) identification of additional biomarkers to improve the detection and diagnosis of the target condition. These two objectives require multifaceted investigation strategies that combine both targeted and global analyses. MS-based proteomics allows for quantitation of a wide breadth of proteins in a highly sensitive and precise manner when analysis strategies address the difficulties associated with human proteomics, as well as the specific target populations, tissues or cell types. MS has the capacity to identify physiologically significant proteins from tissues with limits of quantitation in the picomolar range, well within the protein content of biofluids[136]. Mass spectrometry can also survey deep into the human proteome,

with precise preparation and data acquisition quantifying more than 14,000 proteins in brain tissue[137]. However, proteomic depth and quantitation must also be balanced with throughput to allow for the analysis of large sample numbers. Methods must be developed, and studies designed, to allow for increased proteomic depth, controlled variability, and precise quantitation if these pathologically relevant proteomic shifts are to be detected in humans. Several strategies and technologies in pursuit of these goals will be utilized or advanced in this dissertation.

Additional separation and simplification One strategy for increased proteomic depth comes from simplification by additional separation of the peptide mixture, either in solution, using offline liquid chromatography, or in the gas phase, using rapid ion mobility. In solution, a complex peptide sample is separated into a series of fractions using a chromatographic separation orthogonal to that utilized on-line with the mass spectrometer, in a process called 2D(2D-LC) or multidimensional liquid chromatography. Each one of these fractions is then run as an individual acquisition experiment. Fractions with distal retention times can also be concatenated to reduce acquisition time. For example, if 16 fractions are collected, one is combined with eight, two is combined with 9, and so on. This simplification reduces the number of precursors in each survey scan, increasing the instrument's capacity to identify, isolate and sequence low abundance precursors[138] (**Figure 1.10B**). Although higher fraction numbers increase simplification of the mixture, and in turn the benefits to proteomic depth, this results in steadily diminishing returns, while increasing

run time and material requirements[139]. In the gas phase, high field asymmetric waveform ion mobility spectrometry (FAIMS) can be utilized to similar effect. Whereas formation of liquid fractions separates complex peptide mixtures into simplified liquid fractions, FAIMS acts as a post-ionization filter, transmitting only a specific subpopulation of ions into the instrument at a particular compensating voltage (CV) setting (**Figure 1.6**). The compensating voltage-specificity of peptides is not directly related to their hydrophobicity (which controls online reverse-phase chromatographic separation), allowing filtration to occur sequentially and orthogonally to the preceding chromatography (**Figure 1.10C**). This filter is applied throughout the duration of the acquisition experiment, before repeating using a different CV setting, allowing the selective transmission of another gas-phase fraction (**Figure 1.10C**). The increased proteomic depth derived from FAIMS operates on a similar concept to liquid fractionation, with fewer precursors improving identification, isolation, and sequencing of lower abundance peptide ions[140,141]. FAIMS also filters out singly charged contaminants, reducing chemical noise and enriching for productive multiply charged tryptic peptides[142]. Although solution-based fractionation at high numbers (>8) provides greater benefits to proteomic depth[139,143], these benefits come at the expense of additional preparation time not required in the application of FAIMS. Liquid fractionation was utilized in Chapters 2 and 3, while FAIMS was applied in this manner in Chapters 3 and 5.

Parallel reaction monitoring (PRM) Modifications to acquisition strategy can also allow for increased quantitative precision and reproducibility. One targeted analysis strategy

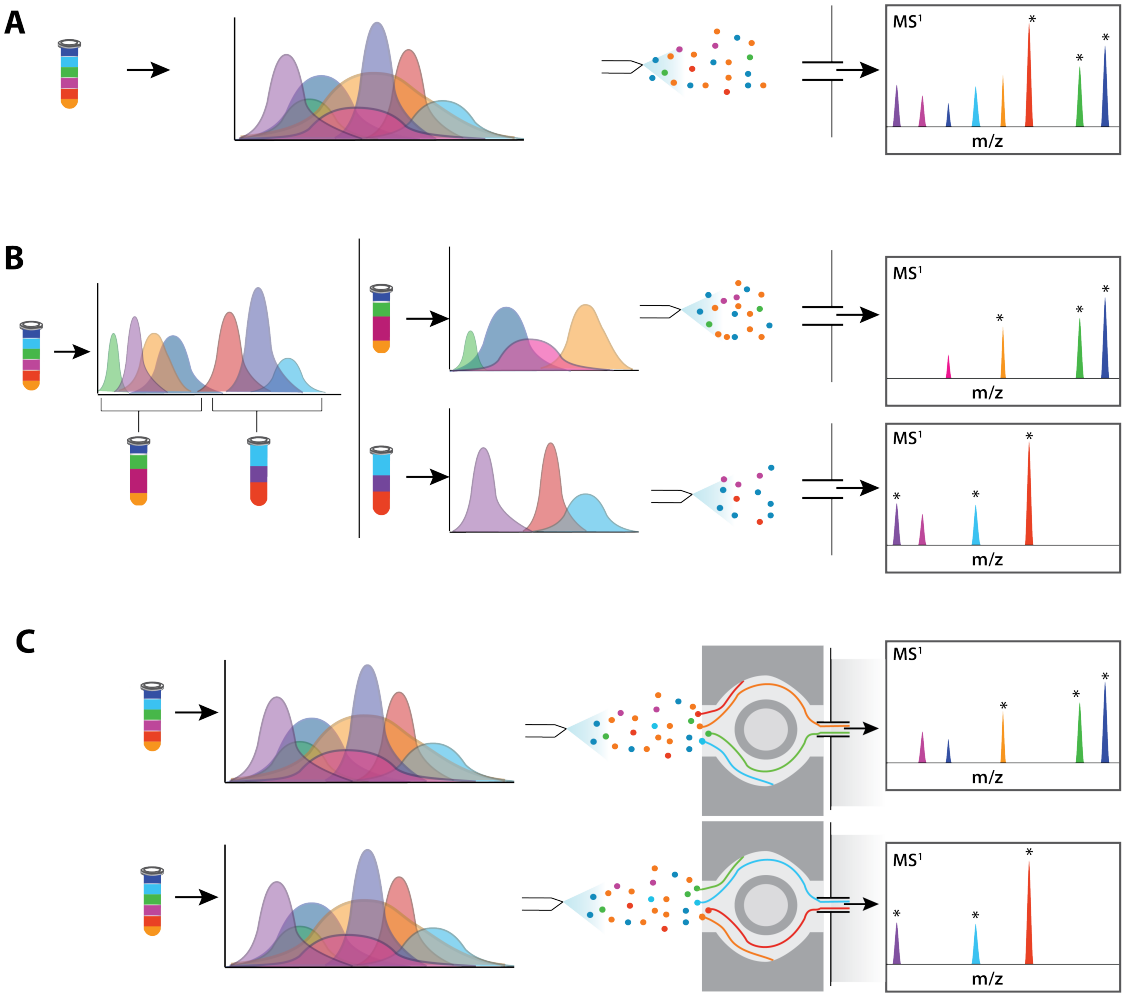


Figure 1.10: Increased depth from additional separations. (A) With complex peptide mixtures and limited time only the some precursors are identified and selected for sequencing, as indicated with an asterisk (B) Splitting the original complex mixture into simplified fractions leads to reduced chromatographic and MS1 complexity, allowing isolation and sequencing of low abundance ions (C) Simplification using two compensating voltages in FAIMS post-ionization provides a similar result, without changing the chromatogram or additional sample preparation upfront.

utilized here- called parallel reaction monitoring (PRM)[144]- relies on a priori knowledge about peptides of interest. Data-dependent experiments are used to build a database of target precursor masses and their associated fragment spectra. These are then combined into an inclusion list of precursor m/z values. During acquisition, the instrument repeatedly cycles through the list, isolating windows centered on each target, collecting ions and fragmenting them before scanning the fragments. In contrast to the data-dependent acquisition described above, these fragment scans are performed in the Orbitrap mass analyzer to allow for the collection of high-resolution spectra and simultaneous scanning of the full mass range[144]. Peptide precursors are then quantified based on the intensity peaks of their fragment ions. This strategy allows for highly robust and reproducible measurements of individual peptide precursors. Quantitation of individual peptides also has substantial biological relevance in tissues rich in signaling peptides and active proteases, such as cerebrospinal fluid and saliva, as discussed in Chapters 3 and 4 respectively.

Study design and statistics Variability in the human proteome also can be partially accounted for in study design and statistical modeling. Distribution of characteristics known to influence the proteome can be matched between cases and controls when examining disease. Even greater proteomic variability can be controlled for by collection of longitudinal samples from the same individual after exposure to a stimulus or development of disease. Statistical models can be built to reflect this style of experiment, such as mixed effect models, treating individual proteomic variation as a random systematic effect[145,146]. A variety of

statistical corrections for multiple hypotheses exist and should be utilized for the separation of true biological effects from proteomic noise[147–149], especially as the number of proteins quantified in a single experiment continues to grow. False conclusions can also be reduced by contextualizing proteomic findings within existing data using public repositories.

Overall mass spectrometry and the associated technologies provide a powerful tool for the investigation of the alterations of proteins in human disease. These technologies allow direct measurements of the protein abundances rather than relying on the inference that occurs in analysis of nucleic acid through genomics and transcriptomics. Although proteomic analysis of human tissues presents many challenges, these can be overcome through specialized sample preparation, acquisition strategies, study design and statistical inquiry. This dissertation includes analysis of two health conditions: (1) Alzheimer's disease; and (2) physical and psychological stress of a simulated combat exercise. I discuss findings in those studies as well as introduce advances in analysis strategies for human tissues, and more specifically biofluids using the FAIMS technology in tandem with LC-MS/MS.

References

[1] International Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, pp. 931–945, Oct. 2004.

[2] T. G. Wolfsberg, J. McEntyre, and G. D. Schuler, "Guide to the draft human genome,"

Nature, vol. 409, pp. 824–826, Feb. 2001.

[3] B. L. Aken, P. Achuthan, W. Akanni, M. R. Amode, F. Bernsdorff, J. Bhai, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, T. Juettemann, S. Keenan, M. R. Laird, I. Lavidas, T. Maurel, W. McLaren, B. Moore, D. N. Murphy, R. Nag, V. Newman, M. Nuhn, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, S. P. Wilder, A. Zadissa, M. Kostadima, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, F. Cunningham, A. Yates, D. R. Zerbino, and P. Flicek, "Ensembl 2017," *Nucleic Acids Res.*, vol. 45, pp. D635–D642, Jan. 2017.

[4] T. Reynolds, "For proteomics research, a new race has begun," 2002.

[5] G. S. Omenn, L. Lane, C. M. Overall, I. M. Cristea, F. J. Corrales, C. Lindskog, Y.-K. Paik, J. E. Van Eyk, S. Liu, S. R. Pennington, M. P. Snyder, M. S. Baker, N. Bandeira, R. Aebersold, R. L. Moritz, and E. W. Deutsch, "Research on the human proteome reaches a major milestone: >90% of predicted human proteins now credibly detected, according to the HUPO human proteome project," *J. Proteome Res.*, vol. 19, pp. 4735–4746, Dec. 2020.

[6] M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, H. Wernerus, L. Björling, and F. Pon-

ten, "Towards a knowledge-based human protein atlas," *Nat. Biotechnol.*, vol. 28, pp. 1248–1250, Dec. 2010.

[7] P. Gaudet, P.-A. Michel, M. Zahn-Zabal, A. Britan, I. Cusin, M. Domagalski, P. D. Duek, A. Gateau, A. Gleizes, V. Hinard, V. Rech de Laval, J. Lin, F. Nikitin, M. Schaeffer, D. Teixeira, L. Lane, and A. Bairoch, "The neXtProt knowledgebase on human proteins: 2017 update," *Nucleic Acids Res.*, vol. 45, pp. D177–D182, Jan. 2017.

[8] UniProt Consortium, "UniProt: the universal protein knowledgebase in 2021," *Nucleic Acids Res.*, vol. 49, pp. D480–D489, Jan. 2021.

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

[10] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg,

S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Pontén, "Proteomics. tissue-based map of the human proteome," *Science*, vol. 347, p. 1260419, Jan. 2015.

[11] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nat. Genet.*, vol. 40, pp. 1413–1415, Dec. 2008.

[12] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge, "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 456, pp. 470–476, Nov. 2008.

[13] G. A. Khoury, R. C. Baliban, and C. A. Floudas, "Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database," *Sci. Rep.*, vol. 1, Sept. 2011.

[14] A. M. N. Silva, R. Vitorino, M. R. M. Domingues, C. M. Spickett, and P. Domingues, "Post-translational modifications and mass spectrometry detection," *Free Radic. Biol. Med.*, vol. 65, pp. 925–941, Dec. 2013.

[15] D. R. Robinson, Y. M. Wu, and S. F. Lin, "The protein tyrosine kinase family of the human genome," *Oncogene*, vol. 19, pp. 5548–5557, Nov. 2000.

[16] S. S. H. Weng, F. Demir, E. K. Ergin, S. Dirnberger, A. Uzozie, D. Tuscher, L. Nierves, J. Tsui, P. F. Huesgen, and P. F. Lange, "Sensitive determination of proteolytic proteoforms in limited microscale proteome samples," *Mol. Cell. Proteomics*, vol. 18, pp. 2335–2347, Nov. 2019.

[17] T. Klein, U. Eckhard, A. Dufour, N. Solis, and C. M. Overall, "Proteolytic Cleavage-Mechanisms, function, and "omic" approaches for a Near-Ubiquitous posttranslational modification," *Chem. Rev.*, vol. 118, pp. 1137–1168, Feb. 2018.

[18] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, pp. 198–207, Mar. 2003.

[19] B. Z. Zmudzka, A. Fomace, J. Collins, and S. H. Wilson, "Characterization of DNA polymerase β mRNA: cell-cycle and growth response in cultured human cells," 1988.

[20] R. D. Kornberg, "Eukaryotic transcriptional control," *Trends Cell Biol.*, vol. 9, pp. M46–9, Dec. 1999.

[21] S. Klinge, F. Voigts-Hoffmann, M. Leibundgut, S. Arpagaus, and N. Ban, "Crystal structure of the eukaryotic 60S ribosomal subunit in complex with initiation factor 6," *Science*, vol. 334, pp. 941–948, Nov. 2011.

[22] A. Ben-Shem, N. G. de Loubresse, S. Melnikov, L. Jenner, G. Yusupova, and M. Yusupov, "The structure of the eukaryotic ribosome at 3.0 a resolution. this entry contains proteins of the 40S subunit, ribosome B," 2011.

[23] B. L. Stoddard, "Homing endonuclease structure and function," *Q. Rev. Biophys.*, vol. 38, pp. 49–95, Feb. 2005.

[24] A. I. Jonckheere, J. A. M. Smeitink, and R. J. T. Rodenburg, "Mitochondrial ATP synthase: architecture, function and pathology," 2012.

[25] X. Zhang, M. H. Chen, X. Wu, A. Kodani, J. Fan, R. Doan, M. Ozawa, J. Ma, N. Yoshida, J. F. Reiter, D. L. Black, P. V. Kharchenko, P. A. Sharp, and C. A. Walsh, "Cell-Type-Specific alternative splicing governs cell fate in the developing cerebral cortex," *Cell*, vol. 166, pp. 1147–1162.e15, Aug. 2016.

[26] T. H. Chae, S. Kim, K. E. Marz, P. I. Hanson, and C. A. Walsh, "The hyh mutation uncovers roles for αSnap in apical protein localization and control of neural cell fate," 2004.

[27] H. Kennedy, R. Douglas, K. Knoblauch, and C. Dehay, "Self-organization and pattern formation in primate cortical networks," *Novartis Found. Symp.*, vol. 288, pp. 178–94 discussion 195–8, 276–81, 2007.

[28] X. Li, Y. Tao, R. Bradley, Z. Du, Y. Tao, L. Kong, Y. Dong, J. Jones, Y. Yan, C. R. K. Harder, L. M. Friedman, M. Bilal, B. Hoffmann, and S.-C. Zhang, "Fast generation of functional subtype astrocytes from human pluripotent stem cells," *Stem Cell Reports*, vol. 11, pp. 998–1008, Oct. 2018.

[29] C. Dehay and H. Kennedy, "Cell-cycle control and cortical development," *Nat. Rev. Neurosci.*, vol. 8, pp. 438–450, June 2007.

[30] B. Huang, X. Li, X. Tu, W. Zhao, D. Zhu, Y. Feng, X. Si, and J.-G. Chen, "OTX1 regulates cell cycle progression of neural progenitors in the developing cerebral cortex," 2018.

[31] J. M. Berg, J. L. Tymoczko, G. J. Gatto, Jr, and L. Stryer, *Biochemistry*. W. H. Freeman, Apr. 2015.

[32] B. Pan, G. S. Géléoc, Y. Asai, G. C. Horwitz, K. Kurima, K. Ishikawa, Y. Kawashima, A. J. Griffith, and J. R. Holt, "TMC1 and TMC2 are components of the mechanotransduction channel in hair cells of the mammalian inner ear," *Neuron*, vol. 79, pp. 504–515, Aug. 2013.

[33] B. Coste, J. Mathur, M. Schmidt, T. J. Earley, S. Ranade, M. J. Petrus, A. E. Dubin, and A. Patapoutian, "Piezo1 and piezo2 are essential components of distinct mechanically activated cation channels," *Science*, vol. 330, pp. 55–60, Oct. 2010.

[34] S. S. Ranade, S.-H. Woo, A. E. Dubin, R. A. Moshourab, C. Wetzel, M. Petrus, J. Mathur, V. Bégay, B. Coste, J. Mainquist, A. J. Wilson, A. G. Francisco, K. Reddy, Z. Qiu, J. N. Wood, G. R. Lewin, and A. Patapoutian, "Piezo2 is the major transducer of mechanical forces for touch sensation in mice," *Nature*, vol. 516, pp. 121–125, Dec. 2014.

[35] F. Sanger, "Some chemical investigations on the structure of insulin," *Cold Spring Harb. Symp. Quant. Biol.*, vol. 14, pp. 153–160, 1950.

[36] A. P. Ryle, F. Sanger, L. F. Smith, and R. Kitai, "The disulphide bonds of insulin," *Biochem. J*, vol. 60, pp. 541–556, Aug. 1955.

[37] W. Konigsberg, "[51] subtractive edman degradation," 1967.

[38] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold, "Correlation between protein and mRNA abundance in yeast," *Mol. Cell. Biol.*, vol. 19, pp. 1720–1730, Mar. 1999.

[39] T. J. Griffin, S. P. Gygi, T. Ideker, B. Rist, J. Eng, L. Hood, and R. Aebersold, "Complementary profiling of gene expression at the transcriptome and proteome levels in saccharomyces cerevisiae," *Mol. Cell. Proteomics*, vol. 1, pp. 323–333, Apr. 2002.

[40] V. C. Wasinger, S. J. Cordwell, A. Cerpa-Poljak, J. X. Yan, A. A. Gooley, M. R. Wilkins, M. W. Duncan, R. Harris, K. L. Williams, and I. Humphery-Smith, "Progress with gene-product mapping of the mollicutes: Mycoplasma genitalium," *Electrophoresis*, vol. 16, pp. 1090–1094, July 1995.

[41] M. R. Wilkins, C. Pasquali, R. D. Appel, K. Ou, O. Golaz, J. C. Sanchez, J. X. Yan, A. A. Gooley, G. Hughes, I. Humphery-Smith, K. L. Williams, and D. F. Hochstrasser, "From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis," *Biotechnology*, vol. 14, pp. 61–65, Jan. 1996.

[42] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse, "Electrospray ionization for mass spectrometry of large biomolecules," *Science*, vol. 246, pp. 64–71, Oct. 1989.

[43] K. Tanaka, H. Waki, Y. Ido, S. Akita, Y. Yoshida, T. Yoshida, and T. Matsuo, "Protein and polymer analyses up tom/z 100 000 by laser ionization time-of-flight mass spectrometry," 1988.

[44] M. Karas, D. Bachmann, U. Bahr, and F. Hillenkamp, "Matrix-assisted ultraviolet laser desorption of non-volatile compounds," 1987.

[45] J. B. Fenn, "Electrospray wings for molecular elephants (nobel lecture)," *Angew. Chem. Int. Ed Engl.*, vol. 42, pp. 3871–3894, Aug. 2003.

[46] A. Kumar, "Faculty opinions recommendation of proteomics. tissue-based map of the human proteome," 2016.

[47] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "The one hour yeast proteome," *Mol. Cell. Proteomics*, vol. 13, pp. 339–347, Jan. 2014.

[48] J. A. Stefely, N. W. Kwiecien, E. C. Freiberger, A. L. Richards, A. Jochem, M. J. P. Rush, A. Ulbrich, K. P. Robinson, P. D. Hutchins, M. T. Veling, X. Guo, Z. A. Kemmerer, K. J. Connors, E. A. Trujillo, J. Sokol, H. Marx, M. S. Westphall, A. S. Hebert, D. J.

Pagliarini, and J. J. Coon, "Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling," *Nat. Biotechnol.*, vol. 34, pp. 1191–1197, Nov. 2016.

[49] J. M. Proffitt, J. Glenn, A. J. Cesnik, A. Jadhav, M. R. Shortreed, L. M. Smith, K. Kavanagh, L. A. Cox, and M. Olivier, "Proteomics in non-human primates: utilizing RNA-Seq data to improve protein identification by mass spectrometry in vervet monkeys," *BMC Genomics*, vol. 18, p. 877, Nov. 2017.

[50] A. L. Richards, A. E. Merrill, and J. J. Coon, "Proteome sequencing goes deep," *Curr. Opin. Chem. Biol.*, vol. 24, pp. 11–17, Feb. 2015.

[51] R. E. Pedder, "Practical quadrupole theory: graphical theory," *Excel Core Mass Spectrometers, Pittsburgh, PA, Extrel Application Note RA_2010 A*, 2001.

[52] J. P. Savaryn, T. K. Toby, and N. L. Kelleher, "A researcher's guide to mass spectrometry-based proteomics," *Proteomics*, vol. 16, pp. 2435–2443, Sept. 2016.

[53] L. A. McDonnell, A. E. Giannakopulos, P. J. Derrick, Y. O. Tsybin, and P. Håkansson, "A theoretical investigation of the kinetic energy of ions trapped in a Radio-Frequency hexapole ion trap," 2002.

[54] N. Konenkov, F. Londry, C. Ding, and D. J. Douglas, "Linear quadrupoles with added hexapole fields," *J. Am. Soc. Mass Spectrom.*, vol. 17, pp. 1063–1073, Aug. 2006.

[55] X. Zhao, O. Granot, and D. J. Douglas, "Quadrupole excitation of ions in linear quadrupole ion traps with added octopole fields," 2008.

[56] J. C. Schwartz, M. W. Senko, and J. E. P. Syka, "A two-dimensional quadrupole ion trap mass spectrometer," *J. Am. Soc. Mass Spectrom.*, vol. 13, pp. 659–669, June 2002.

[57] R. E. March, "Quadrupole ion trap mass spectrometer," 2006.

[58] T. Kind and O. Fiehn, "Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry," *BMC Bioinformatics*, vol. 8, p. 105, Mar. 2007.

[59] A. Makarov, E. Denisov, O. Lange, and S. Horning, "Dynamic range of mass accuracy in LTQ orbitrap hybrid mass spectrometer," *J. Am. Soc. Mass Spectrom.*, vol. 17, pp. 977–982, July 2006.

[60] M. B. Comisarow and A. G. Marshall, "Frequency-sweep fourier transform ion cyclotron resonance spectroscopy," 1974.

[61] A. G. Marshall, C. L. Hendrickson, and G. S. Jackson, "Fourier transform ion cyclotron resonance mass spectrometry: a primer," *Mass Spectrom. Rev.*, vol. 17, pp. 1–35, Jan. 1998.

[62] S. Eliuk and A. Makarov, "Evolution of orbitrap mass spectrometry instrumentation," *Annu. Rev. Anal. Chem.*, vol. 8, pp. 61–80, 2015.

[63] R. A. Zubarev and A. Makarov, "Orbitrap mass spectrometry," *Anal. Chem.*, vol. 85, pp. 5288–5296, June 2013.

[64] R. H. Perry, R. G. Cooks, and R. J. Noll, "Orbitrap mass spectrometry: instrumentation, ion motion and applications," *Mass Spectrom. Rev.*, vol. 27, pp. 661–699, Nov. 2008.

[65] A. Makarov, "Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis," *Anal. Chem.*, vol. 72, pp. 1156–1162, Mar. 2000.

[66] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex fourier series," 1965.

[67] O. Lange, E. Damoc, A. Wieghaus, and A. Makarov, "Enhanced fourier transform for orbitrap mass spectrometry," 2014.

[68] F. W. McLafferty, "Tandem mass spectrometry (MS/MS): a promising new analytical technique for specific component determination in complex mixtures," 1980.

[69] J. S. Brodbelt, "Photodissociation mass spectrometry: new tools for characterization of biological molecules," *Chem. Soc. Rev.*, vol. 43, pp. 2757–2783, Apr. 2014.

[70] J. M. Wells and S. A. McLuckey, "Collision-induced dissociation (CID) of peptides and proteins," *Methods Enzymol.*, vol. 402, pp. 148–185, 2005.

[71] S. A. McLuckey and J. L. Stephenson, Jr, "Ion/ion chemistry of high-mass multiply charged ions," *Mass Spectrom. Rev.*, vol. 17, pp. 369–407, Nov. 1998.

[72] H. J. Cooper, "To what extent is FAIMS beneficial in the analysis of proteins?," 2016.

[73] R. Guevremont, "High-field asymmetric waveform ion mobility spectrometry: a new tool for mass spectrometry," *J. Chromatogr. A*, vol. 1058, pp. 3–19, Nov. 2004.

[74] A. A. Shvartsburg, T. Bryskiewicz, R. W. Purves, K. Tang, R. Guevremont, and R. D. Smith, "Field asymmetric waveform ion mobility spectrometry studies of proteins: Dipole alignment in ion mobility spectrometry?," *J. Phys. Chem. B*, vol. 110, pp. 21966–21980, Nov. 2006.

[75] A. A. Shvartsburg, K. Tang, and R. D. Smith, "FAIMS operation for realistic gas flow profile and asymmetric waveforms including electronic noise and ripple," *J. Am. Soc. Mass Spectrom.*, vol. 16, pp. 1447–1455, Sept. 2005.

[76] I. A. Buryakov, E. V. Krylov, E. G. Nazarov, and U. Kh. Rasulev, "A new method of separation of multi-atomic ions by mobility at atmospheric pressure using a high-frequency amplitude-asymmetric strong electric field," 1993.

[77] R. W. Purves, S. Prasad, M. Belford, A. Vandenberg, and J.-J. Dunyach, "Optimization of a new aerodynamic cylindrical FAIMS device for small molecule analysis," *J. Am. Soc. Mass Spectrom.*, vol. 28, pp. 525–538, Mar. 2017.

[78] S. Prasad, M. W. Belford, J.-J. Dunyach, and R. W. Purves, "On an aerodynamic mechanism to enhance ion transmission and sensitivity of FAIMS for nano-electrospray ionization-mass spectrometry," *J. Am. Soc. Mass Spectrom.*, vol. 25, pp. 2143–2153, Dec. 2014.

[79] A. Macklin, S. Khan, and T. Kislinger, "Recent advances in mass spectrometry based clinical proteomics: applications to cancer research," *Clin. Proteomics*, vol. 17, p. 17, May 2020.

[80] R. J. C. Slebos, J. W. C. Brock, N. F. Winters, S. R. Stuart, M. A. Martinez, M. Li, M. C. Chambers, L. J. Zimmerman, A. J. Ham, D. L. Tabb, and D. C. Liebler, "Evaluation of strong cation exchange versus isoelectric focusing of peptides for multidimensional liquid chromatography-tandem mass spectrometry," *J. Proteome Res.*, vol. 7, pp. 5286–5294, Dec. 2008.

[81] D. J. Morrison, K. Taylor, and T. Preston, "Strong anion-exchange liquid chromatography coupled with isotope ratio mass spectrometry using a liquiface interface," *Rapid Commun. Mass Spectrom.*, vol. 24, pp. 1755–1762, June 2010.

[82] I. K. Ventouri, D. B. A. Malheiro, R. L. C. Voeten, S. Kok, M. Honing, G. W. Somsen, and R. Haselberg, "Probing protein denaturation during Size-Exclusion chromatography using native mass spectrometry," *Anal. Chem.*, vol. 92, pp. 4292–4300, Mar. 2020.

[83] S. Camerini, M. L. Polci, U. Restuccia, V. Usuelli, A. Malgaroli, and A. Bachi, "A novel approach to identify proteins modified by nitric oxide: the HIS-TAG switch method," *J. Proteome Res.*, vol. 6, pp. 3224–3231, Aug. 2007.

[84] Y. Zhang, B. R. Fonslow, B. Shan, M.-C. Baek, and J. R. Yates, "Protein analysis by Shotgun/Bottom-up proteomics," 2013.

[85] M. Wilm, "Principles of electrospray ionization," *Mol. Cell. Proteomics*, vol. 10, p. M111.009407, July 2011.

[86] M. W. Senko, V. V. Kovtoun, P. R. Atherton, J. J. Dunyach, E. R. Wouters, M. Splendore, and W. Siebert, "Ion transport device and modes of operation thereof," Aug. 24 2010. US Patent 7,781,728.

[87] J. E. Elias and S. P. Gygi, "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry," *Nat. Methods*, vol. 4, pp. 207–214, Mar. 2007.

[88] J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, and M. Mann, "Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ," *Mol. Cell. Proteomics*, vol. 13, pp. 2513–2526, Sept. 2014.

[89] S. Tyanova, T. Temu, and J. Cox, "The MaxQuant computational platform for mass spectrometry-based shotgun proteomics," *Nat. Protoc.*, vol. 11, pp. 2301–2319, Dec. 2016.

[90] J. Cox and M. Mann, "MaxQuant enables high peptide identification rates, individ-

ualized p.p.b.-range mass accuracies and proteome-wide protein quantification," 2008.

[91] B. Bogdanov and R. D. Smith, "Proteomics by FTICR mass spectrometry: top down and bottom up," *Mass Spectrom. Rev.*, vol. 24, pp. 168–200, Mar. 2005.

[92] L. Jiang, M. Wang, S. Lin, R. Jian, X. Li, J. Chan, H. Fang, G. Dong, H. Tang, M. P. Snyder, and GTEx Consortium, "A quantitative proteome map of the human body."

[93] K. Suhre, M. I. McCarthy, and J. M. Schwenk, "Genetics meets proteomics: perspectives for large population-based studies," *Nat. Rev. Genet.*, vol. 22, pp. 19–37, Jan. 2021.

[94] N. A. Karp and K. S. Lilley, "Design and analysis issues in quantitative proteomics studies," *Proteomics*, vol. 7 Suppl 1, pp. 42–50, Sept. 2007.

[95] P. D. Piehowski, V. A. Petyuk, D. J. Orton, F. Xie, R. J. Moore, M. Ramirez-Restrepo, A. Engel, A. P. Lieberman, R. L. Albin, D. G. Camp, R. D. Smith, and A. J. Myers, "Sources of technical variability in quantitative LC-MS proteomics: human brain tissue sample analysis," *J. Proteome Res.*, vol. 12, pp. 2128–2137, May 2013.

[96] C. Zhou, K. L. Simpson, L. J. Lancashire, M. J. Walker, M. J. Dawson, R. D. Unwin, A. Rembielak, P. Price, C. West, C. Dive, and A. D. Whetton, "Statistical considerations of optimal study design for human plasma proteomics and biomarker discovery," *J. Proteome Res.*, vol. 11, pp. 2103–2113, Apr. 2012.

[97] S. R. Langley and M. Mayr, "Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics," *J. Proteomics*, vol. 129, pp. 83–92, Nov. 2015.

[98] F. Robert and J. Pelletier, "Exploring the impact of Single-Nucleotide polymorphisms on translation," *Front. Genet.*, vol. 9, p. 507, Oct. 2018.

[99] I. M. Stylianou, J. P. Affourtit, K. R. Shockley, R. Y. Wilpan, F. A. Abdi, S. Bhardwaj, J. Rollins, G. A. Churchill, and B. Paigen, "Applying gene expression, proteomics and single-nucleotide polymorphism analysis for complex trait gene identification," *Genetics*, vol. 178, pp. 1795–1805, Mar. 2008.

[100] R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, D. Altshuler, and International SNP Map Working Group, "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," *Nature*, vol. 409, pp. 928–933, Feb. 2001.

[101] G. A. Churchill, D. M. Gatti, S. C. Munger, and K. L. Svenson, "The diversity outbred mouse population," *Mamm. Genome*, vol. 23, pp. 713–718, Oct. 2012.

[102] I. M. C. M. Rietjens, J. Vervoort, A. Maslowska-Górnicz, N. Van den Brink, and K. Beekmann, "Use of proteomics to detect sex-related differences in effects of toxicants: implications for using proteomics in toxicology," *Crit. Rev. Toxicol.*, vol. 48, pp. 666–681, Sept. 2018.

[103] A. M. Billing, D. Revets, C. Hoffmann, J. D. Turner, S. Vernocchi, and C. P. Muller, "Proteomic profiling of rapid non-genomic and concomitant genomic effects of acute restraint stress on rat thymocytes," *J. Proteomics*, vol. 75, pp. 2064–2079, Apr. 2012.

[104] A. M. Cooksey, N. Momen, R. Stocker, and S. C. Burgess, "Identifying blood biomarkers and physiological processes that distinguish humans with superior performance under psychological stress," *PLoS One*, vol. 4, p. e8371, Dec. 2009.

[105] L. Wu, S. I. Candille, Y. Choi, D. Xie, L. Jiang, J. Li-Pook-Than, H. Tang, and M. Snyder, "Variation and genetic control of protein abundance in humans," *Nature*, vol. 499, pp. 79–82, July 2013.

[106] R. Baetta, M. Pontremoli, A. Martinez Fernandez, C. M. Spickett, and C. Banfi, "Proteomics in cardiovascular diseases: Unveiling sex and gender differences in the era of precision medicine," *J. Proteomics*, vol. 173, pp. 62–76, Feb. 2018.

[107] E. Gianazza, I. Miller, U. Guerrini, L. Palazzolo, C. Parravicini, and I. Eberini, "Gender proteomics i. which proteins in non-sexual organs," *J. Proteomics*, vol. 178, pp. 7–17, Apr. 2018.

[108] J. Tiihonen, M. Koskuvi, M. Storvik, I. Hyötyläinen, Y. Gao, K. A. Puttonen, R. Giniatullina, E. Poguzhelskaya, I. Ojansuu, O. Vaurio, T. D. Cannon, J. Lönnqvist, S. Therman, J. Suvisaari, J. Kaprio, L. Cheng, A. F. Hill, M. Lähteenvuo, J. Tohka, R. Giniatullin, Š. Lehtonen, and J. Koistinaho, "Sex-specific transcriptional and proteomic signatures in schizophrenia," *Nat. Commun.*, vol. 10, p. 3933, Sept. 2019.

[109] Y. Wang, L. Song, M. Liu, R. Ge, Q. Zhou, W. Liu, R. Li, J. Qie, B. Zhen, Y. Wang, F. He, J. Qin, and C. Ding, "A proteomics landscape of circadian clock in mouse liver," *Nat. Commun.*, vol. 9, p. 1553, Apr. 2018.

[110] A. S. Kononikhin, N. L. Starodubtseva, L. K. Pastushkova, D. N. Kashirina, K. Y. Fedorchenko, A. G. Brhozovsky, I. A. Popov, I. M. Larina, and E. N. Nikolaev, "Space-flight induced changes in the human proteome," *Expert Rev. Proteomics*, vol. 14, pp. 15–29, Jan. 2017.

[111] T. Tanaka, A. Biancotto, R. Moaddel, A. Z. Moore, M. Gonzalez-Freire, M. A. Aon, J. Candia, P. Zhang, F. Cheung, G. Fantoni, CHI consortium, R. D. Semba, and L. Ferrucci, "Plasma proteomic signature of age in healthy humans," *Aging Cell*, vol. 17, p. e12799, Oct. 2018.

[112] T. W. Rhoads, M. S. Burhans, V. B. Chen, P. D. Hutchins, M. J. P. Rush, J. P. Clark, J. L. Stark, S. J. McIlwain, H. R. Eghbalnia, D. M. Pavelec, I. M. Ong, J. M. Denu, J. L. Markley, J. J. Coon, R. J. Colman, and R. M. Anderson, "Caloric restriction engages

hepatic RNA processing mechanisms in rhesus monkeys," *Cell Metab.*, vol. 27, pp. 677–688.e5, Mar. 2018.

[113] U. Distler, S. Schumann, H.-G. Kesseler, R. Pielot, K.-H. Smalla, M. Sielaff, M. J. Schmeisser, and S. Tenzer, "Proteomic analysis of brain region and Sex-Specific synaptic protein expression in the adult mouse brain," *Cells*, vol. 9, Jan. 2020.

[114] K. S. Kosik, C. L. Joachim, and D. J. Selkoe, "Microtubule-associated protein tau (tau) is a major antigenic component of paired helical filaments in alzheimer disease," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 83, pp. 4044–4048, June 1986.

[115] G. G. Glenner and C. W. Wong, "Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein," *Biochem. Biophys. Res. Commun.*, vol. 120, pp. 885–890, May 1984.

[116] C. Lynch, "World alzheimer report 2019: Attitudes to dementia, a global survey," 2020.

[117] M. Fricker, A. M. Tolkovsky, V. Borutaite, M. Coleman, and G. C. Brown, "Neuronal cell death," *Physiol. Rev.*, vol. 98, pp. 813–880, Apr. 2018.

[118] P. Theofilas, A. J. Ehrenberg, A. Nguy, J. M. Thackrey, S. Dunlop, M. B. Mejia, A. T. Alho, R. E. Paraizo Leite, R. D. Rodriguez, C. K. Suemoto, C. F. Nascimento, M. Chin, D. Medina-Cleghorn, A. M. Cuervo, M. Arkin, W. W. Seeley, B. L. Miller, R. Nitrini, C. A. Pasqualucci, W. J. Filho, U. Rueb, J. Neuhaus, H. Heinsen, and L. T. Grinberg,

"Probing the correlation of neuronal loss, neurofibrillary tangles, and cell death markers across the alzheimer's disease braak stages: a quantitative study in humans," *Neurobiol. Aging*, vol. 61, pp. 1–12, Jan. 2018.

[119] C. Böttcher, S. Schlickeiser, M. A. M. Sneeboer, D. Kunkel, A. Knop, E. Paza, P. Fidzinski, L. Kraus, G. J. L. Snijders, R. S. Kahn, A. R. Schulz, H. E. Mei, NBB-Psy, E. M. Hol, B. Siegmund, R. Glauben, E. J. Spruth, L. D. de Witte, and J. Priller, "Human microglia regional heterogeneity and phenotypes determined by multiplexed single-cell mass cytometry," *Nat. Neurosci.*, vol. 22, pp. 78–90, Jan. 2019.

[120] M. Y. Batiuk, A. Martirosyan, J. Wahis, F. de Vin, C. Marneffe, C. Kusserow, J. Koeppen, J. F. Viana, J. F. Oliveira, T. Voet, C. P. Ponting, T. G. Belgard, and M. G. Holt, "Identification of region-specific astrocyte subtypes at single cell resolution," *Nat. Commun.*, vol. 11, p. 1220, Mar. 2020.

[121] B. B. Lake, R. Ai, G. E. Kaeser, N. S. Salathia, Y. C. Yung, R. Liu, A. Wildberg, D. Gao, H.-L. Fung, S. Chen, R. Vijayaraghavan, J. Wong, A. Chen, X. Sheng, F. Kaper, R. Shen, M. Ronaghi, J.-B. Fan, W. Wang, J. Chun, and K. Zhang, "Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain," *Science*, vol. 352, pp. 1586–1590, June 2016.

[122] J. A. Miller, S. Horvath, and D. H. Geschwind, "Divergence of human and mouse brain transcriptome highlights alzheimer disease pathways," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, pp. 12698–12703, July 2010.

[123] S. Lin, Y. Lin, J. R. Nery, M. A. Urich, A. Breschi, C. A. Davis, A. Dobin, C. Zaleski, M. A. Beer, W. C. Chapman, T. R. Gingeras, J. R. Ecker, and M. P. Snyder, "Comparison of the transcriptional landscapes between human and mouse tissues," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, pp. 17224–17229, Dec. 2014.

[124] L. Muzio and G. G. Consalez, "Modeling human brain development with cerebral organoids," *Stem Cell Res. Ther.*, vol. 4, no. 6, p. 154, 2013.

[125] S. P. Pașca, "The rise of three-dimensional human brain cultures," 2018.

[126] L. A. Echan, H.-Y. Tang, N. Ali-Khan, K. Lee, and D. W. Speicher, "Depletion of multiple high-abundance proteins improves protein profiling capacities of human serum and plasma," *Proteomics*, vol. 5, pp. 3292–3303, Aug. 2005.

[127] C. Tu, P. A. Rudnick, M. Y. Martinez, K. L. Cheek, S. E. Stein, R. J. C. Slebos, and D. C. Liebler, "Depletion of abundant plasma proteins and limitations of plasma proteomics," *J. Proteome Res.*, vol. 9, pp. 4982–4991, Oct. 2010.

[128] N. K. Magdalinou, A. J. Noyce, R. Pinto, E. Lindstrom, J. Holmén-Larsson, M. Holtta, K. Blennow, H. R. Morris, T. Skillbäck, T. T. Warner, A. J. Lees, I. Pike, M. Ward, H. Zetterberg, and J. Gobom, "Identification of candidate cerebrospinal fluid biomarkers in parkinsonism using quantitative proteomics," *Parkinsonism Relat. Disord.*, vol. 37, pp. 65–71, Apr. 2017.

[129] K. Hansson, R. Dahlén, O. Hansson, E. Pernevik, R. Paterson, J. M. Schott, N. Magdalinou, H. Zetterberg, K. Blennow, and J. Gobom, "Use of the tau protein-to-peptide ratio in CSF to improve diagnostic classification of alzheimer's disease," 2019.

[130] A. J. Percy, J. Yang, A. G. Chambers, R. Simon, D. B. Hardie, and C. H. Borchers, "Multiplexed MRM with internal standards for cerebrospinal fluid candidate protein biomarker quantitation," *J. Proteome Res.*, vol. 13, pp. 3733–3747, Aug. 2014.

[131] D. Ferreira, L.-O. Wahlund, and E. Westman, "The heterogeneity within alzheimer's disease," *Aging*, vol. 10, pp. 3058–3060, Nov. 2018.

[132] W. M. V. d. Flier and W. M. Van der Flier, "Clinical heterogeneity in familial alzheimer's disease," 2016.

[133] N. L. Komarova and C. J. Thalhauser, "Calculating stage duration statistics in multistage diseases," *PLoS One*, vol. 6, p. e28298, Dec. 2011.

[134] G. Devi and P. Scheltens, "Heterogeneity of alzheimer's disease: consequence for drug trials?," *Alzheimers. Res. Ther.*, vol. 10, p. 122, Dec. 2018.

[135] K.-L. Huang, S. Li, P. Mertins, S. Cao, H. P. Gunawardena, K. V. Ruggles, D. R. Mani, K. R. Clauser, M. Tanioka, J. Usary, S. M. Kavuri, L. Xie, C. Yoon, J. W. Qiao, J. Wrobel, M. A. Wyczalkowski, P. Erdmann-Gilmore, J. E. Snider, J. Hoog, P. Singh, B. Niu, Z. Guo, S. Q. Sun, S. Sanati, E. Kawaler, X. Wang, A. Scott, K. Ye, M. D. McLellan, M. C. Wendl, A. Malovannaya, J. M. Held, M. A. Gillette, D. Fenyö, C. R. Kinsinger,

M. Mesri, H. Rodriguez, S. R. Davies, C. M. Perou, C. Ma, R. Reid Townsend, X. Chen, S. A. Carr, M. J. Ellis, and L. Ding, "Proteogenomic integration reveals therapeutic targets in breast cancer xenografts," *Nat. Commun.*, vol. 8, p. 14864, Mar. 2017.

[136] M. Zhou, D. M. Duong, E. C. B. Johnson, J. Dai, J. J. Lah, A. I. Levey, and N. T. Seyfried, "Mass Spectrometry-Based quantification of tau in human cerebrospinal fluid using a complementary tryptic peptide standard," *J. Proteome Res.*, vol. 18, pp. 2422–2432, June 2019.

[137] B. Bai, X. Wang, Y. Li, P. Chen, J. M. Yarbro, T. G. Beach, and J. Peng, "Deep multi-layer brain proteomics identifies molecular networks and netrin-1 accumulation in alzheimer's disease progression," 2020.

[138] F. Yang, Y. Shen, D. G. Camp, and R. D. Smith, "High-pH reversed-phase chromatography with fraction concatenation for 2D proteomic analysis," 2012.

[139] E. Shishkova, *Sample Preparation and Instrumental Strategies for In-Depth Studies of Complex Mammalian Proteomes*. The University of Wisconsin-Madison, 2017.

[140] S. Pfammatter, E. Bonneil, F. P. McManus, and P. Thibault, "Accurate quantitative proteomic analyses using metabolic labeling and high field asymmetric waveform ion mobility spectrometry (FAIMS)," *J. Proteome Res.*, vol. 18, pp. 2129–2138, May 2019.

[141] S. Pfammatter, E. Bonneil, and P. Thibault, "Improvement of quantitative measurements in multiplex proteomics using High-Field asymmetric waveform spectrometry," 2016.

[142] S. Pfammatter, E. Bonneil, F. P. McManus, S. Prasad, D. J. Bailey, M. Belford, J.-J. Dunyach, and P. Thibault, "A novel differential ion mobility device expands the depth of proteome coverage and the sensitivity of multiplex proteomic measurements," 2018.

[143] A. S. Hebert, S. Prasad, M. W. Belford, D. J. Bailey, G. C. McAlister, S. E. Abbatiello, R. Huguet, E. R. Wouters, J.-J. Dunyach, D. R. Brademan, M. S. Westphall, and J. J. Coon, "Comprehensive Single-Shot proteomics with FAIMS on a hybrid orbitrap mass spectrometer," *Anal. Chem.*, vol. 90, pp. 9529–9537, Aug. 2018.

[144] A. C. Peterson, J. D. Russell, D. J. Bailey, M. S. Westphall, and J. J. Coon, "Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics," *Mol. Cell. Proteomics*, vol. 11, pp. 1475–1488, Nov. 2012.

[145] R. C. Hendrickson, A. Y. H. Lee, Q. Song, A. Liaw, M. Wiener, C. P. Paweletz, J. L. Seeburger, J. Li, F. Meng, E. G. Deyanova, M. T. Mazur, R. E. Settlage, X. Zhao, K. Southwick, Y. Du, D. Holder, J. R. Sachs, O. F. Laterza, A. Dallob, D. L. Chappell, K. Snyder, V. Modur, E. King, C. Joachim, A. Y. Bondarenko, M. Shearman, K. A. Soper, A. D. Smith, W. Z. Potter, K. S. Koblan, A. B. Sachs, and N. A. Yates, "High resolution discovery proteomics reveals candidate disease progression markers of

alzheimer's disease in human cerebrospinal fluid," *PLoS One*, vol. 10, p. e0135365, Aug. 2015.

[146] D. S. Daly, K. K. Anderson, E. A. Panisko, S. O. Purvine, R. Fang, M. E. Monroe, and S. E. Baker, "Mixed-effects statistical model for comparative LC-MS proteomics studies," *J. Proteome Res.*, vol. 7, pp. 1209–1217, Mar. 2008.

[147] W. S. Noble, "How does multiple testing correction work?," 2009.

[148] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," 1995.

[149] J. D. Storey, "A direct approach to false discovery rates," 2002.

Chapter 2

PROTEOMIC ATLAS OF THE HUMAN BRAIN IN ALZHEIMER'S DISEASE

JM designed and conducted mass spectrometry experiments, interpreted results, analyzed data, and wrote the manuscript.

This chapter is adapted from a manuscript published in the Journal of Proteome Research with the permission of:

McKetney J, Runde R, Hebert AS, Salamat S, Roy S, Coon JJ. *Proteomic Atlas of the Human Brain in Alzheimer's Disease.* Journal of Proteome Research. **2018**.

Abstract

The brain represents one of the most divergent and critical organs in the human body. Yet, it can be afflicted by a variety of neurodegenerative diseases specifically linked to aging, about which we lack a full biomolecular understanding of onset and progression, such as Alzheimer's Disease (AD). Here we provide a proteomic resource comprising nine anatomically distinct sections from three aged individuals, across a spectrum of disease progression, categorized by quantity of neurofibrillary tangles. Using state-of-the-art mass spectrometry, we identify a core brain proteome that exhibits only small variance in expression, accompanied by a group of proteins that are highly differentially expressed in individual sections and broader regions. AD affected tissue exhibited slightly elevated levels of tau protein with similar relative expression to factors associated with the AD pathology. Substantial differences were identified between previous proteomic studies of mature adult brains and our aged cohort. Our findings suggest considerable value in examining specifically the brain proteome of aged human populations from a multiregional perspective. This resource can serve as a guide, as well as a point of reference for how specific regions of the brain are affected by aging and neurodegeneration.

Introduction

As evinced by its highly variable protein expression profile, the brain embodies one of the most divergent and specialized organs in the body[1,2]. A recent study comparing 13

human tissues found that after the testes, the brain had the second highest occurrence of tissue-enriched genes[3]. And due to diverse cell types and functions, expression levels vary between different areas of the mammalian brain, creating a heterogeneous environment of protein expression in a single organ[4,5]. Given its unique role, it is not surprising that the brain is specifically susceptible to a host of diseases associated with aging, including Alzheimer's disease (AD) and other neurodegenerative disorders. AD makes up more than half of dementias with more than 40 million people suffering from AD globally in 2015[6]. In 2017, an American developed AD every 66 seconds – a rate that is projected to accelerate considerably in the next decades due to increasing life spans and an aging population[7]. Although the pathology originates in the transentorhinal region and spreads through the hippocampal formation and hippocampus proper on to the neocortex, there is little known about the sequential effects on cellular function in different regions of the brain[8,9].

The pathological decline associated with AD is driven by protein, specifically the aggregation of amyloid and tau protein into neuritic plaques and neurofibrillary tangles, respectively. Current hypotheses suggest that the abnormal processing of amyloid precursor leads to the development of plaques, while abnormal phosphorylation of tau can promote aggregation[10,11]. Other proteins play roles in the associated increase in neuronal damage; both as drivers of aggregation[12] and aggravating damage caused by the aggregates[13]. Cellular protein production and function is deeply intertwined with the pathological effect of AD decline within the cell and in the extracellular space. We conclude that to improve our understanding of neurodegeneration and AD progression, a

comprehensive view of brain protein expression is required.

Several large-scale projects have examined protein expression in the mammalian brain. Mouse models can be leveraged to study neurodegenerative diseases such as Parkinson's[14] and Alzheimer's[14–16]. However, animal models are often incomplete and fail to encompass the full variety of protein changes that occur in the human brain with pathological decline and aging, considering substantial differences in lifespan[17,18]. As RNA-based technologies have improved, several projects such as PsychEncode[19], BrainSpan[20] and the Allen Brain Atlas[21] have emerged that quantified transcripts as a measure of protein expression directly from post-mortem human tissue. Transcriptomic studies have also been performed that specifically target Alzheimer's disease[22]. Despite the impressive scope of these studies, several analyses have implicated that protein expression can be regulated at the translational level with different half-lives between mRNA transcripts and proteins[23,24]. This causes a discrepancy in predicting certain protein abundances from transcripts, a phenomenon that has been observed in the brain previously[17 18,25]. Recently, a multi-regional analysis of protein expression at a variety of developmental time points was performed, including individuals with ages spanning from less than one to forty years old[17]. These time points precede the onset of neurodegeneration in the human brain, and though proteomic experiments have been performed that specifically target neurodegeneration with age, they have quantified levels of hundreds of proteins rather than the thousands that are present[26,27] or have examined only a small number of brain regions[18,25]. Many of these studies have focused on the parietal and frontal cortex[18,25] and the hippocampal area[28–30]

despite identification of differentially expressed proteins in AD in a variety of regions of the brain, even those that are often tangle-free[27]. We posit that a more comprehensive atlas of protein expression in the aged human brain would accelerate our understanding of neurodegeneration and dementia.

Here we present a resource cataloging the expression of thousands of proteins in nine distinct sections of the aged human brain. This protein compendium is based on nine sections from three individual brains (**Figure 2.1**) The sections include: Amygdala (AMY), Caudate Nucleus (CNC), Cerebellum (CBM), Entorhinal Cortex (ECX), Inferior Parietal Lobule (IPL), Middle Frontal Gyrus (MFG), Superior Temporal Gyrus (STG), Thalamus (THA), and Visual Cortex (VCX). The three individuals were chosen to cover a spectrum of AD decline, as determined by Braak staging[31]. Braak staging categorizes AD progression based on the spread of neurofibrillary tangles from tau on a scale of 1-6. The "no tangle" brain (0) contained no observable tangles upon dissection and is not afflicted with Alzheimer's disease. The "intermediate tangle" brain (+) was categorized as stage III with tangles identified primarily in the entorhinal region[31]. The "severe tangle" brain (++) falls into the stage VI category with tau aggregates widespread and likely resulting in isocortical destruction[31]. Utilizing state-of-the-art peptide fractionation, liquid chromatography, and tandem mass spectrometry we have collected a high-resolution human brain protein compendium featuring nearly 10,000 unique proteins.

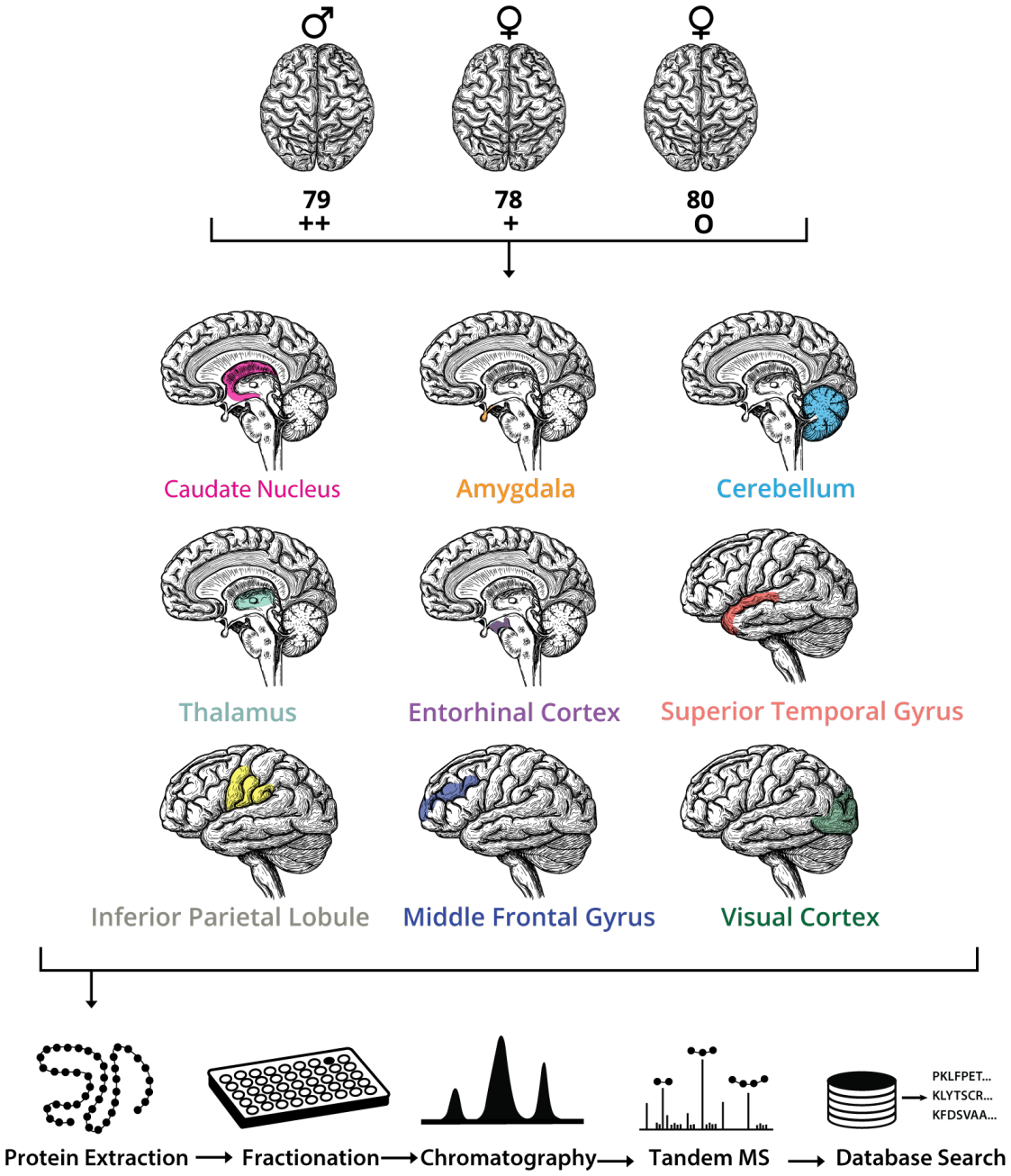


Figure 2.1: Experimental Design. Gender, age and tangle severity is shown for each brain **0**, **+**, **++** representing no tangles, intermediate tangle development, and severe tangles, respectively. The nine sections used in this experiment are displayed in the anatomical illustration below. The analysis workflow is pictured at the bottom including protein extraction, fractionation, mass spectrometry and data base searching.

Experimental Methods

Tissue Extraction. Individual brains were extracted and rapidly flash frozen with a maximum post-mortem interval of five hours for the atlas and fifteen hours for the secondary cohort (**Table 2.1**). Brains were stored in -80° C freezers, taken out and given to the clinical neuropathologist to collect precise brain samples, which averaged about 1g. After collection they were immediately put on dry ice and sent to the proteomics lab.

Tisue Staining. Slices were taken from the anterior hippocampal region and sectioned using a microtome. Sections were warmed to incubation temperature before being stained with α-phosphorylated tau protein antibody for 15 mins. Tissue sections were then stained with hematoxylin before being imaged.

Extraction and Digestion. Each tissue sample was suspended in 1.2 ml of lysis buffer (6M Guanadine HCl, 100mM Tris pH 8) before being probe sonicated (Misonix XL-2000) to lyse the cells. After sonication, a protein BCA assay (Thermo Scientific) was performed to determine protein concentrations of the lysate. Sample lysates ranged in concentration from approximately 6-13 mg/ml. 500ug of protein was aliquoted from each sample lysate. Each aliquot was brought up to 90% methanol before being centrifuged at 14,000 g for 5 min. Supernate was disposed and the precipitate was resuspended in 240ul of reducing and alkylating buffer (8M urea, 10mM TCEP, 40mM CAA, 100mM Tris pH 8). The sample solution was then diluted to 25% concentration with 100mM Tris, pH 8. Trypsin was added

Case	Gender	Age	PMI	Neuropath CERAD	Neurological diagnosis
383	M	79	4:08	2	High AD
440	F	78	4:35	3	High AD, CAA 3
372	F	84	5:00	3	High AD, CAA 3
403	M	74	5:47	3	High AD
343	F	78	4:50	3	High AD
405	F	78	3:36	0	A0B2C0
384	M	81	15:12	0	Old infarcts
254	F	86	6:21	0	Astrocytoma
199	F	85	9:20	0	Infarct, HS
146	F	80	5:00	0	Old minute infarct

Table 2.1: Sample Cohort Metadata. Metadata for each set of tissue samples by individual including age, gender and neuropathological information. Highlighted individuals indicate individuals where all nine sections were used. CAA= cerebral amyloid angiopathy HS=hippocampal sclerosis

to the protein lysate sample at a ratio of 50:1 w/v and digested overnight. Digested samples were desalted using Strata-X Polymeric Reverse Phase column (Phenomenex). Samples were then dried in SpeedVac Concentrator.

Fractionation Dried samples were resuspended in 0.2% formic acid before fractionation using an HPLC (Agilent, Infinity 2000) with a C18 reverse-phase column (Waters, XBridge Peptide BEH, particle size 3.5um). Mobile phase buffer A was a fresh-made mixture of 10mM ammonium acetate, pH 10, while mobile phase buffer B was composed of 10mM ammonium acetate, 80% methanol, pH 10. Each sample was separated into 32 fractions. Fractions were collected directly in round-bottom 96-well plates, allowing three samples to be contained in each plate. Fractions were concatenated by combining fractions 1-8 with 18-25 and combining fractions 9 -17 with 26-32, to yield 16 fractions. In the interest of time, the fraction number was again reduced to 12 by pooling 15 & 16, 13 & 14, 1 & 2, and 3 & 4. Plates were dried down in the SpeedVac Concentrator. Samples were resuspended in 0.2% formic acid for instrument injection.

LC-MS/MS Online reverse-phase columns were prepared in house. A laser puller was used to generate tips on 35cm long silica columns with an inner diameter of 75um and an outer diameter of 360um. Columns were filled with 1.7um, 130 Å pore size, Bridged Ethylene Hybrid C18 particles. Column was heated and maintained at a temperature of 50° C and connected to the instrument by an embedded emitter. A Waters UHPLC was used

for online chromatography with mobile phase buffer A consisting of 0.2% formic acid and mobile phase buffer B consisting of 0.2% formic acid with 70% acetonitrile. Fractionated Samples were loaded onto the column for 12 min at a flow rate of 0.35 ul/minute. Mobile phase B increased to 4% in the first min then increased on a gradient to 55% B at 75 minutes. The method increased percent B to 100% by 76 minutes. Method ended with 3 min wash at 100% B and 10 min wash at 0% B.

Single shot samples were analyzed using a 120 min LC method where Samples were loaded onto the column for 7 min at a flow rate of 0.33 ul/minute. Mobile phase B increased to 7% in the first 6 min then increased on a gradient to 50% B at 104 minutes. The method increased percent B to 100% by 105 minutes. Method ended with 5 min wash at 100% B and 10 min wash at 0% B.

Data Searching & Analysis All raw files from the fractionated samples were searched together in the software MaxQuant(version 1.5.2.8)[32] with each sample input as an experiment made up of 12 fractions. Spectra were searched using fast LFQ against a full human proteome with isoforms downloaded from Uniprot (June 14, 2017). Carbamidomethyl was set as a fixed modification. Matching between runs was used with a retention time window of 0.7 mins. Searches were performed using a protein FDR of 1%, a minimum peptide length of 7, and 0.5 Da MS2 match tolerance. Protein data was then extracted from the "ProteinGroups.txt" file of the Maxquant output after decoy, contaminants, and reverse sequences were removed. All single shot samples were searched together using

the same parameters, with the only difference being the use of a more recent human proteome from Uniprot (November 29, 2018). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[33] partner repository(`http://www.ebi.ac.uk/pride/archive/`) with the dataset identifier PXD010603.

Data imputation and hierarchical clustering were performed using an in-house data development tool(coonlabdatadev.com). Imputation was performed for proteins observed in at least 50% of the samples from each section, by replacing missing values with values selected from the lowest 3% of the distribution of log transformed intensities. Boxplots were generated using R and Boxplotr[34]. Pearson correlations were performed in R. To ensure we were not altering trends with imputation the Pearson analyses were also performed using only proteins observed in all samples and the results were indistinguishable from the confidently imputed dataset. Gene ontology enrichment was performed using the National Cancer Institute Database for Annotation, Visualization and Integrated Discovery (DAVID).

Results

Overall Protein Statistics Tissue sections were lysed by probe sonication and proteins were extracted, denatured, and digested with trypsin[35]. Digested peptides were then fractionated using high-pH reverse-phase liquid chromatography, before being pooled and injected onto our liquid chromatography setup online with the Fusion Lumos Mass Spectrometer[36,37]. This procedure was performed on all nine sections: Amygdala (AMY),

Caudate Nucleus (CNC), Cerebellum (CBM), Entorhinal Cortex (ECX), Inferior Parietal Lobule (IPL), Middle Frontal Gyrus (MFG), Superior Temporal Gyrus (STG), Thalamus (THA), and Visual Cortex (VCX). Overall, we identified 9,735 proteins in at least one sample with 5,098 proteins identified in all 26 samples (**Supplemental Table 2**). On average, 7,387 proteins were quantified in each sample, with the fewest observed in the caudate nucleus and the most observed in the middle frontal gyrus. This equates to greater than 50% of the gene products found to be expressed in the brain by the Human Protein Atlas (14518 transcripts)[38,39] and 60% of proteins identified in the deepest proteomic study of the brain to date (11,840 proteins)[25,40] (**Table 2.2**). 6,256 proteins were quantified in at least 50% of samples from each anatomical region, meaning at least one CBM sample and two or more samples from all other regions (**Supplemental Table 3**). 7,706 proteins were identified in all samples from at least one section of the brain, reflecting a consistency between individuals' expression within regions. **Figure 2.2a** shows the distribution of proteins between the sections. More than 70% of proteins were identified in all nine sections, with 12% identified in only one section and 18% identified in 2-8 sections. The 9,735 proteins identified overall correspond to 129,050 identified unique peptides. On average, proteins were identified by greater than 11 peptides comprising greater than 23% sequence coverage. Only 211 proteins were identified by a single peptide, while the largest number of peptides attributed to a single protein was neuroblast differentiation-associated protein AHNAK with 358 peptides.

Section	♀ 80 **O**	♂ 78 **+**	♀ 79 **++**
Visual Cortex	7448	7138	7224
Entorhinal Cortex	7785	7680	7212
Amygdala	7647	7584	7041
Temporal Lobe	7839	7664	7337
Thalamus	7690	7281	7296
Caudate Nucleus	7408	6363	6753
Parietal Lobe	7811	7703	7341
Cerebellum	6953	------	7088
Frontal Gyrus	7901	7602	7282

Table 2.2: Proteins Quantified. Proteins identified in each sample organized by both section and individual brain

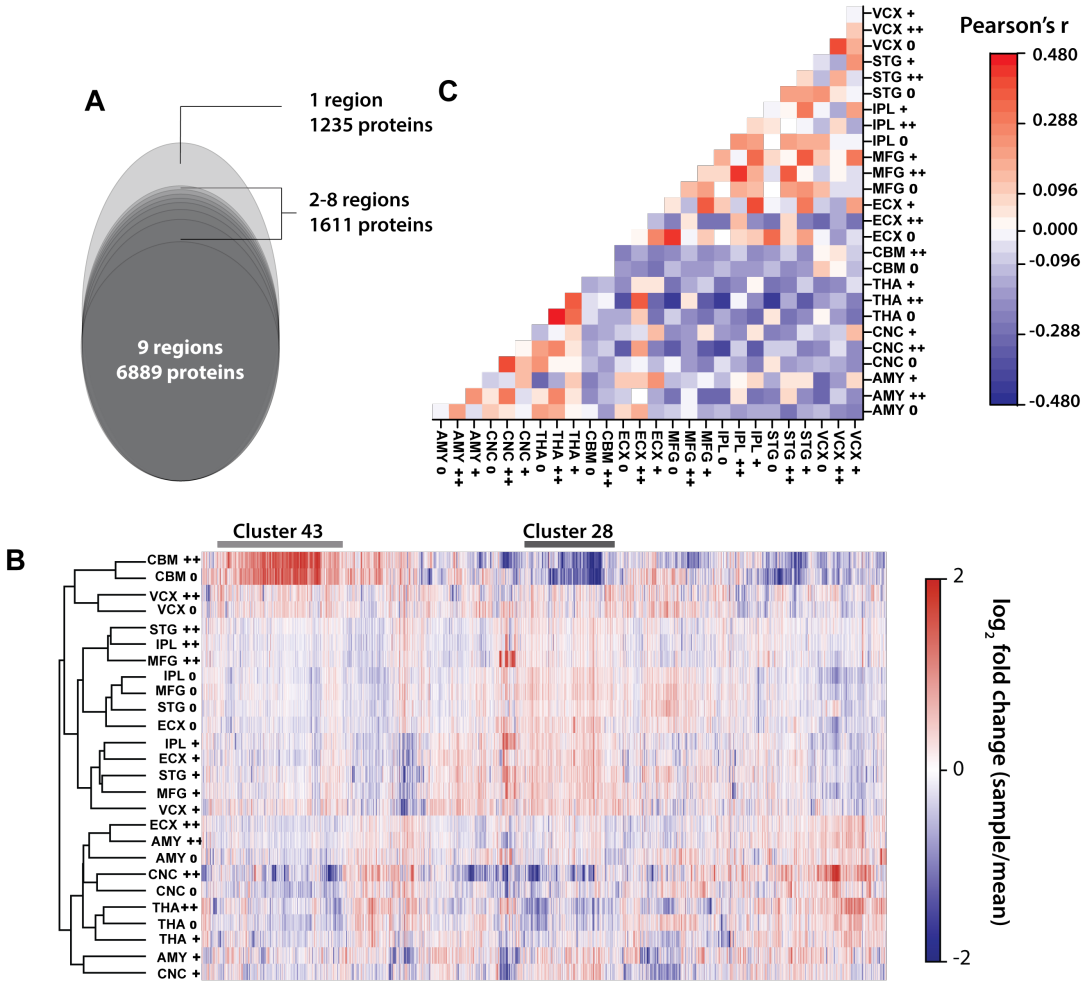
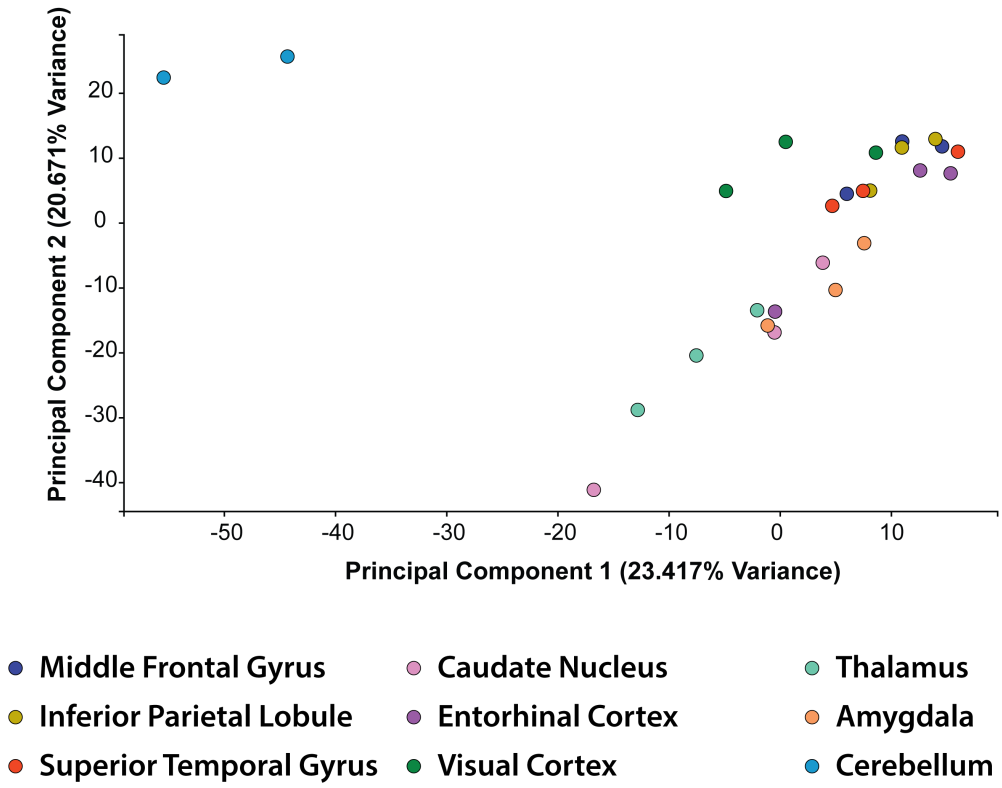


Figure 2.2: Similarity Among Sections and Individual Brains. (A) total of 9735 proteins were identified overall with 70% identified in all nine sections. (B) Heat map showing mean-normalized protein fold changes for all 6256 proteins with confident imputation (>50% of sections). (C) Pairwise Pearson correlation between all samples of relative protein levels for confidently imputed proteins. Samples are labelled based on the sections three letter code followed by the tangle severity (0/+/++).

Sectional Similarity With the global map of protein expression in hand we first performed a hierarchical clustering of the 6,250 proteins that were identified in at least half of the samples from each section. Proteins were clustered by Pearson correlation while tissue samples were clustered using Spearman correlation. We observed that the two CBM (cerebellum) samples exhibit a strong divergence from the all other sections (**Figure 2.2b**). A similar result was reflected in our principal component analysis (**Supplemental Figure 2.1**) with both CBM samples greatly separated from all others by component 1 which encompasses 23.4% of variance. The heat map in **Figure 2.2b** shows that this difference is driven heavily by two protein clusters: cluster 43 and cluster 28 (**Figure 2.2b**).

When Gene Ontology (GO) enrichment was performed on cluster 43 the most significant associated biological processes were related to "transcription" and "mRNA processing". Cluster 43 also included an abundance of proteins related to chromatin maintenance and nucleosome assembly, all of which are processes associated with the cell body and the nucleus, an observation corroborated by previous analysis of the brain proteome[17]. This data suggests that the cerebellum (CBM) samples may be converging due to an increased nuclei density relative to other sections of the brain. Enrichment analysis on cluster 28 shows decreased expression of proteins associated with the membrane and the process of "synaptic transmission". These expression differences support previous findings[41] suggesting the cerebellum is defined by increased cell density as well as decreased participation in specialized signaling relative to the rest of the brain.

Protein levels also exhibited similarity between regions found in the same area more



Supplementary Figure 2.1: Principal Component Analysis. Principal Component Analysis performed on protein expression data set of proteins identified in all 26 samples. Points are colored by section, with the cerebellum sections widely separated from the other sections of the brain.

broadly, with the highest order groupings in our clustering separating the sections in the limbic system and those contained in the cerebral cortex (**Figure 2.2b**). One group contains all limbic sections including the THA, CNC, AMY as well as one ECX sample, while the other group includes all the cerebral cortex sections, which includes the MFG, IPL, STG, VCX, as well as the CBM samples. These groupings fall generally into the interior and exterior of the brain (**Figure 2.2b & Figure 2.1**).

To examine similarities in protein expression we performed a pairwise Pearson Correlation between all samples. We observed the strongest positive correlations between the same sections in different individuals, even more so in relatively isolated regions such as the thalamus (**Figure 2.2c**). A weaker positive correlation was shown among sections located in the same basic region with THA, CNC, and AMY showing some positive correlation. This same weak positive correlation is seen among the sections of the cerebral cortex in the relationships of MFG, IPL and STG in the upper right of **Figure 2.2c**. A corresponding broadly negative correlation is seen in the center of **Figure 2.2c** when comparing the protein profiles of more distant sections of the brain across these different broad regions.

Differentially expressed proteins We examined more closely three sections that had particularly strong correlations: MFG, CNC, and IPL. The samples from each of these sections were averaged for each protein and then expression differences were compared (**Supplemental Table 4**). The protein profile of the frontal gyrus and the caudate nucleus show an anti-correlation overall (Pearson's r = -0.42) that becomes quite strong (Pearson's

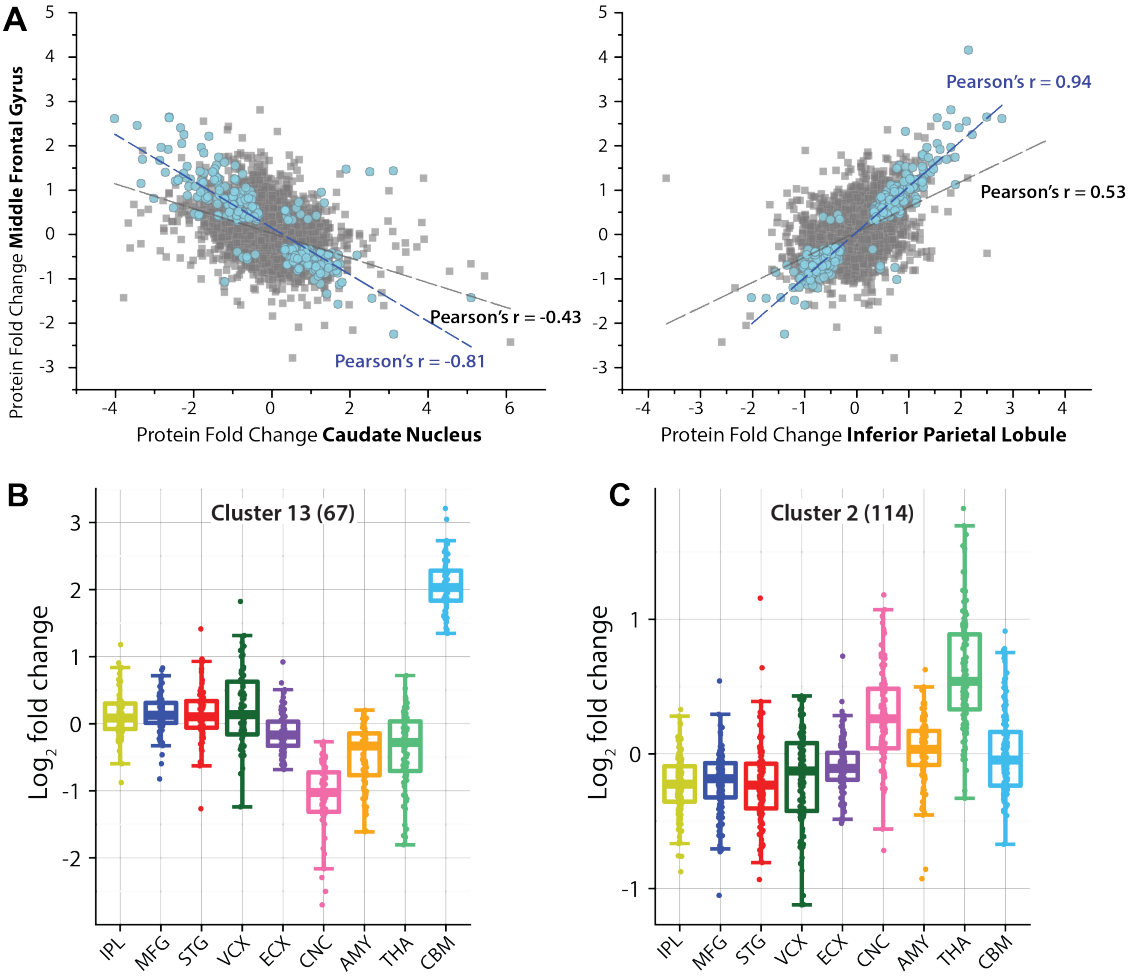


Figure 2.3: Differential Expression. (A) Plot comparing protein fold change in Middle Frontal Gyrus (MFG) region as compared to the Inferior Parietal Lobule (IPL) and Caudate Nucleus (CNC). Both correlations are strongly driven by small group of proteins with greater variation (>30%) and significance ($p < 0.05$) pictured in blue (B, C) Two clusters of differentially expressed proteins averaged by region and showing log2 fold changes. Numbers in parenthesis indicate proteins contained in that cluster. Larger clusters reflect the divergence of the cerebellum (CBM) while smaller clusters show expression contrast between inner and outer brain.

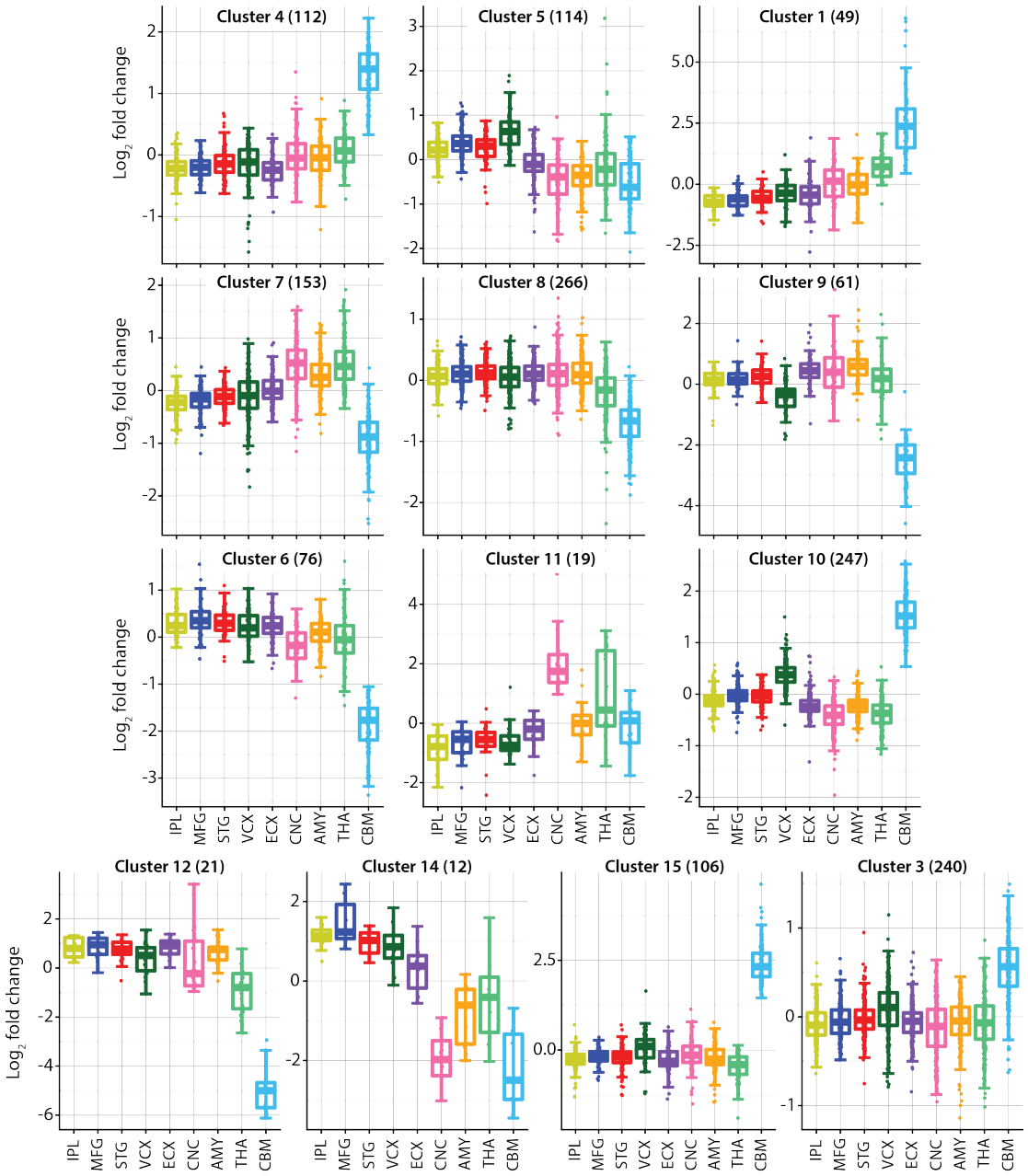
r = -0.81) if we focus on proteins with greater variation (>20% change) and slightly greater significance (p-value < 0.05) (**Figure 2.3a**). The opposite is true of the relationship between MFG and IPL, adjacent lobes of the cerebral cortex, which starts with a positive correlation (Pearson's r = 0.52) that becomes stronger when selecting for more variable and significant proteins (Pearson's r = 0.94). If this grouping is reversed and only proteins with minimal variations (<20% expression change) and significance (p > 0.05) are used the correlation between MFG and CNC is largely eliminated (Pearson's r = -0.13). The gene products within this low correlation group show gene ontology enrichment for "cell-cell adhesion" and "vesicle transport", biological processes ubiquitous to the brain. This leads to the hypothesis that there may be a core brain proteome that exhibits little variation throughout the different anatomical features, as well as a group of variably expressed proteins that account for regional deviation in the proteome.

1411 proteins showing significant differential expression (Bonferroni adj. P < 0.05) in at least one section and were grouped into 15 clusters using Ward's method (**Figure 2.3b & c, Supplemental Figure 2.2 & Supplemental Table 4**). As was observed in other analyses the sectional differences were dominated by the cerebellum, leading to mRNA processing proteins and those related to nuclear function driving many of the larger clusters such as Cluster 13 shown in **Figure 2.3b**. Cluster 2 includes proteins with low expression in the cerebral cortex regions (MFG, IPL and STG) and elevated expression in the caudate nucleus and amygdala with the highest expression in the thalamus (**Figure 3.3**). This group shows GO enrichment for proteins associated with the oxidation-reduction process

with 14 proteins included in that grouping and many others involved in other aspects of metabolism. It has been shown previously that the functionality of antioxidants and therefore mediation of oxidative stress can be associated with AD[42–44] and that these effects can vary significantly by region with age and potentially neurodegeneration[45].

A collection of differentially expressed gene products between healthy, aged brains and those with Alzheimer's disease was assembled using a wide range of previous literature[18,25–29,46,47]. Of the more than 1400 proteins with altered expression associated with pathological decline, 326 of them were identified in our region-specific expression group, with all regions but the IPL containing at least one of these gene products (**Figure 2.4a**). This suggests that many of the same proteins that are defining the expression profiles of a region, whether due to function or cell population, are also significantly connected to the pathology of AD.

We performed a series of single shot experiments focusing on two contrasting regions of the brain: the entorhinal cortex and the neocortex. We prepared and analyzed tissue from 7 additional individuals, 3 healthy age-matched controls, and 4 AD afflicted individuals, in parallel with our corresponding original samples, MFG, IPL, STG (neocortex) and ECX (entorhinal cortex). The 5 AD individuals included fresh neocortex and entorhinal cortex tissue from case 383, the severe tangle brain used in the atlas. We identified 5520 proteins overall but focused our analysis on the 3244 proteins identified in all experiments. We grouped these data into either neocortex or entorhinal cortex and diseased or healthy tissue and performed a two-way ANOVA with the two variables, "region" and "disease". We



Supplementary Figure 2.2: Region-specific Protein Clusters. Boxplots of all other clusters of regional differentially expressed proteins. Protein levels reflect $\log_2$ fold changes normalized to the mean LFQ intensity, and color coded accordingly: orange-amygdala, pink- caudate nucleus, turquoise – cerebellum, purple – entorhinal cortex, royal blue – middle frontal gyrus, yellow – inferior parietal lobule, red – superior temporal gyrus, teal – thalamus, and forest green – visual cortex. Proteins within each cluster can be seen in Supplemental Table 2. Clusters are labelled with number of proteins included in parentheses.

found 594 proteins that were significantly regulated after multiple hypothesis correction (Benjamini-Hochberg, 5% FDR) for one or both variables (**Figure 2.4b & c**). 227 out of the 588 proteins found to be regionally expressed in this expanded study were also identified in the previous region-specific group of the atlas. 9 proteins out of 17 disease-associated proteins from our expanded study were also identified from the Alzheimer's literature. 13 significantly regulated proteins from our expanded study show a 2-fold larger shift in abundance in one region than the other with the addition of AD (Figure 4b). These proteins are associated with the formation of the extracellular matrix (EPB41L1, HAPLN4, TENA), vesicle signaling (CAYP1, SYT2,) and the immune system (S100B, ICAM1, CD44,, ANXA1) all of which are cell functions affected in the development of Alzheimer's disease[48–51] and help define the differences between brain regions due to architecture, cell populations and function[52–55]. All of these proteins were flagged by region rather than disease suggesting they would not have been identified if all regions had been treated as a single Alzheimer's disease group. When comparing AD and healthy controls of the two regions independently we found no overlap in the 30 proteins most significantly associated with disease (**Supplemental Figure 2.3**) further supporting different pathological effects on these two regions of the brain.

Expression of Microtubule-Associated Protein Tau The individual subjects in our study were categorized using Braak staging, which grades pathological progression based on spread and density of neurofibrillary tangles. Upon dissection, brain tissue was stained

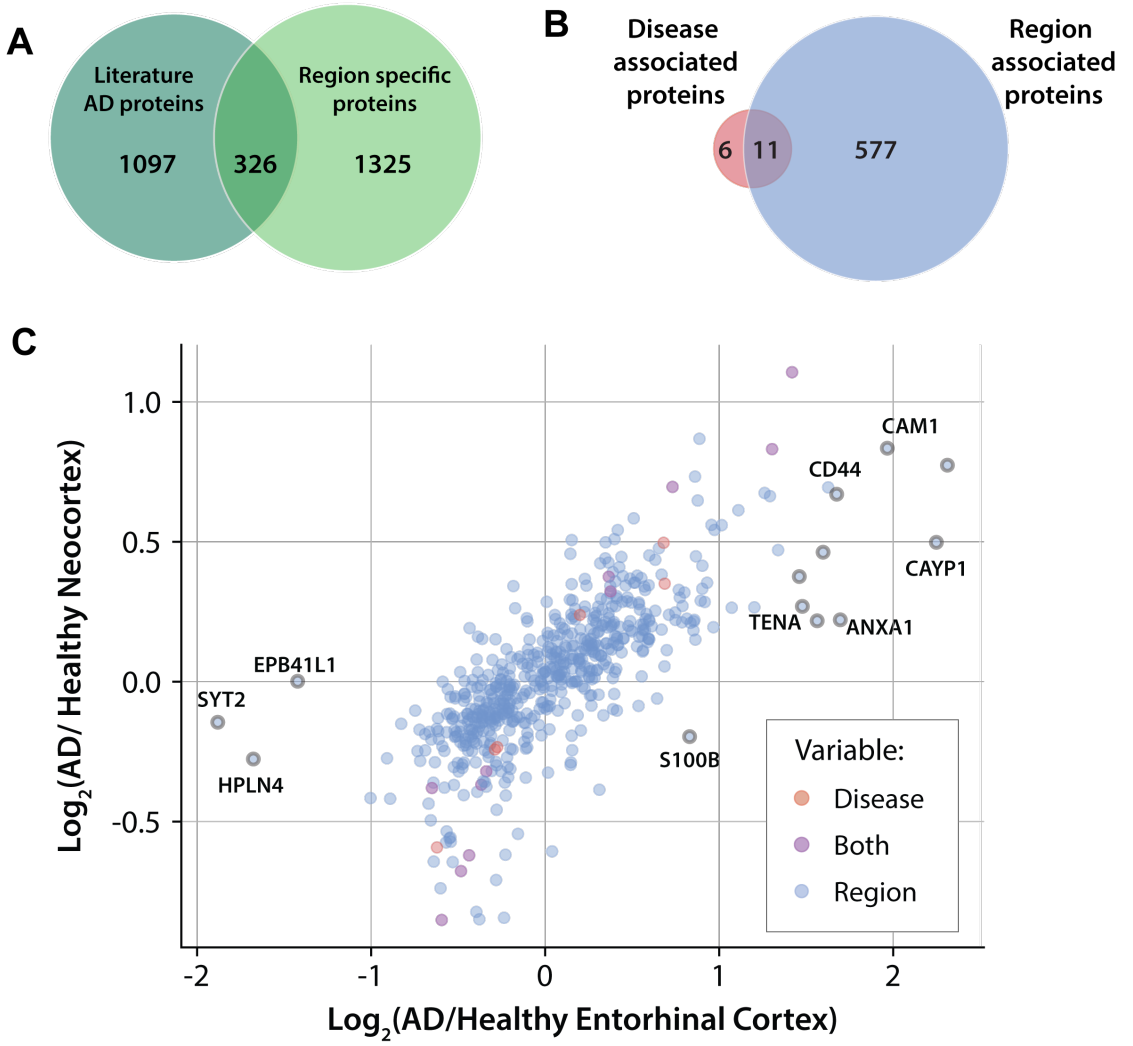
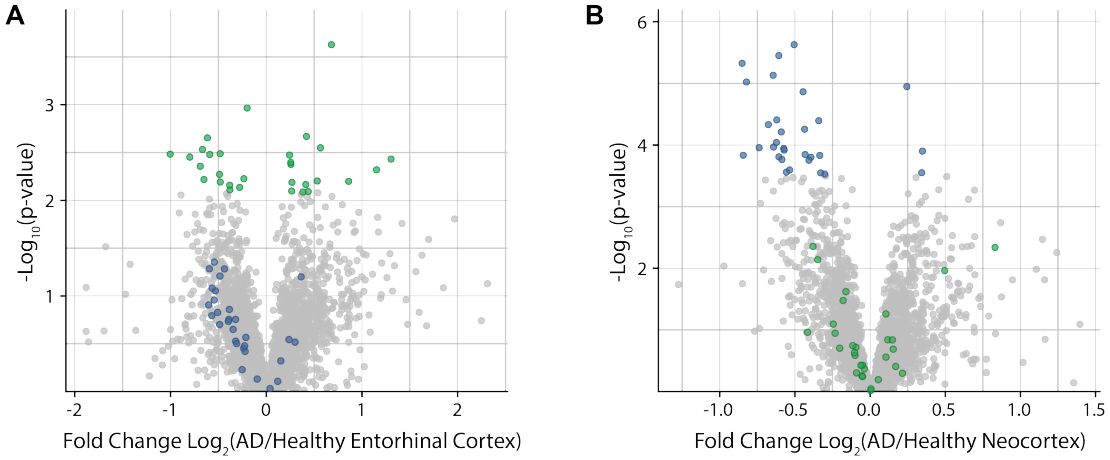


Figure 2.4: Region and Disease Specific Proteins. (A) Overlap of AD associated proteins from previous large-scale proteomic analyses and region-specific proteins from this study. (B) Overlap of significant proteins in expanded analysis. (C) Plot of significantly different protein abundances' fold change between disease and healthy tissue. Entorhinal cortex is on the x-axis and neocortex on the y-axis. Proteins are colored by variable they were found to be significant in. Points with black borders have 2-fold greater change with disease in one section over the other.



Supplementary Figure 2.3: Disease Fold Changes for Two Regions. Plot of p-values and fold change when comparing Alzheimer (AD) and healthy tissue samples from the (A) entorhinal cortex and (B) neocortex. Both plots show data for 30 proteins with most significant change in abundance from each section, indicated in blue and green for the neocortex and entorhinal cortex, respectively. No overlap exists between the most significantly regulated proteins from these two regions.

for aggregates of microtubule associated protein tau (MAPT) using an α-phosphorylated tau antibody (**Figure 2.5a**). Tau expression was compared between the "no tangle" brain and both the intermediate and severe tangle brains (**Figure 2.5b**). The severe tangle brain exhibited greater levels of tau protein in all but one section, VCX. The intermediate tangle brain showed elevated tau protein in more than half of the sections with AMY exhibiting more than double the levels found in the control brain. The tau protein ratios of the severe and intermediate AMY represent p-values of 0.063 and 0.080, respectively, based on the abundance ratios of all proteins present in the three samples from that region. The samples in our expanded study showed on average slightly elevated tau protein levels between the AD positive samples and the controls.

A subgroup of proteins was identified that had highly similar expression profiles to those of microtubule associated protein tau across these samples (**Figure 2.5c**). Generally, these proteins are related to cell growth, development and the metabolism required for division. All these proteins continued to show an association with MAPT in our expanded study except for RAPGEF6 and ACAT2, although there was a positive correlation between MAPT in and ACAT1 (**Supplemental Figure 2.4**). Several of the associated gene products, such as MAP1A and MAPRE2, are involved directly in microtubule formation and stability, linking to the healthy function of MAPT. MAP1A works in a compensatory fashion in microtubule assembly in MAPT knockout animals[56] and so may be similarly expressed to replace aggregated tau. Many of the non-structural proteins have been linked to AD directly or are known to be expressed in the brain. Mitochondrial carrier homolog (MTCH1), has a

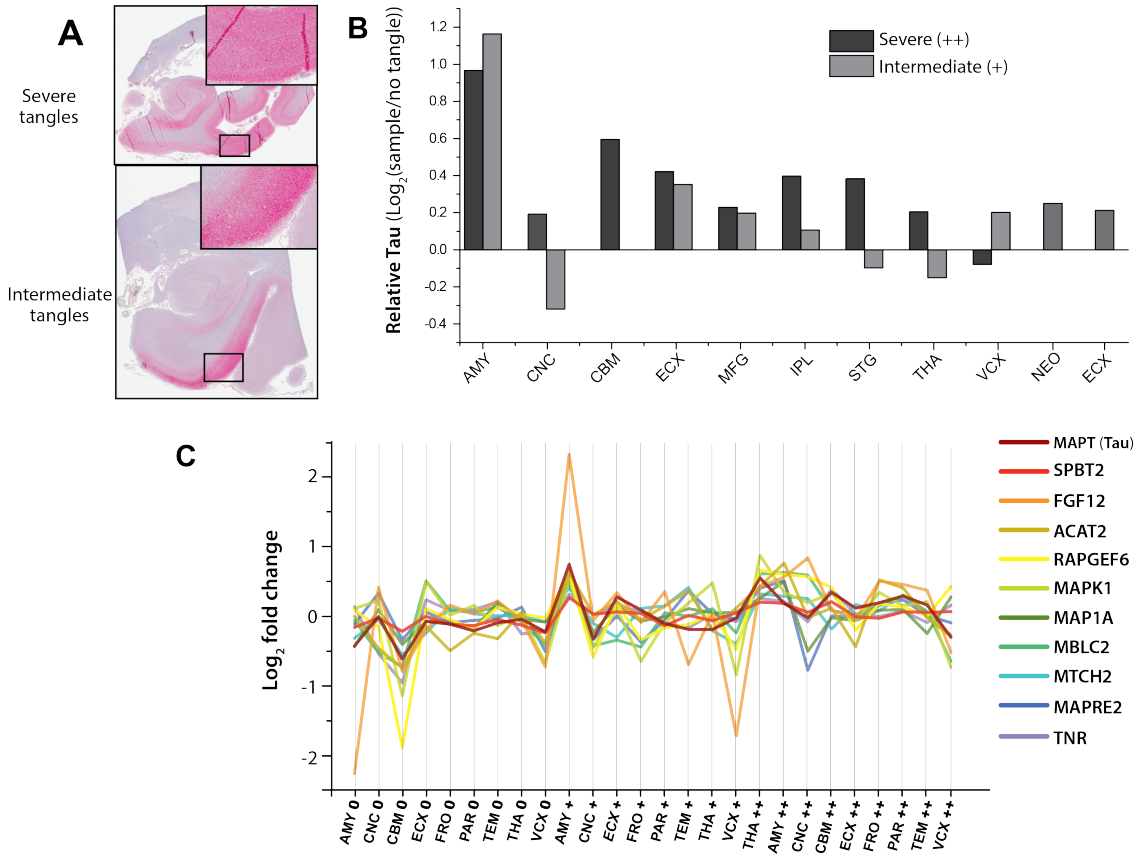
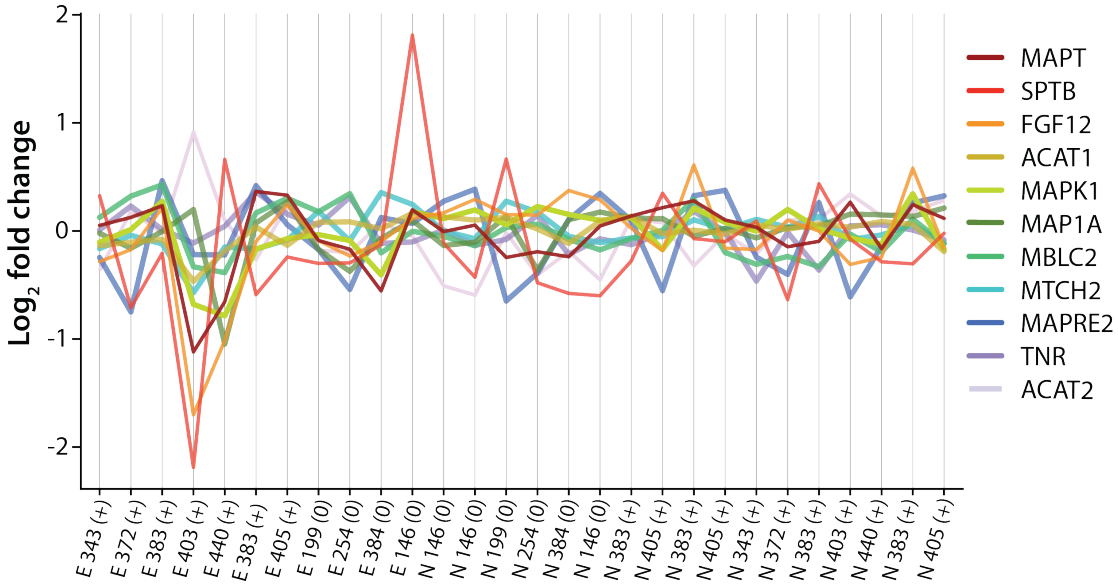


Figure 2.5: Similarity Among Sections and Individual Brains. (A) The development of tau protein aggregates can be seen in tissue staining of both brains that contained tangles, with significantly denser, and more extensive aggregate formation in the severe individual. (B) Elevated tau protein levels were observed relative to the tangle-free brain in many of the sections for both intermediate and severe tangle brains. NEO and ECX samples show average tau protein abundance in AD relative to healthy tissue from that same region. (C) Several proteins that clustered with tau proteins exhibited positively correlated expression profiles across the different samples (Pearson's $r > 0.50$). Many of these proteins are associated with healthy structural modification or have been implicated in the aggravation of the AD pathology.

quantitative trait locus correlation with cases of AD[57]. Tenascin-R (TNR) contributes to the alteration of the extracellular matrix during the development of the plaque and tangle pathology[49]. Both Acetyl-CoA acetyltransferase (ACAT2) and Mitogen activated protein kinase 1(MAPK1) are candidate therapeutic targets for AD[58,59]. ACAT-inhibitors have slowed the development of amyloid plaques while MAPK1-inhibition is hypothesized to reduce hyperphosphorylation of the tau proteins which can lead to aggregation[59]. While fibroblast growth factor 12 (FGF12) and metallo-beta-lactamase domain-containing protein (MBLAC) are expressed in the human brain, they have not been linked to either the structural or pathological function of tau protein[60–62]. Spectrin beta chain (SPBT2) interacts with alpha-synuclein, one of the hallmarks of another neurodegenerative disease, Parkinson's[63].

Comparison to other Human Brain Proteome Datasets Recently a study examined protein expression in seven sections of the brain at developmental time points ranging from less than one year to 40 years old[17]. This experiment relied on a peptide library for each section pooled from the five "adult" samples (23-40 years old). When this protein library was compared to our 26 samples, 3682 proteins were quantified in both studies. Comparing the mean-normalized expression, similar profiles were observed between many of the shared and proximal sections in the two experiments (**Figure 2.6a**) with a positive correlation between all of our samples in the cerebellum, visual cortex, amygdala and thalamus and the two healthier samples in the frontal lobe and caudate nucleus, two regions known to be affected in AD.



Supplementary Figure 2.4: Key Protein Fold Change Across All Samples. Mean-normalized $\log_2$ fold changes of proteins pictured in Figure 5c across expanded set of entorhinal cortex(E) and neocortex(N) samples. Numbers indicate case number with samples grouped into diseased (+) and healthy(0). All proteins except RAPGEF6, and ACAT2 showed positive correlations with MAPT across the expanded sample set. RAPGEF6 was not identified in all samples so the associated profile was not included.

Raw data files for shared sections that were believed to be strongly affected by aging were searched together using the software MaxQuant. These included VCX, AMY, MFG, and CNC samples. The "proteome ruler" calculation in Perseus was then used to estimate copy number per cell[64]. All samples fell within the range of 0.9-3% for the proportion of total signal attributed to histones. Although this is slightly lower than the typical proportion of 2-4%[64] it could possibly be attributed to lower cell body density of neuronal tissue. Protein counts were then compared between the older individuals in our study and the "adult" age group from the previous study, which included five individuals aged 23-40 years. The protein counts were analyzed using a paired tail t-test with correction for multiple hypotheses using the Benjamini-Hochberg method with a false discovery rate (FDR) of 5%. Overall, 621 proteins increased in the older individuals while only 37 were found to decrease with age. The frontal lobe contained the most of these differentially expressed proteins with 478 proteins increasing with age (**Figure 2.66b & c**). The proteins identified here do not seem to have a unifying process or function but many of them are localized to similar cellular compartments, namely the "membrane" and the "extracellular exosome". The disparity between proteins increasing with age as compared to those decreasing should not be particularly surprising given the slow turnover of proteins in the brain[65]. The proteins that show correlation with age exhibit very little overlap between regions, with only THA and MFG sharing a substantial number of proteins. This furthers our hypothesis that regional proteomic differences play a role in the effects of aging and neurodegeneration.

Focusing on the large group of proteins identified in the frontal lobe, we performed

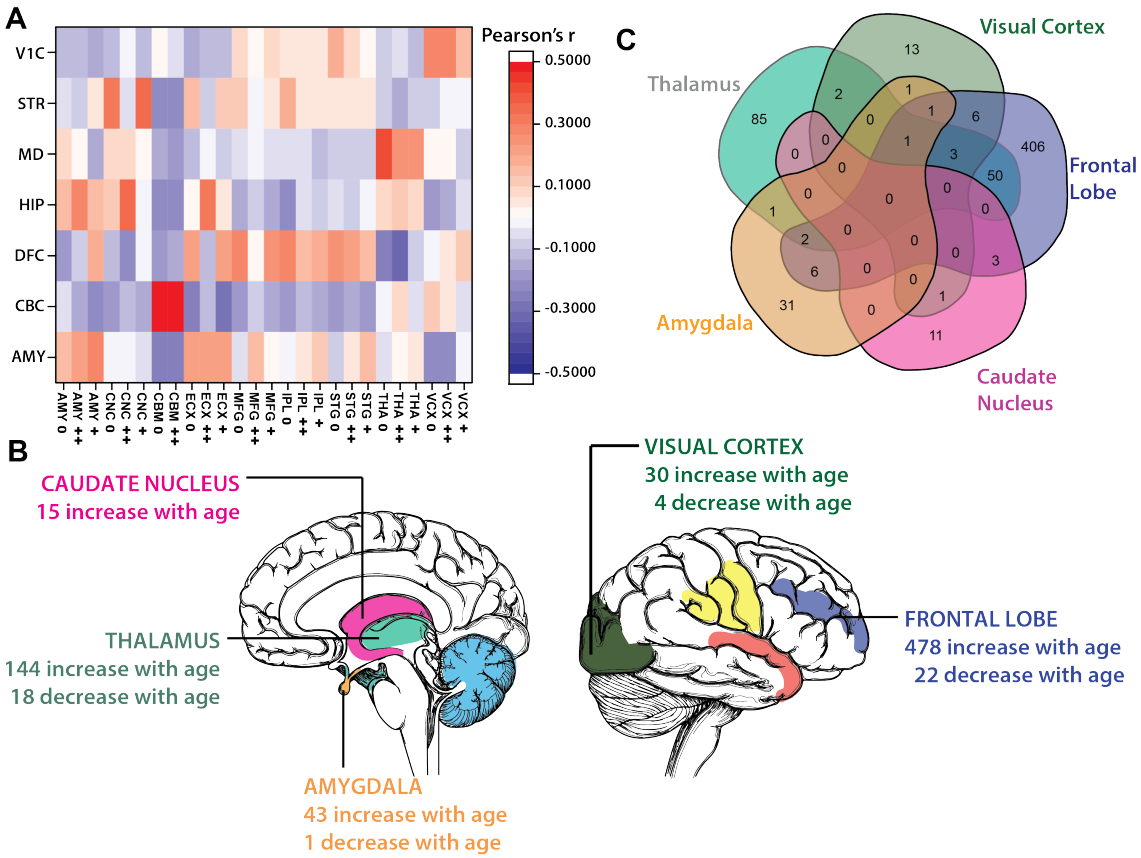


Figure 2.6: Comparison to Younger Brain Proteome . (A) Comparison between older samples analyzed here and "pooled adult" sample shows similar expression patterns between common sections sampled in the two experiments. Samples from previous analyses labelled as V1C: Visual Cortex, STR: Caudate Nucleus, MD: Thalamus, HIP: Hippocampus, DFC: Frontal Lobe, CBC: Cerebellum, AMY: Amygdala. (B) Proteins showing significantly different counts per cell. Proteins primarily increase with age, and the two most divergent sections are the Frontal Lobe and the Thalamus. (C) Diagram of proteins increasing with age shows very little overlap outside of group shared by both thalamus and frontal lobe.

this same analysis with the inclusion of our expanded panel of neocortex samples from our single-shot analysis. Protein counts were compared between young, healthy-aged, and diseased-aged neocortex samples using a one-way ANOVA. Of the 478 proteins with abundance changes in our original analysis of the frontal lobe, 399 of them were found to be significant with an expanded sample size. Unsurprisingly, a majority of proteins (301) were found to be significant when comparing young to both healthy aged and diseased samples, while 57 proteins were significant in the young to healthy-aged comparison only and 41 in the young to diseased comparison only.

Discussion

We anticipate that this resource will prove highly informative and useful to both the human proteomics and neuroscience community. Overall, we quantified 9,748 proteins from 129,680 peptides, with an average of 11 peptides per protein. This is, to our knowledge, the largest dataset of multiregional proteomics focused on aged subjects or Alzheimer's disease[26]. The number of sections and depth acquired in this proteomic analysis richly illustrates the dynamic nature of the human brain proteome.

Examining the expression of the 6,258 proteins we observed sectional similarity of expression across individuals and disease states. This proteomic similarity reflects a distinct expression pattern that may be related to cell populations or functionality within these anatomical features. More importantly for this resource, it begins to validate comparison of sectional proteomes from different individuals. As we widen our analysis, larger-scale

trends appear, separating the major regions of the brain by protein expression, with the limbic system, the cerebral cortex and the cerebellum all showing distinct profiles. A large proportion of proteins, even those involved in neuronal function are expressed with minimal variation across samples. Differentially expressed proteins can be used to glean insight into cell populations and cellular function as we saw in the cerebellum sections in this study and has been seen in mouse models[4,14]. There is no reason to believe that this regional difference in expression is not reflected in the effects of AD and in fact, many of the proteins identified here as regionally expressed have also been identified in association with Alzheimer's disease in previous large-scale proteomics studies. This supports the hypothesis that different regions of the brain experience varying effects with pathological decline in terms of the affected proteins, magnitude and direction of changes in abundance. Upon analysis of a larger set of samples with a limited number of regions, we identified a substantial number of proteins that were significantly different due to both disease and region, reflecting this theory experimentally. In addition, several of the regional differentially expressed proteins showed much larger fold changes with disease in one region than the other indicating a variation in impact of AD. Additional multiregional analysis is required to identify the possible variety of effects Alzheimer's on the different areas of the brain.

Possibly reflecting this diaspora of regional AD effects, our label-free quantitation method detected elevated levels of tau protein in samples from the intermediate and severe tangle individuals relative to the control for a majority of sections. Fluctuating in parallel with tau were several other gene products that play a role in formation of

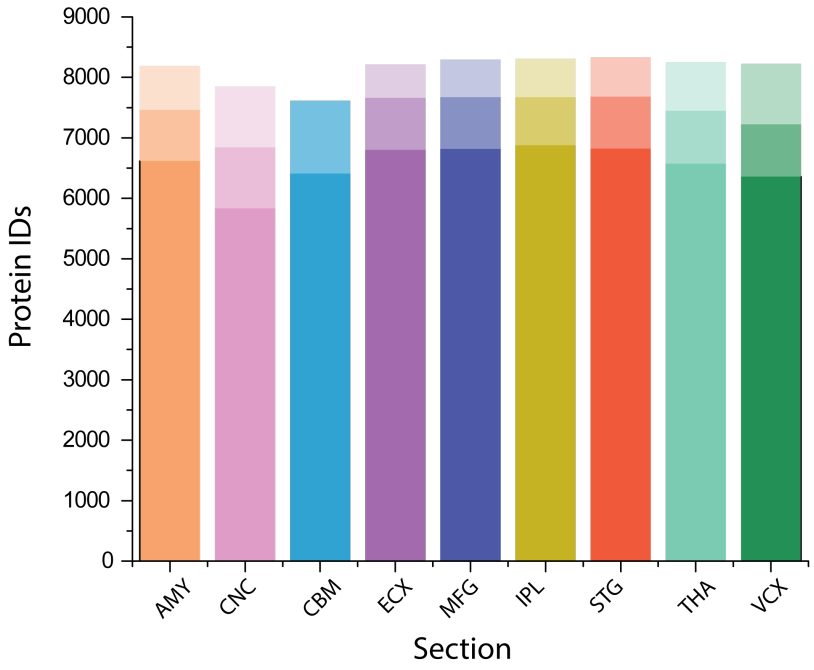
cytoarchitecture by way of microtubules and have been linked to the progression of the AD pathology. In addition, we identified several correlating proteins that are expressed in the brain but have no previous link to tauopathies or healthy function. These proteins could represent leads indicative of the changes in cellular function that occur during Alzheimer's decline or more general dementia. Investigating covariation with tau is simply the most obvious and accessible lead to pursue, but others have already developed large libraries of genes and proteins associated with AD, any of which could be used as a focal point to survey this data with an eye toward regional expression of disease-associated proteins. We observe a strong positive correlation between the global protein expression of regions examined here and the expression in similar regions from previous analyses. In addition, we identified a small group of proteins that exhibit significant differences in abundance between these two studies, suggesting a possible connection to the process of aging or neurodegeneration. Some of these proteins were identified as being variable by age in this previous study[17]. These proteins had largely positive correlations to age, with older individual brains exhibiting higher normalized counts. These age-correlated proteins also exhibited a regional association, with most showing differences only in a single region. When we performed an expanded analysis on the proteins identified within the frontal lobe, we found that most were associated with aging, with a small subset connected specifically to disease. Whether these trends in other regions are related to brain maturity or neurodegeneration would require an expanded study, but our analysis suggests that a non-insignificant number of these expression changes are linked specifically to disease.

This resource is not intended to be the final authority but rather the next step in building a foundational atlas of protein expression in the aged brain. As proteomic technologies improve, the depth, coverage and speed of collection for these proteomes will increase as well. We believe that this dataset adds regional nuance and breadth to several excellent resources already in existence, and that multiregional analysis can help fill gaps in our understanding of the progression of aging and neurodegeneration. This analysis and other multiregional studies support the concept of the variable effects of both age and neurodegeneration on the different regions of the human brain. This atlas of the brain likely reflects many of the shifts that this complex organ undergoes during these processes. An expanded study is likely required to fully disentangle those factors from lifestyle and gender. Yet even given those confounding factors here, protein groups surface with notable functional connections and regions show distinct profiles of expression with age and disease. For this reason, we expect this resource to empower the proteomics and neuroscience community in investigating neurodegenerative disease and aging.

Supplement

Supplemental tables are accessible through the Journal of Proteome Research at the link `https://doi.org/10.1021/acs.jproteome.9b00004`

Supplemental Table 1: Meta data including age, gender and post mortem interval and neuropathological information for each individual brain **Supplemental Table 2:** Full LFQ Intensities for all identified proteins in overall experiment



Supplementary Figure 2.5: Protein Identifications. Bar graph showing number of proteins identified in one, two or three samples from each section, indicated by lightest, darker and darkest colors, respectively. Sections are color coded similarly to Supplemental Figures 1 & 2.

Supplemental Table 3: LFQ intensities, fold changes and z-scores for 6,256 proteins used in bulk of analysis

Supplemental Table 4: Full list of differentially expressed proteins including sectional fold changes and p-values.

Supplemental Table 5: LFQ intensities and log2 LFQ intensities for all proteins identified in additional experiments

Supplemental Table 6: Fold changes between regions and disease states along with p-values for both region and disease variables

References

[1] M.-S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, J. K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N. A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L. D. N. Selvan, A. H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S. K. Sreenivasamurthy, A. Marimuthu, G. J. Sathe, S. Chavan, K. K. Datta, Y. Subbannayya, A. Sahu, S. D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K. R. Murthy, N. Syed, R. Goel, A. A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T.-C. Huang, J. Zhong, X. Wu, P. G. Shaw, D. Freed, M. S. Zahari, K. K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C. J. Mitchell, S. K. Shankar, P. Satishchandra, J. T. Schroeder, R. Sirdeshmukh, A. Maitra, S. D. Leach, C. G. Drake, M. K. Halushka, T. S. K. Prasad, R. H. Hruban, C. L. Kerr, G. D. Bader, C. A. Iacobuzio-

Donahue, H. Gowda, and A. Pandey, "A draft map of the human proteome," *Nature*, vol. 509, pp. 575–581, May 2014.

[2] M. Wilhelm, J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J.-H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, and B. Kuster, "Mass-spectrometry-based draft of the human proteome," *Nature*, vol. 509, pp. 582–587, May 2014.

[3] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Pontén, "Proteomics. tissue-based map of the human proteome," *Science*, vol. 347, p. 1260419, Jan. 2015.

[4] K. Sharma, S. Schmitt, C. G. Bergner, S. Tyanova, N. Kannaiyan, N. Manrique-Hoyos, K. Kongi, L. Cantuti, U.-K. Hanisch, M.-A. Philips, M. J. Rossner, M. Mann, and M. Simons, "Cell type- and brain region-resolved mouse brain proteome," *Nat. Neurosci.*, vol. 18, pp. 1819–1831, Dec. 2015.

[5] E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A. F. Boe, M. S. Boguski, K. S. Brockway, E. J. Byrnes, L. Chen, L. Chen, T.-M. Chen, M. C. Chin,

J. Chong, B. E. Crook, A. Czaplinska, C. N. Dang, S. Datta, N. R. Dee, A. L. Desaki, T. Desta, E. Diep, T. A. Dolbeare, M. J. Donelan, H.-W. Dong, J. G. Dougherty, B. J. Duncan, A. J. Ebbert, G. Eichele, L. K. Estin, C. Faber, B. A. Facer, R. Fields, S. R. Fischer, T. P. Fliss, C. Frensley, S. N. Gates, K. J. Glattfelder, K. R. Halverson, M. R. Hart, J. G. Hohmann, M. P. Howell, D. P. Jeung, R. A. Johnson, P. T. Karr, R. Kawal, J. M. Kidney, R. H. Knapik, C. L. Kuan, J. H. Lake, A. R. Laramee, K. D. Larsen, C. Lau, T. A. Lemon, A. J. Liang, Y. Liu, L. T. Luong, J. Michaels, J. J. Morgan, R. J. Morgan, M. T. Mortrud, N. F. Mosqueda, L. L. Ng, R. Ng, G. J. Orta, C. C. Overly, T. H. Pak, S. E. Parry, S. D. Pathak, O. C. Pearson, R. B. Puchalski, Z. L. Riley, H. R. Rockett, S. A. Rowland, J. J. Royall, M. J. Ruiz, N. R. Sarno, K. Schaffnit, N. V. Shapovalova, T. Sivisay, C. R. Slaughterbeck, S. C. Smith, K. A. Smith, B. I. Smith, A. J. Sodt, N. N. Stewart, K.-R. Stumpf, S. M. Sunkin, M. Sutram, A. Tam, C. D. Teemer, C. Thaller, C. L. Thompson, L. R. Varnam, A. Visel, R. M. Whitlock, P. E. Wohnoutka, C. K. Wolkey, V. Y. Wong, M. Wood, M. B. Yaylaoglu, R. C. Young, B. L. Youngstrom, X. F. Yuan, B. Zhang, T. A. Zwingman, and A. R. Jones, "Genome-wide atlas of gene expression in the adult mouse brain," *Nature*, vol. 445, pp. 168–176, Jan. 2007.

[6] M. J. Prince, *World Alzheimer Report 2015: the global impact of dementia: an analysis of prevalence, incidence, cost and trends*. Alzheimer's Disease International, 2015.

[7] A. Association *et al.*, "2017 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 13, no. 4, pp. 325–373, 2017.

[8] H. Braak, I. Alafuzoff, T. Arzberger, H. Kretzschmar, and K. Del Tredici, "Staging of alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry," *Acta Neuropathol.*, vol. 112, pp. 389–404, Oct. 2006.

[9] C. Bancher, H. Braak, P. Fischer, and K. A. Jellinger, "Neuropathological staging of alzheimer lesions and intellectual status in alzheimer's and parkinson's disease patients," *Neurosci. Lett.*, vol. 162, pp. 179–182, Nov. 1993.

[10] I. Grundke-Iqbal, K. Iqbal, Y. C. Tung, M. Quinlan, H. M. Wisniewski, and L. I. Binder, "Abnormal phosphorylation of the microtubule-associated protein tau (tau) in alzheimer cytoskeletal pathology," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 83, pp. 4913–4917, July 1986.

[11] G. V. De Ferrari and N. C. Inestrosa, "Wnt signaling function in alzheimer's disease," *Brain Res. Brain Res. Rev.*, vol. 33, pp. 1–12, Aug. 2000.

[12] P. J. Dolan and G. V. W. Johnson, "The role of tau kinases in alzheimer's disease," *Curr. Opin. Drug Discov. Devel.*, vol. 13, pp. 595–603, Sept. 2010.

[13] D. Bonneh-Barkay and C. A. Wiley, "Brain extracellular matrix in neurodegeneration," *Brain Pathol.*, vol. 19, pp. 573–585, Oct. 2009.

[14] S. Y. Jung, J. M. Choi, M. W. C. Rousseaux, A. Malovannaya, J. J. Kim, J. Kutzera, Y. Wang, Y. Huang, W. Zhu, S. Maity, H. Y. Zoghbi, and J. Qin, "An anatomically

resolved mouse brain proteome reveals parkinson disease-relevant pathways," *Mol. Cell. Proteomics*, vol. 16, pp. 581–593, Apr. 2017.

[15] E. D. Roberson, K. Scearce-Levie, J. J. Palop, F. Yan, I. H. Cheng, T. Wu, H. Gerstein, G.-Q. Yu, and L. Mucke, "Reducing endogenous tau ameliorates amyloid beta-induced deficits in an alzheimer's disease mouse model," *Science*, vol. 316, pp. 750–754, May 2007.

[16] J. N. Savas, Y.-Z. Wang, L. A. DeNardo, S. Martinez-Bartolome, D. B. McClatchy, T. J. Hark, N. F. Shanks, K. A. Cozzolino, M. Lavallée-Adam, S. N. Smukowski, S. K. Park, J. W. Kelly, E. H. Koo, T. Nakagawa, E. Masliah, A. Ghosh, and J. R. Yates, 3rd, "Amyloid accumulation drives proteome-wide alterations in mouse models of alzheimer's disease-like pathology," *Cell Rep.*, vol. 21, pp. 2614–2627, Nov. 2017.

[17] B. C. Carlyle, R. R. Kitchen, J. E. Kanyo, E. Z. Voss, M. Pletikos, A. M. M. Sousa, T. T. Lam, M. B. Gerstein, N. Sestan, and A. C. Nairn, "A multiregional proteomic survey of the postnatal human brain," *Nat. Neurosci.*, vol. 20, pp. 1787–1795, Dec. 2017.

[18] N. T. Seyfried, E. B. Dammer, V. Swarup, D. Nandakumar, D. M. Duong, L. Yin, Q. Deng, T. Nguyen, C. M. Hales, T. Wingo, J. Glass, M. Gearing, M. Thambisetty, J. C. Troncoso, D. H. Geschwind, J. J. Lah, and A. I. Levey, "A multi-network approach identifies Protein-Specific co-expression in asymptomatic and symptomatic alzheimer's disease," *Cell Syst*, vol. 4, pp. 60–72.e4, Jan. 2017.

[19] PsychENCODE Consortium, S. Akbarian, C. Liu, J. A. Knowles, F. M. Vaccarino, P. J. Farnham, G. E. Crawford, A. E. Jaffe, D. Pinto, S. Dracheva, D. H. Geschwind, J. Mill, A. C. Nairn, A. Abyzov, S. Pochareddy, S. Prabhakar, S. Weissman, P. F. Sullivan, M. W. State, Z. Weng, M. A. Peters, K. P. White, M. B. Gerstein, A. Amiri, C. Armoskus, A. E. Ashley-Koch, T. Bae, A. Beckel-Mitchener, B. P. Berman, G. A. Coetzee, G. Coppola, N. Francoeur, M. Fromer, R. Gao, K. Grennan, J. Herstein, D. H. Kavanagh, N. A. Ivanov, Y. Jiang, R. R. Kitchen, A. Kozlenkov, M. Kundakovic, M. Li, Z. Li, S. Liu, L. M. Mangravite, E. Mattei, E. Markenscoff-Papadimitriou, F. C. P. Navarro, N. North, L. Omberg, D. Panchision, N. Parikshak, J. Poschmann, A. J. Price, M. Purcaro, T. E. Reddy, P. Roussos, S. Schreiner, S. Scuderi, R. Sebra, M. Shibata, A. W. Shieh, M. Skarica, W. Sun, V. Swarup, A. Thomas, J. Tsuji, H. van Bakel, D. Wang, Y. Wang, K. Wang, D. M. Werling, A. J. Willsey, H. Witt, H. Won, C. C. Y. Wong, G. A. Wray, E. Y. Wu, X. Xu, L. Yao, G. Senthil, T. Lehner, P. Sklar, and N. Sestan, "The PsychENCODE project," *Nat. Neurosci.*, vol. 18, pp. 1707–1712, Dec. 2015.

[20] H. J. Kang, Y. I. Kawasawa, F. Cheng, Y. Zhu, X. Xu, M. Li, A. M. M. Sousa, M. Pletikos, K. A. Meyer, G. Sedmak, T. Guennel, Y. Shin, M. B. Johnson, Z. Krsnik, S. Mayer, S. Fertuzinhos, S. Umlauf, S. N. Lisgo, A. Vortmeyer, D. R. Weinberger, S. Mane, T. M. Hyde, A. Huttner, M. Reimers, J. E. Kleinman, and N. Sestan, "Spatio-temporal transcriptome of the human brain," *Nature*, vol. 478, pp. 483–489, Oct. 2011.

[21] M. J. Hawrylycz, E. S. Lein, A. L. Guillozet-Bongaarts, E. H. Shen, L. Ng, J. A. Miller,

L. N. van de Lagemaat, K. A. Smith, A. Ebbert, Z. L. Riley, C. Abajian, C. F. Beckmann, A. Bernard, D. Bertagnolli, A. F. Boe, P. M. Cartagena, M. M. Chakravarty, M. Chapin, J. Chong, R. A. Dalley, B. David Daly, C. Dang, S. Datta, N. Dee, T. A. Dolbeare, V. Faber, D. Feng, D. R. Fowler, J. Goldy, B. W. Gregor, Z. Haradon, D. R. Haynor, J. G. Hohmann, S. Horvath, R. E. Howard, A. Jeromin, J. M. Jochim, M. Kinnunen, C. Lau, E. T. Lazarz, C. Lee, T. A. Lemon, L. Li, Y. Li, J. A. Morris, C. C. Overly, P. D. Parker, S. E. Parry, M. Reding, J. J. Royall, J. Schulkin, P. A. Sequeira, C. R. Slaughterbeck, S. C. Smith, A. J. Sodt, S. M. Sunkin, B. E. Swanson, M. P. Vawter, D. Williams, P. Wohnoutka, H. R. Zielke, D. H. Geschwind, P. R. Hof, S. M. Smith, C. Koch, S. G. N. Grant, and A. R. Jones, "An anatomically comprehensive atlas of the adult human brain transcriptome," *Nature*, vol. 489, pp. 391–399, Sept. 2012.

[22] N. A. Twine, K. Janitz, M. R. Wilkins, and M. Janitz, "Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by alzheimer's disease," *PLoS One*, vol. 6, p. e16266, Jan. 2011.

[23] J. J. Li, P. J. Bickel, and M. D. Biggin, "System wide analyses have underestimated protein abundances and the importance of transcription in mammals," *PeerJ*, vol. 2, p. e270, Feb. 2014.

[24] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach, "Global quantification of mammalian gene expression control," *Nature*, vol. 473, pp. 337–342, May 2011.

[25] L. Ping, D. M. Duong, L. Yin, M. Gearing, J. J. Lah, A. I. Levey, and N. T. Seyfried, "Global quantitative analysis of the human brain proteome in alzheimer's and parkinson's disease," *Sci Data*, vol. 5, p. 180036, Mar. 2018.

[26] S. Musunuri, M. Wetterhall, M. Ingelsson, L. Lannfelt, K. Artemenko, J. Bergquist, K. Kultima, and G. Shevchenko, "Quantification of the brain proteome in alzheimer's disease using multiplexed mass spectrometry," *J. Proteome Res.*, vol. 13, pp. 2056–2068, Apr. 2014.

[27] S. J. Schonberger, P. F. Edgar, R. Kydd, R. L. Faull, and G. J. Cooper, "Proteomic analysis of the brain in alzheimer's disease: molecular phenotype of a complex disease process," *Proteomics*, vol. 1, pp. 1519–1528, Dec. 2001.

[28] I. Begcevic, H. Kosanam, E. Martínez-Morillo, A. Dimitromanolakis, P. Diamandis, U. Kuzmanov, L.-N. Hazrati, and E. P. Diamandis, "Semiquantitative proteomic analysis of human hippocampal tissues from alzheimer's disease and age-matched control brains," *Clin. Proteomics*, vol. 10, p. 5, May 2013.

[29] D. C. Hondius, P. van Nierop, K. W. Li, J. J. M. Hoozemans, R. C. van der Schors, E. S. van Haastert, S. M. van der Vies, A. J. M. Rozemuller, and A. B. Smit, "Profiling the human hippocampal proteome at all pathologic stages of alzheimer's disease," *Alzheimers. Dement.*, vol. 12, pp. 654–668, June 2016.

[30] P. F. Edgar, S. J. Schonberger, B. Dean, R. L. Faull, R. Kydd, and G. J. Cooper, "A com-

parative proteome analysis of hippocampal tissue from schizophrenic and alzheimer's disease individuals," *Mol. Psychiatry*, vol. 4, pp. 173–178, Mar. 1999.

[31] H. Braak and E. Braak, "Evolution of the neuropathology of alzheimer's disease," *Acta Neurol. Scand. Suppl.*, vol. 165, pp. 3–12, 1996.

[32] J. Cox and M. Mann, "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification," *Nat. Biotechnol.*, vol. 26, pp. 1367–1372, Dec. 2008.

[33] J. A. Vizcaíno, A. Csordas, N. Del-Toro, J. A. Dianes, J. Griss, I. Lavidas, G. Mayer, Y. Perez-Riverol, F. Reisinger, T. Ternent, Q.-W. Xu, R. Wang, and H. Hermjakob, "2016 update of the PRIDE database and its related tools," *Nucleic Acids Res.*, vol. 44, p. 11033, Dec. 2016.

[34] M. Spitzer, J. Wildenhain, J. Rappsilber, and M. Tyers, "BoxPlotR: a web tool for generation of box plots," *Nat. Methods*, vol. 11, pp. 121–122, Feb. 2014.

[35] J. A. Stefely, N. W. Kwiecien, E. C. Freiberger, A. L. Richards, A. Jochem, M. J. P. Rush, A. Ulbrich, K. P. Robinson, P. D. Hutchins, M. T. Veling, X. Guo, Z. A. Kemmerer, K. J. Connors, E. A. Trujillo, J. Sokol, H. Marx, M. S. Westphall, A. S. Hebert, D. J. Pagliarini, and J. J. Coon, "Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling," *Nat. Biotechnol.*, vol. 34, pp. 1191–1197, Nov. 2016.

[36] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "The one hour yeast proteome," *Mol. Cell. Proteomics*, vol. 13, pp. 339–347, Jan. 2014.

[37] A. L. Richards, A. S. Hebert, A. Ulbrich, D. J. Bailey, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "One-hour proteome analysis in yeast," *Nat. Protoc.*, vol. 10, pp. 701–714, May 2015.

[38] M. Beck, A. Schmidt, J. Malmstroem, M. Claassen, A. Ori, A. Szymborska, F. Herzog, O. Rinner, J. Ellenberg, and R. Aebersold, "The quantitative proteome of a human cell line," *Mol. Syst. Biol.*, vol. 7, p. 549, Nov. 2011.

[39] N. Nagaraj, J. R. Wisniewski, T. Geiger, J. Cox, M. Kircher, J. Kelso, S. Pääbo, and M. Mann, "Deep proteome and transcriptome mapping of a human cancer cell line," *Mol. Syst. Biol.*, vol. 7, p. 548, Nov. 2011.

[40] E. Sjöstedt, L. Fagerberg, B. M. Hallström, A. Häggmark, N. Mitsios, P. Nilsson, F. Pontén, T. Hökfelt, M. Uhlén, and J. Mulder, "Defining the human brain proteome using transcriptomics and Antibody-Based profiling with a focus on the cerebral cortex," *PLoS One*, vol. 10, p. e0130028, June 2015.

[41] S. Walløe, B. Pakkenberg, and K. Fabricius, "Stereological estimation of total cell numbers in the human cerebral and cerebellar cortex," *Front. Hum. Neurosci.*, vol. 8, p. 508, July 2014.

[42] E. Niedzielska, I. Smaga, M. Gawlik, A. Moniczewski, P. Stankowicz, J. Pera, and M. Filip, "Oxidative stress in neurodegenerative diseases," *Molecular neurobiology*, vol. 53, no. 6, pp. 4094–4125, 2016.

[43] C. Ramassamy, D. Averill, U. Beffert, S. Bastianetto, L. Theroux, S. Lussier-Cacan, J. S. Cohn, Y. Christen, J. Davignon, R. Quirion, *et al.*, "Oxidative damage and protection by antioxidants in the frontal cortex of alzheimer's disease is related to the apolipoprotein e genotype," *Free Radical Biology and Medicine*, vol. 27, no. 5-6, pp. 544–553, 1999.

[44] A. M. Swomley and D. A. Butterfield, "Oxidative stress in alzheimer disease and mild cognitive impairment: evidence from human data provided by redox proteomics," *Archives of toxicology*, vol. 89, no. 10, pp. 1669–1680, 2015.

[45] C. Venkateshappa, G. Harish, A. Mahadevan, M. S. Bharath, and S. Shankar, "Elevated oxidative stress and decreased antioxidant function in the human hippocampus and frontal cortex with increasing age: implications for neurodegeneration in alzheimer's disease," *Neurochemical research*, vol. 37, no. 8, pp. 1601–1614, 2012.

[46] J. F. Loring, J. G. Porter, J. Seilhammer, M. R. Kaser, and R. Wesselschmidt, "A gene expression profile of embryonic stem cells and embryonic stem cell-derived neurons," *Restor. Neurol. Neurosci.*, vol. 18, no. 2-3, pp. 81–88, 2001.

[47] V. P. Andreev, V. A. Petyuk, H. M. Brewer, Y. V. Karpievitch, F. Xie, J. Clarke, D. Camp, R. D. Smith, A. P. Lieberman, R. L. Albin, Z. Nawaz, J. El Hokayem, and A. J. Myers,

"Label-free quantitative LC-MS proteomics of alzheimer's disease and normally aged human brains," *J. Proteome Res.*, vol. 11, pp. 3053–3067, June 2012.

[48] J. M. Ajmo, L. A. Bailey, M. D. Howell, L. K. Cortez, K. R. Pennypacker, H. N. Mehta, D. Morgan, M. N. Gordon, and P. E. Gottschall, "Abnormal post-translational and extracellular processing of brevican in plaque-bearing mice over-expressing APPsw," *J. Neurochem.*, vol. 113, pp. 784–795, May 2010.

[49] M. Morawski, G. Brückner, C. Jäger, G. Seeger, R. T. Matthews, and T. Arendt, "Involvement of perineuronal and perisynaptic extracellular matrix in alzheimer's disease neuropathology," *Brain Pathol.*, vol. 22, pp. 547–561, July 2012.

[50] E. J. Mufson, S. E. Counts, and S. D. Ginsberg, "Gene expression profiles of cholinergic nucleus basalis neurons in alzheimer's disease," *Neurochem. Res.*, vol. 27, pp. 1035–1048, Oct. 2002.

[51] A. Olmos-Alonso, S. T. T. Schetters, S. Sri, K. Askew, R. Mancuso, M. Vargas-Caballero, C. Holscher, V. H. Perry, and D. Gomez-Nicola, "Pharmacological targeting of CSF1R inhibits microglial proliferation and prevents the progression of alzheimer's-like pathology," *Brain*, vol. 139, pp. 891–907, Mar. 2016.

[52] A. Dityatev, C. I. Seidenbecher, and M. Schachner, "Compartmentalization from the outside: the extracellular matrix and functional microdomains in the brain," *Trends Neurosci.*, vol. 33, pp. 503–512, Nov. 2010.

[53] G. Glavan, R. Schliebs, and M. Zivin, "Synaptotagmins in neurodegeneration," *Anat. Rec.*, vol. 292, pp. 1849–1862, Dec. 2009.

[54] P. R. Huttenlocher and A. S. Dabholkar, "Regional differences in synaptogenesis in human cerebral cortex," *J. Comp. Neurol.*, vol. 387, pp. 167–178, Oct. 1997.

[55] E. Furube, S. Kawai, H. Inagaki, S. Takagi, and S. Miyata, "Brain region-dependent heterogeneity and dose-dependent difference in transient microglia population increase during lipopolysaccharide-induced inflammation," *Sci. Rep.*, vol. 8, p. 2203, Feb. 2018.

[56] A. Harada, K. Oguchi, S. Okabe, J. Kuno, S. Terada, T. Ohshima, R. Sato-Yoshitake, Y. Takei, T. Noda, and N. Hirokawa, "Altered microtubule organization in small-calibre axons of mice lacking tau protein," *Nature*, vol. 369, pp. 488–491, June 1994.

[57] C. M. Karch, L. A. Ezerskiy, S. Bertelsen, Alzheimer's Disease Genetics Consortium (ADGC), and A. M. Goate, "Alzheimer's disease risk polymorphisms regulate gene expression in the ZCWPW1 and the CELF1 loci," *PLoS One*, vol. 11, p. e0148717, Feb. 2016.

[58] H. J. Huttunen, D. Havas, C. Peach, C. Barren, S. Duller, W. Xia, M. P. Frosch, B. Hutter-Paier, M. Windisch, and D. M. Kovacs, "The acyl-coenzyme a: cholesterol acyltransferase inhibitor CI-1011 reverses diffuse brain amyloid pathology in aged amyloid

precursor protein transgenic mice," *J. Neuropathol. Exp. Neurol.*, vol. 69, pp. 777–788, Aug. 2010.

[59] S. Le Corre, H. W. Klafki, N. Plesnila, G. Hübinger, A. Obermeier, H. Sahagún, B. Monse, P. Seneci, J. Lewis, J. Eriksen, C. Zehr, M. Yue, E. McGowan, D. W. Dickson, M. Hutton, and H. M. Roder, "An inhibitor of tau hyperphosphorylation prevents severe motor impairments in tau transgenic mice," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, pp. 9673–9678, June 2006.

[60] M. C. Oldham, S. Horvath, and D. H. Geschwind, "Conservation and evolution of gene coexpression networks in human and chimpanzee brains," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, pp. 17973–17978, Nov. 2006.

[61] G. Kempermann, E. J. Chesler, L. Lu, R. W. Williams, and F. H. Gage, "Natural variation and genetic covariance in adult hippocampal neurogenesis," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, pp. 780–785, Jan. 2006.

[62] C. Cesaretti, L. Spaccini, A. Righini, C. Parazzini, G. Conte, F. Crosti, S. Redaelli, G. Bulfamante, L. Avagliano, and M. Rustico, "Prenatal detection of 5q14.3 duplication including MEF2C and brain phenotype," *Am. J. Med. Genet. A*, vol. 170A, pp. 1352–1357, May 2016.

[63] H. J. Lee, K. Lee, and H. Im, "α-Synuclein modulates neurite outgrowth by interacting with SPTBN1," *Biochem. Biophys. Res. Commun.*, vol. 424, pp. 497–502, Aug. 2012.

[64] J. R. Wiśniewski, "Label-Free and Standard-Free absolute quantitative proteomics using the "total protein" and "proteomic ruler" approaches," *Methods Enzymol.*, vol. 585, pp. 49–60, 2017.

[65] M. Rahman, S. F. Previs, T. Kasumov, and R. G. Sadygov, "Gaussian process modeling of protein turnover," *J. Proteome Res.*, vol. 15, pp. 2115–2122, July 2016.

Chapter 3

PILOT PROTEOMIC ANALYSIS OF CEREBROSPINAL FLUID IN ALZHEIMER'S DISEASE

JM designed and conducted mass spectrometry experiments, interpreted results, analyzed data, and wrote the manuscript.

This chapter is adapted from a manuscript published in the Proteomics: Clinical Applications with the permission of:

McKetney J, Panyard DJ, Johnson SC, Carlsson C, Engelman CD, Coon JJ. *Pilot Proteomic Analysis of Cerebrospinal Fluid in Alzheimer's Disease.* Proteomics: Clinical Applications. **2021**.

Abstract

Proteomic analysis of cerebrospinal fluid (CSF) holds great promise in understanding the progression of neurodegenerative diseases, including Alzheimer's disease (AD). As one of the primary reservoirs of neuronal biomolecules, CSF provides a window into the biochemical and cellular aspects of the neurological environment. CSF can be drawn from living participants allowing the potential alignment of clinical changes with these biochemical markers. Using cutting-edge mass spectrometry technologies, we perform a streamlined proteomic analysis of CSF. We quantify greater than 700 proteins across 10 pairs of age- and sex-matched participants in approximately one hour of analysis time each. Using the paired participant study structure, we identify a small group of biologically relevant proteins that show substantial changes in abundance between cognitive normal and AD participants, which were then analyzed at the peptide level using parallel reaction monitoring experiments. Our findings suggest the utility of fractionating a single sample and using matching to increase proteomic depth in cerebrospinal fluid, as well as the potential power of an expanded study.

Introduction

Alzheimer's disease (AD) is the sixth leading cause of death in the United States[1] and affects tens of millions worldwide[2]. Much remains to be understood about the onset and progression of AD, and no effective therapeutics to significantly alter its course currently

exist[3]. Proteomic analysis of brain tissue across age[4], cell type[5], and brain region[6–8] has been extensive, but brain-focused studies require post-mortem tissue samples and thus offer limited insight into the molecular timeline of disease progression. Proteomic analysis of cerebrospinal fluid (CSF), in contrast, allows for detection of molecular changes that occur during pathological decline. This approach holds great potential to discover additional biomarkers for AD and to increase understanding of the biological factors that lead to the diverse neurological effects observed across individuals. The unique benefits of CSF may be tempered by two competing objectives: targeting specific AD proteins[9–11] at the expense of discovery capacity vs. generating extensive catalogues of human CSF proteins[12–14] at the expense of preparation and analysis speed, which can impede the ability to make large-scale comparisons. In humans, large scale comparisons are often required to overcome population heterogeneity due to factors such as age and sex, both of which have been shown to have substantial effects on the protein abundances in CSF[15–19]. In this pilot study, we sought to determine the most effective solution for larger analyses of global protein expression in CSF by comparing three different proteomics methodologies: single-shot experiments, experiments with addition of a high field asymmetric waveform ion mobility spectrometry (FAIMS) interface, and experiments analyzed in parallel with commercial CSF fractions. The capacity of these three strategies to quantify proteins and to capture variation in protein abundances was compared using a cohort of 20 individual samples. This sample set was comprised of ten age- and sex-matched AD case-control pairs, evenly split between males and females in an attempt to control for these sources

of variability. We avoided individual sample fractionation and high-abundance protein depletion (common steps in CSF proteomics) to increase precision and ease of preparation while decreasing preparation time[20]. Despite the statistical limits of our sample size, this analysis quantified over 700 proteins that were detected across all 20 participants, including several found to be significantly associated with AD. These proteins included multiple 14-3-3 proteins, which have been previously colocalized to the neurofibrillary tangles and hypothesized to function in sequestering misfolded proteins, representing a potentially beneficial reaction to pathological protein aggregation. Our analysis also identifies more than 30 AD-associated proteins at lower significance that have been previously observed in large-scale studies[15,16,21] and metareviews[22,23]. We believe this analysis can provide a valuable framework for large-scale global proteomic analysis of CSF, both in AD and in other neurodegenerative diseases.

Materials and Methods

Participant Selection. Subjects came from the Wisconsin Alzheimer's Disease Research Center (WADRC) clinical core, a longitudinal cohort study of middle-aged and older adults. Ten AD case/control pairs of subjects, matched for sex and age at lumbar puncture, were selected from the WADRC cohort for this study. The mean age of the AD case group was 72.1 with a maximum of 80.4 and a minimum of 62.1 while the control group ages ranged from 62.1 to 80.4 with a mean of 72.1 as well (**Table 3.1**). AD cases were defined as subjects who met all three of the following criteria: 1) diagnosed with dementia due to AD by consensus

conference using the National Institute on Aging—Alzheimer's Association criteria[24]; 2) amyloid positive status, defined here as having either a CSF Aβ42 measurement less than 471.54 pg/mL (Innotest method) or having a CSF Aβ42:Aβ40 ratio less than 0.09 (Triplex method)[25]; and 3) tau positive status, defined as having either a total CSF tau level greater than 461.26 pg/mL or a phosphorylated tau greater than 59.5 pg/mL[25]. Controls were defined as subjects meeting the following criteria: 1) cognitively normal according to the consensus conference; 2) amyloid negative status (defined as above); 3) tau negative status. For each subject, the clinical diagnosis used was the diagnosis closest in time to the date of the lumbar puncture that generated the CSF sample used in this study, and no diagnosis was more than six months removed from that date.

CSF Sample Collection. CSF was collected via lumbar puncture in the morning after a 12-hour fast[26]. Briefly, after gentle extraction, mixing, and centrifugation, supernatants were flash frozen and stored at -80 degrees until the time of preparation.

Extraction and Digestion Samples were brought to 90% methanol and centrifuged, with the precipitate pellet then resuspended in 8M urea, 10mM TCEP, 40mM CAA, 100mM Tris pH 8. The solution was then diluted to 25% concentration with 100mM Tris, pH 8, and trypsin was added at a ratio of 50:1 w/w and digested overnight at ambient temperature. Digested peptides were desalted using Strata-X Polymeric Reverse Phase column (Phenomenex).

		Age		CSF AB42 (pg/mL)		CSF AB42:AB40		CSF ptau (pg/mL)	
Diagnosis	Sex	Mean	SD	Mean	SD	Mean	SD	Mean	SD
AD	Female	72.0	7.2	395.0	58.3	0.06	0.006	76.6	16.0
Control	Female	72.1	7.3	757.2	134.0	0.11	0.012	38.8	11.5
AD	Male	72.1	7.3	505.8	287.1	0.06	0.012	86.0	28.2
Control	Male	72.0	7.0	773.8	150.2	0.11	0.013	42.2	12.5

Table 3.1: Sample Characteristics. Mean and standard deviation of age, AB42 concentration, amyloid ratio and phosphorylated tau concentration for each sex-diagnosis group

Chromatographic Columns. Online reverse-phase columns were prepared in-house using a high-pressure packing apparatus[27]. In brief, 1.5 μm Bridged Ethylene Hybrid C18 particles were packed at 30,000 psi into a New Objective PicoTip™ emitter (Stock# PF360-75-10-N-5) with an inner diameter of 75 μm and an outer diameter of 360 μm.

Experimental Strategy 1: Single-Shot Experiments. Mobile phase A consisted of 0.2% formic acid in water and mobile phase B consisted of 0.2% formic acid in 70% acetonitrile. Samples were loaded onto the column for 8.6 minutes at 300 nL/min. Mobile phase B was increased to 5% in the first 8.6 min then increased to 50% by 51 min. The method finished with a wash stage of 100% B from 52-54 minutes and an equilibration step of 1% B from 55-60 minutes. Eluting ions were analyzed on a Thermo Orbitrap Fusion Lumos in a data-dependent manner with survey scans taken in the orbitrap at 240,000 resolution and MS2 scans taken in the ion trap using the "rapid" setting. Samples were analyzed in duplicate.

Experimental Strategy 2: FAIMS Experiments. Identical chromatographic methods were used for the FAIMS experiments as detailed above for the single-shot experiments. Peptide ions passing through the FAIMS interface are destabilized by an alternating electric field, the effect of which is countered by a peptide-specific compensating voltage (CV). This stabilization allows for the selection of a specific subpopulation of peptide ions when using a specific compensation voltage setting. We performed three analyses on each participant

sample, one for each of three CV settings: -60, -75, and -90. Survey scans of precursors were taken in the orbitrap at 120,000 resolution while MS2 scans were taken in the ion trap using the "rapid" setting.

Experimental Strategy 3: Experiments Run with Fractions. Non-designated BioIVT cerebrospinal fluid was denatured, digested, and desalted as detailed above. Dried samples were resuspended in 0.2% formic acid and fractionated using an HPLC (Agilent, Infinity 2000) with a 150 mm C18 reverse phase column (Waters, XBridge Peptide BEH, particle size 3.5 μm). Mobile phase buffer A was a freshly prepared mixture of 10mM ammonium bicarbonate, pH 9.5; mobile phase B was composed of 10mM ammonium bicarbonate, 80% methanol, pH 9.5. The gradient method was 20 minutes in length with 32 fractions collected from 5 to 20 minutes and a flow rate of 800 nL/min throughout. Mobile phase B was increased from 5% to 35% in the first 2 minutes before increasing to 100% B by 13 minutes. 32 fractions were concatenated into 16 fractions by combining every other column in the sample collection plate. The chromatographic method was lengthened slightly for fractions and participant samples run alongside fractions in order to accommodate the lower peptide concentration of fractions. Samples were loaded onto the column for 12 minutes at 350 nL/min. Mobile phase B increased to 12% in the first 12 min then increased to 65% by 55 min. The method finished with a wash stage of 100% B from 56-59 minutes and an equilibration step of 0% B from 60-70 minutes. These experiments used the same instrument acquisition method as the single-shot experiments.

Protein Quantification. The resulting spectra were searched in MaxQuant (1.6.0.13) using fast LFQ against a full human proteome with isoforms downloaded from Uniprot (October 29, 2018). Each set of experimental strategies was searched separately. Carbamidomethylation of cysteine was set as fixed modification. Matching between runs was used with a retention time window of 0.7 min. Searches were performed using a protein FDR of 1%, a minimum peptide length of 7, and a 0.5 Da MS2 match tolerance. Protein data were then extracted from the "ProteinGroups.txt" file of the MaxQuant output after decoy, contaminants, and reverse sequences were removed. The protein counts were based on protein groups with an LFQ Intensity > 0.

Protein Variation, Pairwise Correlations, and Differential Abundance in AD. Single-shot analysis of participant samples alongside a fractionated commercial CSF sample searched using match-between-runs, was found to yield the greatest number of proteins quantified across all participant samples. This experimental strategy was then used to generate pairwise correlations between samples and to test the association with AD (paired t-test, unequal variance). A paired t-test was used due to the matched nature of the study design. Data analysis was performed in R (3.6.1) using the base package. Plots were generated using lattice plotting and ggplot2 in R.

Parallel Reaction Monitoring Experiments. The MaxQuant (1.6.0.13) output tables for the fractionated CSF were used to build a spectral library in Skyline (19.1.0.193). Proteins

of interest were then selected from the library based on significance (t-test, p-value < 0.02) in the data-dependent analysis. Scheduled PRM experiments monitored 27 peptide ions from seven proteins using 4-min retention time windows. Peptides were isolated with a 1.6 m/z window before being scanned in the orbitrap at 60,000 resolution. Peak areas and spectral traces for parallel reaction monitoring experiments were extracted from Skyline and processed in R. The MS proteomics files have been deposited to the publicly available Chorus Project repository (chorusproject.org) under the project title "Proteomic Analysis of Cerebrospinal Fluid in Alzheimer's Disease".

Results

Protein Quantification and Variation. Proteins in the CSF samples were extracted, denatured, desalted, and then digested with trypsin[27,28]. Peptides resulting from trypsin digestion of the commercially available CSF were fractionated using high-pH reverse phase fractionation. All samples were injected separately onto an online liquid chromatography system and analyzed with a quadrupole-orbitrap dual-cell linear ion trap hybrid mass spectrometer (**Figure 3.1**) for all three strategies described above: single shot (40 total hours of instrument time), single shot with FAIMS added (60 hours), and single-shot analysis of participant samples run in parallel with a fractionated commercial CSF sample (65 hours).

The latter strategy led to the highest number of identified proteins with a total of 2,118. Of these proteins, 939 were quantified in greater than 50% of each sex-disease group and 776 were quantified across all 20 participants (**Supplemental Table 3**). These numbers

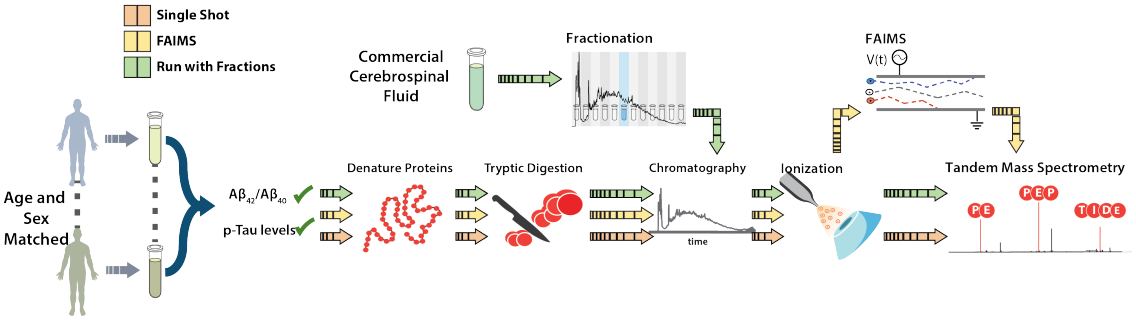


Figure 3.1: Experimental Design. 20 individuals were age- and sex-matched to form 10 AD case-control pairs. CSF samples were collected by lumbar puncture, after which proteins were extracted, denatured, and digested before being analyzed by tandem mass spectrometry. Three different analysis strategies were utilized: single-shot (orange), addition of FAIMS (yellow), and single-shot analyses run in parallel with a fractionated commercial CSF sample (green).

are comparable to those in recent DIA analyses of CSF, which relied on substantially more fractionation and longer chromatographic gradients[15,29]. Analyses that included fractions showed a 35% increase in the number of proteins quantified in all participant samples over the FAIMS analysis (**Figure 3.2b**) and a 56% increase over the single-shot analyses alone. These additional proteins included several well-characterized AD-related proteins, including neuroligin[30]. While FAIMS also increased the number of proteins quantified compared to single-shot, non-FAIMS experiments, the increase was not proportional to the additional analysis time, unlike the experiments with fractions (50% increase in instrument time, 50% increase in proteins quantified).

The label-free abundances (LFQ intensity) from experiments run alongside the commercial fractions were used for the remaining analyses in this study. To eliminate erroneous characterization[30,31], the following analyses were performed using only proteins for which an LFQ intensity was produced across all 20 participant samples, with no missing values(776 proteins referenced above). Like many other body fluids, CSF is highly dynamic in molecular content[32,33], leading to higher relative standard deviations (RSDs) for protein abundance than would be expected for other tissues[34]. Indeed, large variations in protein levels have been one of bottlenecks of proteomic analysis in CSF, with relative standard deviations > 1.00 previously reported[35]. We observed median RSDs in AD samples for both sexes that were higher than those of the healthy controls (21% vs. 17% for females and 29% vs. 20% for males in our experiments with fractions). Due to the consistency of this RSD pattern across all three methodologies (**Figure 3.2c**), we infer that these differences

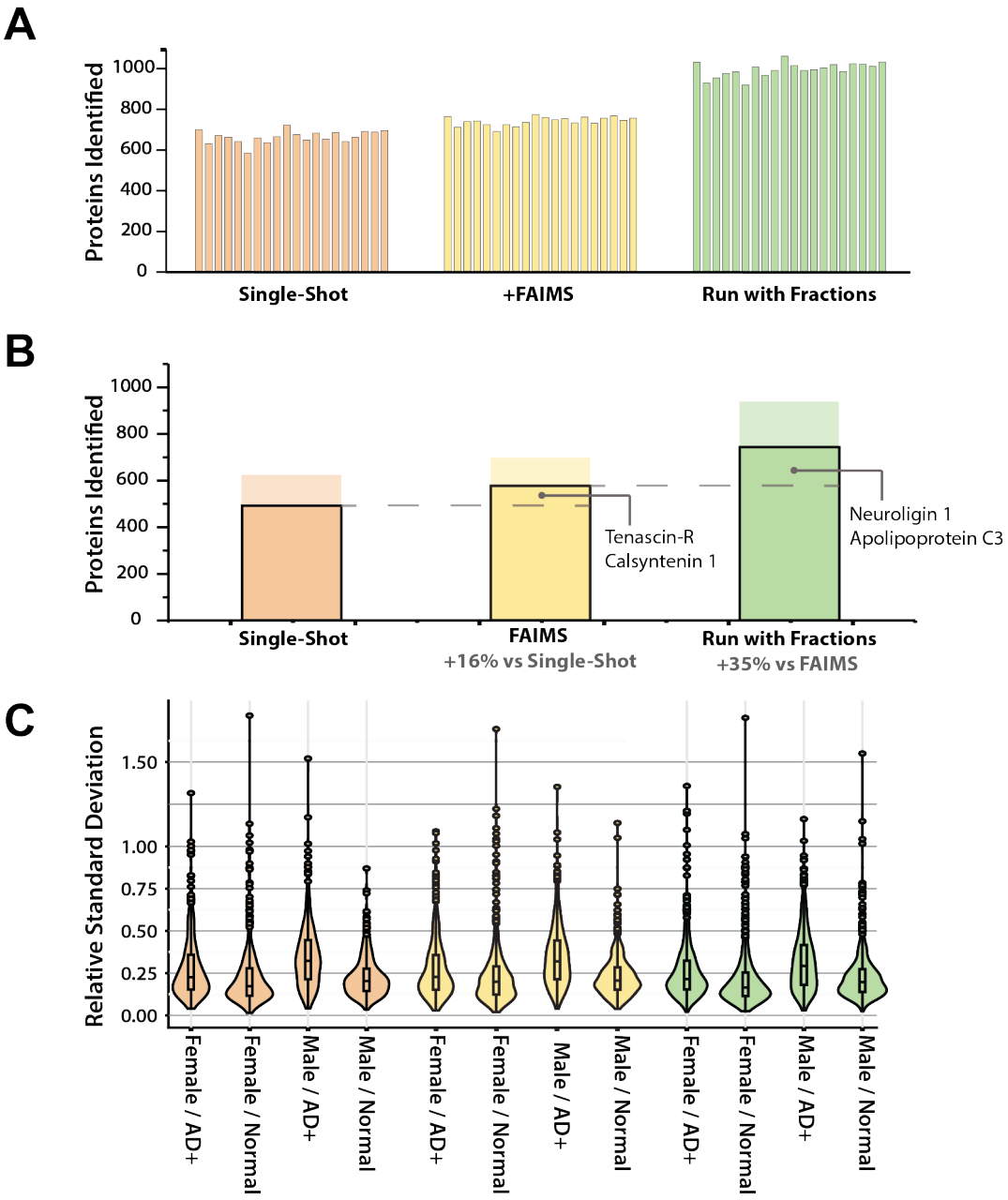


Figure 3.2: Distribution of proteins quantified across three methods. (A) Number of proteins quantified in each sample for each analysis method. (B) Proteins quantified overall for each analysis method. Solid bar shows proteins quantified in all 20 samples. Translucent bar shows number of proteins quantified in at least 50% of each of the four sex-disease groups. Additional proteins quantified include several previously implicated in AD. (C) Relative standard deviation (RSD) distribution for each sex-disease group. AD groups exhibit higher median RSDs than healthy normal counterparts for all sexes and analysis strategies.

stemmed from the innate characteristics of the samples rather than artifacts of the analysis method.

Pairwise Correlation. When comparing protein abundance across all samples using a pairwise Pearson correlation, we observe patterns that vary in direction and magnitude even within sex and disease state groups, with several unexpected correlations (**Figure 3.3a**). The strongest positive correlation (R=0.64) compares the protein abundances of the youngest samples from both the male and female AD+ group. The strongest anticorrelation (R = -0.64) compares two male, AD+ samples, the youngest and the oldest. These two correlations would initially seem counterintuitive given the misalignment of sex and the alignment of disease state, respectively. Although our sample population is not explicitly structured to examine age, we explored it as a possible explanation of these unexpected correlations by building two regression models: one relied on only disease state and sex as explanatory variables; the other also included age. Protein abundances were fit to each model, and the strength of the two models were compared using an analysis of variance (ANOVA). The p-value associated with that ANOVA was used to establish the significance of age in the expression profile for our sample set (**Supplemental Table 1**). We plotted normalized protein abundances for the two comparisons described above (**Figure 3.3a**) with proteins colored by significance of age as an explanatory variable in the linear model (**Figure 3.3b and 3.3c**). In the positive correlation(**Figure 3.3b**), we see a dense clustering of age-associated proteins in quadrant I, strengthening the correlation by increasing the

slope; while in the anticorrelated comparison (**Figure 3.3c**), we see these same proteins in quadrant IV, decreasing the slope, and in turn magnifying the negative correlation. Previous work has shown a substantial effect of age on protein abundance in CSF[15], as well as in plasma[15,21] and the brain in Alzheimer's[16]. Significant protein expression shifts have also been identified in CSF in the normal aging process[17,36]. These patterns indicate the importance of age in studying the human CSF proteome and neurodegeneration.

Differentially Expressed Disease Proteins We next examined protein expression differences in CSF between AD and healthy samples using a paired t-test, accounting for unequal variance. Although several proteins in the female cohort exhibited high statistical significance (p-value < 0.001), they failed to meet our significance threshold after correction for multiple comparisons (Benjamini-Hochberg, 5% FDR)[37]. When examining the male sample group, one protein was significantly different after correction, N-acetylglucosamine-6-sulfatase (GSN), which decreased in the presence of AD (**Figure 3.4b**). GSN plays an important role in the regulation of the extracellular matrix in the brain by hydrolyzing heparan sulfate[38,39]. Decreased activity of GSN can lead to mucopolysaccharidosis, a condition associated with neurodegeneration[40]. Although previous work has shown alterations in the CSF[41] and plasma[19] proteome between sexes in mammals, differences observed may reflect limited statistical power due to small sample size, and a larger analysis would be required to fully disentangle these factors. When examining all ten sample pairs together, our analysis identified three proteins as significantly associated with disease: 14-3-3 protein

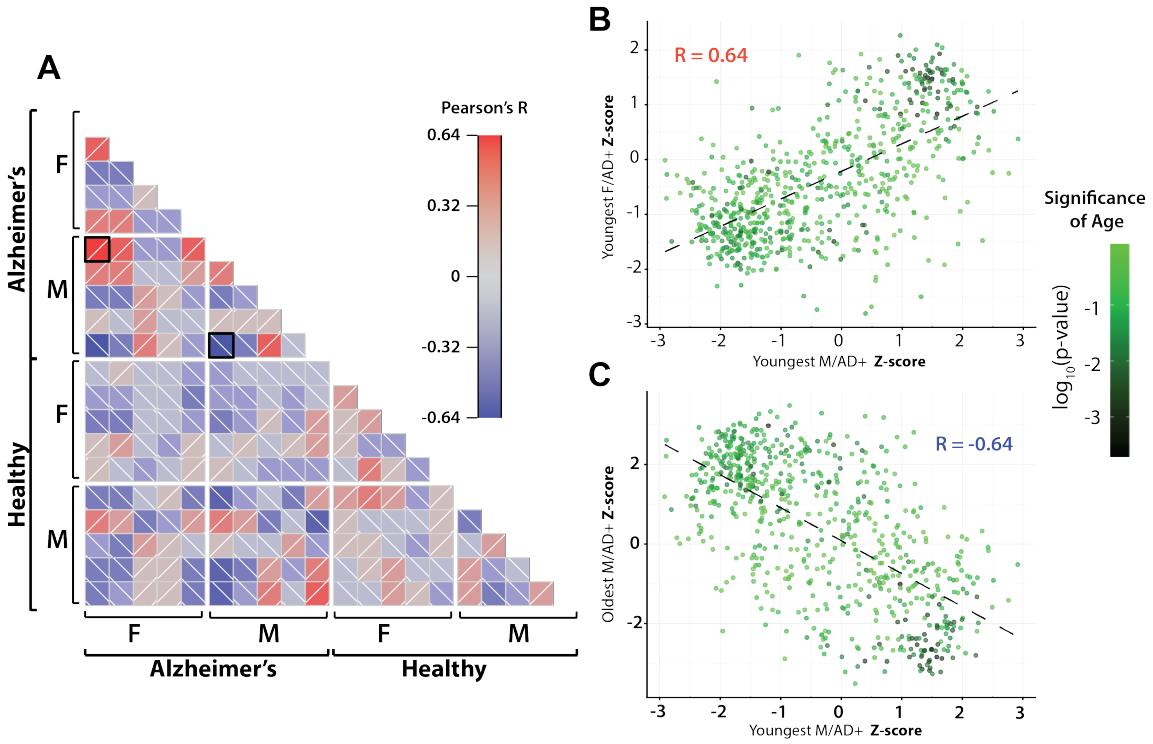


Figure 3.3: Pairwise Pearson correlations. (A) Pairwise Pearson correlation across all samples. Samples are separated by sex and disease state. Within disease/sex groups, samples are ordered by age with youngest samples at left on the x-axis and at top on y-axis. Black boxes indicate comparisons expanded in B and C. (B) Scatterplot of z-score normalized protein expression between youngest male AD+ sample and youngest female AD+ sample. Points are colored by significance when including age as explanatory variable in linear model (C) Scatterplot of z-score normalized expression between youngest male AD+ sample and oldest male AD+ sample. A high number of age-related proteins are present in the quadrants driving protein correlations in both B and C.

zeta (1433Z), 14-3-3 protein gamma (1433G), C-X-C motif chemokine 16 (CXCL16)(**Figure 3.4c, Supplemental Table 2**). These three proteins were also found to be significant when using an unpaired t-test with the same correction. When comparing linear models with the addition of disease as an explanatory variable, similar to the above age analysis, only 1433Z and 1433G were found to be significant after correction. 14-3-3 proteins co-localize with neurofibrillary tangles within the brain[42]. One hypothesis for the role of these proteins in AD suggests that the proteins may operate in a similar capacity to their role as chaperones[43] by sequestering aggregated tau protein to reduce cytotoxicity[44,45]. 14-3-3 proteins are also associated with the development and maturation of synapses, where they function as signal transducers and recognition molecules[44–46]. Synaptic degradation is a sign of the neurodegeneration in AD[47]. We also observed differential expression to a lesser degree of significance (p-value < 0.05) of hypoxanthine phosphoribosyltransferase 1(HPRT1) and myristoylated alanine-rich protein C-kinase substrate (MARCKS), which play a role in the development of neurons[48] and synapses[49,50], respectively. These proteins as well as the 14-3-3 proteins, and CXCL16, have been identified as biomarkers of AD in several large-scale studies[15,16,21,23,51].

The upregulation of CXCL16 among the AD cases may reflect cellular signaling events occurring due to general neuroinflammation. This general neuroinflammation can occur as an effect of both normal aging and AD. CXCL16 functions as a chemokine signaler, helping to promote macrophage chemotaxis and endocytosis[52,53]. CXCL16 has both a transmembrane and soluble form, potentially increasing the chances of detecting elevated CXCL16 levels in

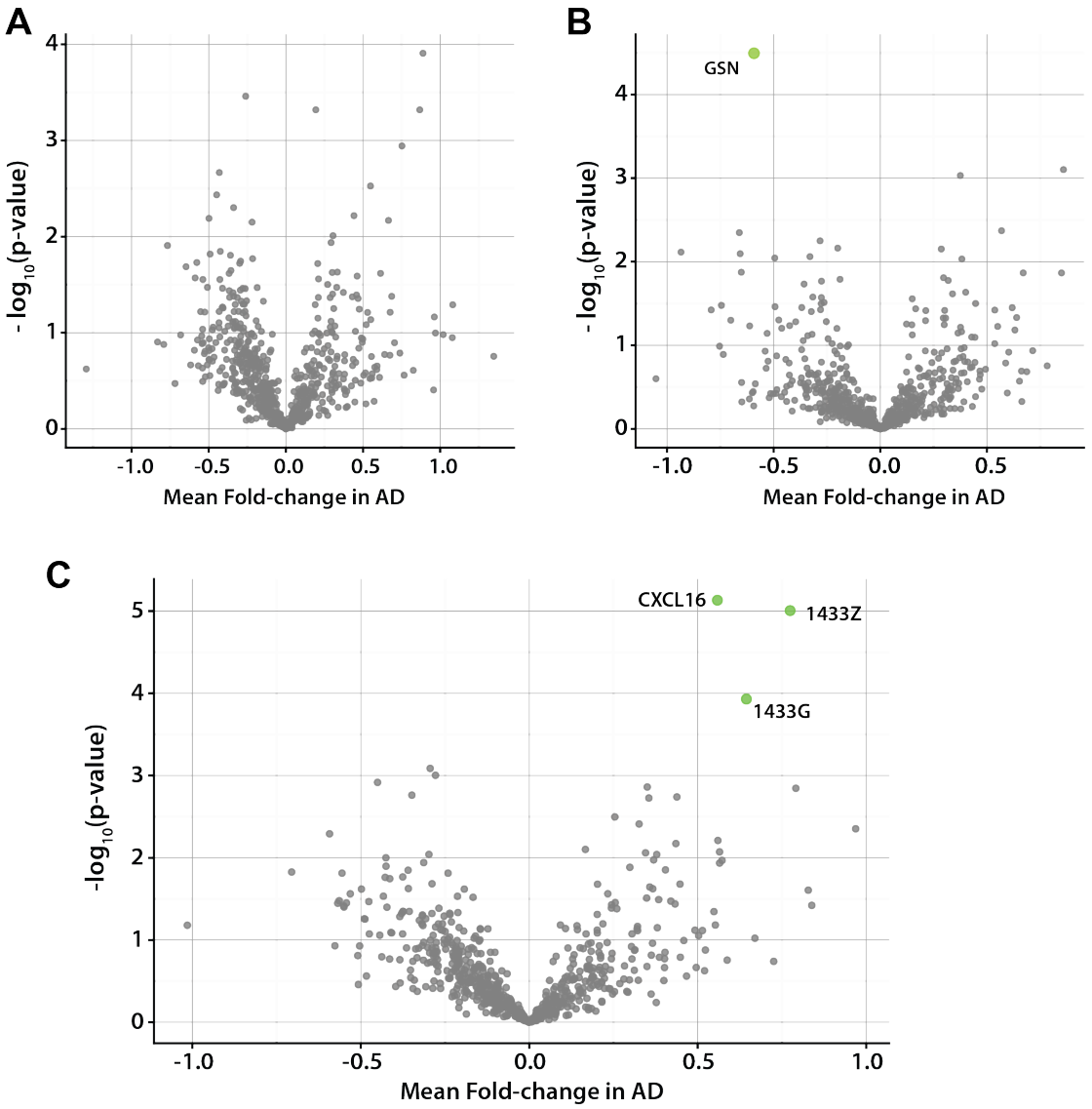


Figure 3.4: Differentially expressed proteins in disease. (A) Volcano plot showing average fold change and significance for protein expression differences between female AD+/normal pairs. No proteins were found to be significant after correction for multiple comparisons. (B) Volcano plot showing fold change and significance for protein expression differences between male AD+/normal pairs. Only N-acetylglucosamine-6-sulfatase (GNS) was found to be significant after correction for multiple comparisons. (C) Volcano plot showing fold change and significance for protein expression across all AD+/normal pairs. Three proteins were identified to be significantly associated with disease 1433G, 1433Z and CXCL16.

a CSF analysis[54]. Two recent proteomic studies of CSF identified CXCL16 as a candidate biomarker for AD, observing increased abundance with disease[51,55]. Although only three proteins met our stringent significance cutoff, more than 80 proteins exhibited p-values < 0.05 (**Supplemental Table 2**), representing multiple biological processes previously associated with AD in CSF, including several proteins recently identified as part of a 40-protein AD signature[15]. We observe substantial upregulation of proteins involved in glycolysis and sugar metabolism including PML-RARA-regulated adapter molecule 1 (PRAM1), lactate dehydrogenase A (LDHA), aldolase fructose bisphosphate (ALDOA), malate dehydrogenase (MDH1), and hypoxanthine phosphoribosyltransferase 1(HPRT1) which have found to be elevated in the cerebrospinal fluid of Alzheimer's patients[15,16,56–58]. Altered abundances were also observed for proteins involved in development and regulation of the extracellular matrix including serine protease HTRA1 (HTRA1), Pastin-2 (LCP1), and collagen alpha-2(IV) chain (COL4A). Previous work had grouped these proteins together into a co-expression module significantly correlated with AD[16].

We also assessed the role of known AD-related proteins in our data set, including amyloid precursor protein (APP), chitinase 3-like 1 protein (CHI3L1 or YKL-40), triggering receptor expressed on myeloid cells 2 protein (TREM-2), amyloid-like protein 1 (APLP1), and amyloid-like protein 2 (APLP2). Although these analyses did not find statistically significant differences between cases and controls, the method was still able to quantify these proteins across our samples. In future studies with larger sample sizes, the changes in these proteins may be better identified. Although other biomarkers such as NfL were

sequenced in our fractionated samples, the associated peptides were not quantified in the participant samples, potentially reflecting the limitations of foregoing high-abundance depletion.

Targeted Analysis using Parallel Reaction Monitoring The relatively low abundance of cells in the CSF poses a challenge to proteomic analysis, as *in vivo* deficiency leads to low protein content and diversity in our decellularized samples. Many of the peptides measured come from extracellular proteins[59] or those released by apoptosis[60–62]. Previous work has shown the potential utility of quantifying endogenous peptides processed by CSF native proteases[63,64]. These peptide subpopulations may experience abundance changes independent from overall protein expression shifts. To allow relative quantitation of specific peptides in a sensitive and accurate manner, we performed parallel reaction monitoring (PRM)[65], a targeted mass spectrometry technique that samples and quantifies a specific list of peptides using the peptides' fragmentation spectra. Our PRM experiment targeted 26 peptides that included 31 precursor ions derived from seven proteins. Proteins were chosen due to their differential abundance in the untargeted study and functional similarity to identified disease-associated proteins.

Summed fragment intensities were compared between case/control sample pairs. Of the 31 peptide ions, 24 had good quality spectral transitions across all 20 samples. Of those peptides, eight had p-values < 0.05 in at least one sex, four derived from the 14-3-3 proteins, three derived from Lysozyme C(LYZ), and one derived from heat shock cognate protein

(HSPA8). Three out of the four precursor ions monitored from lysozyme C had p-values < 0.05 in male participants (**Figure 3.5**). Although previous proteomic studies in CSF have observed different magnitude fold changes of disease-associated proteins between males and females[15], here this difference may reflect statistical power limitations due to samples size. Lysozyme C plays a bacteriolytic role in humans, stemming from macrophages, which supports the continued theme of immune activation and inflammation signals being elevated in participants with AD. Evidence for the coincidence of AD with bacterial infection of the brain[66,67] and other tissues exists[68,69], which could promote upregulation of bactericidal pathways. Lysozyme C also increases the activity of other inflammatory signaling molecules[70].

Discussion

Using a streamlined approach, we quantify more than 700 proteins across all 20 samples of our participant cohort, laying the groundwork for future large-scale proteomic analyses of CSF. This protein number encompasses > 20% of all proteins identified in the deepest CSF proteome analysis to date using a fraction of the preparation and analysis time[13]. A large collection of the proteins quantified have been previously associated with AD and neurodegeneration. Increased depth was achieved both with the addition of FAIMS and a parallel analysis of fractionated CSF. This proteomic depth was acquired without the use of immune depletion of high-abundance proteins or fractionation of the participant samples, allowing for efficient sample preparation and accurate reflection of protein abundances.

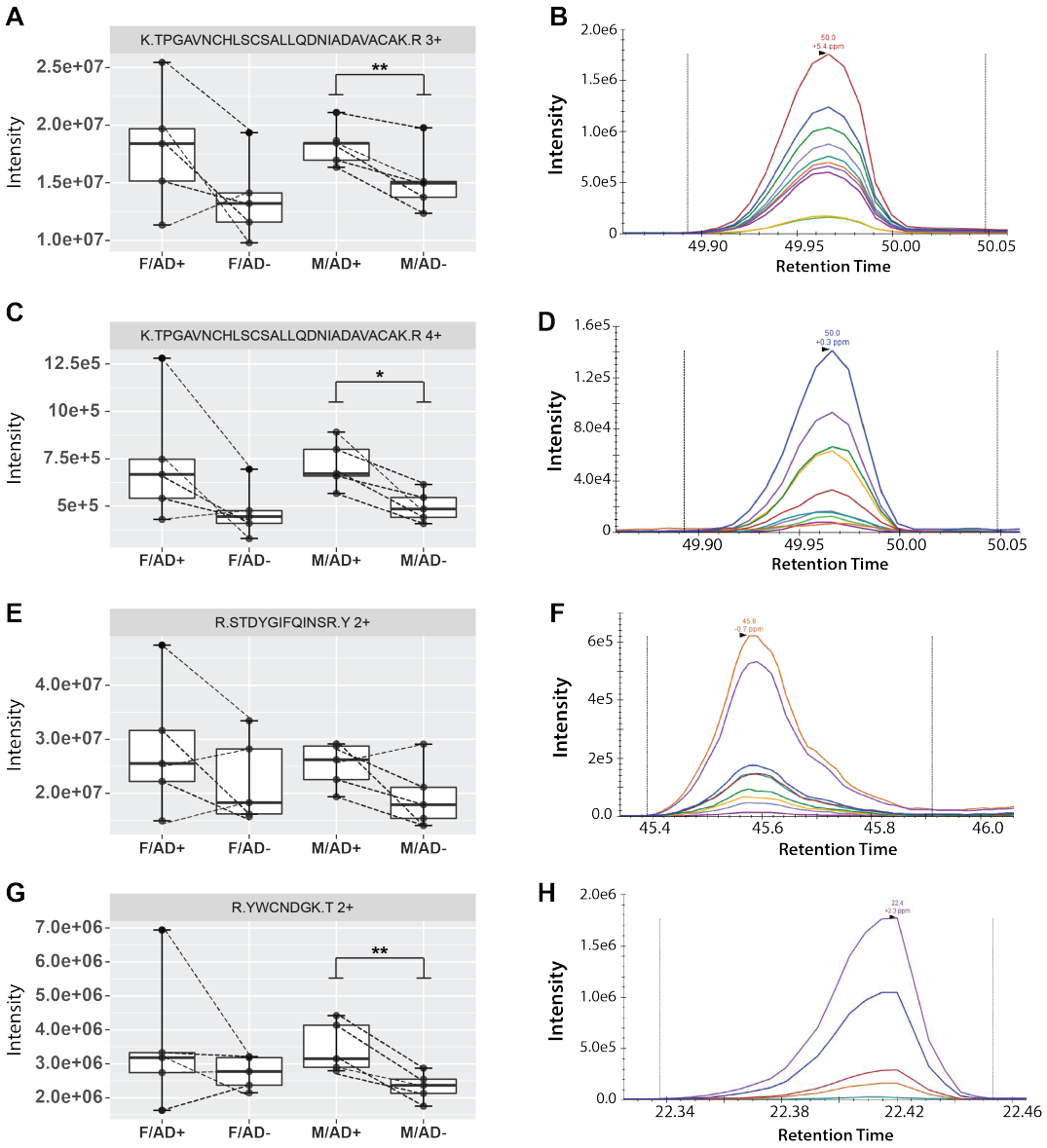


Figure 3.5: Intensities and example transitions for Lysozyme C peptides targeted in parallel reaction monitoring experiments. (A, B) TPGAVNCHLSCSALLQDNIADAVACAK, 3+ charge state (C,D) TPGAVNCHLSCSALLQDNIADAVACAK, 4+ charge state (E,F) STDYGIFQINSR, 2+ charge state (G,H) YWCNDGK, 2+ charge state. Dotted lines connect sample pairs. All four peptide ions exhibited increased intensity with addition of AD across a majority of disease/normal pairs. P-values were determined by paired sample t-test. * = **p-value** < 0.05, ** = **p-value** < 0.01. **M** = male, **F** = female, **AD+** = Alzheimer's disease, **AD-** = cognitively healthy control.

Methods ran nearly one hour in total, providing a technique with excellent scaling potential to larger sample sizes. This rapid analysis proved both efficient and effective, identifying multiple disease-associated proteins despite statistical limitations of sample size and large variations in protein abundance that occur in CSF. When comparing proteomic profiles of individual samples, several strong correlations arose not completely explained by sex or disease state groups. Upon closer examination we observed a possible effect of age, although an altered study would be required to confirm these findings. When comparing across case-control pairs, proteins were identified as significantly disease-associated in a single sex (GSN in males) and both sexes (1433G, 1433Z and CXCL16), showing the interconnected effects of sex and disease state on CSF protein composition. PRM allowed for comparison of the relative abundance of specific peptides with high significance or AD-related functions. Several peptide ions from Lysozyme C exhibited significant shifts in abundance in males, with an increase in AD+ samples. Identification of Lysozyme C in follow-up analyses using PRM indicates the potential of this global profiling followed by targeted analysis to identify additional disease-associated proteins when equipped with greater statistical power. Although we controlled here for distribution of age and sex using our paired study design, these are only two of the myriad sources of between-individual variation in protein abundances in CSF. Alternative study structure would be required to identify the interaction between age, sex, and disease in an expanded study. This study highlights the advantages of CSF as an incredibly reactive and dynamic fluid that undergoes significant changes with neurodegeneration. Here, we present a next step in

utilizing proteomic analysis of this tissue to gain a better understanding of the biochemical and cellular environment in AD. Given the promising findings detailed in this small pilot study, we expect that a coming study with expanded statistical power utilizing the same approach will identify more nuanced changes in this biological fluid.

References

[1] A. Association, "2019 alzheimer's disease facts and figures," *Alzheimer's & dementia*, vol. 15, no. 3, pp. 321–387, 2019.

[2] M. J. Prince, *World Alzheimer Report 2015: the global impact of dementia: an analysis of prevalence, incidence, cost and trends*. Alzheimer's Disease International, 2015.

[3] U. D. of Health, H. Services, *et al.*, "National plans to address alzheimer's disease: 2017 update," 2017.

[4] B. C. Carlyle, R. R. Kitchen, J. E. Kanyo, E. Z. Voss, M. Pletikos, A. M. M. Sousa, T. T. Lam, M. B. Gerstein, N. Sestan, and A. C. Nairn, "A multiregional proteomic survey of the postnatal human brain," *Nat. Neurosci.*, vol. 20, pp. 1787–1795, Dec. 2017.

[5] N. T. Seyfried, E. B. Dammer, V. Swarup, D. Nandakumar, D. M. Duong, L. Yin, Q. Deng, T. Nguyen, C. M. Hales, T. Wingo, J. Glass, M. Gearing, M. Thambisetty, J. C. Troncoso, D. H. Geschwind, J. J. Lah, and A. I. Levey, "A multi-network approach identifies Protein-Specific co-expression in asymptomatic and symptomatic alzheimer's disease," *Cell Syst*, vol. 4, pp. 60–72.e4, Jan. 2017.

[6] J. McKetney, R. M. Runde, A. S. Hebert, S. Salamat, S. Roy, and J. J. Coon, "Proteomic atlas of the human brain in alzheimer's disease," *J. Proteome Res.*, vol. 18, pp. 1380–1391, Mar. 2019.

[7] C. F. Mendonça, M. Kuras, F. C. S. Nogueira, I. Plá, T. Hortobágyi, L. Csiba, M. Palkovits, É. Renner, P. Döme, G. Marko-Varga, G. B. Domont, and M. Rezeli, "Proteomic signatures of brain regions affected by tau pathology in early and late stages of alzheimer's disease," *Neurobiol. Dis.*, vol. 130, p. 104509, Oct. 2019.

[8] J. Xu, S. Patassini, N. Rustogi, I. Riba-Garcia, B. D. Hale, A. M. Phillips, H. Waldvogel, R. Haines, P. Bradbury, A. Stevens, R. L. M. Faull, A. W. Dowsey, G. J. S. Cooper, and R. D. Unwin, "Regional protein expression in human alzheimer's brain correlates with disease severity," *Commun Biol*, vol. 2, p. 43, Feb. 2019.

[9] A. Andersson, J. Remnestål, B. Nellgård, H. Vunk, D. Kotol, F. Edfors, M. Uhlén, J. M. Schwenk, L. L. Ilag, H. Zetterberg, K. Blennow, A. Månberg, P. Nilsson, and C. Fredolini, "Development of parallel reaction monitoring assays for cerebrospinal fluid proteins associated with alzheimer's disease," *Clin. Chim. Acta*, vol. 494, pp. 79–93, July 2019.

[10] F. Brosseron, A. Traschütz, C. N. Widmann, M. P. Kummer, P. Tacik, F. Santarelli, F. Jessen, and M. T. Heneka, "Characterization and clinical use of inflammatory cerebrospinal fluid protein markers in alzheimer's disease," *Alzheimers. Res. Ther.*, vol. 10, p. 25, Feb. 2018.

[11] A. Quaranta, M. Spasova, E. Passarini, I. Karlsson, L. Ndreu, G. Thorsén, and L. L. Ilag, "N-Glycosylation profiling of intact target proteins by high-resolution mass spectrometry (MS) and glycan analysis using ion mobility-MS/MS," *Analyst*, vol. 145, pp. 1737–1748, Mar. 2020.

[12] A. C. Kroksveen, A. Guldbrandsen, M. Vaudel, R. R. Lereim, H. Barsnes, K.-M. Myhr, Ø. Torkildsen, and F. S. Berven, "In-Depth cerebrospinal fluid quantitative proteome and deglycoproteome analysis: Presenting a comprehensive picture of pathways and processes affected by multiple sclerosis," *J. Proteome Res.*, vol. 16, pp. 179–194, Jan. 2017.

[13] C. Macron, L. Lane, A. Núñez Galindo, and L. Dayon, "Deep dive on the proteome of human cerebrospinal fluid: A valuable data resource for biomarker discovery and missing protein identification," *J. Proteome Res.*, vol. 17, pp. 4113–4126, Dec. 2018.

[14] A. Guldbrandsen, H. Vethe, Y. Farag, E. Oveland, H. Garberg, M. Berle, K.-M. Myhr, J. A. Opsahl, H. Barsnes, and F. S. Berven, "In-depth characterization of the cerebrospinal fluid (CSF) proteome displayed through the CSF proteome resource (CSF-PR)," *Mol. Cell. Proteomics*, vol. 13, pp. 3152–3163, Nov. 2014.

[15] J. M. Bader, P. E. Geyer, J. B. Müller, M. T. Strauss, M. Koch, F. Leypoldt, P. Koertvelyessy, D. Bittner, C. G. Schipke, E. I. Incesoy, O. Peters, N. Deigendesch, M. Simons, M. K. Jensen, H. Zetterberg, and M. Mann, "Proteome profiling in cerebrospinal fluid reveals novel biomarkers of alzheimer's disease," *Mol. Syst. Biol.*, vol. 16, p. e9356, June 2020.

[16] E. C. B. Johnson, E. B. Dammer, D. M. Duong, L. Ping, M. Zhou, L. Yin, L. A. Higginbotham, A. Guajardo, B. White, J. C. Troncoso, M. Thambisetty, T. J. Montine, E. B. Lee, J. Q. Trojanowski, T. G. Beach, E. M. Reiman, V. Haroutunian, M. Wang, E. Schadt, B. Zhang, D. W. Dickson, N. Ertekin-Taner, T. E. Golde, V. A. Petyuk, P. L. De Jager, D. A. Bennett, T. S. Wingo, S. Rangaraju, I. Hajjar, J. M. Shulman, J. J. Lah, A. I. Levey, and N. T. Seyfried, "Large-scale proteomic analysis of alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation," *Nat. Med.*, vol. 26, pp. 769–780, May 2020.

[17] G. S. Baird, S. K. Nelson, T. R. Keeney, A. Stewart, S. Williams, S. Kraemer, E. R. Peskind, and T. J. Montine, "Age-dependent changes in the cerebrospinal fluid proteome by slow off-rate modified aptamer array," *Am. J. Pathol.*, vol. 180, pp. 446–456, Feb. 2012.

[18] C. Parrado-Fernández, K. Blennow, M. Hansson, V. Leoni, A. Cedazo-Minguez, and I. Björkhem, "Evidence for sex difference in the CSF/plasma albumin ratio in 20 000 patients and 335 healthy volunteers," *J. Cell. Mol. Med.*, vol. 22, pp. 5151–5154, Oct. 2018.

[19] B. Lehallier, D. Gate, N. Schaum, T. Nanasi, S. E. Lee, H. Yousef, P. Moran Losada, D. Berdnik, A. Keller, J. Verghese, S. Sathyan, C. Franceschi, S. Milman, N. Barzilai, and T. Wyss-Coray, "Undulating changes in human plasma proteome profiles across the lifespan," *Nat. Med.*, vol. 25, pp. 1843–1850, Dec. 2019.

[20] A. Núñez Galindo, M. Kussmann, and L. Dayon, "Proteomics of cerebrospinal fluid: Throughput and robustness using a scalable automated analysis pipeline for biomarker discovery," *Anal. Chem.*, vol. 87, pp. 10755–10761, Nov. 2015.

[21] C. D. Whelan, N. Mattsson, M. W. Nagle, S. Vijayaraghavan, C. Hyde, S. Janelidze, E. Stomrud, J. Lee, L. Fitz, T. A. Samad, G. Ramaswamy, R. A. Margolin, A. Malarstig, and O. Hansson, "Multiplex proteomics identifies novel CSF and plasma biomarkers of early alzheimer's disease," *Acta Neuropathol Commun*, vol. 7, p. 169, Nov. 2019.

[22] K. E. J. Wesenhagen, C. E. Teunissen, P. J. Visser, and B. M. Tijms, "Cerebrospinal fluid proteomics and biological heterogeneity in alzheimer's disease: A literature review," *Crit. Rev. Clin. Lab. Sci.*, vol. 57, pp. 86–98, Mar. 2020.

[23] C. M. Pedrero-Prieto, S. García-Carpintero, J. Frontiñán-Rubio, E. Llanos-González, C. Aguilera García, F. J. Alcaín, I. Lindberg, M. Durán-Prado, J. R. Peinado, and Y. Rabanal-Ruiz, "A comprehensive systematic review of CSF proteins and peptides that define alzheimer's disease," *Clin. Proteomics*, vol. 17, p. 21, June 2020.

[24] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack, Jr, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, R. C. Mohs, J. C. Morris, M. N. Rossor, P. Scheltens, M. C. Carrillo, B. Thies, S. Weintraub, and C. H. Phelps, "The diagnosis of dementia due to alzheimer's disease: recommendations from the national institute on Aging-Alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimers. Dement.*, vol. 7, pp. 263–269, May 2011.

[25] L. R. Clark, S. E. Berman, D. Norton, R. L. Koscik, E. Jonaitis, K. Blennow, B. B. Bendlin, S. Asthana, S. C. Johnson, H. Zetterberg, and C. M. Carlsson, "Age-accelerated cognitive decline in asymptomatic adults with CSF β-amyloid," *Neurology*, vol. 90, pp. e1306–e1315, Apr. 2018.

[26] B. F. Darst, R. L. Koscik, A. M. Racine, J. M. Oh, R. A. Krause, C. M. Carlsson, H. Zetterberg, K. Blennow, B. T. Christian, B. B. Bendlin, O. C. Okonkwo, K. J. Hogan, B. P. Hermann, M. A. Sager, S. Asthana, S. C. Johnson, and C. D. Engelman, "Pathway-Specific polygenic risk scores as predictors of Amyloid-β deposition and cognitive function in a sample at increased risk for alzheimer's disease," *J. Alzheimers. Dis.*, vol. 55, no. 2, pp. 473–484, 2017.

[27] E. Shishkova, A. S. Hebert, M. S. Westphall, and J. J. Coon, "Ultra-High pressure (>30,000 psi) packing of capillary columns enhancing depth of shotgun proteomic analyses," *Anal. Chem.*, vol. 90, pp. 11503–11508, Oct. 2018.

[28] J. A. Stefely, N. W. Kwiecien, E. C. Freiberger, A. L. Richards, A. Jochem, M. J. P. Rush, A. Ulbrich, K. P. Robinson, P. D. Hutchins, M. T. Veling, X. Guo, Z. A. Kemmerer, K. J. Connors, E. A. Trujillo, J. Sokol, H. Marx, M. S. Westphall, A. S. Hebert, D. J. Pagliarini, and J. J. Coon, "Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling," *Nat. Biotechnol.*, vol. 34, pp. 1191–1197, Nov. 2016.

[29] K. Barkovits, A. Linden, S. Galozzi, L. Schilde, S. Pacharra, B. Mollenhauer, N. Stoepel, S. Steinbach, C. May, J. Uszkoreit, M. Eisenacher, and K. Marcus, "Characterization of

cerebrospinal fluid via Data-Independent acquisition mass spectrometry," *J. Proteome Res.*, vol. 17, pp. 3418–3430, Oct. 2018.

[30] I. A. Sindi, R. K. Tannenberg, and P. R. Dodd, "Role for the neurexin-neuroligin complex in alzheimer's disease," *Neurobiol. Aging*, vol. 35, pp. 746–756, Apr. 2014.

[31] M. Y. Lim, J. A. Paulo, and S. P. Gygi, "Evaluating false transfer rates from the Match-between-Runs algorithm with a Two-Proteome model," *J. Proteome Res.*, vol. 18, pp. 4020–4026, Nov. 2019.

[32] Z. Lin, Q. Gong, C. Wu, J. Yu, T. Lu, X. Pan, S. Lin, and X. Li, "Dynamic change of serum FGF21 levels in response to glucose challenge in human," *J. Clin. Endocrinol. Metab.*, vol. 97, pp. E1224–8, July 2012.

[33] Y. Yasui, M. Pepe, M. L. Thompson, B.-L. Adam, G. L. Wright, Jr, Y. Qu, J. D. Potter, M. Winget, M. Thornquist, and Z. Feng, "A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection," *Biostatistics*, vol. 4, pp. 449–463, July 2003.

[34] B. A. Trombetta, B. C. Carlyle, A. M. Koenig, L. M. Shaw, J. Q. Trojanowski, D. A. Wolk, J. J. Locascio, and S. E. Arnold, "The technical reliability and biotemporal stability of cerebrospinal fluid biomarkers for profiling multiple pathophysiologies in alzheimer's disease," *PLoS One*, vol. 13, p. e0193707, Mar. 2018.

[35] L. M. Schilde, S. Kösters, S. Steinbach, K. Schork, M. Eisenacher, S. Galozzi, M. Turewicz, K. Barkovits, B. Mollenhauer, K. Marcus, and C. May, "Protein variability in cerebrospinal fluid and its possible implications for neurological protein biomarker research," *PLoS One*, vol. 13, p. e0206478, Nov. 2018.

[36] J. Zhang, D. R. Goodlett, E. R. Peskind, J. F. Quinn, Y. Zhou, Q. Wang, C. Pan, E. Yi, J. Eng, R. H. Aebersold, and T. J. Montine, "Quantitative proteomic analysis of age-related changes in human cerebrospinal fluid," *Neurobiol. Aging*, vol. 26, pp. 207–227, Feb. 2005.

[37] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[38] S. R. Hanson, M. D. Best, and C.-H. Wong, "Sulfatases: structure, mechanism, biological activity, inhibition, and synthetic utility," *Angew. Chem. Int. Ed Engl.*, vol. 43, pp. 5736–5763, Nov. 2004.

[39] G. Parenti, G. Meroni, and A. Ballabio, "The sulfatase gene family," *Current opinion in genetics & development*, vol. 7, no. 3, pp. 386–391, 1997.

[40] A. Mok, H. Cao, and R. A. Hegele, "Genomic basis of mucopolysaccharidosis type IIID (MIM 252940) revealed by sequencing of GNS encoding n-acetylglucosamine-6-sulfatase," *Genomics*, vol. 81, pp. 1–5, Jan. 2003.

[41] T. Quintela, H. Marcelino, M. J. Deery, R. Feret, J. Howard, K. S. Lilley, T. Albuquerque, I. Gonçalves, A. C. Duarte, and C. R. A. Santos, "Sex-Related differences in rat choroid plexus and cerebrospinal fluid: A cDNA microarray and proteomic analysis," *J. Neuroendocrinol.*, vol. 28, Jan. 2016.

[42] T. Umahara, T. Uchihara, K. Tsuchiya, A. Nakamura, T. Iwamoto, K. Ikeda, and M. Takasaki, "14-3-3 proteins and zeta isoform containing neurofibrillary tangles in patients with alzheimer's disease," *Acta Neuropathol.*, vol. 108, pp. 279–286, Oct. 2004.

[43] Z. Xu, K. Graham, M. Foote, F. Liang, R. Rizkallah, M. Hurt, Y. Wang, Y. Wu, and Y. Zhou, "14-3-3 protein targets misfolded chaperone-associated proteins to aggresomes," *J. Cell Sci.*, vol. 126, pp. 4173–4186, Sept. 2013.

[44] K. Kaneko and N. S. Hachiya, "The alternative role of 14-3-3 zeta as a sweeper of misfolded proteins in disease conditions," *Med. Hypotheses*, vol. 67, pp. 169–171, Mar. 2006.

[45] C. Mackintosh, "Dynamic interactions between 14-3-3 proteins and phosphoproteins regulate diverse cellular processes," *Biochem. J*, vol. 381, pp. 329–342, July 2004.

[46] L. Strochlic, A. Cartaud, A. Mejat, R. Grailhe, L. Schaeffer, J.-P. Changeux, and J. Cartaud, "14-3-3 gamma associates with muscle specific kinase and regulates synaptic gene transcription at vertebrate neuromuscular synapse," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, pp. 18189–18194, Dec. 2004.

[47] O. Preische, S. A. Schultz, A. Apel, J. Kuhle, S. A. Kaeser, C. Barro, S. Gräber, E. Kuder-Buletta, C. LaFougere, C. Laske, J. Vöglein, J. Levin, C. L. Masters, R. Martins, P. R. Schofield, M. N. Rossor, N. R. Graff-Radford, S. Salloway, B. Ghetti, J. M. Ringman, J. M. Noble, J. Chhatwal, A. M. Goate, T. L. S. Benzinger, J. C. Morris, R. J. Bateman, G. Wang, A. M. Fagan, E. M. McDade, B. A. Gordon, M. Jucker, and Dominantly Inherited Alzheimer Network, "Serum neurofilament dynamics predicts neurodegeneration and clinical progression in presymptomatic alzheimer's disease," *Nat. Med.*, vol. 25, pp. 277–283, Feb. 2019.

[48] G.-H. Guibinga, S. Hsu, and T. Friedmann, "Deficiency of the housekeeping gene hypoxanthine-guanine phosphoribosyltransferase (HPRT) dysregulates neurogenesis," *Mol. Ther.*, vol. 18, pp. 54–62, Jan. 2010.

[49] Y. Liu, P. Zhang, Y. Zheng, C. Yang, T. Du, M. Ge, X. Chang, R. Duan, and G. Ma, "Effects of NMDAR antagonist on the regulation of P-MARCKS protein to Aβ oligomers induced neurotoxicity," *Neurochem. Res.*, vol. 43, pp. 2008–2015, Oct. 2018.

[50] R. K. McNamara, R. J. Hussain, E. J. Simon, D. J. Stumpo, P. J. Blackshear, T. Abel, and R. H. Lenox, "Effect of myristoylated alanine-rich C kinase substrate (MARCKS) overexpression on hippocampus-dependent learning and hippocampal synaptic plasticity in MARCKS transgenic mice," *Hippocampus*, vol. 15, no. 5, pp. 675–683, 2005.

[51] G. Sathe, C. H. Na, S. Renuse, A. K. Madugundu, M. Albert, A. Moghekar, and A. Pandey, "Quantitative proteomic profiling of cerebrospinal fluid to identify candi-

date biomarkers for alzheimer's disease," *Proteomics Clin. Appl.*, vol. 13, p. e1800105, July 2019.

[52] M. Matloubian, A. David, S. Engel, J. E. Ryan, and J. G. Cyster, "A transmembrane CXC chemokine is a ligand for HIV-coreceptor bonzo," *Nat. Immunol.*, vol. 1, pp. 298–304, Oct. 2000.

[53] A. Lleó, R. Núñez-Llaves, D. Alcolea, C. Chiva, D. Balateu-Paños, M. Colom-Cadena, G. Gomez-Giro, L. Muñoz, M. Querol-Vilaseca, J. Pegueroles, L. Rami, A. Lladó, J. L. Molinuevo, M. Tainta, J. Clarimón, T. Spires-Jones, R. Blesa, J. Fortea, P. Martínez-Lage, R. Sánchez-Valle, E. Sabidó, À. Bayés, and O. Belbin, "Changes in synaptic proteins precede neurodegeneration markers in preclinical alzheimer's disease cerebrospinal fluid," *Mol. Cell. Proteomics*, vol. 18, pp. 546–560, Mar. 2019.

[54] A. Wilbanks, S. C. Zondlo, K. Murphy, S. Mak, D. Soler, P. Langdon, D. P. Andrew, L. Wu, and M. Briskin, "Expression cloning of the STRL33/BONZO/TYMSTRligand reveals elements of CC, CXC, and CX3C chemokines," *J. Immunol.*, vol. 166, pp. 5145–5154, Apr. 2001.

[55] T. Skillbäck, N. Mattsson, K. Hansson, E. Mirgorodskaya, R. Dahlén, W. van der Flier, P. Scheltens, F. Duits, O. Hansson, C. Teunissen, K. Blennow, H. Zetterberg, and J. Gobom, "A novel quantification-driven proteomic strategy identifies an endogenous peptide of pleiotrophin as a new biomarker of alzheimer's disease," *Sci. Rep.*, vol. 7, p. 13333, Oct. 2017.

[56] C. Liguori, A. Stefani, G. Sancesario, G. M. Sancesario, M. G. Marciani, and M. Pierantozzi, "CSF lactate levels, τ proteins, cognitive decline: a dynamic relationship in alzheimer's disease," *J. Neurol. Neurosurg. Psychiatry*, vol. 86, pp. 655–659, June 2015.

[57] M. Bélanger, I. Allaman, and P. J. Magistretti, "Brain energy metabolism: focus on astrocyte-neuron metabolic cooperation," *Cell Metab.*, vol. 14, pp. 724–738, Dec. 2011.

[58] L. Higginbotham, L. Ping, E. B. Dammer, D. M. Duong, M. Zhou, M. Gearing, C. Hurst, J. D. Glass, S. A. Factor, E. C. Johnson, *et al.*, "Integrated proteomics reveals brain-based cerebrospinal fluid biomarkers in asymptomatic and symptomatic alzheimer's disease," *Science advances*, vol. 6, no. 43, p. eaaz9360, 2020.

[59] M. P. Lun, E. S. Monuki, and M. K. Lehtinen, "Development and functions of the choroid plexus-cerebrospinal fluid system," *Nat. Rev. Neurosci.*, vol. 16, pp. 445–457, Aug. 2015.

[60] M. Uzan, H. Erman, T. Tanriverdi, G. Z. Sanus, A. Kafadar, and H. Uzun, "Evaluation of apoptosis in cerebrospinal fluid of patients with severe head injury," *Acta Neurochir.*, vol. 148, pp. 1157–64; discussion, Nov. 2006.

[61] Z. Nemes, L. Fésüs, A. Egerházi, A. Keszthelyi, and I. M. Degrell, "N(epsilon)(gamma-glutamyl)lysine in cerebrospinal fluid marks alzheimer type and vascular dementia," *Neurobiol. Aging*, vol. 22, pp. 403–406, May 2001.

[62] I. Vermes, E. N. H. J. Steur, G. F. Jirikowski, and C. Haanen, "Elevated concentration of cerebrospinal fluid tissue transglutaminase in parkinson's disease indicating apoptosis," *Mov. Disord.*, vol. 19, pp. 1252–1254, Oct. 2004.

[63] K. T. Hansson, T. Skillbäck, E. Pernevik, S. Kern, E. Portelius, K. Höglund, G. Brinkmalm, J. Holmén-Larsson, K. Blennow, H. Zetterberg, and J. Gobom, "Expanding the cerebrospinal fluid endopeptidome," *Proteomics*, vol. 17, Mar. 2017.

[64] K. Hansson, R. Dahlén, O. Hansson, E. Pernevik, R. Paterson, J. M. Schott, N. Magdalinou, H. Zetterberg, K. Blennow, and J. Gobom, "Use of the tau protein-to-peptide ratio in csf to improve diagnostic classification of alzheimer's disease," *Clinical Mass Spectrometry*, vol. 14, pp. 74–82, 2019.

[65] A. C. Peterson, J. D. Russell, D. J. Bailey, M. S. Westphall, and J. J. Coon, "Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics," *Mol. Cell. Proteomics*, vol. 11, pp. 1475–1488, Nov. 2012.

[66] P. Maheshwari and G. D. Eslick, "Bacterial infection and alzheimer's disease: a meta-analysis," *J. Alzheimers. Dis.*, vol. 43, no. 3, pp. 957–966, 2015.

[67] J. Miklossy, *Handbook of Infection and Alzheimer's Disease*. IOS Press, Mar. 2017.

[68] S. S. Dominy, C. Lynch, F. Ermini, M. Benedyk, A. Marczyk, A. Konradi, M. Nguyen, U. Haditsch, D. Raha, C. Griffin, L. J. Holsinger, S. Arastu-Kapur, S. Kaba, A. Lee, M. I. Ryder, B. Potempa, P. Mydel, A. Hellvard, K. Adamowicz, H. Hasturk, G. D.

Walker, E. C. Reynolds, R. L. M. Faull, M. A. Curtis, M. Dragunow, and J. Potempa, "in alzheimer's disease brains: Evidence for disease causation and treatment with small-molecule inhibitors," *Sci Adv*, vol. 5, p. eaau3333, Jan. 2019.

[69] N. Ishida, Y. Ishihara, K. Ishida, H. Tada, Y. Funaki-Kato, M. Hagiwara, T. Ferdous, M. Abdullah, A. Mitani, M. Michikawa, *et al.*, "Periodontitis induced by bacterial infection exacerbates features of alzheimer's disease in transgenic mice," *NPJ aging and mechanisms of disease*, vol. 3, no. 1, pp. 1–7, 2017.

[70] S. Reitamo, M. Klockars, M. Adinolfi, and E. F. Osserman, "Human lysozyme (origin and distribution in health and disease)," *Ric. Clin. Lab.*, vol. 8, pp. 211–231, Oct. 1978.

Chapter 4

INTEGRATED PROTEOMIC AND METABOLOMIC PROFILING OF STRESS EVENTS ASSOCIATED WITH MILITARY EXERCISES

JM designed and conducted proteomics experiments, analyzed proteomics data and drafted manuscript. JM also developed figures and performed data integration between proteomics and metabolomics along with discriminatory models for both datasets.

This chapter is adapted from a manuscript in preparation

McKetney J, Jenkins CC, Mach PM, Glaros TG, Hussey EK, Coon JJ, Dhummakupt ES. *Proteomic and Metabolomic Profiling of Acute and Chronic Stress Events Associated with Military Exercises.* **2021**.

Abstract

A multiomic analysis of warfighter stress response helps to understand the performance impacts that specifically affect the men and women of the Armed Forces in a mission environment. Biological system wide fluctuations were captured by metabolomic and proteomic techniques at discrete time points of high cognitive and physical load, and analysis was performed to understand how the perturbations from basal state impact the overall performance and health of the warfighter. It was observed in this study that the warfighter experienced perturbations in the immune, metabolic, and protein manufacturing and processing systems during stressful events. This observed shift can lead to a greater understanding of proper treatment and supplementation to enhance warfighter performance.

Introduction

Human stress response serves an important evolutionary purpose in preparing and mobilizing physiological systems to react to altered homeostasis or a perceived threat[1,2]. When these stressor events occur, sympathetic stimulation of the autonomic nervous system releases adrenaline and noradrenaline[1,3]. These signaling molecules trigger many high stimulation functions associated with acute stress, like vasoconstriction, rapid glucose breakdown and elevated heart rate. A slower, secondary response stems from the stimulation of the hypothalamic-pituitary-adrenal (HPA) axis. This stimulation releases glucocorti-

coids, such as cortisol, in humans. Glucocorticoid receptors are present on a variety of tissues including brain, heart and liver, which regulate metabolism, energy availability, inflammation and cognitive function[4].

Human beings experience a variety of stressors in their daily lives. They fluctuate in type, timing and severity and have biological effects ranging from alterations in homeostasis to life-threatening conditions[5]. Individuals in particular occupations, such as warfighters, medics, airline pilots, and athletes encounter what are deemed "high-stress" events on a regular basis[6–11]. Identification of when and how acute and chronic stress impact these individuals and the resulting change in homeostasis is worthy of investigation. Stress can inhibit optimal performance and negatively impact physiology and health; especially as chronic stress can cause the development of long-term issues such as post-traumatic stress disorder[12–17], metabolic syndromes, cardiovascular disease, and immune dysfunction[18–20].

Members of the Armed Forces encounter multiple types of stress (i.e. acute, chronic) depending on particular situations (i.e. skirmishes, infantry tactics, and training)[8]. Understanding how these different stressors affect warfighter performance and their long-term health outcomes are top priorities for sustaining forces and ensuring health post-service. This requires not only identification of biomarkers for these conditions but also a deeper knowledge of how cellular systems are altered when experiencing stressors. Previous work has identified peptides[21], proteins[22] and small molecules[23] associated with physical and mental fatigue in a variety of contexts, from physicians to athletes[24]. Insight into the molecular basis of stress and fatigue can play a vital role in strategic decisions regarding

warfighters in the field, as well as potential preventative action that can reduce the negative health consequences associated with high-stress events.

Saliva presents an accessible biofluid that can be obtained in a non-invasive manner[25,26] without specialized training or clinical setting. In addition to ease of collection, the risk of infection to sampling personnel is much lower than blood or other biofluids[27,28]. The molecular composition of saliva is highly dynamic and impacted by both physical and psychological factors including age, circadian rhythm, pain level and stress[27,28]. The dynamic composition of saliva along with its ease of collection make it an ideal diagnostic and investigatory tool when individual variability can be accounted for. Proteomic and metabolomic analysis of saliva have identified biomarkers for cancer, multiple sclerosis, diabetes and heart disease[29–31]. Additionally, salivary compounds associated with fetal development[32], hypoxic conditions[33], and stress[33] have also been identified.

For this study, the researchers had the unique opportunity to analyze saliva collected from members of the 82nd Airborne's 2nd Battalion, 505th Parachute Infantry Regiment of the U.S. Army. The saliva was collected at discrete time points (**Figure 4.1**) before, during, and after a typical 72-hour field exercise in which the warfighters experience multiple instances of acute stress, as well as the chronic stress of the general exercise. From these salivary samples, proteomic and metabolomic analyses were performed to elucidate biomarkers and biochemical pathways that are most affected from these two different categories of stress. Repeated measurements of the same individual allowed for control of individual variability. While a number of bioindicators such as dopamine, cortisol, and

serotonin have been shown to be perturbed during stress[34], here we identify a set of proteins and small molecules dynamically changing in conjunction with stress events.

Materials and Methods

Saliva Collection Salivary samples were collected from 30 participants at discrete times during a training event. Up to 5 mL of saliva was collected by passive drool and frozen at -80 °C until analysis. Participants refrained from eating, drinking, and nicotine consumption thirty minutes prior to each collection. The collection protocol was approved for human subjects' research by the local designated Combat Capabilities Development Command Review Board (IRB). In accordance with the Declaration of Helsinki, all participants provided written informed consent before the completion of any study procedures.

Metabolomics - Sample Preparation Samples were prepared according to Mcbride et al.[26]. Immediately prior to analysis by liquid chromatography mass spectrometry (LC-MS), sample was reconstituted in 100 µl of Optima grade water with 0.1% formic acid (Fisher Scientific, LS118) and vortexed briefly. Samples were placed at 4 °C for 10 minutes to allow for resuspension. Finally, each sample is centrifuged at 20,000 g for 10 minutes and transferred to glass autosampler vials (Agilent Technologies, Santa Clara, CA) for analysis.

Metabolomics - Data Acquisition Each sample was analyzed on a Thermo Fisher (Waltham, MA) Orbitrap Q Exactive Plus mass spectrometer coupled to a Thermo Fisher Ultimate 3000 uHPLC system. Injections of each sample (2 µl) were resolved with the analytical pump

Figure 4.1: Timeline of sample collection from field study. Collection time points for salivary samples during field study. Each rectangle represents one day, and the collections are broadly categorized into three time categories: blue – pre-mission, red – mission, yellow – post-mission.

(350 µl/min) on a 100 µm x 2.1 mm id ACE Excel 1.7 um C18-PFP (Mac-Mod Analytical, Chadds Ford, PA) using a 22.5 min flow gradient. The A buffer was Optima grade water with 0.1% formic acid and the B buffer was 100% Optima grade acetonitrile. For the first 3 minutes, mobile phase A was held at 100%. Mobile phase B was increased to 80% from minute 3 to 13 and then held for 2 minutes. The method finished with an equilibration step made up of a 4 minute gradient back to initial conditions with a 2.5 minute hold at 0% B. MS1 scans were acquired with a resolution of 70,000 with a scan range of m/z 70–1000. AGC target was set to 3E6 with a maximum injection time of 100 ms. All metabolomics data were acquired in positive ionization mode using the heated electrospray source ionization (HESI). The source settings were as follows: spray voltage: 63.7 kV, capillary temperature: 325°C, sheath gas (N2): 30 arbitrary units (AU), auxiliary gas (N2): 10 AU, and probe heater: 350°C.

Metabolomics - Data Analysis Data was searched using Compound Discoverer (CD) 3.1 (Thermo Fisher Scientific) using the workflow detailed in Supplementary Figure 5. The quantitation was normalized to the total intensity (Constant Sum) and MS1 based identifications were performed with an in-house generated library, Chemspider (searching the human metabolome database) and Metabolika modules. Secondary analysis of results were performed by in-house python scripts to clean and process the CD output. The Web hosted version of Metaboanalyst 4.0 (`https://www.metaboanalyst.ca/MetaboAnalyst/home.xhtml`) was utilized for pathway enrichment analysis of small molecules.

Proteomics - Sample Preparation Saliva was aliquoted into 96 well plates and dried down in a speed-vacuum concentrator before being resuspended in 4 M guanidine hydrochloride (GnHCl), 100 mM Tris, pH 8. Protein concentration was determined via protein BCA (Pierce, Thermo Fisher). Resuspended samples were incubated at 100 °C for 6 minutes three times with 6 minutes of rest at room temperature between incubations. Sample wells were then brought to 90% methanol before centrifugation at 4,000 rpm for 40 minutes. Supernatant was disposed of.

Protein precipitate was resuspended in reducing and alkylating buffer (10 mM TCEP, 40 mM CAA, 8 M urea, 100 mM Tris, pH 8) to a total protein concentration between 2-4 mg/ml. Endoproteinase Lys-C was added to each well at an approximate ratio of 50:1 w/w total protein/protease. Plates were incubated at room temperature for 4 hours with slow rocking. Reducing and alkylating buffer was diluted to 25% concentration with 100 mM Tris, pH 8. Trypsin was added to sample wells at an approximate ratio of 50:1 w/w total protein/protease. Samples were incubated overnight at room temperature.

Digestion reactions were quenched when samples were brought to 0.5% TFA. Digested peptides were desalted using Phenomenex Strata C18 96-well plates (8E-S001-BGB) following manufacturers' instructions before being dried in a speed-vacuum concentrator (Thermo Scientific).

Online reverse-phase columns were prepared in-house using a high-pressure packing apparatus previously described[35]. In brief, 1.5 μm Bridged Ethylene Hybrid C18 particles were packed at 30,000 psi into a New Objective PicoTipTM emitter (Stock# PF360-75-10-N-5)

with an inner diameter of 75 μm and an outer diameter of 360 μm. During separations, the column was heated to a temperature of 50° C inside an in-house heater and interfaced with the mass spectrometer via an embedded emitter.

Proteomics - Data Acquisition An UltiMate 3000 RSLCnanoSystem (Thermo Fisher Scientific) was used for online chromatography with mobile phase buffer A consisting of 0.2% formic acid in water and mobile phase buffer B consisting of Optima grade water with 0.2% formic acid in 70% Optima grade acetonitrile. Samples were loaded onto the column for 4 minutes at 300 nL/min. Mobile phase B was increased to 9% in the first 4 minutes then increased to 52% by 59 minutes. The method finished with a wash stage of 100% B from 60-69 minutes and an equilibration step of 0% B from 70-80 minutes.

Eluted peptides were ionized by electrospray ionization and analyzed on a Thermo Orbitrap Eclipse. Survey scans of precursors were taken from 300 to 1400 m/z at 240,000 resolution while using Advanced Precursor Determination[36] with an AGC target of 1E6 and a maximum injection time of 50 ms. Tandem MS was performed using an isolation window of 0.5 Da with 20 ppm mass tolerance and a dynamic exclusion time of 10 s. Selected precursors were fragmented using HCD with a normalized collision energy of 27%. The MS2 AGC target was set at 3E4 with a maximum injection time of 20 ms. Scans were taken in the ion trap using the turbo setting, and only peptides with a charge state of +2 or greater were selected for fragmentation. Samples were analyzed in duplicate.

Proteomics - Data Searching The resulting spectra were searched in MaxQuant (1.6.0.13) using fast LFQ against a full human proteome with isoforms downloaded from Uniprot (October 29, 2019). Carbamidomethylation of cysteine was set as fixed modification. Matching between runs was used with a retention time window of 0.7 min. Searches were performed using a protein FDR of 1%, a minimum peptide length of 7, and a 0.5 Da MS2 match tolerance. Protein data were then extracted from the "ProteinGroups.txt" file of the Maxquant output after decoy, contaminants, and reverse sequences were removed. The protein counts were based on protein groups with an LFQ Intensity > 0.

Post Processing Post processing was performed primarily in R (version 3.6.3). Protein gene ontologies were retrieved from the NCBI DAVID database with significance established using the Fisher exact test function in R or the built-in function of the NCBI website. Protein LFQ intensities were normalized by log2 transformation. Paired significance testing and Pearson correlations were performed using base R. Partial least squares discriminant analysis (PLS-DA) was performed using the plsda function from the mixOmics[37] package (version 6.10.9) in R. Linear discriminant analyses were performed using the lda function of the MASS package[38](7.3.51.5), with area under the curve(AUC) of the receiver operator characteristic calculated using the auc or the multiclass.auc function of the pROC package[39,40](1.16.2). Mixed-effect models were generated using the lmer function of the lme4 package[41](1.1.23). Fixed-effect models were generated using lm function in base R. Circular dendrogram was generated using the circlize (0.4.11) and dendextend (1.14.0)

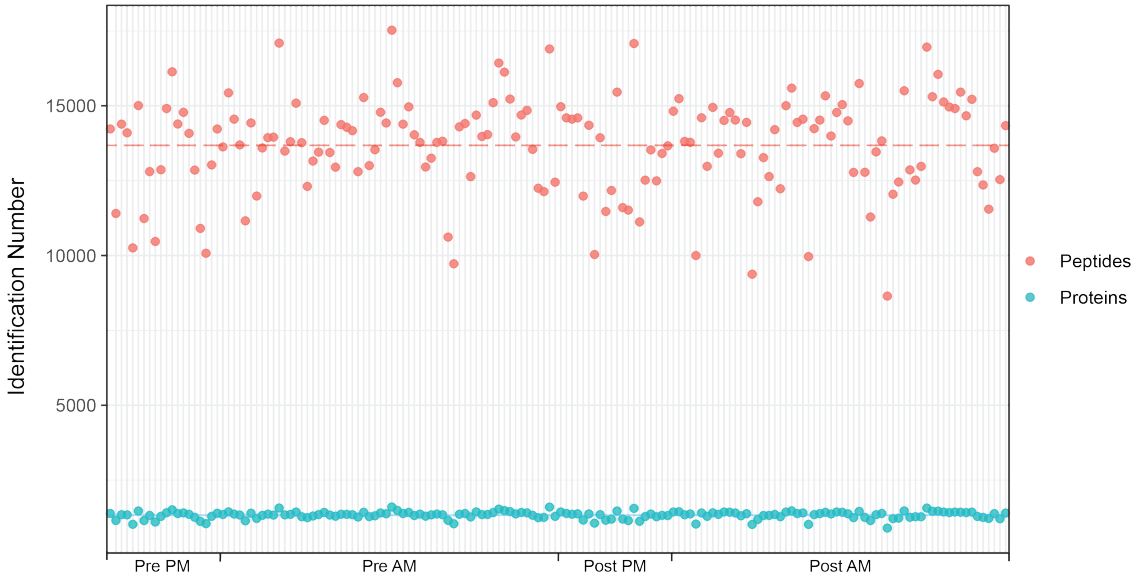
packages in R. Other plots generated using base R and ggplot2 (3.3.2).

Results

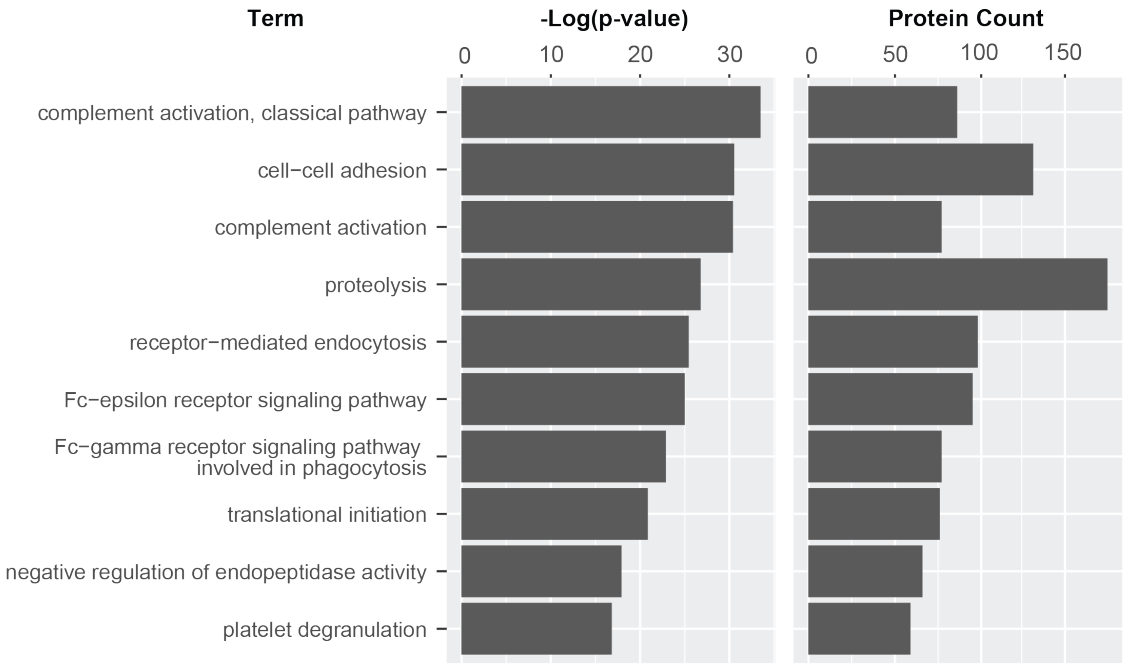
Molecule Identification and Differential Abundance Proteomics was performed on saliva samples from 20 individuals across eight different time points for a total of 160 samples. The time points included three morning and one evening collection before the mission began (Pre-Mission AM1/AM3/AM4/PM), and one evening and three morning collections after the mission initiation (Mission PM, Post-Mission AM1/AM3/AM4). Given this structure, the proteomic samples were organized into four pre-post mission initiation pairs to assess the proteomic effects of the mission-derived stress. Overall, 2087 proteins were identified based on 28,511 peptides (**Supplemental Figure 4.1**). On average, 1,332 proteins and 13,680 peptides were quantified in each sample. Although we observe high variability in our peptide identifications, outliers do not appear systematically related to other variables and the protein identifications remain relatively consistent. Our analysis quantified 545 protein groups across all 160 samples, with 720 proteins consistently quantified in >50% of pre-post mission initiation pairs. When examining the gene ontology enrichment using a full human proteome as background, we see the highest enrichment for the biological processes proteolysis, cell-cell adhesion and complement activation, (**Supplemental Figure 4.2**), which would be expected given the role of saliva as a host-microbe interface and its proximity to the epithelial cells of the mouth. We also observe enrichment of proteins involved in protein turnover such as translation initiation and endopeptidase activity. The

720 proteins consistently quantified here represents 20% of the known human salivary proteome[42]. Metabolomic analysis was also performed on 38 samples from each of these same 20 individuals, with nine morning and four evening collections before the mission began, and 20 morning and five evening collections after mission initiation. More than 5,000 compound features were identified overall with more than 4,000 quantified across all samples analyzed by metabolomics. A total of 93 samples were analyzed by both proteomics and metabolomics.

We first investigated the proteomic effect of the simulated combat mission by examining the pair fold change between pre- and post-mission initiation samples for each soldier at each sampling time of day (AM1/AM3/AM4/PM). Overall, proteins tended to show higher abundances post-mission initiation. When performing a paired t-test accounting for unequal variance and multiple hypotheses (Benjamini Hochberg, 5% FDR) for all proteins identified in at least half of all pairs (720), we identified 302 significant proteins (**Figure 4.2A**). This group was highly enriched in proteins associated with functions commonly found in saliva such as the complement and coagulation cascade, and protein processing in the endoplasmic reticulum (ER), as well as containing more specific pathways, such as stress-activated MAPK cascade. Interestingly, the only protein significantly upregulated in the stress-activated MAPK cascade was angiotensinogen, which helps to regulate osmotic balance in the vascular system[43], where changes have been observed previously in cases of stress[2,44]. We observe a consistent increase in abundance for proteins associated with the complement and coagulation system upon mission initiation (**Figure 4.2A**), while ER



Supplementary Figure 4.1: Protein and Peptide Identifications. Number of proteins and peptides identified in each sample with line indicating the mean across all 160 samples. Protein and peptide identifications do not appear systematically linked to mission or time of day.



Supplementary Figure 4.2: Gene Ontology Enrichment. Count and significance of 10 most significant biological processes for all proteins identified in any sample. Significance is calculated using a fisher exact test with full human proteome as background.

processing proteins show varying directions of change, suggesting aspects of the pathway are responding differently to stress.

We performed a pairwise correlation of expression profiles for all 302 proteins significantly associated with mission initiation, from which we identify four well-defined clusters (**Figure 4.2B**). Two of these clusters include proteins that modify structure via the cytoskeleton and extracellular matrix. The other two clusters continue the theme of innate immunity and protein processing. Interestingly, the same cluster containing protein chaperones involved in protein folding also contains several proteins for processing reactive oxygen species, compounds that are typically detrimental to the structure of proteins and known to be generated from consistent physical stress[45] as well as ER stress[46]. When examining the complement and coagulation cluster more closely (**Figure 4.2B, bottom right**), we observe more than 10 complement proteins, including multiple subcomponents of the C1 initiator and the alpha, beta and gamma chain of complement component C8, which forms the membrane attack complex. A comparison of normalized expression across all 160 samples of complement C2, complement C1r, complement 4b, complement C6, complement factor B and complement factor I, reveals highly similar profiles of these proteins reflecting the tight regulation of the innate immune system at this host-pathogen interface (**Figure 4.2C**).

Our analysis attempts to account for several different factors contributing to molecular abundances in saliva including time of day, physical and mental stress, and individual variability. Understanding and controlling for each of these variables is important when at-

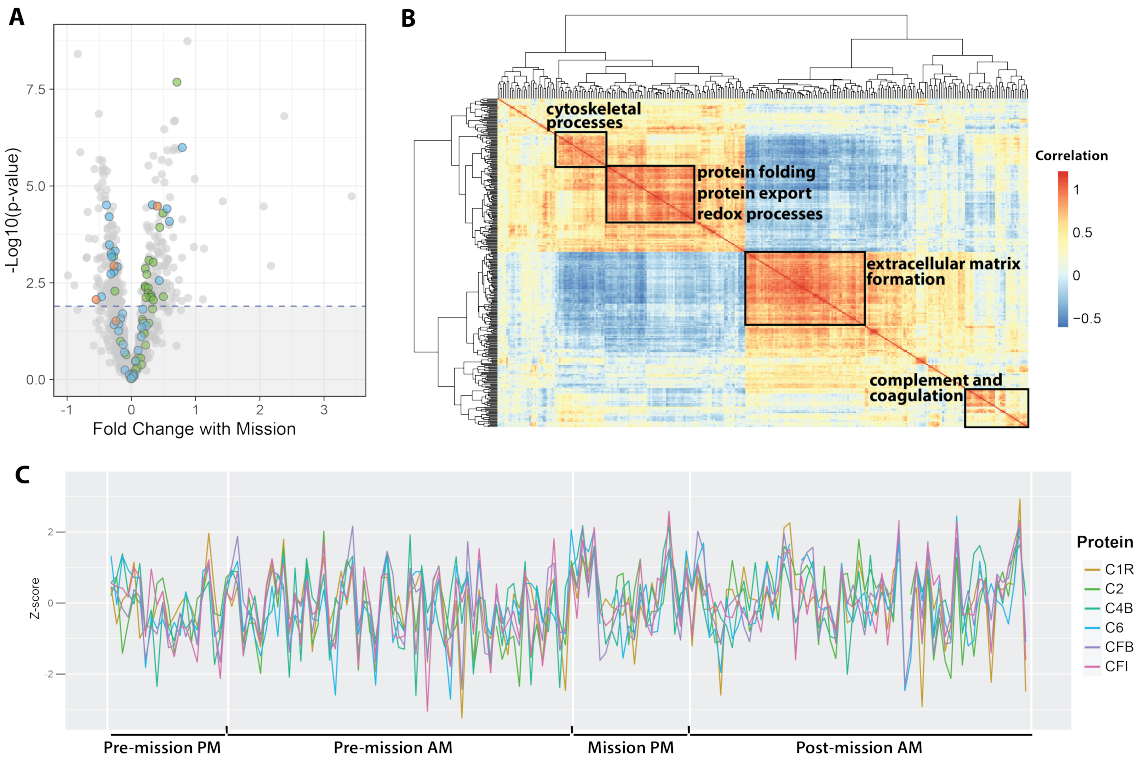
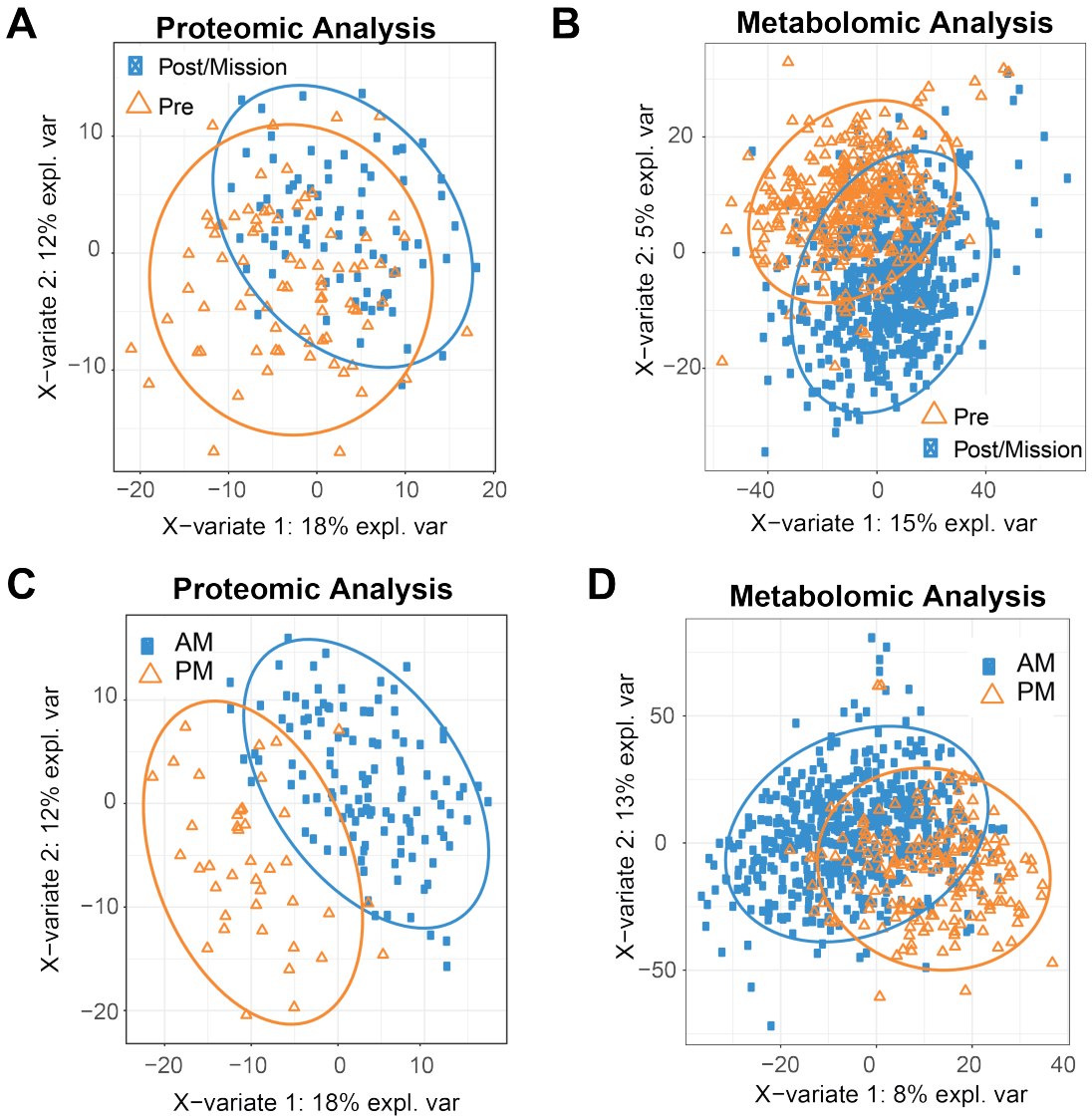


Figure 4.2: Differentially Expressed Proteins with Mission. (A) Volcano plot showing mean fold change and significance for proteins across mission initiation pairs. Dotted line indicates approximate significance cutoff for Benjamini-Hochberg correction with FDR of 5%. Proteins are colored by three KEGG pathways, with protein processing in the ER, complement and coagulation cascade, and stress-activated MAPK cascade depicted in blue, green and orange, respectively. (B) Protein-protein correlation heatmap for all 302 proteins identified as significant in A. Four strongly correlated clusters with their proteins' most common functions are highlighted. (C) Line plot showing normalized expression across all 160 samples of six complement proteins found to be significantly associated with mission: complement C2, complement C1r, complement 4b(C4B), complement C6, complement factor B(CFB) and complement factor I(CFI). All six proteins are present in the bottom right cluster of B.

tempting to quantify stress effects. When applying a partial least squares discriminant analysis (PLS-DA) to our proteomic data, a slight separation can be seen between pre-mission time points and those after mission initiation(**Supplementary Figure 4.3A**), while more substantial separation occurs between AM and PM collection events in(**Supplementary Figure 4.3C**). Metabolomic analyses show less discrepancy between these two variables with some separation based on both axes (**Supplementary Figure 4.3B,D**). Time of day has shown a well-documented effect on the salivary proteome[47] similar to other human biofluids[48], and circadian rhythm is known to control the pulsatile releases of cortisol[49], one of the primary hormonal signalers in stress[50]. The discriminatory power of our data based on time of day (morning vs. evening) may reflect a capacity to capture hormonal homeostatic systems, such as sleep-wake cycle and stress, both of which interact with the hypothalamus-pituitary-adrenal axis in the human brain[4,51].

Discriminatory Molecules for Mission In an effort to isolate these mission discriminant molecules, we performed a linear discriminant analysis for both our proteomic and metabolomic data. We calculated the area-under-the-curve for the receiver operator characteristic (ROC-AUC), utilizing the leave one out strategy. When performing an LDA using single proteins, periplakin (PPL), galactosidase beta 1(GLB1) complement component C9 (C9), ubiquitin fold modifier 1 (UFM1) and prelamin A/C (LMNA) provided the best performance, with AUCs of 0.701, 0.697, 0.683, 0.683, 0.674, respectively (**Figure 4.3A**). Among the metabolites, unidentified compound 2908, pipecolic acid, acetylspermidine,



Supplementary Figure 4.3: Partial Least Squares Discriminant Analysis. (A) Performed supervised clustering of proteomic samples across initiation of mission (Pre vs Post/Mission) based on decomposed values without filtering. (B) Performed supervised clustering of metabolomic samples across initiation of mission (Pre vs Post/Mission) based on decomposed values without filtering. (C) Performed supervised clustering of proteomic samples by time of day (AM vs PM) based on decomposed values without filtering. PLS-DA separates along time of day axis more effectively than along mission initiation axis. (D) Performed supervised clustering of metabolomic samples by time of day (AM vs PM) based on decomposed values without filtering. PLS-DA separates along time of day axis more effectively than along mission initiation axis.

proline and the dipeptide histidine-histidine(his-his) exhibited the greatest discriminatory performance for mission with areas under the curve of 0.835, 0.819, 0.808, 0.801 and 0.797, respectively(**Figure 4.3B**). We note that the discriminatory power of small molecules far outweighs that of even our best performing proteins, with substantially higher AUCs (**Figure 4.3**).

Fatigue likely manifests in a variety of cellular systems due to compounding instances of acute, as well as medium-term physical and mental stress events. Detection of these myriad effects requires the monitoring of multiple compounds and proteins. As such, we pursued an LDA relying on five compounds for both our proteomic and small molecule analysis. We observe a substantial increase in ROC-AUC when utilizing a combination of the top five components listed above for each dataset. However, the best performing five-component model was identified by iteratively testing combinations and optimizing area-under-the-curve, starting with periplakin or unidentified compound 2908, for proteins or small molecules, respectively. The optimal protein combination included periplakin (PPL), heat shock protein alpha family class B member 1 (HSPAB1), myeloperoxidase (MPO), heat shock protein family A member 9 (HSPA9), and transketolase (TKT) (**Figure 4.3A**). While the optimal small molecule combination consisted of compound 2908, γ-butyrobetaine, proline, and proline-glycine, and pipecolic acid.

Although increasing the complexity of a model by adding explanatory variables will always improve performance, we found it interesting that of the nine potentially discriminant proteins we identify, two are components of chaperones that participate in the unfolded

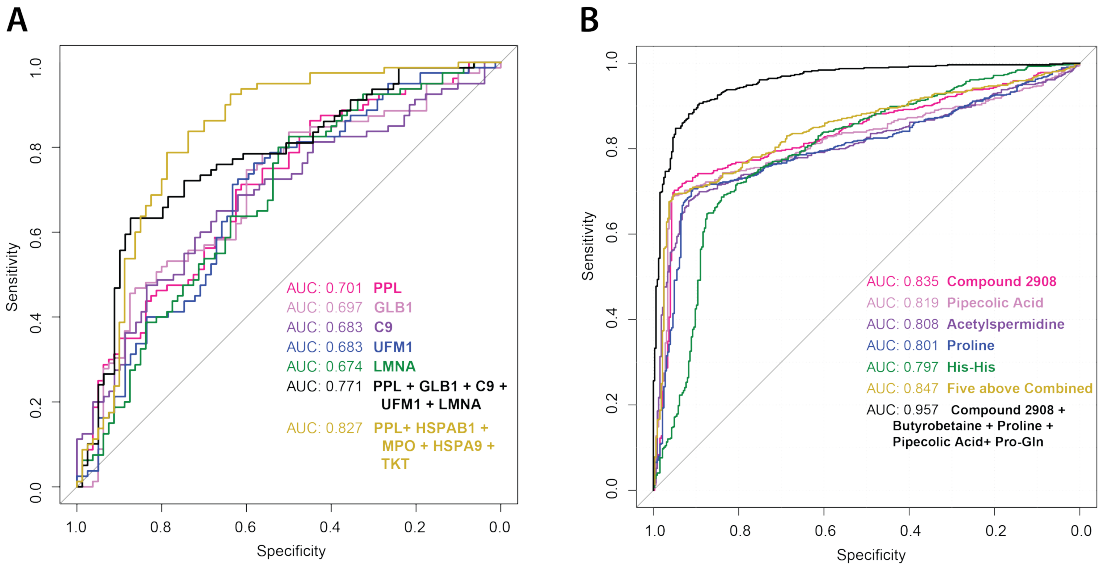
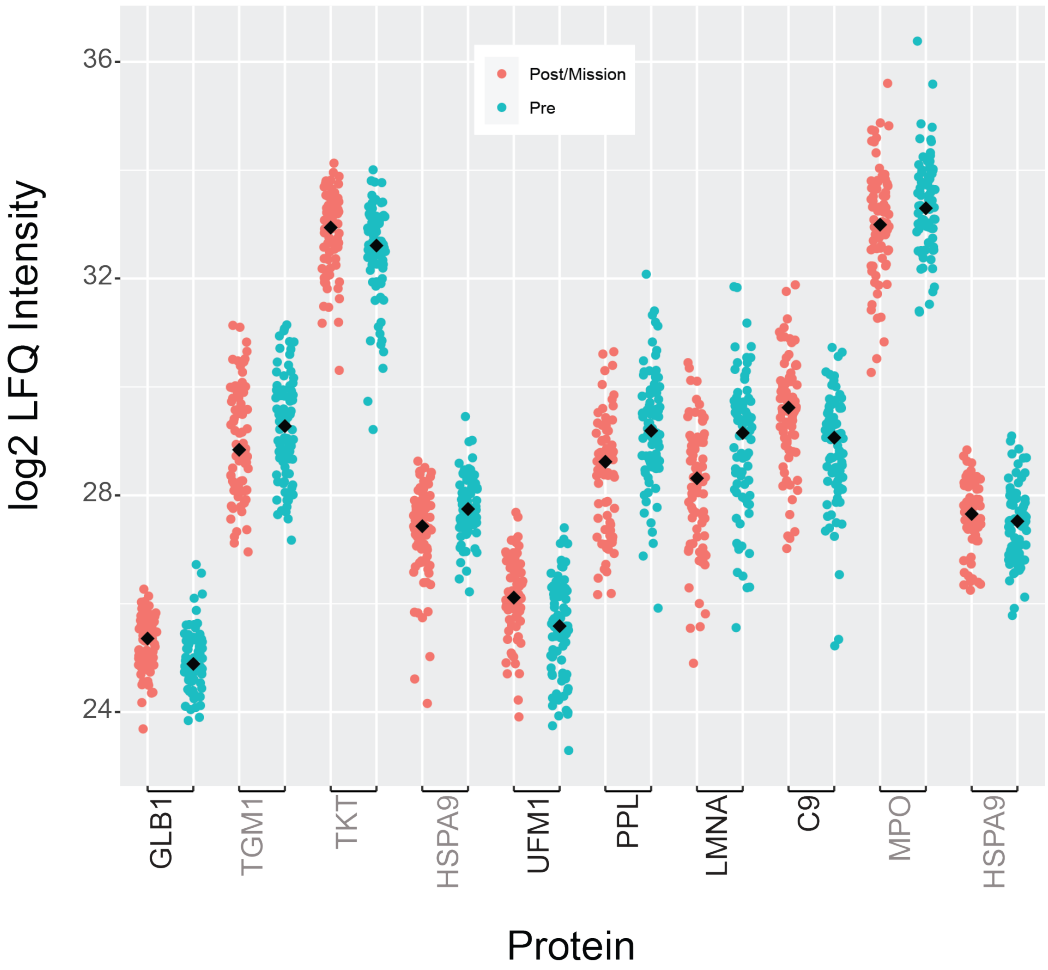


Figure 4.3: Linear Discriminant Analysis. (A) Receiver operator characteristic (ROC) curve for linear discriminant models generated from the top five most discriminant proteins based on area-under-the-curve (AUC) both as single component models and in combination. The model indicated in yellow was generated by iteratively identifying the most complimentary four additional proteins starting from periplakin (PPL). (B) ROC curve for LDA separating samples based on mission initiation (Pre vs. Post/Mission), excluding stress shoots. Single component models are shown for five best performing compound features based on AUC. Five component model from five best performing features is shown in yellow while an optimized five component model is shown in black.

protein response (HSPAB1 and HSPA9) and one mediates the ER stress response (UFM1). In the metabolomic model, γ-butyrobetaine, proline, pipecolic acid, and proline-glycine are all associated with amino acid metabolism. γ-butyrobetaine is involved in lysine degradation and a precursor to L-carnatine[52]. Pipecolic acid also plays a role in the metabolism of lysine. L-Proline is an amino acid itself, with proline-rich proteins playing an important signaling role in human saliva[53]. Proline-glycine is involved in protein catabolism but also plays roles in metabolism and coagulation[54–57]. The proteins identified here are of interest in understanding the biological effects of stress, but more targeted methods would be required for validating protein signatures as even the most discriminatory proteins exhibit somewhat minor shifts in abundance overall (**Supplementary Figure 4.4**).

Proteomics Metabolomics Integration In order to integrate our two -omics analyses, we subset the 93 samples for which both protein and small molecule data had been collected. Significance for each protein across time of day (AM vs. PM) and mission initiation (Pre vs Post/Mission) was established using a mixed effect model in which time of day and mission initiation were treated as fixed effects and individual soldier was treated as a random effect. Linear models were then constructed for each protein and comparison was made using analysis of variance (ANOVA) with and without mission initiation or time of day as explanatory variables. As suggested by our discriminant analysis, substantially more proteins were associated with time of day than mission (**Figure 4.4A**). Small molecule significance was established using analysis of variance on fixed effect linear models, as



Supplementary Figure 4.4: Abundance Shifts for Discriminant Proteins. Log2 transformed LFQ intensities of all samples for five proteins most discriminant for mission initiation[periplakin (PPL), galactosidase beta 1(GLB1) complement component C9 (C9), ubiquitin fold modifier 1 (UFM1) and prelamin A/C (LMNA)] along with proteins from the most discriminant five-protein model[heat shock protein alpha family class B member 1 (HSPAB1), myeloperoxidase (MPO), heat shock protein family A member 9 (HSPA9), and transketolase (TKT)]. Pre-mission samples are colored in blue, while mission and post-mission samples are colored in red. Black diamonds indicate median intensities for each group.

the complexity of the mixed effect model led to overfitting. These metabolomic linear models also focused on time of day and mission initiation. Compound features exhibited a more even distribution in significance between these two variables as compared to proteins (**Figure 4.4B**). Normalized metabolite and protein abundances were then correlated using the 93 overlapping samples. We identified a single large cluster of metabolites that exhibit strong correlations to two protein clusters: one positively correlated and the other negatively correlated (**Figure 4.4C**). When examining the gene ontology terms associated with these two protein clusters, we observe overrepresentation of proteins related to protein processing and degradation, the endoplasmic reticulum(ER), and amino acid metabolism (**Table 4.1 and Table 4.2**). In addition, the metabolite cluster contains more than 40 dipeptides, potentially resulting from protein degradation associated with the proteasome and the ER. The metabolite cluster also includes several isolated canonical amino acids and their precursors.

Our analysis quantified 21 proteins related to protein processing in the ER, with 14 proteins found to be significantly associated with mission initiation. From that significant subset of proteins, five increase with the initiation of the mission, while the other nine decrease with mission (**Figure 4.5A**). Proteins upregulated with mission exhibit a larger magnitude shift with protein OS9, nucleotide exchange factor SIL1(SIL1) and mannosyl-oligosaccharide 1,2-alpha-mannosidase IA (MAN1A1) most effected. MAN1A1 and OS9 both function in shuttling glycoproteins out of the endomembrane system, by maturation of their glycan chain or degradation of misfolded proteins, respectively[58,59]. Neutral alpha-

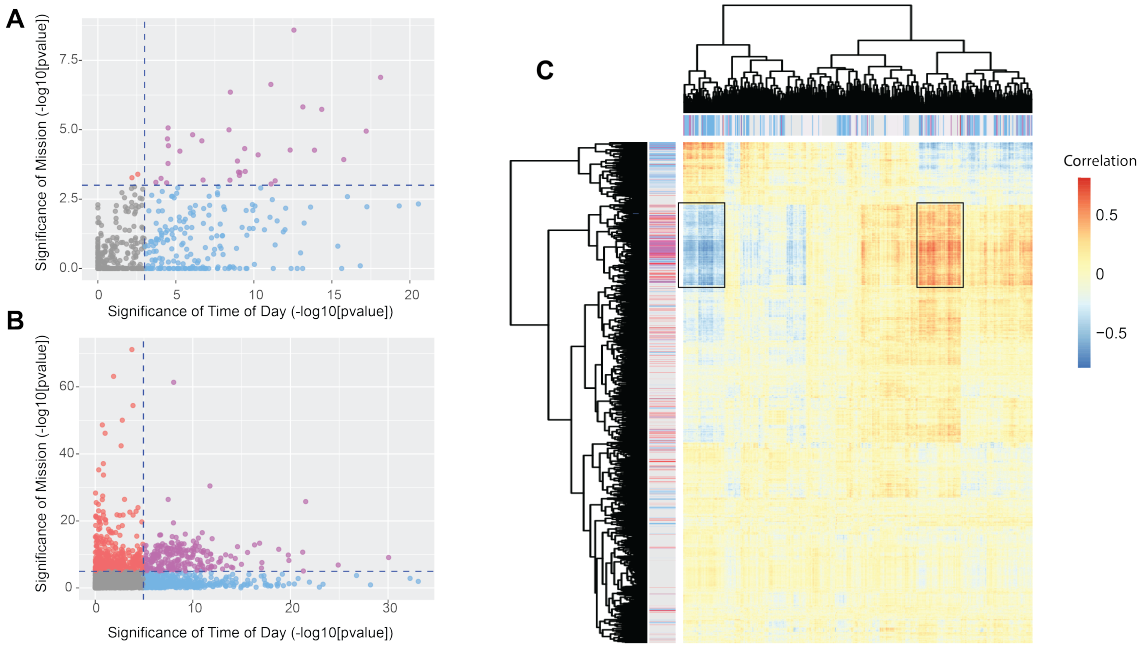


Figure 4.4: Correlation of Proteins, Compounds, mission and Time of Day. (A) Significance of time of day and mission initiation for proteins based on these variables' inclusion in regression models. Cutoff set at p-value < 0.001 (B) Significance of time of day and mission initiation for small molecules based on these variables' inclusion in regression models. Due to increased significance, cutoff for metabolites was set at p-value $< 10^{-5}$ (C) Heatmap showing correlation of proteins (x-axis) and compound features (y-axis) with axis annotations based on significance of time of day, mission or both. Significance across time of day, mission or both is indicated by blue, red and purple respectively. Observe single metabolite cluster and two protein clusters with substantial correlations.

GO term	Count	P-value
Negative regulation of cysteine-type endopeptidase activity	5	8.03e-6
Platelet degranulation	9	1.27e-4
O-glycan processing	4	7.48e-3
Retina homeostasis	4	1.23e-2
Cell activation	2	1.66e-2
IRE1-mediated unfolded protein response	3	1.73e-2

Table 4.1: Associated Processes for Proteins in Anticorrelated Cluster. Six most significant biological processes included in protein cluster negatively correlated with small molecule cluster.

GO Term	Count	P-value
Proteolysis involved in cellular protein catabolic processes	10	2.15e-7
Protein polyubiquitination	11	5.00e-7
MAPK cascade	12	1.40e-6
Regulation of mRNA stability	11	1.45e-6
Stimulatory C-type lectin receptor signaling pathway	10	2.19e-6
Proteasome	9	3.67e-6
T cell receptor signaling pathway	10	5.20e-6
Regulation of cellular amino acid metabolic processes	9	5.43e-6

Table 4.2: Associated Processes for Proteins in Positively Correlated Cluster. Eight most significant biological processes included in protein cluster positively correlated with small molecule cluster.

glucosidase AB (GANAB1) also functions in the maturation process of glycoproteins and increases with mission[60]. SIL1 functions as a cochaperone to ER chaperone BiP (HSPA5), allowing HSPA5 to exchange ADP for ATP. SIL1 has been shown to play a role in cellular sensing of ER stress, with aberrant SIL1 associated with accumulation of ubiquitinated proteins[61]. We observe a decrease in abundance of proteins such as PDIA6 and HSPA5 that can repress the unfolded protein response (UPR) along with increases in DNAJC3 which is activated by ER stress and the UPR[62–64]. Many of the proteins decreasing with mission function partially in the refolding of misfolded proteins including DNAJB1, CANX1, HSPA1B, HSPA8[65–67] although these changes are of a smaller magnitude.

Our analysis quantified 46 proteins and 120 metabolites related to amino acid biosynthesis or metabolism with 14 and 25 mission-significant proteins and metabolites, respectively (**Figure 4.5B**). When examining this significant protein subset, we noticed that many of the proteins had their primary function in cellular metabolism and ATP generation rather than acting directly on amino acids or amino acid precursors. Three key enzyme in glycolysis exhibited significantly increased abundance upon initiation of the mission, pyruvate kinase (PKM), phosphoglycerate kinase(PGK1) and glyceraldehyde-3-phosphate dehydrogenase (GPDH) as well as lactate dehydrogenase (LDHB), the enzymatic driver of anaerobic respiration. Although not reaching significance, we also observed elevated abundance with mission of seven other enzymes involved in glycolysis: hexokinase(HK1), phosphofructokinase(PFKL), fructose bisphosphate aldolase(ALDOA), triosephosphate isomerase(TPI), enolase(ENO1), and phosphoglycerate mutase(PGAM) (**Figure 4.5C**). These proteins also

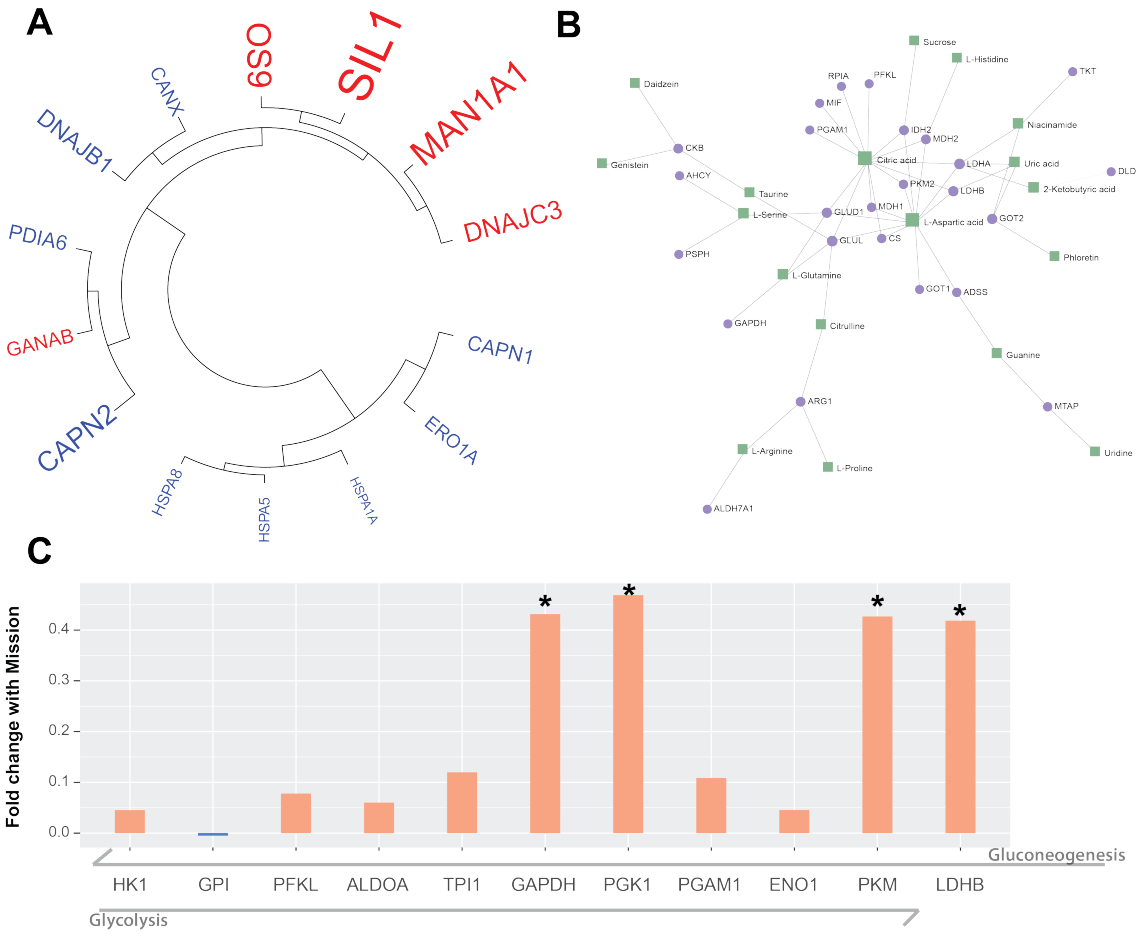


Figure 4.5: Altered Expression in Protein Processing and Metabolism. (A) Hierarchical clustering of proteins significantly affected by mission associated with protein processing in the ER. Gene name label size indicates magnitude of mean fold change with red proteins exhibiting increased expression with mission and blue proteins exhibiting decreased expression (B) Network plot showing proteins and small molecules involved in biosynthesis and metabolism of amino acids identified in our analysis. Metabolites and protein are indicated by green squares and purple circles, respectively (C) Bar plot showing fold change with mission of enzymes involved in glycolysis and gluconeogenesis. Proteins that increase upon mission initiation are colored in red while those that decrease are colored in blue. Asterisk indicates significance. Although only a subset meets our significance threshold, nearly all enzymes increase with mission.

play a key role in the reverse reaction of glycolysis, gluconeogenesis, which has been shown to be upregulated in situations of acute psychological stress[68,69], with elevated blood glucose occurring as one of the fastest reactions to stressors[2]. Increased expression of glycolysis proteins and those associated with anaerobic respiration could reflect cellular expression changes in response to continued energetic stress.

Detecting Acute Stress Events Using Metabolomics Given the high variability of small molecule abundances and the promising results from the associated linear models, it may be possible for metabolomics to detect acute stress events. In an effort to test that capacity, we expanded our metabolomic analysis in terms of both the number of individuals and the number of time events. Saliva samples from 30 individuals were collected at five time point events: pre-mission(Pre), stress-shoot pre-mission(SS1), during mission(Mission), post-mission(Post), stress-shoot post-mission(SS2). The stress shoot simulates both an acute physical and cognitive stress event, in which subjects must learn a series of signs and symbols before the activity that relate to how they must execute encountered tasks (i.e. recognizing friend versus foe, aim points, accuracy, position: standing, kneeling, or prone). These tasks were then performed while the subjects were timed.

Normalized peak areas for all compound features were averaged for each time point group with groups and compounds then clustered based on Euclidean distance. **Figure 4.6A** shows the distinct grouping of metabolites that occurs among compounds when comparing these five groups. We specifically observe increased abundance in several

compound clusters in the two stress shoot samples, while the other three time point groups cluster together, with the baseline event (pre-mission) located most distally (**Figure 4.6A**). Based on this preliminary result, 200 metabolites were determined to experience significant changes in abundance between the five time point groups, with p-values $\leqslant 0.05$. These compounds included small molecules involved in a variety of biological pathways largely related to metabolism of amino acids, continuing this theme. A Metaboanalyst[70] pathway enrichment analysis (**Figure 4.6B**) reveals enrichment occurring in 30 different pathways including arginine biosynthesis, histidine metabolism, beta-alanine and alanine, aspartate and glutamate metabolism.

Compound features from two significantly impacted pathways, arginine biosynthesis and histidine metabolism, were extracted and run through a similar clustering analysis. Compounds related to arginine biosynthesis cluster into two high level groups: one with large variability in mean abundance across the different time point events and one with minimal variability (**Figure 4.6C**). The high variance compounds drive separation of the two stress shoot samples from the other three time point groups (**Figure 4.6C**). For compounds related to histidine metabolism we see a much larger proportion of compound features with minimal variation in mean abundance across the different time point groups. These minimally altered compounds form one high level cluster while two other compound clusters are composed of features with greater variance. Although the two stress shoot time points remain the most distal, the time point events do not form any meaningful clusters based on these histidine metabolism compounds (**Figure 4.6D**).

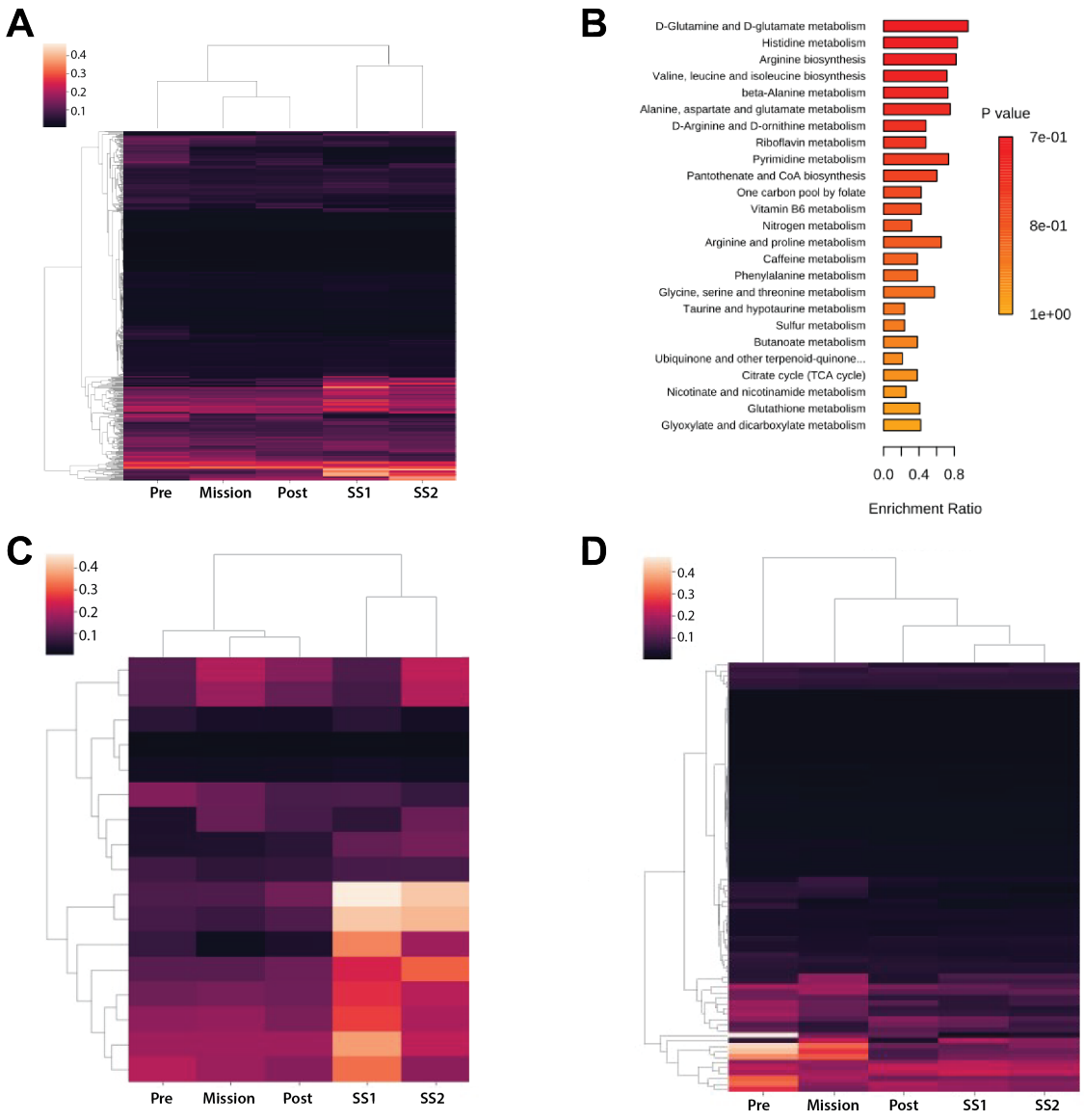


Figure 4.6: Differential Metabolites Across Five Time Point Events. (A) Hierarchical clustering of all compounds based on mean normalized peak areas of each of five time point event groups (Pre = pre-mission, Mission = during mission, Post = post-mission, SS1 = pre-mission stress shoot, SS2 = post-mission stress shoot). (B) Barplot showing significance and enrichment ratio for pathways associated with compounds found to be significantly different between the five time point groups. (C) Heatmap showing hierarchical clustering of compound features associated with arginine biosynthesis. (D) Heatmap showing hierarchical clustering of compounds features associated with histidine metabolism.

We performed a linear discriminant analysis with the objective of separating the five time point groups. Our LDA model experiences difficulty in distinguishing between the five event groups as compared to the two groups centered around mission initiation, especially when relying on a single discriminant compound, unidentified compound 2616. The model shows especially poor performance in identifying the "pre-mission" and "pre-mission stress shoot" groups (**Figure 4.7**). When relying on the best performing five-compound combination, these groups shift to some of the most accurate classifications, suggesting a compound feature that associates with the initiation of the mission. This shift is driven by the lysine metabolite, pipecolic acid which shows substantial increase in the pre-mission and pre-mission stress shoot time point samples (**Supplemental Figure 4.5**), continuing the theme of amino acid metabolism. In future studies we hope to integrate these more granular metabolomic analyses with proteomics or perhaps other quantitative metrics of stress such as heart rate and cortisol levels.

Discussion

We present here a multi-omics analysis of raw saliva focusing on identifying the molecular and cellular components of physical and mental stress. We observe significant abundance changes in molecules related to three cellular systems: (1)innate immunity, (2)protein cycling and processing, and (3) metabolism of sugars and amino acids. We identify increased expression of proteins involved in the complement system upon initiation of the mission, indicating immune activation and inflammation. We observe altered abundance of

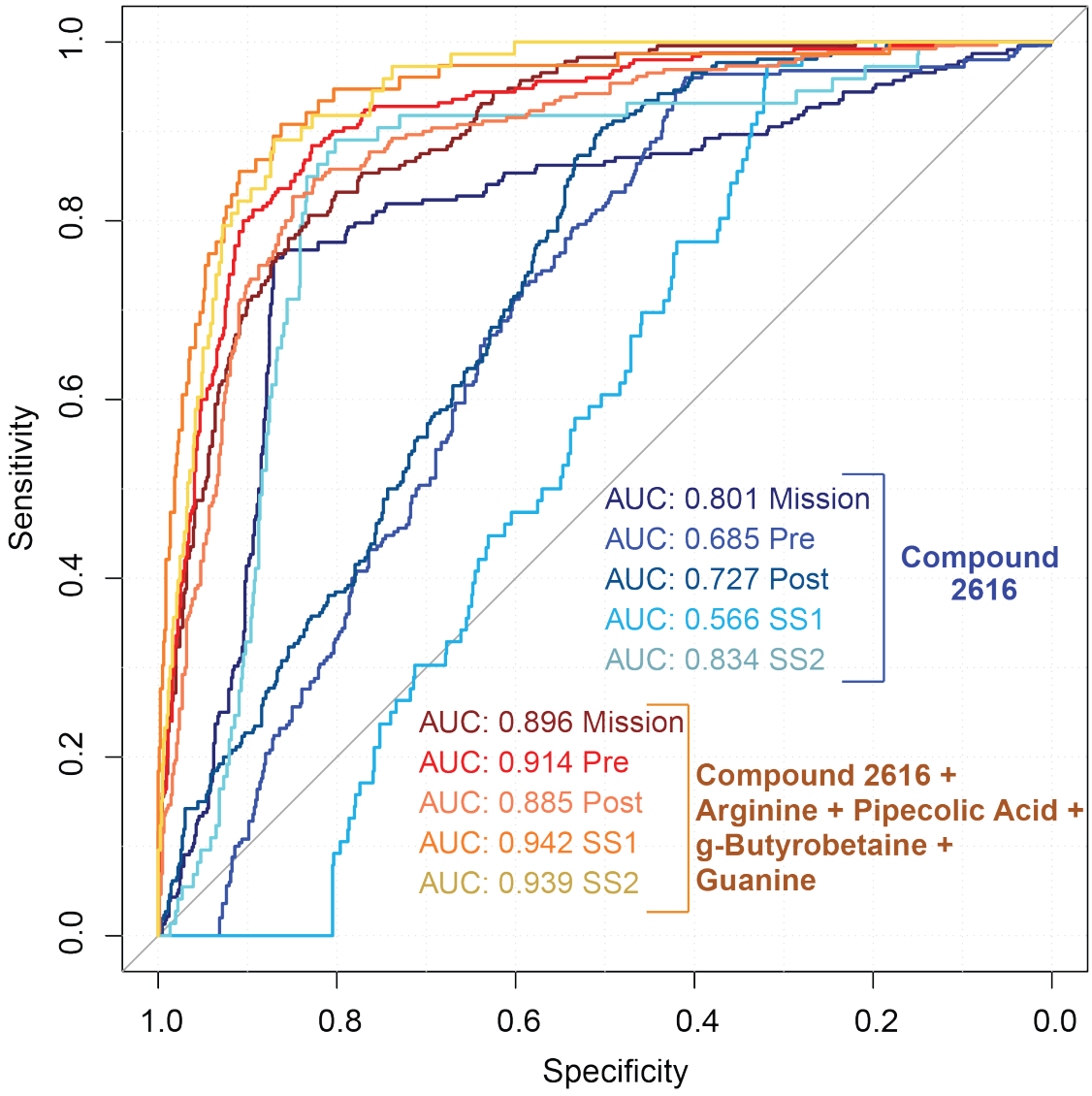
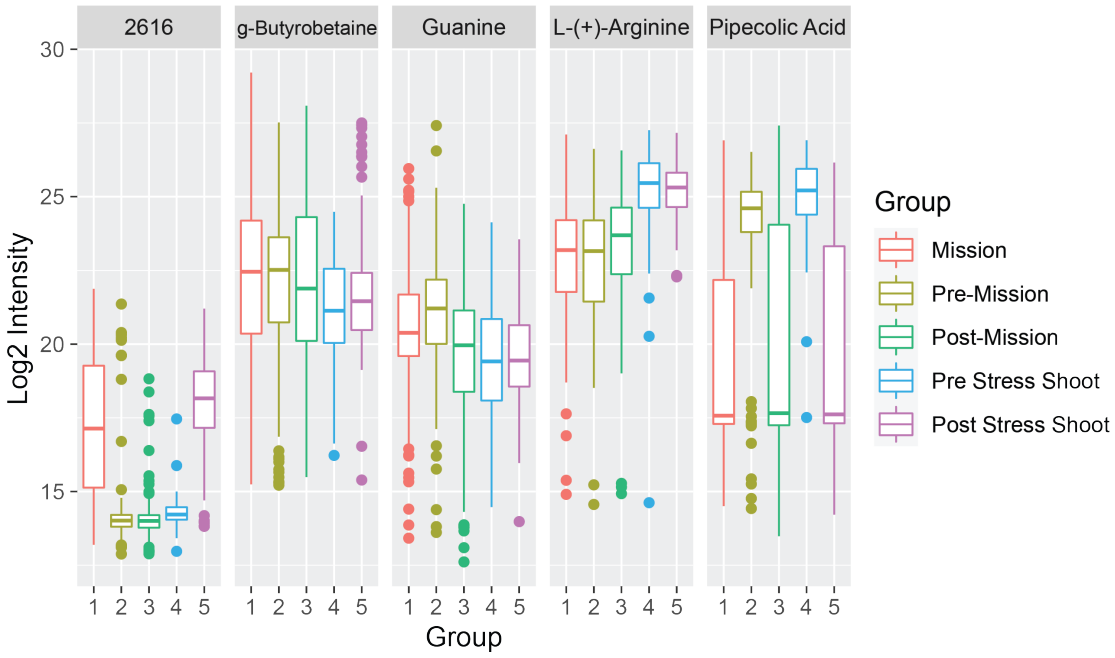


Figure 4.7: Metabolite Linear Discriminant Analysis for Five Time Points. ROC curve for LDA separating samples into five time point groups (Pre = pre-mission, Mission = during mission, Post = post-mission, SS1 = pre-mission stress shoot, SS2 = post-mission stress shoot). Model relying on the single best performing compound feature, unidentified compound 2616, is shown in shades of blue, while an optimized five feature model is shown in shades of red/orange.



Supplementary Figure 4.5: Abundance Shifts for Discriminant Small Molecules. Boxplot of log2 transformed intensities of five small molecules components of most discriminant model for separation of five time point groups (pre-mission, pre-mission stress shoot, mission, post mission, post-mission stress shoot). These five compound features are arginine, pipecolic acid, γ-butyrobetaine and guanine.

endoplasmic reticulum processing proteins, such as chaperones and those involved in the unfolded protein response (UPR). These changes are accompanied by increased abundance of metabolites associated with amino acid metabolism and translation. Changes are also observed in proteins functioning in sugar metabolism.

Our proteomic analysis identifies more than 300 proteins affected by the mission. We observe relatively uniform upregulation of complement proteins upon mission initiation, with tightly correlated shifts from sample to sample. The complement system has been shown to be stimulated in situations of both short and long duration exercise[71–73], with moderate physical training providing a beneficial immunomodulatory effect[74], while extreme physical exertion can increase susceptibility to infection[75] by depressing immune cell populations and signaling molecules. Similar changes have been observed in cases of psychological stress with increased expression of proinflammatory cytokines in acute psychological stress[76,77] and immune dysfunction linked to long-term psychological health impacts such as post-traumatic stress disorder[78,79]. In contrast to the complement system, two other pathways of interest, protein processing in the endoplasmic reticulum and the stress-activated MAPK cascade exhibit varying directions in their abundance changes. In the stress activated MAPK cascade, angiotensinogen exhibits increased abundance with mission, likely in response to the associated physical stress. When attempting to construct a discriminant proteomic model, several of the proteins identified were related to innate immunity or protein processing, including complement component C9 and multiple heat shock cognate chaperones.

When correlating the proteomic and metabolomic datasets, we observed the strongest correlations among molecules related to protein cycling and metabolism. A single strongly protein-correlated cluster of metabolites is comprised of isolated amino acids and short polypeptides, products of protein degradation. We observe positive correlation between these small molecules and a protein cluster enriched for components of the proteasome, as well as the polyubiquitination process, protein catabolism and regulation of amino acid metabolism. These same compounds also exhibit an anticorrelation with a second protein cluster containing the ER-associated processes of glycosylation and the unfolded protein response.

When examining the ER-associated proteins, we observe altered abundance of many key chaperones associated with protein quality control, managing ER stress and the unfolded protein response (UPR) such as SIL1, PDIA, DNAJC1, DNAJB1, OS9, HSPA5, HSPA1B, and HSPA8[58,59,65–67,80]. ER stress and the UPR can be induced by both physical and mental stress. Accelerated metabolic activity in skeletal muscle in response to physical exertion generates increased reactive oxygen species and stimulates the UPR[81,82]. Stress-induced stimulation of the hypothalamic-pituitary-adrenal axis increases free radicals in the brain and has been associated with altered chaperone expression[83–85]. Similar to immunity, moderate triggering of these quality control system can be beneficially adaptive[86] but constant or extreme pressure on the endoplasmic reticulum will lead to autophagy or apoptosis[46], and can lead to a pathogenic reduction in response such as in diabetes[87].

When we scrutinized the affected metabolic pathways more closely, we observed

mission-associated increases of the glycolysis enzymes glyceraldehyde-3-phosphate dehydrogenase, phosphoglycerate kinase, pyruvate kinase, as well as lactate dehydrogenase the key enzyme in anaerobic respiration. We also observed less significant increases in seven other glycolysis-related enzymes: hexokinase, phosphofructokinase, fructose bisphosphate aldolase, triosephosphate isomerase, enolase, and phosphoglycerate mutase. These shifts suggest increased glycolytic activity and anaerobic energy generation. Metabolic changes of this type have been observed in muscle in both endurance training[88] as well as short interval, high intensity training[89–91]. Psychological stress can also lead to stimulation of lactate dehydrogenase[92,93] as well as induction of genes that stimulate glycolysis[94–96], with chronic stress leading to a variety of metabolic diseases, including obesity and diabetes[97,98]. These proteins also participate in the reverse reaction, gluconeogenesis, generating glucose. Glucose serves as the only usable fuel for the central nervous system, and elevated blood glucose through gluconeogenesis activity has been observed in acute stress in animals[3].

Our metabolomic analysis identified more than 100 compound features associated with mission initiation, but perhaps more intriguing was the discriminatory ability for acute stress events, with more than 300 features significantly associated with one of the five event groups (pre-mission stress shoot, post-mission stress shoot, pre-mission, post-mission, during mission). Metabolites associated with specific time point events were related to amino acid metabolism again, with arginine, histidine, aspartate, and glutamate especially impacted. Histidine and arginine metabolites exhibited increased abundance specifically in samples from the two stress shoot events, indicating a possible physiological effect of acute

stress in these systems. Previous work has indicated a role of L-arginine as a component molecule of the arginine nitric oxide pathway, which can be stimulated by psychological stress[99]. Some evidence also exists for L-arginine as a stress reducing molecule when ingested as a dietary supplement[100]. Alternatively, or potentially in tandem, amino acid and aminoacyl-tRNA biosynthesis may be increased in order to modify translation. Perturbations of translation via aminoacyl-tRNA biosynthesis have been observed as a response to a variety of stressors, including temperature, oxidative environment and nutrient deprivation[101,102]. When building a discriminatory model for mission initiation or individual time events we observe pipecolic acid, arginine, and γ-butyrobetaine, as highly discriminant for stress events, all compounds involved in amino acid biosynthesis and translation.

Stress responses play a complex role in human physiology, effecting a variety of systems and requiring a multifaceted approach for their study. Integrating analysis of both proteins and small molecules from raw saliva provides a highly accessible diagnostic tool for understanding the effects of both acute and prolonged stress. Our findings reflect tissues that are experiencing a variety of potential stressors, at both the molecular and systems level. At an organism level, release of stress hormones activates the muscles, the liver, and the cardiovascular, immune, and central nervous system to allow rapid reaction to perceived threats. At the molecular level, cells are tasked with producing and maintaining the energy, materials, and protein machinery that power these organ systems in a rapid and adaptable manner, often in chemically stressful environments. Understanding the functions, and limitations of these interconnected systems plays a crucial role in identifying stress and

mitigating its long-term impact on human health.

References

[1] S. F. De Boer, S. J. Koopmans, J. L. Slangen, and J. Van der Gugten, "Plasma catecholamine, corticosterone and glucose responses to repeated stress in rats: effect of interstressor interval length," *Physiol. Behav.*, vol. 47, pp. 1117–1124, June 1990.

[2] V. A. Viblanc, Q. Schull, T. Cornioley, A. Stier, J.-J. Ménard, R. Groscolas, and J.-P. Robin, "An integrative appraisal of the hormonal and metabolic changes induced by acute stress using king penguins as a model," *Gen. Comp. Endocrinol.*, vol. 269, pp. 1–10, Dec. 2018.

[3] R. M. Sapolsky, L. M. Romero, and A. U. Munck, "How do glucocorticoids influence stress responses? integrating permissive, suppressive, stimulatory, and preparative actions," *Endocr. Rev.*, vol. 21, pp. 55–89, Feb. 2000.

[4] G. Russell and S. Lightman, "The human stress response," *Nat. Rev. Endocrinol.*, vol. 15, pp. 525–534, Sept. 2019.

[5] H. Yaribeygi, Y. Panahi, H. Sahraei, T. P. Johnston, and A. Sahebkar, "The impact of stress on body function: A review," *EXCLI J.*, vol. 16, pp. 1057–1072, July 2017.

[6] C. Santone, V. Dinallo, M. Paci, S. D'Ottavio, G. Barbato, and S. Bernardini, "Saliva

metabolomics by NMR for the evaluation of sport performance," *J. Pharm. Biomed. Anal.*, vol. 88, pp. 441–446, Jan. 2014.

[7] J. Langan-Fox, M. J. Sankey, and J. M. Canty, "Human factors measurement for future air traffic control systems," *Hum. Factors*, vol. 51, pp. 595–637, Oct. 2009.

[8] C. A. Morgan, 3rd, S. Wang, S. M. Southwick, A. Rasmusson, G. Hazlett, R. L. Hauger, and D. S. Charney, "Plasma neuropeptide-y concentrations in humans exposed to military survival training," *Biol. Psychiatry*, vol. 47, pp. 902–909, May 2000.

[9] J. F. O'Hanlon, *Boredom: Practical Consequences and a Theory*. 1980.

[10] M. B. Weinger, "Vigilance, boredom, and sleepiness," *J. Clin. Monit. Comput.*, vol. 15, pp. 549–552, Dec. 1999.

[11] M. L. Cummings, C. Mastracchio, K. M. Thornburg, and A. Mkrtchyan, "Boredom and distraction in multiple unmanned vehicle supervisory control," *Interacting with computers*, vol. 25, no. 1, pp. 34–47, 2013.

[12] M. L. Pacella, B. Hruska, and D. L. Delahanty, "The physical health consequences of PTSD and PTSD symptoms: a meta-analytic review," *J. Anxiety Disord.*, vol. 27, pp. 33–46, Jan. 2013.

[13] R. Yang, A. Gautam, D. Getnet, B. J. Daigle, S. Miller, B. Misganaw, K. R. Dean, R. Kumar, S. Muhie, K. Wang, I. Lee, D. Abu-Amara, J. D. Flory, PTSD Systems Biology Consortium, L. Hood, O. M. Wolkowitz, S. H. Mellon, F. J. Doyle, 3rd,

R. Yehuda, C. R. Marmar, K. J. Ressler, R. Hammamieh, and M. Jett, "Epigenetic biotypes of post-traumatic stress disorder in war-zone exposed veteran and active duty males," *Mol. Psychiatry*, Dec. 2020.

[14] A. Gautam, P. D'Arpa, D. E. Donohue, S. Muhie, N. Chakraborty, B. T. Luke, D. Grapov, E. E. Carroll, J. L. Meyerhoff, R. Hammamieh, and M. Jett, "Acute and chronic plasma metabolomic and liver transcriptomic stress effects in a mouse model with features of post-traumatic stress disorder," *PLoS One*, vol. 10, p. e0117092, Jan. 2015.

[15] E. M. Blessing, V. Reus, S. H. Mellon, O. M. Wolkowitz, J. D. Flory, L. Bierer, D. Lindqvist, F. Dhabhar, M. Li, M. Qian, D. Abu-Amara, I. Galatzer-Levy, R. Yehuda, and C. R. Marmar, "Biological predictors of insulin resistance associated with post-traumatic stress disorder in young military veterans," *Psychoneuroendocrinology*, vol. 82, pp. 91–97, Aug. 2017.

[16] S. Maguen, B. Cohen, L. Ren, J. Bosch, R. Kimerling, and K. Seal, "Gender differences in military sexual trauma and mental health diagnoses among iraq and afghanistan veterans with posttraumatic stress disorder," *Womens. Health Issues*, vol. 22, pp. e61–6, Jan. 2012.

[17] A. O'Donovan, B. E. Cohen, K. H. Seal, D. Bertenthal, M. Margaretten, K. Nishimi, and T. C. Neylan, "Elevated risk for autoimmune disorders in iraq and afghanistan veterans with posttraumatic stress disorder," *Biol. Psychiatry*, vol. 77, pp. 365–374, Feb. 2015.

[18] R. Glaser and J. K. Kiecolt-Glaser, "Stress-induced immune dysfunction: implications for health," *Nat. Rev. Immunol.*, vol. 5, pp. 243–251, Mar. 2005.

[19] J. E. Dimsdale, "Psychological stress and cardiovascular disease," *J. Am. Coll. Cardiol.*, vol. 51, pp. 1237–1246, Apr. 2008.

[20] W.-C. Kuo, L. C. Bratzke, L. D. Oakley, F. Kuo, H. Wang, and R. L. Brown, "The association between psychological stress and metabolic syndrome: A systematic review and meta-analysis," *Obes. Rev.*, vol. 20, pp. 1651–1664, Nov. 2019.

[21] D. J. Michael, S. Daugherty, A. Santos, B. C. Ruby, and J. E. Kalns, "Fatigue biomarker index: an objective salivary measure of fatigue level," *Accid. Anal. Prev.*, vol. 45 Suppl, pp. 68–73, Mar. 2012.

[22] D. J. Michael, B. Valle, J. Cox, J. E. Kalns, and D. L. Fogt, "Salivary biomarkers of physical fatigue as markers of sleep deprivation," *J. Clin. Sleep Med.*, vol. 9, pp. 1325–1331, Dec. 2013.

[23] Y.-L. Xu, Y.-N. Gong, D. Xiao, C.-X. Zhao, X.-H. Gao, X.-H. Peng, A.-P. Xi, L.-H. He, L.-P. Lu, M. Ding, Y. Li, H. Jianjun, X.-H. Su, F.-L. Liu, J.-Z. Wang, Z.-J. Liu, and J.-Z. Zhang, "Discovery and identification of fatigue-related biomarkers in human saliva," *Eur. Rev. Med. Pharmacol. Sci.*, vol. 22, pp. 8519–8536, Dec. 2018.

[24] S.-G. Ra, S. Maeda, R. Higashino, T. Imai, and S. Miyakawa, "Metabolomics of

salivary fatigue markers in soccer players after consecutive games," *Appl. Physiol. Nutr. Metab.*, vol. 39, pp. 1120–1126, Oct. 2014.

[25] K. Ngamchuea, K. Chaisiwamongkhol, C. Batchelor-McAuley, and R. G. Compton, "Chemical analysis in saliva and the search for salivary biomarkers - a tutorial review," *Analyst*, vol. 143, pp. 81–99, Dec. 2017.

[26] E. M. McBride, R. J. Lawrence, K. McGee, P. M. Mach, P. S. Demond, M. W. Busch, J. W. Ramsay, E. K. Hussey, T. Glaros, and E. S. Dhummakupt, "Rapid liquid chromatography tandem mass spectrometry method for targeted quantitation of human performance metabolites in saliva," *J. Chromatogr. A*, vol. 1601, pp. 205–213, Sept. 2019.

[27] B. L. Schulz, J. Cooper-White, and C. K. Punyadeera, "Saliva proteome research: current status and future outlook," *Crit. Rev. Biotechnol.*, vol. 33, pp. 246–259, Sept. 2013.

[28] I. Messana, T. Cabras, F. Iavarone, B. Manconi, L. Huang, C. Martelli, A. Olianas, M. T. Sanna, E. Pisano, M. Sanna, M. Arba, A. D'Alessandro, C. Desiderio, A. Vitali, D. Pirolli, C. Tirone, A. Lio, G. Vento, C. Romagnoli, M. Cordaro, A. Manni, P. Gallenzi, A. Fiorita, E. Scarano, L. Calò, G. C. Passali, P. M. Picciotti, G. Paludetti, V. Fanos, G. Faa, and M. Castagnola, "Chrono-proteomics of human saliva: variations of the salivary proteome during human development," *J. Proteome Res.*, vol. 14, pp. 1666–1677, Apr. 2015.

[29] L. M. R. B. Arantes, A. C. De Carvalho, M. E. Melendez, and A. Lopes Carvalho, "Serum, plasma and saliva biomarkers for head and neck cancer," *Expert Rev. Mol. Diagn.*, vol. 18, pp. 85–112, Jan. 2018.

[30] B. Manconi, B. Liori, T. Cabras, F. Vincenzoni, F. Iavarone, L. Lorefice, E. Cocco, M. Castagnola, I. Messana, and A. Olianas, "Top-down proteomic profiling of human saliva in multiple sclerosis patients," *J. Proteomics*, vol. 187, pp. 212–222, Sept. 2018.

[31] M. Bayani, M. Pourali, and M. Keivan, "Possible interaction between visfatin, periodontal infection, and other systemic diseases: A brief review of literature," *Eur. J. Dent.*, vol. 11, pp. 407–410, July 2017.

[32] A. K. Dey, B. Kumar, A. K. Singh, P. Ranjan, R. Thiruvengadam, B. K. Desiraju, P. Kshetrapal, N. Wadhwa, S. Bhatnagar, F. Rashid, D. Malakar, D. M. Salunke, T. K. Maiti, and GARBH-Ini Study Group*, "Salivary proteome signatures in the early and middle stages of human pregnancy with term birth outcome," *Sci. Rep.*, vol. 10, p. 8022, May 2020.

[33] S. Jain, Y. Ahmad, and K. Bhargava, "Salivary proteome patterns of individuals exposed to high altitude," *Arch. Oral Biol.*, vol. 96, pp. 104–112, Dec. 2018.

[34] A. J. Steckl and P. Ray, "Stress biomarkers in biological fluids and their Point-of-Use detection," *ACS Sens*, vol. 3, pp. 2025–2044, Oct. 2018.

[35] E. Shishkova, A. S. Hebert, M. S. Westphall, and J. J. Coon, "Ultra-High pressure (>30,000 psi) packing of capillary columns enhancing depth of shotgun proteomic analyses," *Anal. Chem.*, vol. 90, pp. 11503–11508, Oct. 2018.

[36] A. S. Hebert, C. Thöing, N. M. Riley, N. W. Kwiecien, E. Shiskova, R. Huguet, H. L. Cardasis, A. Kuehn, S. Eliuk, V. Zabrouskov, M. S. Westphall, G. C. McAlister, and J. J. Coon, "Improved precursor characterization for Data-Dependent mass spectrometry," *Anal. Chem.*, vol. 90, pp. 2333–2340, Feb. 2018.

[37] F. Rohart, B. Gautier, A. Singh, and K.-A. Lê Cao, "mixomics: An R package for 'omics feature selection and multiple data integration," *PLoS Comput. Biol.*, vol. 13, p. e1005752, Nov. 2017.

[38] F. Kemp, "Modern applied statistics with s," 2003.

[39] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, "pROC: an open-source package for R and s+ to analyze and compare ROC curves," *BMC Bioinformatics*, vol. 12, p. 77, Mar. 2011.

[40] D. J. Hand and R. J. Till, "A simple generalisation of the area under the roc curve for multiple class classification problems," *Machine learning*, vol. 45, no. 2, pp. 171–186, 2001.

[41] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *arXiv preprint arXiv:1406.5823*, 2014.

[42] P. Sivadasan, M. K. Gupta, G. J. Sathe, L. Balakrishnan, P. Palit, H. Gowda, A. Suresh, M. A. Kuriakose, and R. Sirdeshmukh, "Human salivary proteome–a resource of potential biomarkers for oral cancer," *J. Proteomics*, vol. 127, pp. 89–95, Sept. 2015.

[43] C. Delles, E. Carrick, D. Graham, and S. A. Nicklin, "Utilizing proteomics to understand and define hypertension: where are we and where do we go?," *Expert Rev. Proteomics*, vol. 15, pp. 581–592, July 2018.

[44] Y.-J. Chen, F. Huang, M. Zhang, and H.-Y. Shang, "Psychological stress alters ultrastructure and energy metabolism of masticatory muscle in rats," *J. Biomed. Biotechnol.*, vol. 2010, p. 302693, Oct. 2010.

[45] K. Sahlin, I. G. Shabalina, C. M. Mattsson, L. Bakkman, M. Fernström, Z. Rozhdestvenskaya, J. K. Enqvist, J. Nedergaard, B. Ekblom, and M. Tonkonogi, "Ultraendurance exercise increases the production of reactive oxygen species in isolated mitochondria from human skeletal muscle," *J. Appl. Physiol.*, vol. 108, pp. 780–787, Apr. 2010.

[46] J. D. Malhotra, H. Miao, K. Zhang, A. Wolfson, S. Pennathur, S. W. Pipe, and R. J. Kaufman, "Antioxidants reduce endoplasmic reticulum stress and improve protein secretion," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, pp. 18525–18530, Nov. 2008.

[47] S. K. Al-Tarawneh, M. B. Border, C. F. Dibble, and S. Bencharit, "Defining salivary biomarkers using mass spectrometry-based proteomics: a systematic review," *OMICS*, vol. 15, pp. 353–361, June 2011.

[48] S. Schenk, S. C. Bannister, F. J. Sedlazeck, D. Anrather, B. Q. Minh, A. Bileck, M. Hartl, A. von Haeseler, C. Gerner, F. Raible, and K. Tessmar-Raible, "Combined transcriptome and proteome profiling reveals specific molecular brain signatures for sex, maturation and circalunar clock phase," *Elife*, vol. 8, Feb. 2019.

[49] B. Gibbison, F. Spiga, J. J. Walker, G. M. Russell, K. Stevenson, Y. Kershaw, Z. Zhao, D. Henley, G. D. Angelini, and S. L. Lightman, "Dynamic pituitary-adrenal interactions in response to cardiac surgery," *Crit. Care Med.*, vol. 43, pp. 791–800, Apr. 2015.

[50] S. L. Lightman, "The neuroendocrinology of stress: a never ending story," *J. Neuroendocrinol.*, vol. 20, pp. 880–884, June 2008.

[51] T. Roenneberg and M. Merrow, "The circadian clock and human health," *Curr. Biol.*, vol. 26, pp. R432–43, May 2016.

[52] F. M. Vaz and R. J. A. Wanders, "Carnitine biosynthesis in mammals," *Biochem. J*, vol. 361, pp. 417–429, Feb. 2002.

[53] B. Manconi, M. Castagnola, T. Cabras, A. Olianas, A. Vitali, C. Desiderio, M. T. Sanna, and I. Messana, "The intriguing heterogeneity of human salivary proline-rich proteins: Short title: Salivary proline-rich protein species," *J. Proteomics*, vol. 134, pp. 47–56, Feb. 2016.

[54] M. Mesgari-Abbasi, H. Valizadeh, N. Mirzakhani, and T. Vahdatpour, "Protective effects of di- and tri-peptides containing proline, glycine, and leucine on liver enzymology and histopathology of diabetic mice," *Arch. Physiol. Biochem.*, pp. 1–10, Sept. 2019.

[55] T. Vahdatpour, A. Nokhodchi, P. Zakeri-Milani, M. Mesgari-Abbasi, N. Ahmadi-Asl, and H. Valizadeh, "Leucine-glycine and carnosine dipeptides prevent diabetes induced by multiple low-doses of streptozotocin in an experimental model of adult mice," *J. Diabetes Investig.*, vol. 10, pp. 1177–1188, Sept. 2019.

[56] M. E. Grigorieva, T. Y. Obergan, E. S. Maystrenko, and M. D. Kalugina, "Anticoagulant effects of heparin complexes with Prolyl-Glycine peptide and glycine and proline amino acids," *Bull. Exp. Biol. Med.*, vol. 161, pp. 54–57, May 2016.

[57] M. Zhang, J. Xu, T. Wang, X. Wan, F. Zhang, L. Wang, X. Zhu, P. Gao, G. Shu, Q. Jiang, and S. Wang, "The dipeptide Pro-Gly promotes IGF-1 expression and secretion in HepG2 and female mice via PepT1-JAK2/STAT5 pathway," *Front. Endocrinol.*, vol. 9, p. 424, July 2018.

[58] Y. Wang, X. Fu, S. Gaiser, M. Köttgen, A. Kramer-Zucker, G. Walz, and T. Wegierski, "OS-9 regulates the transit and polyubiquitination of TRPV4 in the endoplasmic reticulum," *J. Biol. Chem.*, vol. 282, pp. 36561–36570, Dec. 2007.

[59] K. Legler, R. Rosprim, T. Karius, K. Eylmann, M. Rossberg, R. M. Wirtz, V. Müller,

I. Witzel, B. Schmalfeldt, K. Milde-Langosch, and L. Oliveira-Ferrer, "Reduced mannosidase MAN1A1 expression leads to aberrant n-glycosylation and impaired survival in breast cancer," *Br. J. Cancer*, vol. 118, pp. 847–856, Mar. 2018.

[60] B. Porath, V. G. Gainullin, E. Cornec-Le Gall, E. K. Dillinger, C. M. Heyer, K. Hopp, M. E. Edwards, C. D. Madsen, S. R. Mauritz, C. J. Banks, S. Baheti, B. Reddy, J. I. Herrero, J. M. Bañales, M. C. Hogan, V. Tasic, T. J. Watnick, A. B. Chapman, C. Vigneau, F. Lavainne, M.-P. Audrézet, C. Ferec, Y. Le Meur, V. E. Torres, Genkyst Study Group, HALT Progression of Polycystic Kidney Disease Group, Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease, and P. C. Harris, "Mutations in GANAB, encoding the glucosidase IIα subunit, cause Autosomal-Dominant polycystic kidney and liver disease," *Am. J. Hum. Genet.*, vol. 98, pp. 1193–1207, June 2016.

[61] L. Zhao, C. Longo-Guess, B. S. Harris, J.-W. Lee, and S. L. Ackerman, "Protein accumulation and neurodegeneration in the woozy mutant mouse is caused by disruption of SIL1, a cochaperone of BiP," *Nat. Genet.*, vol. 37, pp. 974–979, Sept. 2005.

[62] D. Eletto, D. Eletto, D. Dersh, T. Gidalevitz, and Y. Argon, "Protein disulfide isomerase A6 controls the decay of IRE1α signaling via disulfide-dependent association," *Mol. Cell*, vol. 53, pp. 562–576, Feb. 2014.

[63] R. van Huizen, J. L. Martindale, M. Gorospe, and N. J. Holbrook, "P58IPK, a novel

endoplasmic reticulum stress-inducible protein and potential negative regulator of eIF2alpha signaling," *J. Biol. Chem.*, vol. 278, pp. 15558–15564, May 2003.

[64] N. Amin-Wetzel, R. A. Saunders, M. J. Kamphuis, C. Rato, S. Preissler, H. P. Harding, and D. Ron, "A J-Protein co-chaperone recruits BiP to monomerize IRE1 and repress the unfolded protein response," *Cell*, vol. 171, pp. 1625–1637.e13, Dec. 2017.

[65] J. N. Rauch and J. E. Gestwicki, "Binding of human nucleotide exchange factors to heat shock protein 70 (hsp70) generates functionally distinct complexes in vitro," *J. Biol. Chem.*, vol. 289, pp. 1402–1414, Jan. 2014.

[66] V. S. Stronge, Y. Saito, Y. Ihara, and D. B. Williams, "Relationship between calnexin and BiP in suppressing aggregation and promoting refolding of protein and glycoprotein substrates," *J. Biol. Chem.*, vol. 276, pp. 39779–39787, Oct. 2001.

[67] F. Stricher, C. Macri, M. Ruff, and S. Muller, "HSPA8/HSC70 chaperone protein: structure, function, and chemical targeting," *Autophagy*, vol. 9, pp. 1937–1954, Dec. 2013.

[68] J. H. Exton and C. R. Park, "Control of gluconeogenesis in liver. i. general features of gluconeogenesis in the perfused livers of rats," *J. Biol. Chem.*, vol. 242, pp. 2622–2636, June 1967.

[69] S. J. Pilkis and D. K. Granner, *Molecular Physiology of the Regulation of Hepatic Gluconeogenesis and Glycolysis*. 1992.

[70] J. Chong, O. Soufan, C. Li, I. Caraus, S. Li, G. Bourque, D. S. Wishart, and J. Xia, "MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis," *Nucleic Acids Res.*, vol. 46, pp. W486–W494, July 2018.

[71] B. Dufaux and U. Order, "Complement activation after prolonged exercise," *Clin. Chim. Acta*, vol. 179, pp. 45–49, Jan. 1989.

[72] J. Smith, D. Chi, S. Salazar, G. Krish, S. Berk, S. Reynolds, and G. Cambron, "Effect of moderate exercise on proliferative responses of peripheral blood mononuclear cells," *J. Sports Med. Phys. Fitness*, vol. 33, pp. 152–158, June 1993.

[73] B. Dufaux, U. Order, and H. Liesen, "Effect of a short maximal physical exercise on coagulation, fibrinolysis, and complement system," *Int. J. Sports Med.*, vol. 12 Suppl 1, pp. S38–42, June 1991.

[74] E. Ortega, "Neuroendocrine mediators in the modulation of phagocytosis by exercise: physiological implications," *Exerc. Immunol. Rev.*, vol. 9, pp. 70–93, 2003.

[75] D. Nieman, *Exercise Testing & Prescription*. McGraw-Hill Humanities/Social Sciences/Languages, 2007.

[76] V. E. Burns, K. M. Edwards, C. Ring, M. Drayson, and D. Carroll, "Complement cascade activation after an acute psychological stress task," *Psychosom. Med.*, vol. 70, pp. 387–396, May 2008.

[77] M. S. Breen, N. Beliakova-Bethell, L. R. Mujica-Parodi, J. M. Carlson, W. Y. Ensign, C. H. Woelk, and B. K. Rana, "Acute psychological stress induces short-term variable immune response," *Brain Behav. Immun.*, vol. 53, pp. 172–182, Mar. 2016.

[78] M. S. Breen, A. X. Maihofer, S. J. Glatt, D. S. Tylee, S. D. Chandler, M. T. Tsuang, V. B. Risbrough, D. G. Baker, D. T. O'Connor, C. M. Nievergelt, and C. H. Woelk, "Gene networks specific for innate immunity define post-traumatic stress disorder," *Mol. Psychiatry*, vol. 20, pp. 1538–1545, Dec. 2015.

[79] L. P. Hovhannisyan, G. M. Mkrtchyan, S. H. Sukiasian, and A. S. Boyajyan, "Alterations in the complement cascade in post-traumatic stress disorder," *Allergy Asthma Clin. Immunol.*, vol. 6, p. 3, Feb. 2010.

[80] T. Liu, C. K. Daniels, and S. Cao, "Comprehensive review on the HSC70 functions, interactions with related molecules and involvement in clinical diseases and therapeutic potential," *Pharmacol. Ther.*, vol. 136, pp. 354–374, Dec. 2012.

[81] J. Wu, J. L. Ruas, J. L. Estall, K. A. Rasbach, J. H. Choi, L. Ye, P. Boström, H. M. Tyra, R. W. Crawford, K. P. Campbell, D. T. Rutkowski, R. J. Kaufman, and B. M. Spiegelman, "The unfolded protein response mediates adaptation to exercise in skeletal muscle through a PGC-1α/ATF6α complex," *Cell Metab.*, vol. 13, pp. 160–169, Feb. 2011.

[82] D. I. Ogborn, B. R. McKay, J. D. Crane, G. Parise, and M. A. Tarnopolsky, "The

unfolded protein response is triggered following a single, unaccustomed resistance-exercise bout," *Am. J. Physiol. Regul. Integr. Comp. Physiol.*, vol. 307, pp. R664–9, Sept. 2014.

[83] S. Currie, S. LeBlanc, M. A. Watters, and K. M. Gilmour, "Agonistic encounters and cellular angst: social interactions induce heat shock proteins in juvenile salmonid fish," *Proc. Biol. Sci.*, vol. 277, pp. 905–913, Mar. 2010.

[84] D. Vassilopoulos and D. Mantzoukis, "Dialogue between the brain and the immune system in inflammatory arthritis," *Ann. N. Y. Acad. Sci.*, vol. 1088, pp. 132–138, Nov. 2006.

[85] T. Hayashi, "Conversion of psychological stress into cellular stress response: roles of the sigma-1 receptor in the process," *Psychiatry Clin. Neurosci.*, vol. 69, pp. 179–191, Apr. 2015.

[86] M. Ost, V. Coleman, J. Kasch, and S. Klaus, "Regulation of myokine expression: Role of exercise and cellular stress," *Free Radic. Biol. Med.*, vol. 98, pp. 78–89, Sept. 2016.

[87] M. Cnop, F. Foufelle, and L. A. Velloso, "Endoplasmic reticulum stress, obesity and diabetes," *Trends Mol. Med.*, vol. 18, pp. 59–68, Jan. 2012.

[88] A. S. Deshmukh, D. E. Steenberg, M. Hostrup, J. B. Birk, J. K. Larsen, A. Santos, R. Kjøbsted, J. R. Hingst, C. C. Schéele, M. Murgia, B. Kiens, E. A. Richter, M. Mann, and J. F. P. Wojtaszewski, "Deep muscle-proteomic analysis of freeze-dried human

muscle biopsies reveals fiber type-specific adaptations to exercise training," *Nat. Commun.*, vol. 12, p. 304, Jan. 2021.

[89] I. Jacobs, M. Esbjörnsson, C. Sylvén, I. Holm, and E. Jansson, "Sprint training effects on muscle myoglobin, enzymes, fiber types, and blood lactate," *Med. Sci. Sports Exerc.*, vol. 19, pp. 368–374, Aug. 1987.

[90] M. T. Linossier, D. Dormois, C. Perier, J. Frey, A. Geyssant, and C. Denis, "Enzyme adaptations of human skeletal muscle during bicycle short-sprint training and detraining," *Acta Physiol. Scand.*, vol. 161, pp. 439–445, Dec. 1997.

[91] T. Abe, Y. Kitaoka, D. M. Kikuchi, K. Takeda, O. Numata, and T. Takemasa, "High-intensity interval training-induced metabolic adaptation coupled with an increase in Hif-1α and glycolytic protein expression," *J. Appl. Physiol.*, vol. 119, pp. 1297–1302, Dec. 2015.

[92] B. Cui, Y. Luo, P. Tian, F. Peng, J. Lu, Y. Yang, Q. Su, B. Liu, J. Yu, X. Luo, L. Yin, W. Cheng, F. An, B. He, D. Liang, S. Wu, P. Chu, L. Song, X. Liu, H. Luo, J. Xu, Y. Pan, Y. Wang, D. Li, P. Huang, Q. Yang, L. Zhang, B. P. Zhou, S. Liu, G. Xu, E. W.-F. Lam, K. W. Kelley, and Q. Liu, "Stress-induced epinephrine enhances lactate dehydrogenase a and promotes breast cancer stem-like cells," *J. Clin. Invest.*, vol. 129, pp. 1030–1046, Mar. 2019.

[93] H. Arakawa, H. Kodama, N. Matsuoka, and I. Yamaguchi, "Stress increases plasma

enzyme activity in rats: differential effects of adrenergic and cholinergic blockades," *J. Pharmacol. Exp. Ther.*, vol. 280, pp. 1296–1303, Mar. 1997.

[94] G. Ermak, M. A. Pritchard, S. Dronjak, B. Niu, and K. J. A. Davies, "Do RCAN1 proteins link chronic stress with neurodegeneration?," *FASEB J.*, vol. 25, pp. 3306–3311, Oct. 2011.

[95] Y. Hirakawa, L. J. Nary, and R. D. Medh, "Glucocorticoid evoked upregulation of RCAN1-1 in human leukemic CEM cells susceptible to apoptosis," *J. Mol. Signal.*, vol. 4, p. 6, Sept. 2009.

[96] G. Ermak, S. Sojitra, F. Yin, E. Cadenas, A. M. Cuervo, and K. J. A. Davies, "Chronic expression of RCAN1-1L protein induces mitochondrial autophagy and metabolic shift from oxidative phosphorylation to glycolysis in neuronal cells," *J. Biol. Chem.*, vol. 287, pp. 14088–14098, Apr. 2012.

[97] M. Picard and B. S. McEwen, "Psychological stress and mitochondria: A systematic review," *Psychosom. Med.*, vol. 80, no. 2, pp. 141–153, 2018.

[98] C. Rabasa and S. L. Dickson, "Impact of stress on metabolism and energy balance," *Current Opinion in Behavioral Sciences*, vol. 9, pp. 71–77, 2016.

[99] M. Reimann, M. Hamer, N. T. Malan, M. P. Schlaich, G. W. Lambert, T. Ziemssen, R. H. Boeger, and L. Malan, "Effects of acute and chronic stress on the l-arginine nitric oxide pathway in black and white south africans: the sympathetic activity and

ambulatory blood pressure in africans study," *Psychosom. Med.*, vol. 75, pp. 751–758, Oct. 2013.

[100] M. Smriga, T. Ando, M. Akutsu, Y. Furukawa, K. Miwa, and Y. Morinaga, "Oral treatment with l-lysine and l-arginine reduces anxiety and basal cortisol levels in healthy humans," *Biomed. Res.*, vol. 28, pp. 85–90, Apr. 2007.

[101] N.-C. Han, P. Kelly, and M. Ibba, "Translational quality control and reprogramming during stress adaptation," *Exp. Cell Res.*, vol. 394, p. 112161, Sept. 2020.

[102] T. Pan, "Adaptive translation as a mechanism of stress response and adaptation," *Annu. Rev. Genet.*, vol. 47, pp. 121–137, Aug. 2013.

Chapter 5

PREDICTIVE MODELING OF PEPTIDE TRANSMISSION IN HIGH FIELD ASYMMETRIC WAVEFORM ION MOBILITY (FAIMS) USING MACHINE LEARNING

JM designed and conducted mass spectrometry experiments, cleaned and prepared the data, developed the random forest model, developed figures, and drafted the manuscript. The author would like to specifically acknowledge JGM for development of the neural net model and data searching.

This chapter is adapted from a manuscript in preparation

McKetney J, Meyer JG, Miller IJ, Coon JJ. *Predictive Modeling of Peptide Transmission in High Field Asymmetric Waveform Ion Mobility (FAIMS) using Machine Learning.* **2021**.

Abstract

Peptide ion mobility adds an extra dimension of separation to mass spectrometry-based proteomics. The ability to accurately predict peptide ion mobility would allow assay development to be expedited. There are methods to accurately predict peptide ion mobility through drift tube devices, but there are not yet methods to predict mobility through high field asymmetric waveform ion mobility (FAIMS). Here, we show that prediction of peptide FAIMS ion mobility is not a simple regression problem due to observation of peptide transmission at multiple settings, and successfully model this problem using multi-label classification. We trained two models, a random forest and a long-term short-term memory (LSTM) neural network. Both models had different strengths, and the ensemble average of model predictions produced higher F2 score than either model alone. Finally, we explore cases where the models make mistakes, and demonstrate predictive performance of F2=0.66 (ROC-AUC =0.928) on a new test dataset of nearly 40,000 different *E. coli* peptide precursor ions.

Introduction

Ion mobility spectrometry (IMS) has long played an important role in mass spectrometry (MS), providing an additional dimension of separation complementary to separations by liquid chromatography (LC) and mass-to-charge[1–4]. IMS has allowed for isolation and separation of biomolecules based on structure[5–8], including differentiating isomers[9,10],

which is a challenging task for LC-MS/MS alone. Increases in speed and efficiency of IMS, and integration of IMS with commercial MS systems has led to its widespread application in proteomics[11–16]. Among the several variants of IMS, high field asymmetric-waveform ion mobility spectrometry (FAIMS) with high transmission efficiency was commercially released in 2018 as a source-attached module known as the FAIMS Pro, which allows for rapid separation of peptide ions on timescales compatible with LC-MS experiments[14,17].

FAIMS relies on the differential mobility of ions in an electric field of varying strength[18,19]. Ions ejected from the electrospray emitter follow a roughly parabolic path through the FAIMS module before entering the mass spectrometer. Within the FAIMS module, ions pass between two electrodes where they are destabilized by an electric field with an asymmetric waveform causing their collision with one of the electrodes. An ion-specific compensating voltage (CV), set by the user, stabilizes a subpopulation of ions, allowing their transmission into the instrument. In contrast with other ion mobility techniques, FAIMS therefore acts as a filter that simplifies the mixture of ions entering the MS instrument.

FAIMS filtering simplifies the gas-phase peptide ion mixture, allowing increased isolation, fragmentation and identification of low abundance peptides in proteomic analysis. This increased access to low abundance peptides improves proteomic depth, in a similar fashion to orthogonal chromatographic fractionation[14]. CV can be rapidly changed within a method, in a process known as internal stepping, allowing for the selection and quantification of several "gas-phase fractions", leading to an expanded number of protein identifications with no increased sample preparation. The increase in signal-to-noise and

stepping speed of the FAIMS interface provides substantial benefits in quantitative accuracy to targeted methods[20] and identifications in DIA strategies utilizing short gradients[21]. In fact, FAIMS filtering has enabled fast proteome analysis without the use of LC[22]. Thus, FAIMS is a valuable tool for proteomic analysis.

At present, the optimally transmissive CV for a given peptide through FAIMS must be determined empirically by performing experiments across the entire voltage range. Although previous work has identified some of the most important parameters affecting ion mobility, such as charge, the relationship is not straightforward or linear. Since the inception of IMS, researchers have investigated peptide ion behavior in hopes of developing predictive models[4,23]. Much of the more recent work has focused on identifying drift times[24–26] or collisional cross section[27]. Drift time and collisional cross section are single value properties that lend themselves to regression analysis. In contrast, peptides separated by FAIMS are often found to transmit at several CVs, and CV settings tend to be designed in a discrete stepwise fashion, which makes regression less attractive.

Machine learning models have been applied to virtually all areas of mass spectrometry and associated technologies, seeking predictive information for a broad spectrum of molecules including proteins, lipids[28–30], sugars[31] and nucleic acids[32]. In the proteomics field alone, substantial research has focused on the development of models predicting retention time[32–36], fragmentation spectra[37,38] and high-signal peptides for targeted proteomic methods[39]. Many of the resulting tools contribute to a more streamlined design of optimized methods *in silico*. These machine learning tools enable more rapid and cost-effective

method development for all MS data acquisition modes (e.g. PRM, DDA, and DIA), and could be used as additional constraints for peptide identification scoring.

Here we present a pair of machine learning algorithms, a random forest (RF)[40] and a long-term short-term memory[39,41] artificial neural net (NN), for *in silico* prediction of the optimal compensating voltage settings for any given peptide ion. The RF represents an approach that is both robust and easily interpretable, while the NN represents a state-of-the-art machine learning model that has shown strong performance previously in predicting peptide behavior[27]. We explore peptide properties that determine FAIMS transmission profiles and demonstrate that peptide transmission through FAIMS is more complex than simple collisional cross sections. Although CV is a continuous parameter in earnest, regression or single label classification models cannot accurately encompass the multitudinous nature of peptide transmissive CV. Therefore, we instead framed this problem as multi-label classification. The ensemble based on the average of both model predictions trained on over 100,000 human peptide precursors resulted in a 0.928 ROC-AUC score when predicting FAIMS CVs for a novel test set of over 40,000 *E. coli* peptide precursors.

Materials and Methods

Chemicals and Reagents *Escherichia coli* samples of strain K12 MG1655 donated by the Cox group of the UW-Madison Biochemistry Department, were set to grow overnight at 37 °C. Human lysate was derived from intact mass spec-compatible human (K562 cells) protein extract purchased from Promega (Product # V6941).

Sample preparation *E. coli* pellet was resuspended in 4M guanidine HCl before cells were lysed by probe sonication, boiled, and allowed to cool before being brought to 90% MeOH. Lysate was centrifuged at 15,000 g for 7 minutes and supernatant was disposed, while precipitate was resuspended in reducing and alkylating buffer (8M Urea, 40 mM TCEP, 10 mM CAA, 100 mM Tris pH 8). Lys-C was added at a ratio of 50:1 w/w protein to protease, and the sample was incubated at room temperature for 4 hours. Buffer was diluted to 25% with 100mM Tris pH 8 and trypsin was added at a ratio of 50:1 before the sample was digested at ambient temperature overnight. Human extract was prepared in an identical manner starting with addition of methanol. Digested peptides were desalted using Strata-X Polymeric Reverse Phase column (Phenomenex).

nLC-MS/MS Data collection Online reverse-phase columns were prepared in-house using a high-pressure packing apparatus[42]. In brief, 1.7 µm Bridged Ethylene Hybrid C18 particles were packed at 30,000 psi into a New Objective PicoTipTM emitter (Stock# PF360-75-10-N-5) with an inner diameter of 75 µm and an outer diameter of 360 µm. During separations, the column was heated to a temperature of 50 °C inside a heater (developed in-house) and interfaced with the mass spectrometer via an embedded emitter.

A Dionex UltiMate 3000 nanoflow UHPLC was used for online chromatography with mobile phase buffer A consisting of 0.2% formic acid in water and mobile phase buffer B consisting of 0.2% formic acid in 70% acetonitrile. Mobile phase B was increased to 9% in the first 6.5 min then increased to 43% B at 41 min. The method finished with a wash stage

of 100% B from 44-48 minutes and an equilibration step of 0% B from 50-60 minutes, for a total method length of one hour. Flow rate was 335 nanoliters per minute throughout.

Eluting peptides were ionized by electrospray ionization and passed through the FAIMS Pro module (Thermo Scientific) using a single compensating voltage for the duration of each method. Ions were analyzed on a Thermo Orbitrap Fusion Lumos with survey scans taken from 300 to 1500 m/z at 240,000 resolution while using Advanced Precursor Determination[43] and an AGC target of 1e6 with a maximum injection time of 50 ms. Precursor isolation used a window of 0.7 Th with 15 ppm mass tolerance and a dynamic exclusion time of 20 seconds. Selected precursors were fragmented using HCD with a normalized collision energy of 25%. The MS2 AGC target was set at 3e4 with a maximum injection time of 14 ms and scans taken using the turbo setting. Analysis was performed in technical duplicate.

Peptide Identification and Quantification Raw files were converted to .mzML format using Proteowizard[44] (version 3.0.19039). Peptides were identified by database search against the human proteome including isoforms (downloaded 2019-12-19) or the *E. coli* proteome (downloaded 2020-05-01) using MS-Fragger[45](version 2.2) through the FragPipe GUI (version 12.1). The default search parameters were used except for the precursor tolerance, which was set to +/- 5 ppm and the fragment tolerance was set to 0.5 Daltons. Database search output files in pep.xml format were combined into one file using iProphet[46] within philosopher[47](version 2.0). Peptide identifications were imported into Skyline[48] for quantification using an iProphet score cutoff of 0.99. In an attempt to determine the true

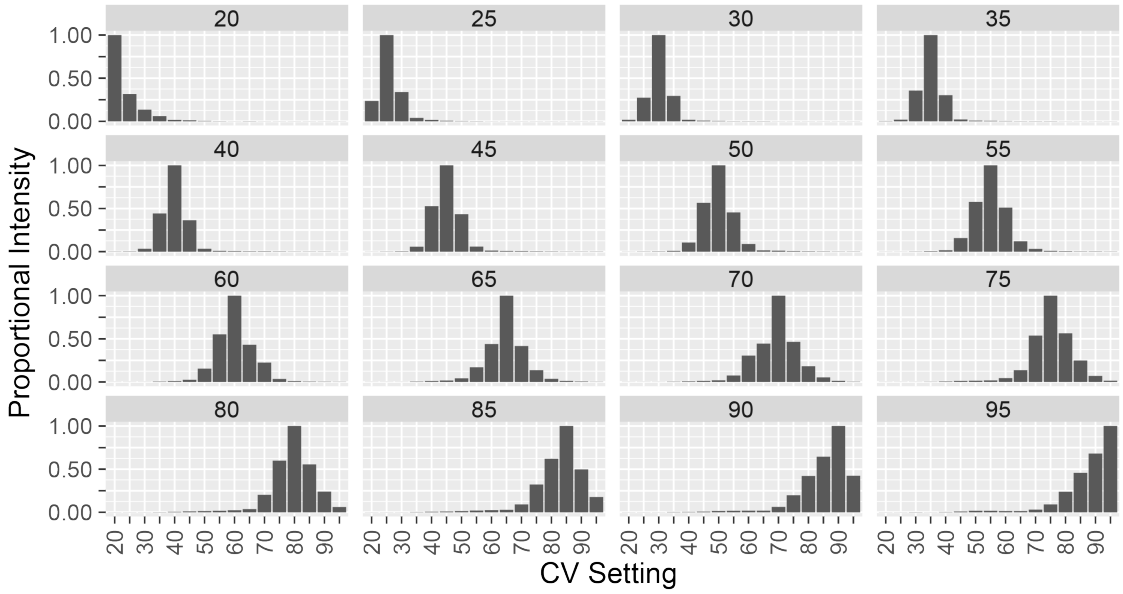
distribution of peptide precursor transmission through the FAIMS device independently of the stochastic DDA identification process, precursor (MS1) signals were extracted from all files corresponding to FAIMS CV values from -20 V to -100V. MS1 signal was extracted within 8 ppm of the predicted precursor mass and 1 minute (*E. coli* data) or 0.5 minutes (human data) of the MS/MS identifications. Precursor signals for decoy peptides were also extracted to allow mProphet determination of statistical significance of peak peaking[49]. Only precursor ion peaks with q-value of < 0.01 were integrated, and a custom Skyline report was output containing the q-values, precursor ion dot product with the theoretical isotope envelope (idotp), and peak area for each FAIMS CV. The skyline report was further processed in R to determine the FAIMS CV transmission profile for each peptide precursor ion.

Determination of Peptide CV transmission profiles Skyline reports were read into R (version 3.6.3) and decoy peptides along with peptides matching reversed proteins were removed. Each precursor's intensities were scaled from 0-1, with 1 representing the CV setting with maximum intensity and therefore, transmission (CV_{max}). Labels were assigned based on a precursor ion's CV_{max}. A CV bin was assigned a positive label if it was greater than both 0.5 and the average proportional intensity of the second-best CV setting for precursors sharing that CV_{max} (**Figure 5.2**). For example, if the max CV setting for peptide precursor X was -45 V, because the second-best setting for these precursors is -50 V with an average of 0.52 (**Supplemental Figure 5.1**), an individual setting would be

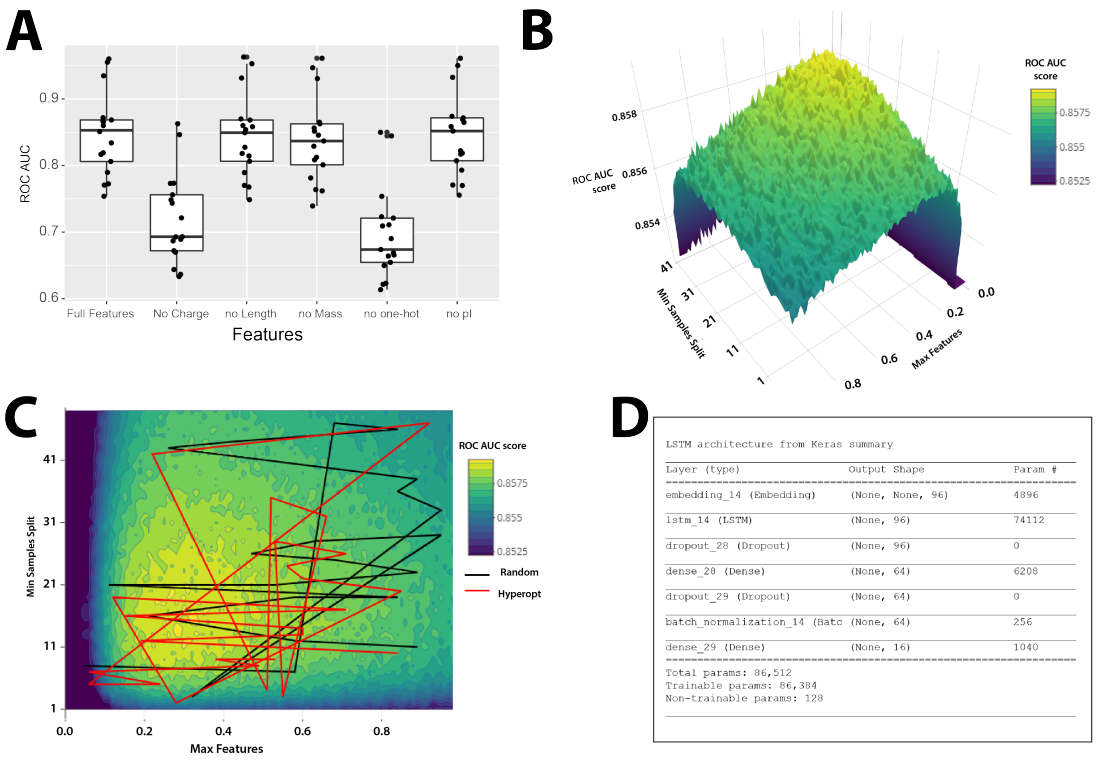
required to be >0.52 to receive an additional positive label. This labelling scheme allowed for more conservative label prediction with a focus on the top 1-3 labels, additional discriminatory ability regarding co-occurring labels, and a compensating effect for peptide precursor groups with broader distributions. Finally, precursors with non-consecutive labels were removed from the dataset, as they were considered reflective of potentially spurious identifications or stochastic events that may undermine the predictive performance of the model. This group included <5% of all precursors. The R script (Data/human/Preprocessing_FeatureMaxCVIsolation.R) for class labeling is available on GitHub.

Three labelled peptide precursor groups were utilized in our analysis. Two were derived from the human precursor dataset, a 70% training set and a 30% holdout set for testing and mitigating overfitting. The third group included all non-overlapping precursors (99.7%) from the *E. coli* dataset for external evaluation of the models' performance. The human dataset was split in an iterative fashion that preserved label distributions between the full and split datasets (**Supplemental Figure 5.2**) using the IterativeStratification function in scikit-multilearn[50] (version 0.2.0). The human 70% training set was used for optimizing the model hyperparameters. After optimizing the models with the 70% human training data, final models were trained using the full human precursor dataset, and the full model was used to make predictions against the true test set of non-overlapping *E. coli* precursors.

Machine Learning - Random Forest The random forest (RF) model received as inputs: peptide length, charge state, mass, isoelectric point (pI), and one-hot encoded sequence.



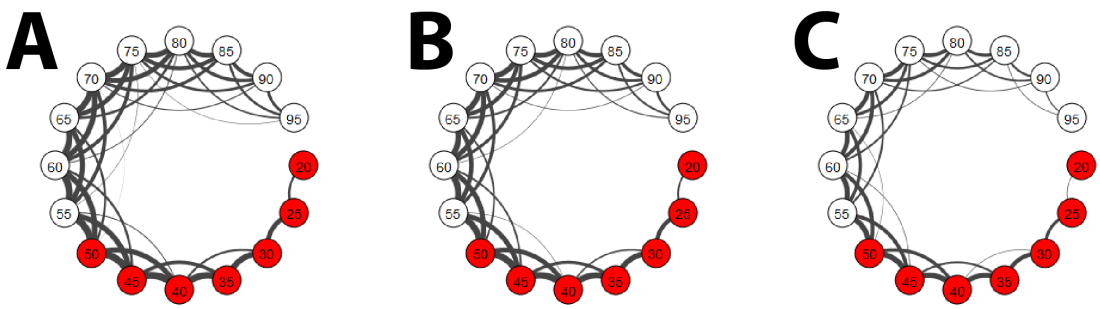
Supplementary Figure 5.1: Average proportional intensity for all precursors grouped by CV_{max}. Precursors with greater magnitude CV_{max} exhibit broader transmission distributions.



Supplementary Figure 5.2: Optimization of Features and Hyperparameters. (A) Boxplot of ROC-AUC score distribution when excluding features from the random forest model. Each point represents performance of single CV label, with bold line indicating mean score of all CVs. (B) Three-dimensional ROC-AUC score landscape for simple random forest model when using 100 estimators and varying two hyperparameters: min_samples_split and max_features (the two hyperparameters contributing the most to performance variability), with yellow indicating higher scores and blue indicating lower scores. Landscape contains many local maxima. (C) Two-dimensional ROC-AUC score landscape with a red and black line segment indicating hyperparameter space explored by the hyperopt and random search optimization functions, respectively. Here we depict every other model tested by each of the two function in the interest of clarity. (D) Summary of the LSTM neural network architecture.

Length and charge were both derived from the original peptide search results and labelling process. Mass and pI were generated using the pyteomics[51,52] (version 4.3.3) package in Python based on the unmodified sequence of the peptide. The sequence was one-hot encoded using Keras (version 2.2.4) with 22-member alphabet that included the 20 standard proteinogenic amino acids as well as the common modifications N-terminal acetylation and oxidation of methionine. The multilabel random forest classifier was developed based on the use of the OneVsRestClassifier wrapper function in tandem with the RandomForestClassifier function in scikit-learn[53] (version 0.23.0) in Python. Although all input features improved performance for some labels, charge and one-hot sequence were crucial to performance of the model (**Supplemental Figure 5.3A**).

Three different strategies were used to optimize the hyperparameters for the random forest: a parallel grid search, a random search using the RandomizedSearchCV function from scikit learn, and a Bayesian search using the hyperopt[54] package (version 0.2.4) in Python. Hyperparameter optimization focused on five parameters: (1) number of trees in the forest("n_estimators"), (2) the maximum depth of the trees("max_depth"), (3) the number of features to use at each split("max_features"), (4) the minimum number of samples required at each leaf node ("min_samples_leaf"), and (5) the minimum number of samples required to split a node ("min_samples_split"). The grid search tested 2191 total hyperparameter combinations (**Supplemental Figure 5.1**) largely in parallel using the compute resources and assistance of the UW-Madison Center for High Throughput Computing (CHTC) in the Department of Computer Sciences. Due to computational limitations and the



Supplementary Figure 5.3: Mapping label combinations.. Network plot connecting any co-occurring CV labels with thickness of connecting line indicating frequency of label combination. (A) Combinations for all human precursors (B) Combinations for training human precursors (C) Combinations for testing human precursors. Both subsets mirror label combination distribution for human precursors overall.

sequential nature of the random and Bayesian search, each was run for 50 iterations, with weighted ROC-AUC used as the scoring metric. Despite the abundance of local maxima and minima even when varying only two hyperparameters the hyperopt search was able to identify fruitful areas of the parameter space (**Supplemental Figure 5.3C**). In the interest of balancing recall and precision, a hybrid set of hyperparameters was used, combining aspects of the Bayesian and grid optimization. The final set of hyperparameters utilize: (1) 450 trees, (2) a maximum depth of 30, (3) 0.3807 of the features at each split, (4) a minimum of 6 samples per leaf node, (5) a minimum of 9 samples to split a node. The given hyperparameters for any machine learning problem can be infinitely iterated, but we have identified here a strong hyperparameter set that significantly increased performance as compared to the default hyperparameters. The Python code for all three hyperparameter optimizations is available on GitHub.

In order to mitigate the label imbalance between the most common CV labels and the rarest, training data for the random forest was labelled using lowered thresholds for positive labels (**Supplemental Figure 5.4**). These thresholds are described in **Table 5.4** and were applied to the 70% human training data when testing against the 30% test set (**Table 5.3**). These thresholds were also applied to the full human data when training the random forest for validation against the *E. coli* precursor dataset **Table 5.4**).

Machine Learning - Neural Network The LSTM NN was built with keras, which is a high-level interface to tensorflow (version 2.1). The exact anaconda environment can be

Training Set	Precision	Recall	ROC AUC	Threshold Accuracy (0.5)	F_2
True	0.562 +/- 0.005	0.747 +/-0.004	0.862 +/- 0.002	0.861 +/-0.001	0.611 +/-0.005
Tuned	0.558 +/- 0.004	0.836 +/-0.003	0.862 +/-0.002	0.835 +/-0.001	0.808 +/-0.003

Table 5.1: 5-fold Cross Validation of Random Forest Using True or Tuned Threshold for Training Labels.. Mean and standard deviation for performance metrics of 5-fold cross validation of the random forest using the 70% human training dataset. Training precursors at each fold used true labels described in Figure 2 or were re-labelled using the lower tuned threshold (Supplemental Figure 5). Using more permissive threshold for training data substantially increased recall with minimal decreases in other performance metrics, despite using true labels for all test sets.

CV Setting	Threshold
-20	0.25
-25	0.25
-30	0.25
-35	0.31
-40	0.28
-45	0.25
-50	0.25
-55	0.05

CV Setting	Threshold
-60	0.02
-65	0
-70	0
-75	0
-80	0
-85	0
-90	0
-95	0

Table 5.2: Random Forest Training Label Thresholds. Proportional intensity threshold required for precursor to be given CV label in the training dataset for the random forest classifier. Label thresholds were lowered for training data in attempt to counteract the label imbalance between more CVs of higher magnitude and those in the center.

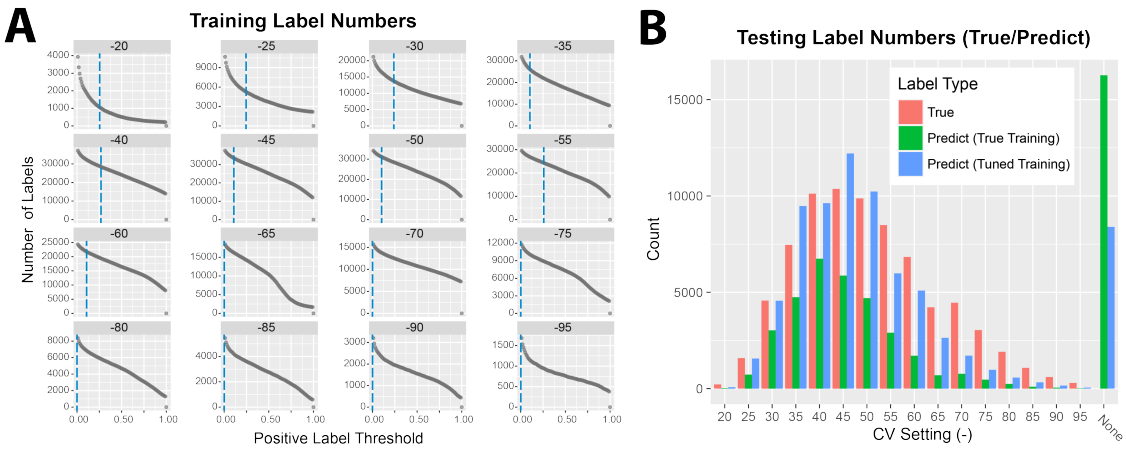
created using the file on the github repository at neural_net/tensorflow21.yml. Precursor charge was appended to the front of the peptide sequence, and the string of charge and sequence was converted to a string of integers for input to the neural network. Peptides with length less than 50 were padded with an integer encoding "end" to ensure all inputs were the same length. Neural network output was a string of 16 probabilities corresponding to each of the possible CV values from -20 volts to -95 volts in steps of -5 V. A positive class was assigned for any prediction > 0.5. The general neural network structure used four layers with two dropout steps and one batch normalization: (1) an embedding layer that converted the length 51 integers into vectors of real-value numbers, (2) the LSTM layer, (3) dropout, (4) dense with ReLU activation, (5) dropout, (6) batch normalization, (7) dense output with sigmoid activation. Binary cross-entropy loss was used with the Adam optimizer. Keras-tuner was used to optimize the LSTM hyperparameters targeting high ROC-AUC on the validation set with the hyperband algorithm. The hyper-parameter optimization was done using 5-fold cross validation within the 70% human training data slice. Possible hyperparameters were: (1) number of outputs from the embedding layer, from 32 to 96 in steps of 32, (2) number of outputs from the LSTM layer from 32 to 96 in steps 32, (3) dropout proportion from 0 to 0.5, (4) number of outputs from dense layer 1 from 32 to 96 in steps of 32, (5) dropout proportion from 0 to 0.5, (6) learning rate from 1e-4 to 1e-2 with log sampling. The optimal parameters were: (1) embedding layer output of 96, (2) LSTM output of 96, (3) dropout #1 of 0.32255, (4) dense output of 64, (5) dropout #2 of 0.02152, (6) learning rate of 0.001758. The best hyperparameters were then used to find

the best number of epochs with early stopping monitoring the validation set ROC-AUC in 5-fold cross validation, and the 52 epochs was best. A model was then trained using all the 70% human training data, and the performance was evaluated using the held-out 30% human test set. A second model was then trained using all 100% of the human precursors, and this model was evaluated with the second external test set of 40,000 precursors from *E. coli*.

Data and Code Availability Raw mass spectrometry data and MS-Fragger search outputs are available from the MassIVE[55] repository (massive.ucsd.edu project MSV000085707, password: faims) or Pride (PXD021174). Skyline documents containing quantitative data for each peptide are available on Panorama[56] (https://panoramaweb.org/AOueYU.url) The skyline reports, processed data, and machine learning code is available on GitHub.

Results

Data Overview Previous work demonstrated the benefit of proteomics data collection methods combining multiple CVs; the greatest increase in peptide identifications was observed when utilizing settings spaced 15V apart[14], suggesting that peptide ions have a FAIMS CV transmission range of less than 15V. In order to better resolve the degree of transmission, we acquired data using the narrow spacing of 5V bins (**Figure 5.1A**). After processing, a total of 128,402 and 42,719 unique peptide ions (redundant peptide sequences with different charge states were counted multiple time) were identified in human and *E.*



Supplementary Figure 5.4: Alternate training label threshold for random forest. (A) Number of positive labels as the labelling threshold is increased for each CV bin. Dotted green line indicates 0.5 threshold. Dotted blue line indicates lowered threshold tuned to better match prediction numbers to true label numbers in the random forest. (B) Number of true (red) and predicted labels that result from training the random forest using the true labelling scheme (green), or a lower, tuned labelling threshold (blue). When training with more conservative true labelling scheme, random forest substantially underpredicts labels at all bins and results in high number of peptide ions with no predicted label ("None")

coli samples, respectively. Most peptides were charge state 2 (56.7%) but many were also 3 (35.1%), 4 (7.1%) and 5 (1.1%) (**Figure 5.1B**).

This high-resolution FAIMS CV data from 5V steps showed that although transmission of individual peptide ions can be quite variable, in aggregate, peptide ions exhibit substantial transmission (>50% of max) across 1-3 of the 5V bins tested here. When grouping peptide ions by their most transmissive CV (CV_{max}), groups exhibit different average transmission distributions. Peptide ions with maximum transmission at lower magnitude CVs close to 0 tend to have narrower shapes, while those that favor higher magnitude voltages exhibit wider distributions. Although in aggregate the peptides seem to have a normal distribution (**Supplemental Figure 5.1**), individual cases are often much more irregular in distribution (**Figure 5.1B**), further motivating our multilabel approach.

Relationship Between, Charge, CV, and Retention Time Unsurprisingly, we observed charge state to be the most important factor in deciding the distribution of peptides across the CV space. This relationship between peptide or protein charge state and differential ion mobility has been widely reported in many types of ion mobility[24–27] including FAIMS[14]. Although there appears to be a broadly inverse relationship between peptide charge and CV, it is not linear, and higher charge states exhibit narrower CV distributions. For example, 90% of charge state 5 peptide ions were observed between CV -45 V and -55 V, with -50 being the most populated bin, whereas only 23% of charge state 2 peptides are observed in this range (**Figure 5.1D**).

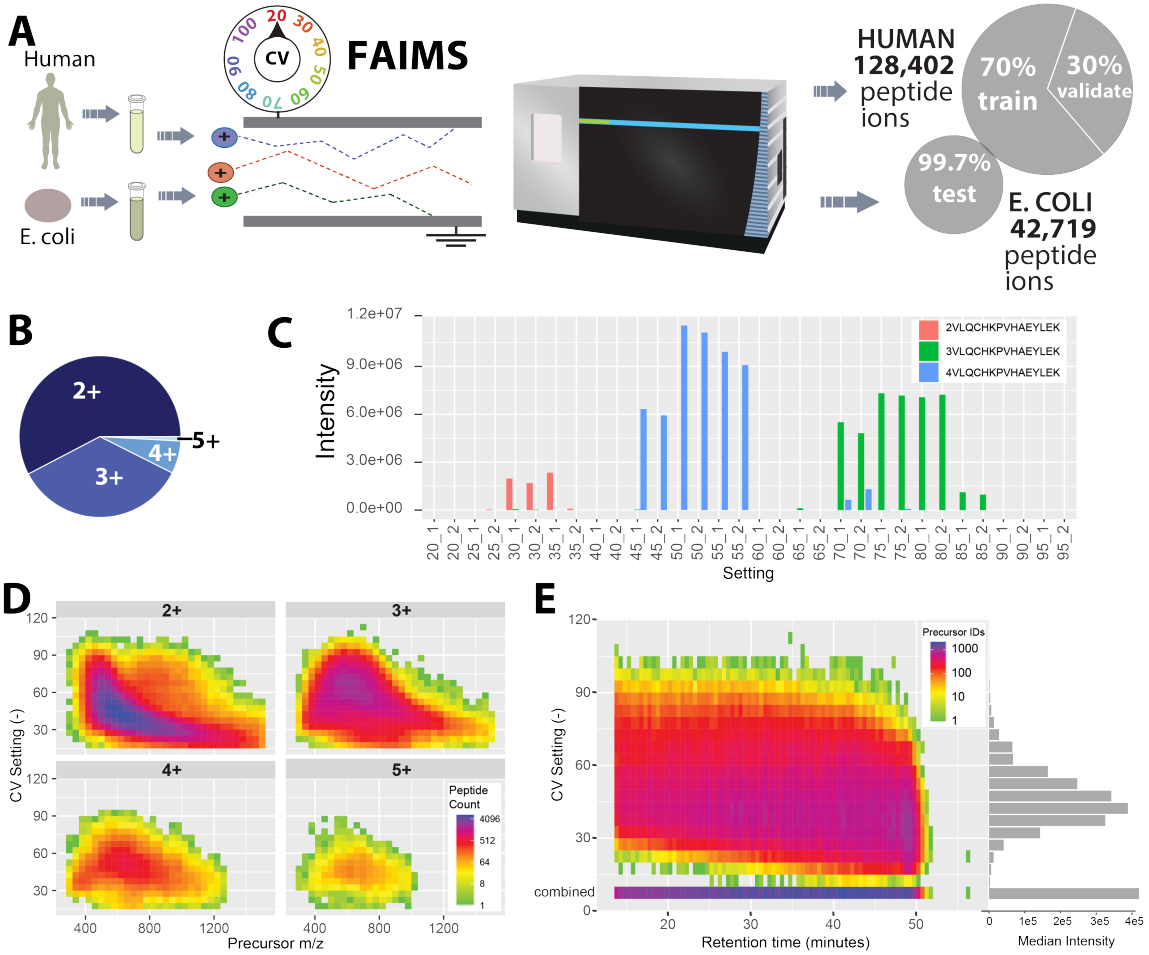


Figure 5.1: Experimental and Data Overview. (A) Experimental Design. Human and *E. coli* protein lysate were denatured and digested with trypsin and lys-C before being analyzed using a series of data-dependent single-CV one-hour analyses. More than 120,000 and 40,000 peptide precursor ions were quantified for Human and *E. coli* samples, respectively. (B) Pie chart depicting the proportion of identified peptide ions from each charge state for all human peptides, a majority of which are 2+. (C) Intensity distribution for three charge states of example peptide VLQCHKPVHAEYLEK across experimental CV range with replicates shown. Different charge states exhibit different CV_{max} as well as different transmission distributions. (D) Density plot of number of human peptides observed at each charge state and compensating voltage across the mass-to-charge range. Observe distinct distributions by charge state. (E) Density plot depicting number of human peptide ions observed at each compensating voltage for the duration of the method, sorted into 0.5 minute RT bins. Horizontal bar plot depicts the median peptide intensity at each compensating voltage. Combined indicates number of unique human precursors identified across all FAIMS CV experiments at each point in the gradient, with the horizontal bar representing the median intensity of all identified human peptide ions.

Given the prominence of mass and charge state as features in our models, one could imagine the utilization of a more interpretable linear regression model based on these two variables alone, as has been used previously[57]. However, when relying on these two features but removing one-hot encoded sequence we observe a dramatic drop in performance (**Supplemental Figure 5.2A**), indicating the necessity of a more complex model allowing incorporation of sequence information.

FAIMS can enhance detection of low abundance peptides by filtering out high abundance components of complex peptide mixtures. At more extreme CV settings, we observed a greater simplification effect, as indicated by reduced peptide observations. When overlaying all precursors from across the full CV range we identify >2000 peptide ions per min for much of the method duration (**Figure 5.1E, "combined"**) indicating the increased depth provided by FAIMS. Further supporting the theory that greater filtering leads to increased detection of low abundance peptides, we observed decreased median intensities at the edges of our measured CV range (**Figure 5.1E**). Median intensity largely mirrored precursor observations across CV space with a maximum at -45 V and minimum at -120 V, the most and least populated CV bins, respectively.

A Model Solution Based on the data shape we aimed to develop a model that would take peptides as input and output transmissive compensating voltages. Although compensating voltage is in reality a continuous variable, its discrete nature here led us to a classification model where peptide ions would be given a label for each compensating voltage included

in our experimental dataset(5V bins from 20 to 95). A positive label(1) would indicate substantial transmission at that CV setting, while a negative label(0) would indicate poor or no transmission. We selected a multilabel classification framing with the objective of identifying the 1-3 most transmissive compensating voltage settings. Two technical challenges arose from this multi-label strategy and objective: (1) converting continuous intensity into a binary label indicating substantive transmission at an individual setting (2) Compensating for large label imbalances commonly found in classification problems[58].

Initially, true positive labels were assigned to all settings with proportional intensity >0.5. This labelling scheme led to common co-labelling of peptide ions transmitted at higher magnitude voltages (-45 to -95) due to their broader transmission distribution (**Supplemental Figure 5.1**),), including ions with seven or more labels. In hopes of increasing discriminatory ability, labels were assigned based on a precursor's highest intensity CV (CV_{max}). A CV bin was assigned an additional positive label if it was greater than both 0.5 and the average proportional intensity of the second-best setting for peptide ions sharing that CV_{max} (**, Figure 5.2**). This labelling scheme allowed for more conservative label prediction, with a focus on the top 1-3 labels discussed above, and a compensating effect for peptide ion groups with broader distributions. This labelling scheme also reduced the number of ions with 6 or more labels by >30%.

Label imbalances make classification tasks difficult[59], which can be compounded in multilabel schemes[60,61] due to the wealth of negative labels. In our full human peptide ion dataset (128,000 peptide ions) our rarest label (CV=-20) includes 727 peptide ions,

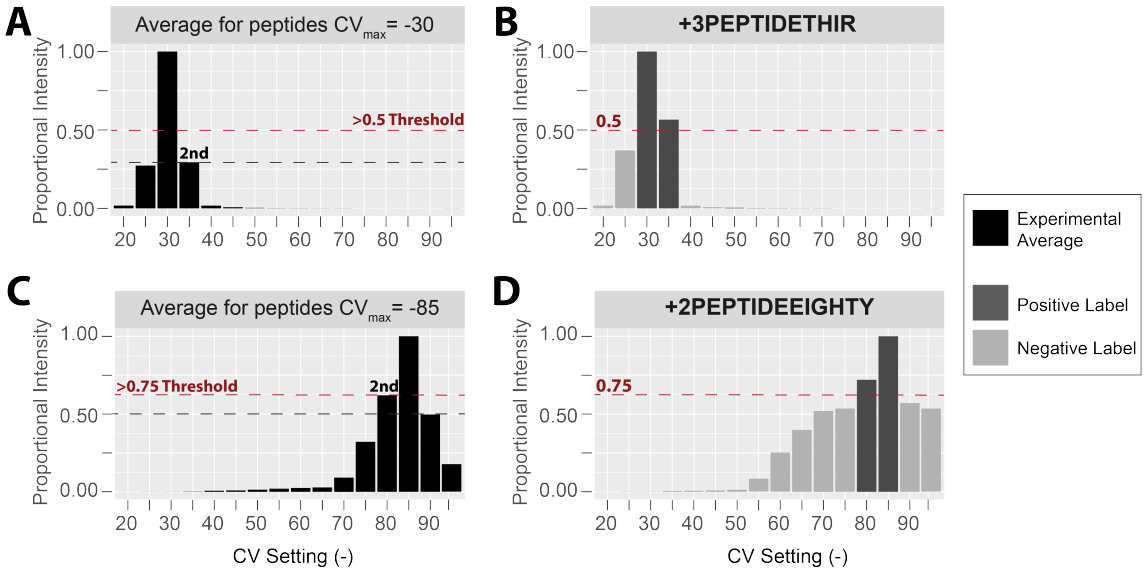


Figure 5.2: True Positive CV_{max}-specific Labelling Scheme. Labelling thresholds are based on the average transmission distribution of peptide ions that share a CV_{max}. The threshold is set at either 0.5 or the average proportional intensity of the second-best compensating voltage, whichever is higher. (A) On average peptide ions with CV_{max} = -30 have a proportional intensity of 0.3 at -35 V, their second-best setting. (B) Because in this case of 0.3 < 0.5, our theoretical example peptide ions receives a positive label for each setting with > 0.5 proportional intensity. (C) In contrast, precursors with CV_{max} = -85 have an average proportional intensity of 0.75 at their second-best setting, -80 V. (D) 0.75 > 0.5 so our second example peptide receives a positive label for each setting with > 0.75 proportional intensity. This strategy leads to the theoretical example +2PEPTIDEEIGHTY receiving the two labels with the greatest transmission rather than five resulting from a 0.5 threshold.

composing <1% of the peptide ion population. These imbalances lead to substantial underprediction from our random forest model across all labels with >40% peptide ions receiving no predicted labels (**Supplemental Figure 5.4**). A common solution in the field involves oversampling positive labels through resampling or addition of synthetic training data[62–64]. To increase the number of predicted labels by the random forest, we allowed the model to see more of each label class during training by lowering the proportional intensity threshold used to assign additional labels in the training data without changing the test set label threshold. This increased training label number, leading to an increase in predicted labels and predictions that more closely matched our true labels to improve recall and F2 score (**Supplemental Figure 5.4**) at the minor expense of thresholded accuracy (**Table 5.1**).

Performance: Human Tests Metrics The human peptide data was split into 70% training and a 30% testing based on multi-class membership with scikit-multilearn[50] (**Figure 5.1A**), which preserved the proportional distribution of labels (**Supplemental Figure 5.2**). Both models were effective at predicting optimal CV settings (**Table 5.3**) with the RF and NN producing binary accuracies of 0.81 and 0.90, respectively, compared to 0.49 for a uniformly random dummy classifier. The NN showed greater precision (NN = 0.66 vs. RF = 0.44) and reduced false positives, resulting in a greater ROC-AUC score (0.94 vs. 0.86). The RF model favored selection of a greater number of relevant elements at the expense of false positives leading to increased recall (RF = 0.83 vs. NN = 0.56) and F2 score (0.70 vs. 0.58).

Metric	Random forest	Neural network	Ensemble
Binary accuracy	0.8133	0.9073	0.8970
precision	0.3900	0.6610	0.5774
recall	0.8271	0.5581	0.7136
F_2	0.6756	0.5760	0.6814
ROC AUC	0.8977	0.9414	0.9344
At least one match	0.9018	0.7535	0.8403

Table 5.3: Performance Metrics when Training and Testing using Human Dataset. Binary accuracy, precision, recall, F2 and ROC-AUC are all calculated when training each model and an averaged ensemble using 70% of the human data and testing against the remaining 30%. At least one match describes the proportion of peptide ions that share at least one positive label between experimental labels and predicted labels.

One utility of predicting FAIMS CV is for targeted method design. Given the continuous nature of compensating voltage, perfect matches between the profiles of predicted and experimental settings are uncommon. However, even a reduction in the possible CV range from 100 V (twenty possible) to 20V (four possible) would be useful in experimental design. With this utility in mind, we quantified the proportion of peptide ions in which at least one label was predicted correctly. We found that 75% and 90% of peptide ions had at least one label in common between the predicted and experimental labels in the neural net and random forest, respectively.

Performance: *E. coli* Test Metrics *E. coli* labels were assigned in the same manner as the human test data labels (Materials and Methods). When both models were trained using the full human peptide ion data set and validated against a group of *E. coli* peptide ions novel to the model, overall prediction metrics were quite similar, with only minor declines in performance (Table 2). All metrics were similar suggesting the recognition of underlying peptide ion properties by both algorithms and the application of these properties to transmission predictions.

ROC curve and CV-specific Performance Labels Receiver operator characteristic curves (ROC-AUC) for *E. coli* label predictions help to visualize the per-CV performance of the models. Both models exhibit variable performance across the different CV labels(**Figure 5.3**). Interestingly, the algorithms exhibit greatest area-under-the-curve for labels with

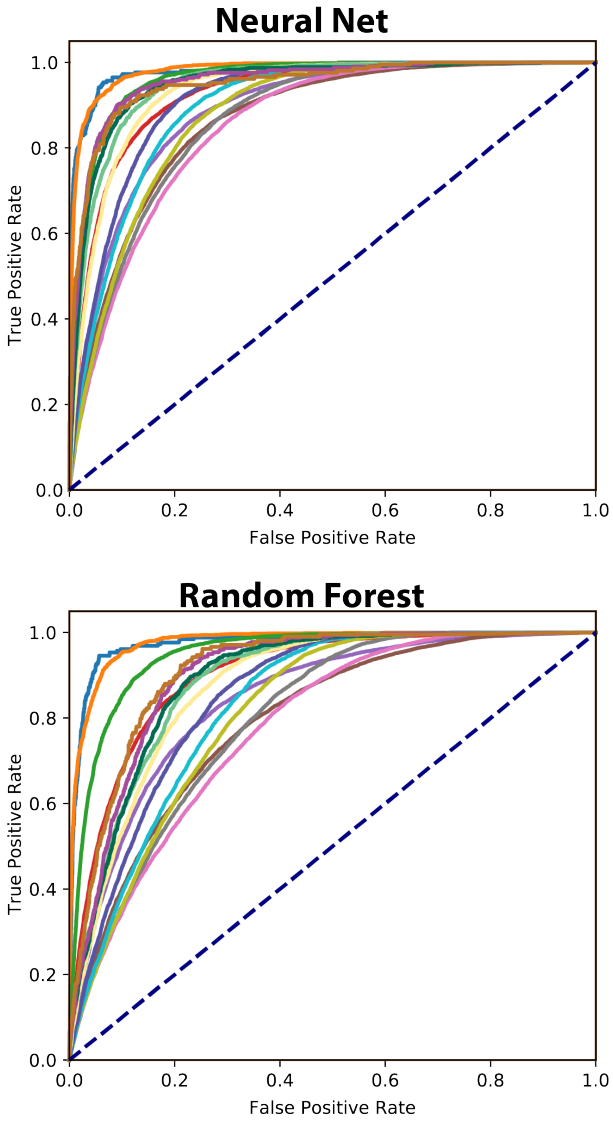
Metric	Random forest	Neural network	Ensemble
Binary accuracy	0.8023	0.9027	0.8950
precision	0.3789	0.6503	0.5830
recall	0.8084	0.5487	0.6837
F_2	0.6590	0.5664	0.6609
ROC AUC	0.8834	0.9368	0.9280
At least one match	0.885	0.7447	0.8195

Table 5.4: Performance Metrics when Training with Human Dataset and Testing against *E. coli* Dataset. Binary accuracy, precision, recall, F2 and ROC-AUC are all calculated when training each model and an averaged ensemble using 100% of the human data and testing against the full *E. coli* precursor dataset. At least one match describes the proportion of peptide ions that share at least one positive label between experimental labels and predicted labels.

fewer observations, approaching unity in the ROC curve for -25 and -20 CV. The models struggle most within the middle voltage range (-45 to -55), despite a wealth of examples in the training data. This may be caused by the peptide ions in this range commonly having >3 labels, leading to difficulty in identifying discriminatory properties. This effect may also be a result of conservative labelling scheme that disfavors middle labels by providing higher thresholds, leading to the appearance of poor performance in this range.

Individual Peptide Examples To better understand the prediction qualities of each model, we picked illustrative examples with true labels spanning the CV space from the *E. coli* test set. This analysis revealed that the distribution of CV prediction probabilities mirrored the true intensity distributions (**Figure 5.4**). These individual peptides examples highlight the differences in metrics - described in aggregate above - between the two models. The neural net provides a more selective set of labels, while the random forest generates wider distributions with higher probabilities, resulting in more false negatives and false positives for the NN and RF, respectively. Despite these differences in label predictions, both models reflect the asymmetrical transmission distributions that broaden as the voltage increases in magnitude. These examples illustrate the capacity for machine learning to parameterize the chemo-physical properties of peptides to predict their ion mobility in CV space.

Based on this observation of complementary predictions, we wondered if an ensemble of the model's predictions would provide improved performance. The class predictions for each model were averaged, and the same metrics were re-computed using the average



CV(-)	AUC_{NN}	AUC_{RF}
20	0.98	0.97
25	0.98	0.98
30	0.96	0.95
35	0.93	0.90
40	0.89	0.84
45	0.87	0.79
50	0.85	0.78
55	0.87	0.80
60	0.88	0.82
65	0.90	0.83
70	0.91	0.85
75	0.93	0.88
80	0.95	0.89
85	0.95	0.89
90	0.96	0.90
95	0.95	0.91

Figure 5.3: Receiver Operator Characteristic Curve for *E. coli* Predictions. Receiver operator characteristic (ROC) curve for each compensating voltage label in each of the two models, the neural net and the random forest when predicting labels for *E. coli* peptide ions. Observe slightly better area-under-the-curve (AUC) from neural net across all CV labels. Both models exhibit best performance on voltages closest to zero.

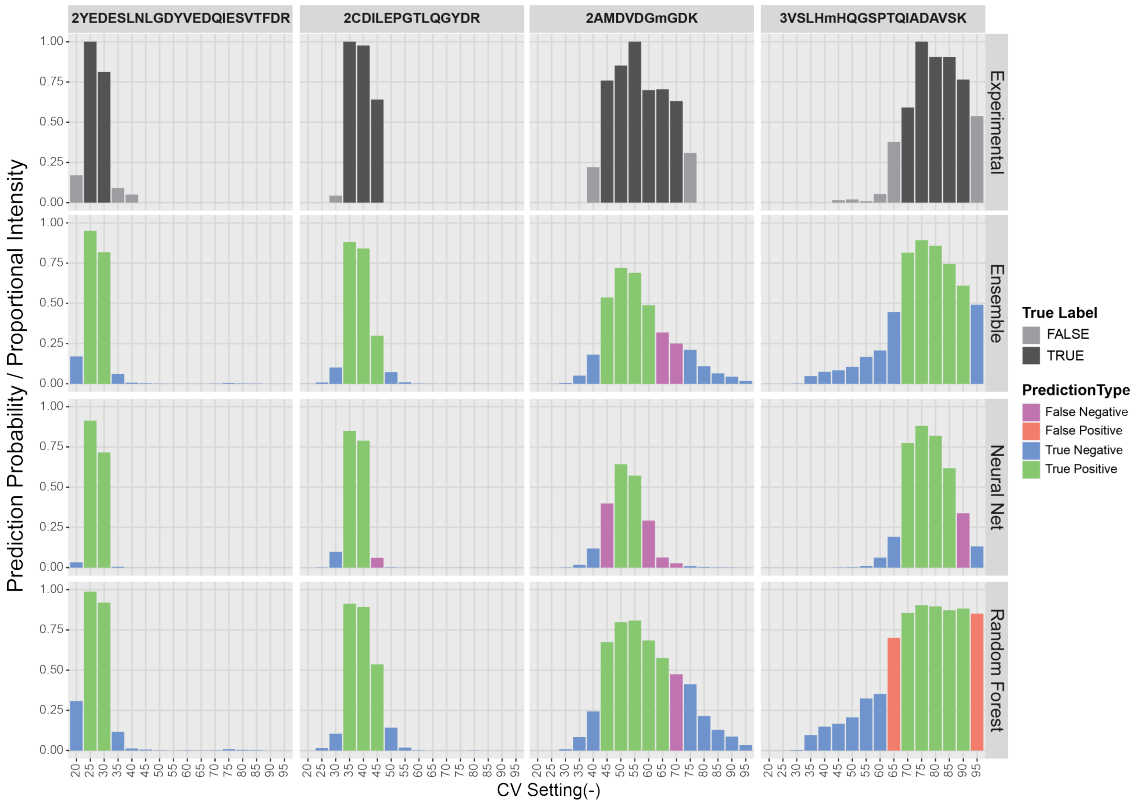


Figure 5.4: Example Peptide Prediction Probabilities. Prediction probabilities and proportional intensities for the tested CV settings for each of four example peptide ions: YEDESLNLGDYVEDQIESVTFDR +2, CDILEPGTLQGYDR +2, AMDVDGM(Ox)GDK +2, and VSLHM(Ox)HQGSPTQIADAVSK +2. For experimental data, dark grey indicates positive labels and light grey indicates negative labels. For predictions, green indicates true positives blue indicates true negatives, red indicates false positives and purple indicates false negatives.

probabilities (**Table 5.3 and Table 5.4**). Nearly all the metrics fell between the original values of the two separate models. However, F2 score, which summarizes both precision and recall as the harmonic mean, was higher for the ensemble than for the either model alone, suggesting a benefit in combining the two separate prediction strategies.

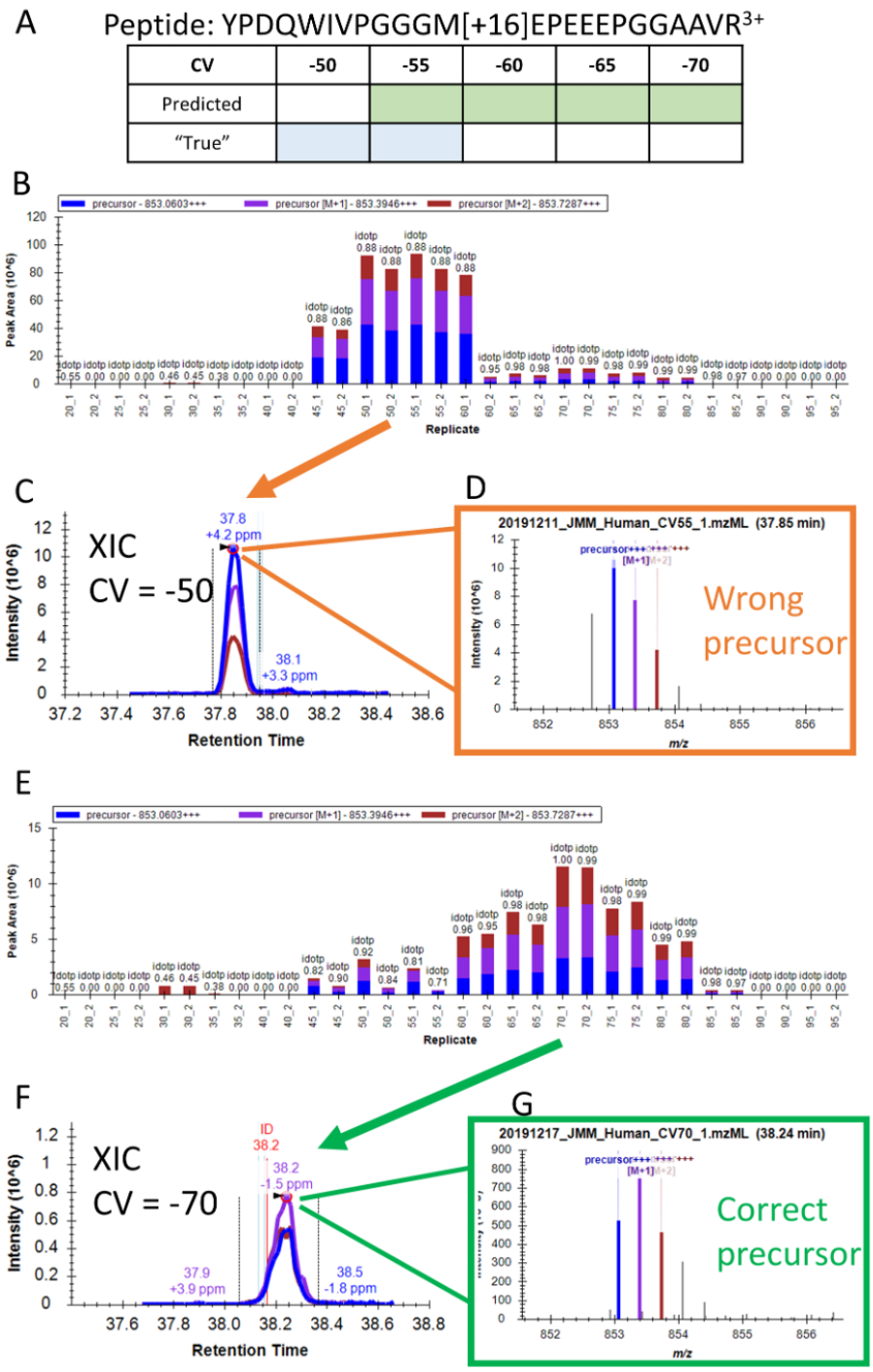
What Drives Incorrect Predictions? To better understand what causes the models to be incorrect, an incorrect prediction was traced back to the raw data (**Supplemental Figure 5.6**). This peptide ion was predicted to transmit through FAIMS at CVs -55, -60, -65, and -70, but was assigned true labels of -50 and -55. Inspection of the precursor ion areas across the individual CV runs appears to show high signal in CV settings of -50 and -55 as the peptide ion was labeled, but there was an apparent low level of precursor signal in some of the predicted channels with higher precursor shape matches to the theoretical envelope pattern (idotp, as generated by Skyline). A view of the extracted ion chromatogram (XIC) from the CV of -55 shows one prominent peak with a much smaller peak in the future by 0.3 minutes. Further inspection of the precursor isotope pattern of the more prominent peak from the XIC in the "true" label CV -55 analyses showed that the incorrect peak was chosen; clearly the M+1, M+2, and M+3 peaks of a different peptide were integrated. After correcting the integration to the second, smaller peak, the distribution of integrated peaks better matched the values predicted by the neural network. Therefore, in this case, the model predictions were more accurate than the "true" labels. This highlights the importance of generating ground truth data for model training, and also highlights the difficulty of data labeling for

this task of peptide ion mobility through FAIMS.

Discussion

FAIMS offers an opportunity to increase proteomic depth and sensitivity without increasing sample preparation time. However, the promise of improved targeted sensitivity, further developments in targeted FAIMS methods, and new ideas for discrimination of true and decoy matches is currently slowed by necessary method optimization experiments. We present here a pair of machine learning models that when applied individually or in tandem allow the circumvention of the empirical FAIMS method development process by predicting optimal compensating voltages for peptides of interest. The models display excellent predictive performance when validated against a naïve collection of *E. coli* peptide ions exhibiting 0.88 and 0.93 ROC-AUC for the random forest and neural net, respectively. The models also show the capacity to approximate the transmission distributions of individual peptides across voltage range used here.

The multilabel classification strategy utilized here represents a substantial divergence from the regression-based algorithms commonly used for ion mobility behavior prediction. This strategy capitalizes on the distinct transmission distribution of individual peptides and interprets and predicts a two-dimensional shape. Although we believe multi-label classification was the optimal approach to frame the problem, this strategy introduced several technical challenges, including variable CV transmission profiles and label imbalances. We adapted our labelling scheme to provide the most useful prediction information, which we



Supplementary Figure 5.5: Example of "wrong" prediction due to incorrect automated peak picking in Skyline.. (A) Peptide sequence and table of predicted and true labels. Panels B-D show the automated peak picking by Skyline without adjustment, and E-G show after adjustment. (B) Peak areas for the peptide precursor across FAIMS CV duplicates. The most intense signal is at CV = -50 and CV = -55 replicates, which were assigned as true labels. The average peak areas of CV = -45 and CV = -60 duplicates would be less than ½ of the maximum and were not chosen as true labels. (C) Extracted ion chromatogram (XIC) of the precursor in one of the CV = -50 replicates showing good peak shape. (D) Peptide precursor ion envelope from the apex of the peak in (C) reveals that the signal comes from M+1, M+2, and M+3 isotopes of a different peptide precursor. (E) Peak areas over FAIMS CV duplicates after adjusting integration to the correct peak. (F) XIC of the peak in the CV = -70 replicate where the peptide was also identified. (G) Peptide precursor ion envelope at the apex of the peak in (F) showing the correct isotope shape.

decided was a small set of optimal CV settings rather than a large range.

The models performed well on a true test of model-naïve *E. coli* peptide ions, with the strongest performance observed for CV bins with the fewest training examples. This performance suggests generalizability of these models to tryptic peptides regardless of sample source without a need for additional data collection. Interestingly, the models performed worst when making peptide transmission predictions for the most populous voltage range of the training data. This decreased performance indicates that although we have the capacity to survey a greater number of peptide ions, it may not be beneficial to predictive accuracy.

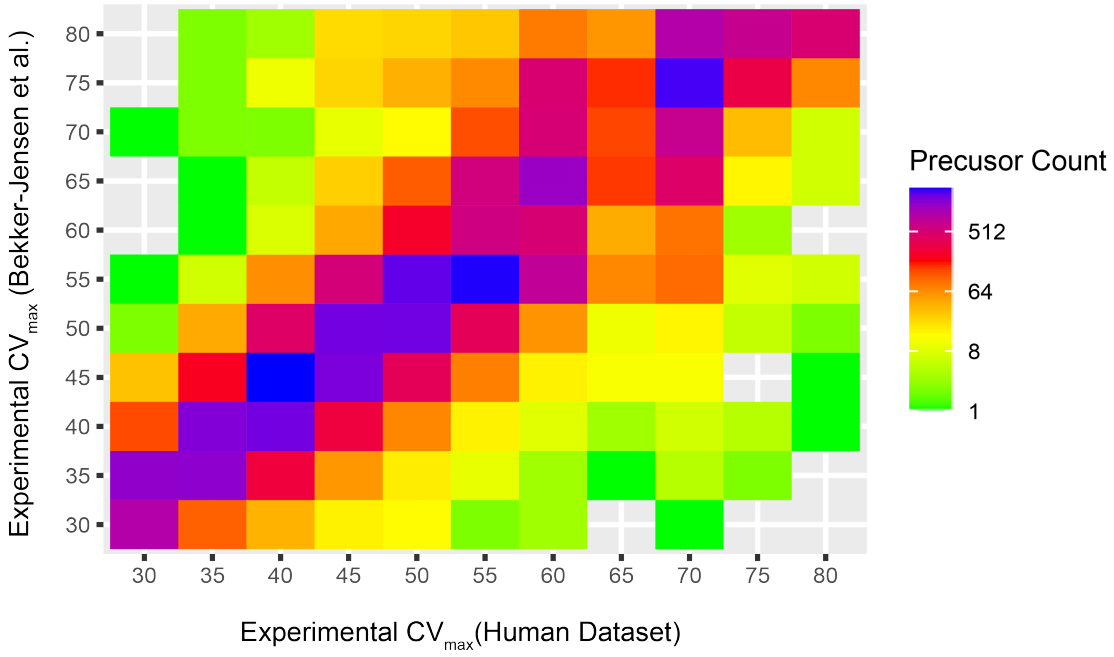
When combining peptide identifications across our entire voltage range, we identified more than 120,000 human peptide precursors using a one-hour LC gradient method. These peptide ions represent the peptide space accessible to a FAIMS-integrated mass spectrometry method. Our CV prediction methods allow for access to this vastly expanded peptide ion space for any organism or any tissue. These models add to the ever-expanding toolkit for developing LC-MS methods *in silico*, which includes predictions for spectral fragmentation and retention time. When used in tandem, these tools allow for greater depth in untargeted experiments (e.g. DIA) and increased sensitivity and quantitative accuracy for targeted experiments (e.g. PRM). This could be especially useful for challenging tissues or cell types, such as plasma or cerebrospinal fluid.

We have also made the prediction models available as a webtool to allow for users to generate transmission predictions for peptides of their choice. As an example, we entered

the coronavirus peptide +2QQTVTLLPAADLDDFSK around which Renuse and colleagues recently developed a targeted parallel reaction monitoring method using FAIMS for rapid diagnosis of COVID[65]. Our ensemble model would have classified voltages -30 and -35 as transmissive settings, overlapping with their chosen CV of -30, without any need for optimization experiments. Increasing development speed in this way can be crucial in diagnostic development especially in the case of infectious diseases.

Many parameters, such as those associated with chromatographic gradients, electrospray ionization, mass spectrometry instruments and data searching, can affect identification and quantification of peptide ions in data-dependent experiments. Despite these parameter differences, when comparing overlapping peptide ions from our human dataset to those recently collected by Bekker-Jensen and colleagues using FAIMS and substantially shorter gradients[21], we observe similar transmission patterns (**Supplemental Figure 5.6**).

Although our analysis here focuses on largely unmodified tryptic peptides, based on our success, we expect ions from modified peptides or metabolites will be amenable to prediction of FAIMS transmission. Our human peptide dataset included N-terminal acetylation and oxidation of methionine, increasing our peptide sequence alphabet to 22. The available residue alphabet could easily be expanded to include other commonly modified residues allowing for predictions of a wider peptide space. Given the base of these effective models, retraining for modified peptides and non-tryptic peptides should be straightforward given a set of at least 10,000 additional training examples. We expect that our success in modeling peptide transmission through FAIMS could potentially be



Supplementary Figure 5.6: Overlap of Max CV. Density plot indicating correlation of CV_{max} setting for 33,420 precursor ions commonly identified between the human dataset used here, and that from Bekker-Jensen et al[21]

replicated for lipids or polar metabolite transmission profiles. Our machine learning modeling results more generally will also inform future machine learning studies that use peptides as input. Altogether, we expect these tools to be widely adopted in studies that utilize FAIMS.

References

[1] S. J. Valentine, A. E. Counterman, C. S. Hoaglund, J. P. Reilly, and D. E. Clemmer, "Gas-phase separations of protease digests," *J. Am. Soc. Mass Spectrom.*, vol. 9, pp. 1213–1216, Nov. 1998.

[2] S. J. Valentine, M. D. Plasencia, X. Liu, M. Krishnan, S. Naylor, H. R. Udseth, R. D. Smith, and D. E. Clemmer, "Toward plasma proteome profiling with ion mobility-mass spectrometry," *J. Proteome Res.*, vol. 5, pp. 2977–2984, Nov. 2006.

[3] E. S. Baker, B. H. Clowers, F. Li, K. Tang, A. V. Tolmachev, D. C. Prior, M. E. Belov, and R. D. Smith, "Ion mobility spectrometry-mass spectrometry performance using electrodynamic ion funnels and elevated drift gas pressures," *J. Am. Soc. Mass Spectrom.*, vol. 18, pp. 1176–1187, July 2007.

[4] S. Valentine, "a. e. counterman, cs hoaglund-hyzer, de clemmer, intrinsic amino acid size parameters from a series of 113 lysine-terminated tryptic digest peptide ions," *J. Phys. Chem. B*, vol. 103, pp. 1203–1207, 1999.

[5] S. J. Allen, R. M. Eaton, and M. F. Bush, "Structural dynamics of Native-Like ions in the gas phase: Results from tandem ion mobility of cytochrome c," *Anal. Chem.*, vol. 89, pp. 7527–7534, July 2017.

[6] O. J. Hale, E. Illes-Toth, T. H. Mize, and H. J. Cooper, "High-Field asymmetric waveform ion mobility spectrometry and native mass spectrometry: Analysis of intact protein assemblies and protein complexes," *Anal. Chem.*, vol. 92, pp. 6811–6816, May 2020.

[7] J. P. Williams, L. J. Morrison, J. M. Brown, J. S. Beckman, V. G. Voinov, and F. Lermyte, "Top-Down characterization of denatured proteins and native protein complexes using electron capture dissociation implemented within a modified ion Mobility-Mass spectrometer," *Anal. Chem.*, vol. 92, pp. 3674–3681, Mar. 2020.

[8] M. Zhou, C. M. Jones, and V. H. Wysocki, "Dissecting the large noncovalent protein complex GroEL with surface-induced dissociation and ion mobility-mass spectrometry," *Anal. Chem.*, vol. 85, pp. 8262–8267, Sept. 2013.

[9] G. Nagy, C. D. Chouinard, I. K. Attah, I. K. Webb, S. V. B. Garimella, Y. M. Ibrahim, E. S. Baker, and R. D. Smith, "Distinguishing enantiomeric amino acids with chiral cyclodextrin adducts and structures for lossless ion manipulations," *Electrophoresis*, vol. 39, pp. 3148–3155, Dec. 2018.

[10] R. Wojcik, G. Nagy, I. K. Attah, I. K. Webb, S. V. B. Garimella, K. K. Weitz, A. Hollerbach, M. E. Monroe, M. R. Ligare, F. F. Nielson, R. V. Norheim, R. S. Renslow, T. O. Metz,

Y. M. Ibrahim, and R. D. Smith, "SLIM ultrahigh resolution ion mobility spectrometry separations of isotopologues and isotopomers reveal mobility shifts due to mass distribution changes," *Anal. Chem.*, vol. 91, pp. 11952–11962, Sept. 2019.

[11] K. Venne, E. Bonneil, K. Eng, and P. Thibault, "Improvement in peptide detection for proteomics analyses using NanoLC-MS and high-field asymmetry waveform ion mobility mass spectrometry," *Anal. Chem.*, vol. 77, pp. 2176–2186, Apr. 2005.

[12] P. V. Shliaha, N. J. Bond, L. Gatto, and K. S. Lilley, "Effects of traveling wave ion mobility separation on data independent acquisition in proteomics studies," *J. Proteome Res.*, vol. 12, pp. 2323–2339, June 2013.

[13] K. E. Swearingen, M. R. Hoopmann, R. S. Johnson, R. A. Saleem, J. D. Aitchison, and R. L. Moritz, "Nanospray FAIMS fractionation provides significant increases in proteome coverage of unfractionated complex protein digests," *Mol. Cell. Proteomics*, vol. 11, p. M111.014985, Apr. 2012.

[14] A. S. Hebert, S. Prasad, M. W. Belford, D. J. Bailey, G. C. McAlister, S. E. Abbatiello, R. Huguet, E. R. Wouters, J.-J. Dunyach, D. R. Brademan, M. S. Westphall, and J. J. Coon, "Comprehensive Single-Shot proteomics with FAIMS on a hybrid orbitrap mass spectrometer," *Anal. Chem.*, vol. 90, pp. 9529–9537, Aug. 2018.

[15] R. W. Purves, S. Prasad, M. Belford, A. Vandenberg, and J.-J. Dunyach, "Optimization

of a new aerodynamic cylindrical FAIMS device for small molecule analysis," *J. Am. Soc. Mass Spectrom.*, vol. 28, pp. 525–538, Mar. 2017.

[16] S. Prasad, M. W. Belford, J.-J. Dunyach, and R. W. Purves, "On an aerodynamic mechanism to enhance ion transmission and sensitivity of FAIMS for nano-electrospray ionization-mass spectrometry," *J. Am. Soc. Mass Spectrom.*, vol. 25, pp. 2143–2153, Dec. 2014.

[17] D. K. Schweppe, S. Prasad, M. W. Belford, J. Navarrete-Perea, D. J. Bailey, R. Huguet, M. P. Jedrychowski, R. Rad, G. McAlister, S. E. Abbatiello, E. R. Woulters, V. Zabrouskov, J.-J. Dunyach, J. A. Paulo, and S. P. Gygi, "Characterization and optimization of multiplexed quantitative analyses using High-Field Asymmetric-Waveform ion mobility mass spectrometry," *Anal. Chem.*, vol. 91, pp. 4010–4016, Mar. 2019.

[18] A. A. Shvartsburg, F. Li, K. Tang, and R. D. Smith, "High-resolution field asymmetric waveform ion mobility spectrometry using new planar geometry analyzers," *Anal. Chem.*, vol. 78, pp. 3706–3714, June 2006.

[19] H. J. Cooper, "To what extent is FAIMS beneficial in the analysis of proteins?," *J. Am. Soc. Mass Spectrom.*, vol. 27, pp. 566–577, Apr. 2016.

[20] Y.-Q. Xia, S. T. Wu, and M. Jemal, "LC-FAIMS-MS/MS for quantification of a peptide

in plasma and evaluation of FAIMS global selectivity from plasma components," *Anal. Chem.*, vol. 80, pp. 7137–7143, Sept. 2008.

[21] D. B. Bekker-Jensen, A. Martínez-Val, S. Steigerwald, P. Rüther, K. L. Fort, T. N. Arrey, A. Harder, A. Makarov, and J. V. Olsen, "A compact Quadrupole-Orbitrap mass spectrometer with FAIMS interface improves proteome coverage in short LC gradients," *Mol. Cell. Proteomics*, vol. 19, pp. 716–729, Apr. 2020.

[22] J. G. Meyer, N. M. Niemi, D. J. Pagliarini, and J. J. Coon, "Quantitative shotgun proteome analysis by direct infusion," *Nat. Methods*, vol. 17, pp. 1222–1228, Dec. 2020.

[23] S. J. Valentine, A. E. Counterman, and D. E. Clemmer, "A database of 660 peptide ion cross sections: use of intrinsic size parameters for bona fide predictions of cross sections," *J. Am. Soc. Mass Spectrom.*, vol. 10, pp. 1188–1211, Nov. 1999.

[24] A. R. Shah, K. Agarwal, E. S. Baker, M. Singhal, A. M. Mayampurath, Y. M. Ibrahim, L. J. Kangas, M. E. Monroe, R. Zhao, M. E. Belov, G. A. Anderson, and R. D. Smith, "Machine learning based prediction for peptide drift times in ion mobility spectrometry," *Bioinformatics*, vol. 26, pp. 1601–1607, July 2010.

[25] B. Wang, S. Valentine, M. Plasencia, S. Raghuraman, and X. Zhang, "Artificial neural networks for the prediction of peptide drift time in ion mobility mass spectrometry," *BMC Bioinformatics*, vol. 11, p. 182, Apr. 2010.

[26] B. Wang, J. Zhang, P. Chen, Z. Ji, S. Deng, and C. Li, "Prediction of peptide drift time in ion mobility mass spectrometry from sequence-based features," *BMC Bioinformatics*, vol. 14 Suppl 8, p. S9, May 2013.

[27] F. Meier, N. D. Köhler, A.-D. Brunner, J.-M. H. Wanka, E. Voytik, M. T. Strauss, F. J. Theis, and M. Mann, "Deep learning the collisional cross sections of the peptide universe from a million experimental values," *Nat. Commun.*, vol. 12, p. 1185, Feb. 2021.

[28] P. D. Hutchins, J. D. Russell, and J. J. Coon, "Accelerating lipidomic method development through simulation," *Anal. Chem.*, vol. 91, pp. 9698–9706, Aug. 2019.

[29] P. D. Hutchins, J. D. Russell, and J. J. Coon, "Mapping lipid fragmentation for tailored mass spectral libraries," *J. Am. Soc. Mass Spectrom.*, vol. 30, pp. 659–668, Apr. 2019.

[30] I. Blaženović, T. Shen, S. S. Mehta, T. Kind, J. Ji, M. Piparo, F. Cacciola, L. Mondello, and O. Fiehn, "Increasing compound identification rates in untargeted lipidomics research with liquid chromatography drift Time-Ion mobility mass spectrometry," *Anal. Chem.*, vol. 90, pp. 10758–10764, Sept. 2018.

[31] C. Sabater, A. Olano, N. Corzo, and A. Montilla, "GC-MS characterisation of novel artichoke (cynara scolymus) pectic-oligosaccharides mixtures by the application of machine learning algorithms and competitive fragmentation modelling," *Carbohydr. Polym.*, vol. 205, pp. 513–523, Feb. 2019.

[32] G. Yamankurt, E. J. Berns, A. Xue, A. Lee, N. Bagheri, M. Mrksich, and C. A. Mirkin, "Exploration of the nanomedicine-design space with high-throughput screening and machine learning," *Nat Biomed Eng*, vol. 3, pp. 318–327, Apr. 2019.

[33] C. Ma, Y. Ren, J. Yang, Z. Ren, H. Yang, and S. Liu, "Improved peptide retention time prediction in liquid chromatography through deep learning," *Anal. Chem.*, vol. 90, pp. 10881–10888, Sept. 2018.

[34] L. Moruz, D. Tomazela, and L. Käll, "Training, selection, and robust calibration of retention time models for targeted proteomics," *J. Proteome Res.*, vol. 9, pp. 5209–5216, Oct. 2010.

[35] N. Pfeifer, A. Leinenbach, C. G. Huber, and O. Kohlbacher, "Statistical learning of peptide retention behavior in chromatographic separations: a new kernel-based approach for computational proteomics," *BMC Bioinformatics*, vol. 8, p. 468, Nov. 2007.

[36] H. Maboudi Afkham, X. Qiu, M. The, and L. Käll, "Uncertainty estimation of predictions of peptides' chromatographic retention times in shotgun proteomics," *Bioinformatics*, vol. 33, pp. 508–513, Feb. 2017.

[37] S. Gessulat, T. Schmidt, D. P. Zolg, P. Samaras, K. Schnatbaum, J. Zerweck, T. Knaute, J. Rechenberger, B. Delanghe, A. Huhmer, U. Reimer, H.-C. Ehrlich, S. Aiche, B. Kuster, and M. Wilhelm, "Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning," *Nat. Methods*, vol. 16, pp. 509–518, June 2019.

[38] S. Tiwary, R. Levy, P. Gutenbrunner, F. Salinas Soto, K. K. Palaniappan, L. Deming, M. Berndl, A. Brant, P. Cimermancic, and J. Cox, "High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis," *Nat. Methods*, vol. 16, pp. 519–525, June 2019.

[39] V. A. Fusaro, D. R. Mani, J. P. Mesirov, and S. A. Carr, "Prediction of high-responding peptides for targeted protein assays by mass spectrometry," *Nat. Biotechnol.*, vol. 27, pp. 190–198, Feb. 2009.

[40] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.

[42] E. Shishkova, A. S. Hebert, M. S. Westphall, and J. J. Coon, "Ultra-High pressure (>30,000 psi) packing of capillary columns enhancing depth of shotgun proteomic analyses," *Anal. Chem.*, vol. 90, pp. 11503–11508, Oct. 2018.

[43] A. S. Hebert, C. Thöing, N. M. Riley, N. W. Kwiecien, E. Shiskova, R. Huguet, H. L. Cardasis, A. Kuehn, S. Eliuk, V. Zabrouskov, M. S. Westphall, G. C. McAlister, and J. J. Coon, "Improved precursor characterization for Data-Dependent mass spectrometry," *Anal. Chem.*, vol. 90, pp. 2333–2340, Feb. 2018.

[44] D. Kessner, M. Chambers, R. Burke, D. Agus, and P. Mallick, "ProteoWizard: open

source software for rapid proteomics tools development," *Bioinformatics*, vol. 24, pp. 2534–2536, Nov. 2008.

[45] A. T. Kong, F. V. Leprevost, D. M. Avtonomov, D. Mellacheruvu, and A. I. Nesvizhskii, "MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics," *Nat. Methods*, vol. 14, pp. 513–520, May 2017.

[46] D. Shteynberg, E. W. Deutsch, H. Lam, J. K. Eng, Z. Sun, N. Tasman, L. Mendoza, R. L. Moritz, R. Aebersold, and A. I. Nesvizhskii, "iprophet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates," *Mol. Cell. Proteomics*, vol. 10, p. M111.007690, Dec. 2011.

[47] F. da Veiga Leprevost, S. E. Haynes, D. M. Avtonomov, H.-Y. Chang, A. K. Shanmugam, D. Mellacheruvu, A. T. Kong, and A. I. Nesvizhskii, "Philosopher: a versatile toolkit for shotgun proteomics data analysis," *Nat. Methods*, vol. 17, pp. 869–870, Sept. 2020.

[48] B. MacLean, D. M. Tomazela, N. Shulman, M. Chambers, G. L. Finney, B. Frewen, R. Kern, D. L. Tabb, D. C. Liebler, and M. J. MacCoss, "Skyline: an open source document editor for creating and analyzing targeted proteomics experiments," *Bioinformatics*, vol. 26, pp. 966–968, Apr. 2010.

[49] L. Reiter, O. Rinner, P. Picotti, R. Hüttenhain, M. Beck, M.-Y. Brusniak, M. O. Hengartner, and R. Aebersold, "mprophet: automated data processing and statistical

validation for large-scale SRM experiments," *Nat. Methods*, vol. 8, pp. 430–435, May 2011.

[50] P. Szymański and T. Kajdanowicz, "A scikit-based python environment for performing multi-label classification," *arXiv preprint arXiv:1702.01460*, 2017.

[51] A. A. Goloborodko, L. I. Levitsky, M. V. Ivanov, and M. V. Gorshkov, "Pyteomics–a python framework for exploratory data analysis and rapid software prototyping in proteomics," *J. Am. Soc. Mass Spectrom.*, vol. 24, pp. 301–304, Feb. 2013.

[52] L. I. Levitsky, J. A. Klein, M. V. Ivanov, and M. V. Gorshkov, "Pyteomics 4.0: Five years of development of a python proteomics framework," *J. Proteome Res.*, vol. 18, pp. 709–714, Feb. 2019.

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[54] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *International conference on machine learning*, pp. 115–123, PMLR, 2013.

[55] M. Wang, J. Wang, J. Carver, B. S. Pullman, S. W. Cha, and N. Bandeira, "Assembling the Community-Scale discoverable human proteome," *Cell Syst*, vol. 7, pp. 412–421.e5, Oct. 2018.

[56] V. Sharma, J. Eckels, G. K. Taylor, N. J. Shulman, A. B. Stergachis, S. A. Joyner, P. Yan, J. R. Whiteaker, G. N. Halusa, B. Schilling, B. W. Gibson, C. M. Colangelo, A. G. Paulovich, S. A. Carr, J. D. Jaffe, M. J. MacCoss, and B. MacLean, "Panorama: a targeted proteomics knowledge base," *J. Proteome Res.*, vol. 13, pp. 4205–4210, Sept. 2014.

[57] A. A. Aksenov, J. Kapron, and C. E. Davis, "Predicting compensation voltage for singly-charged ions in high-field asymmetric waveform ion mobility spectrometry (FAIMS)," *J. Am. Soc. Mass Spectrom.*, vol. 23, pp. 1794–1798, Oct. 2012.

[58] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Springer, Oct. 2018.

[59] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*. John Wiley & Sons, June 2013.

[60] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation," 2015.

[61] M. A. Tahir, J. Kittler, and A. Bouridane, "Multilabel classification using heterogeneous ensemble of multi-label classifiers," 2012.

[62] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new Over-Sampling method in imbalanced data sets learning," 2005.

[63] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008.

[64] Z. Zheng, Y. Cai, and Y. Li, "Oversampling method for imbalanced classification," *COMPUTING AND INFORMATICS*, vol. 34, p. 1017–1037, Mar. 2016.

[65] S. Renuse, P. M. Vanderboom, A. D. Maus, J. V. Kemp, K. M. Gurtner, A. K. Madugundu, S. Chavan, J. A. Peterson, B. J. Madden, K. K. Mangalaparthi, D.-G. Mun, S. Singh, B. R. Kipp, S. Dasari, R. J. Singh, S. K. Grebe, and A. Pandey, "Development of mass spectrometry-based targeted assay for direct detection of novel SARS-CoV-2 coronavirus from clinical specimens."

Chapter 6

CONCLUSIONS AND FUTURE DIRECTIONS

Summary

Mass spectrometry (MS) has become a powerful tool for characterizing the proteomic heterogeneity of healthy and diseased human populations, tissues and cells. MS and associated technologies allow direct measurements of the protein abundances rather than relying on the inference that occurs in analysis of nucleic acids through genomics and transcriptomics. Continued advancements are needed in specialized sample preparation, acquisition strategies, analysis speed, sensitivity, precision, and adaptability to novel sequences to maximize the tool's effectiveness. This dissertation examined existing technologies and techniques, attempted to address challenges in their application to research projects relevant to human health, and proposed future directions to advance and capitalize on the most current technologies. More specifically it provided additional biological insight into the human proteomic perturbations induced by stress, aging and neurodegeneration, as well as advancing technological and methodological strategies for these types of investigations. Characterizing human proteomics *in situ* presents a multifaceted challenge, with proteomic diversity, small fold changes of interest, heterogeneous cell populations, temporal dynamism and large dynamic range all presenting substantial obstacles. To avoid these issues, researchers use model organisms and primary cell culture to study human protein expression in disease. Yet these models often fail to encompass the full complexity of human organisms, which can be crucial in the development of treatments and interventions, especially in highly structured organs like the brain. In the first chapter, these challenges are detailed, along

with the underlying concepts of instrument operation and the basic sample preparation and analysis workflow. Chapter two describes regional protein signatures for nine different regions of the human brain and demonstrates the interaction between these region-specific proteins and protein changes in Alzheimer's disease(AD), as well as those that occur during the normal aging process. The chapter that follows uses an age- and sex-matched sample cohort to develop an efficient and scalable proteomic analysis for cerebrospinal fluid (CSF) in AD. In the fourth chapter, a human proteomic analysis of saliva collected before and after a simulated combat mission identifies three cellular systems affected by the associated stress: innate immunity, protein processing and metabolism. The fifth chapter presents a model for sequence-based prediction of peptides' transmissive compensating voltages in the ion mobility device, FAIMS. The biological insights gleaned from our analyses in these research areas represent a series of contributions to expand the knowledge base of the human proteome in health and disease. These insights can inform future hypotheses and expanded analyses. Future developments in analytical methodology and technology follow a similar path, with advances building on existing tools or utilizing them in novel combinations or applications. The following section details some of these new applications and additional investigations that stem from our findings in saliva and the central nervous system (CNS), and how they align with the future objectives and challenges within the field of human proteomics.

Additional Biological Analysis

The saliva proteome and stress Saliva has historically been an attractive diagnostic tissue, with relatively straightforward, low-risk collection[1,2], recently gaining great popularity in the monitoring of the SARS-COV2 pandemic[3,4]. Extensive work has been done mining the salivary proteome for biomarkers in oral disease such as periodontitis[5–7] and oral cancers[8,9], as well as a variety of non-oral conditions[10–14]. MS-based proteomics has played a foundational role in the characterization of the oral microbiome[15]. Research targeting the salivary proteomic and metabolomic effects of physical and mental fatigue has also been conducted[16,17]. Improvements in analytical technologies have led to better characterization of affected molecules, with increasing numbers of stress-associated peptides and proteins[18,19].

While investigating the proteomic effects of stress caused by a simulated combat mission (as described in chapter 4) the authors observed changes in proteins associated with the complement system, gluconeogenesis and protein processing in the endoplasmic reticulum. The most concerted changes were a stress-associated upregulation of complement proteins and gluconeogenesis enzymes. Although these experiments identified more than 300 proteins as significantly associated with stress, the enrichment of proteins from these three functional pathways show much more promise in understanding the biological effects of stress. Future analyses should focus on proteins involved in these pathways, surveying them by targeted mass spectrometry or other non-MS proteomic methods

Beyond targeted MS experiments, our research findings suggested that several additional proteomic investigations would expand our understanding of stress mechanisms, biomarkers and potential impacts on human behavior and health. Small molecule perturbations during acute stress events were observed in our analysis, and proteomic perturbations within short timeframes($<$ 1 hour after stressor) have also been observed in humans previously[20]. Application of our global protein analysis to samples from acute stress events, as opposed to more medium-term sustained stress of the mission overall, could identify additional stress-associated proteins or changes of larger magnitude. Further, If the acute stress event took the form of a training exercise, protein measurements could be aligned with performance, along with other common metrics of stress/fatigue, such as cortisol[21], alpha-amylase[22], or creatine kinase[23,24] levels in blood. Lastly, given the connection between the gut microbiome and psychological stress in mammals[25,26], and our findings regarding immunity and metabolism, salivary microbial profiling could be useful in monitoring the impact of stress on health.

Although saliva is relatively easy to collect, several preparation challenges need to be addressed to expand the fluid's utility for proteomics. The signaling events in human saliva often rely on the cleavage of specific peptides and proteins by proteases[27]. These enzymes cleave proteins into non-tryptic peptides, which fail to be matched during data searching. These proteases can be denatured by heat, high concentrations of organics and chaotropic agents such as urea, but efficient inactivation will be an important step in bringing any analysis to scale. Alternatively, as MS data searching becomes faster, and

higher resolution spectra are available, spectra can be matched using a search in which a protease is not specified[28]. Another challenge of accessing the salivary proteome comes from variability in protein content and viscosity among individuals. Protein concentrations differing by more than an order of magnitude is common in saliva. This dynamic range can significantly inhibit throughput when samples are prepared manually, and cause difficulties when transferring precise volumes. The field has responded to these challenges in sample preparation by designing specialized preparation traps[29] and methods with reduced liquid transfer[29,30]. Automated liquid handling systems have greatly improved throughput in proteomic analysis of plasma[31], serum[32] and solid tissues[33,34], although presently this technology is largely restricted to nucleic acid analysis in saliva[35,36]. Continued innovation in sample preparation and data searching will allow raw saliva analysis to be performed at an even greater scale.

Proteomics in the study of neurodegeneration roteins play a critical role in the age-associated neurodegenerative condition Alzheimer's disease (AD), yet many aspects of pathological mechanisms remain unclear[37,38]. This fact may explain why no drug currently exists to substantially inhibit or reverse AD, despite the health burden of the disease[39]. MS-based proteomics has assisted in the characterization of two of the key pathological proteins: microtubule-associated protein tau and amyloid precursor protein[40–42]. As technology has improved, mass spectrometry has allowed for the simultaneous monitoring of an ever-growing number of proteins in the nervous system. Improvements in complementary

analytical and biochemistry techniques have permitted protein profiling of increasingly narrow subgroups including synapses[43], cell types[44,45], insoluble populations[46], and neuroanatomical regions[47,48].

We applied these cutting-edge proteomic technologies to investigate the effects of age and neurodegeneration on the central nervous system in both postmortem tissue (Chapter 2) and in living participants (Chapter 3). Both approaches are essential in clarifying the dynamics of proteins in the inception and progression of AD.

Chapter 2 describes the regional protein signatures identified in the brain, with substantial overlap observed between region-associated proteins and those related to AD. Regional protein shifts resulting from the normal aging process were also elucidated by comparing to publicly available proteomic data acquired from middle aged adult brains[48]. Understanding the brain proteome as regionally distinct plays an important role in understanding the large heterogeneity in the symptoms of Alzheimer's and aging[49–53]. At the time of publication, this analysis represented the largest regional proteomic survey of the aged human brain based on the number of sections. Since then, interest has continued to increase in region-specific effects in a variety of biological processes and health conditions[54–58].

Although post-mortem tissue is required to make region-specific conclusions, increased understanding of disease progression and improved detection requires sampling from living individuals. Chapter 3 detailed a favorable proteomic analysis and acquisition strategy for a structured cohort of cerebral spinal fluid (CSF) samples targeting AD. Despite limits on statistical power due to sample size, we identified three highly significant proteins and

several groups of proteins at lesser significance that had been previously associated with AD. The pilot analysis described in this chapter provided a methodological foundation for expanded biological inquiry in CSF, particularly when focusing on Alzheimer's.

Future CSF studies will be able to apply our methodology to increasingly complex samples and targeting intermediate disease stages. Our pilot analysis categorized cases based on positivity for both tau and amyloid biomarkers in order to emphasize contrast between proteomes. However, AD progression is highly nuanced and complex, and many intermediate disease groups exist[59,60], suggesting the value of including multiple AD subtypes. Several recent large-scale analyses of proteins in CSF have included multiple disease groups, indicating differential protein abundances between them[61–63]. The addition of intermediate disease groups attempts to capture the molecular timeline of pathological progression in a somewhat generalized manner. The most informative disease timeline would be developed from a long-term longitudinal study where samples are matched within individuals and paired with clinical monitoring. This would allow for molecular changes in the neurological environment to be linked to clinical symptoms. Recent studies of this type have allowed for the identification of clinically predictive biomarkers such as neurofilament light chain[64].

Our research also identified critical avenues for future study of the regional brain proteome. Continued expansion of the diversity of brain regions analyzed by proteomics will be key in understanding the myriad effects of age and neurodegeneration. The brain serves a multitude of functions with hundreds of distinct structures, substructures, and cell popu-

lations contained within. Several groups have undertaken more extensive regional surveys of the proteome examining more than double the number of regions studied here[65–67]. Their intent is to begin to complement the existing transcriptomic resources for which anatomical regions number in the hundreds[68]. Our analysis correlated experimentally-identified, region-specific proteins with AD-associated proteins from the literature, with some experimental overlap of region and AD. Future experimental endeavors comparing control and diseased tissues could span across a greater number of anatomical regions, identifying additional regional disease effects. Many efforts of this type are already underway, comparing AD effects across more regions[61,69,70].

Continued Technological and Methodological Development

The biological investigations discussed above are possible because of recent advancements in mass spectrometers and their associated technologies. Over the last 20 years, mass spectrometers have become faster, more sensitive, and more accurate. Ion mobility and liquid chromatography instruments have also increased their separation capacity and precision. As one example, in 2013, Zubarev and co-workers identified nearly 5,000 proteins from cultured human cells in 4 hours of analysis[71]. One year later, work in our group would identify 4,000 proteins in just over one hour of analysis[72]. By 2018, with the addition of FAIMS, nearly 5,600 proteins could be identified in one hour of analysis time[73].

As we approach feasible surveillance of a near-comprehensive proteome, many of the remaining challenges relate to quantitation of low abundance proteins and the balance of

proteomic depth with throughput. Addressing them requires the type of methodological and technological innovations detailed in chapters 3 and 5, where sample preparation and data acquisition strategies are applied to maximize the physical capacity of the mass spectrometer. Interestingly, many of the most intriguing future directions for the FAIMS prediction models generated in Chapter 5 can be applied to the topic of neurodegenerative disease and proteomic surveillance of cerebrospinal fluid.

Predictive modeling for mass spectrometry parameters Algorithmic predictions of peptide ion parameters such as mobility were developed long before "machine learning" entered the common research vernacular. Building on Valentine's work assigning intrinsic size parameters to amino acids residues[74,75], two competing groups developed neural nets for predicting ion drift time in 2010[76,77]. Since then, prediction algorithms have been developed for fragmentation spectra[78,79], retention time[80] high-signal peptides[80,81] and collisional cross sections[82]. Chapter 5 described the development of two models, a random forest and a neural net, for predicting transmissive compensating voltages (CV) when using the ion mobility module, FAIMS. Their efficacy, alone and in combination, was demonstrated by training using a dataset of 128,000 human peptide precursor ions before testing prediction performance for a set of 40,000 *E. coli* peptide precursor ions naïve to the models. Both models showed excellent binary accuracy, with the random forest exhibiting greater recall, while the neural net provided enhanced precision.

Expansion of the residue alphabet. The models currently rely on a sequence feature component that accepts all canonical amino acids as single letters, as well as oxidation of methionine (represented as "m") and N-terminal acetylation (represented as "a"). Several recent studies have indicated the FAIMS module's capacity to separate peptides based on post-translational modifications[83–85], suggesting a benefit to predictions regarding transmission of modified peptides. In the future, the acceptable residue alphabet could be expanded to include several of the most common modifications, such as phosphorylation, methylation, and acetylation of different side chains. The models could also be expanded to include more complex modifications, such as glycosylation, although the oligomeric and branched nature of glycans would make this challenging for all but the best characterized glycan motifs.

Synergistic applications with existing prediction technologies Several other existing prediction tools can be used in concert with the models to develop optimized proteomics experiments *in silico*. Targeted experiments are the most straightforward application, as CV predictions could be combined with predictions of retention time and high-signal peptides to create an optimized, scheduled parallel reaction monitoring experiment. However, the largest benefits would result from global proteomics experiments where CV windows are scheduled along predicted retention time to favor additional protein identifications. These *in silico*-designed global proteomic experiments could be tailored to the proteomes of specific tissues or cell populations. Broad hypothetical examples of both experiment

types- targeted and global -are provided below.

Imagine a situation in which we hope to precisely quantify the abundance of four proteins (**Figure 6.1**). These four proteins could be digested *in silico* into their respective peptides, with the peptides generating the greatest ion current selected using the prediction tool developed by Vincent Fusaro and colleagues[81]. Once target peptides had been selected, transmissive CVs would be predicted based on sequence, using the prediction tool described above. Chromatographic retention time(RT) would also be predicted based on sequence using DeepRT[80], a neural net for predicting retention time. With peptide parameters in hand, CV and m/z isolation could be scheduled based on retention time of target peptides (**Figure 6.1**). This would allow for improved separation and isolation of peptides, and in turn improved fragment spectra and quantitation, all without use of sample material or instrument time for optimization.

In another use example, retention time and compensating voltage predictions could be applied to increase protein identifications in global data-dependent proteomics experiments. A whole proteome could be digested into component peptides *in silico*, which could then be mapped across CV-RT space similarly to **Figure 5.1E**. Compensating voltage could then be scheduled across retention time, specifically favoring the sequencing and scanning of peptides derived from previously unidentified proteins or proteins with low sequence coverage. In contrast to the targeted method described above, a series of MS1 and MS/MS scans would be collected in a data-dependent manner in each CV-RT window. This strategy may prove even more effective when applied to specific cell types or tissues, such as CSF,

Protein	Peptide	m/z	RT	CV
Protein A	PeptideA1	217.55	11.8	40
	PeptideA2	1139.75	38.3	40
	PeptideA3	307.95	45.2	80
Protein B	PeptideB1	661.12	28.2	50
	PeptideB2	284.06	5.1	55
	PeptideB3	584.77	21.3	65
Protein C	PeptideC1	1311.68	8.2	30
	PeptideC2	766.29	48.2	25
	PeptideC3	371.14	24.9	40
Protein D	PeptideD1	435.41	35.5	60
	PeptideD2	1278.3	32.1	75
	PeptideD3	667.39	14.8	50
Protein E	PeptideE1	392.03	52.1	35
	PeptideE2	1275.42	18.1	80
	PeptideE3	399.67	41.7	50

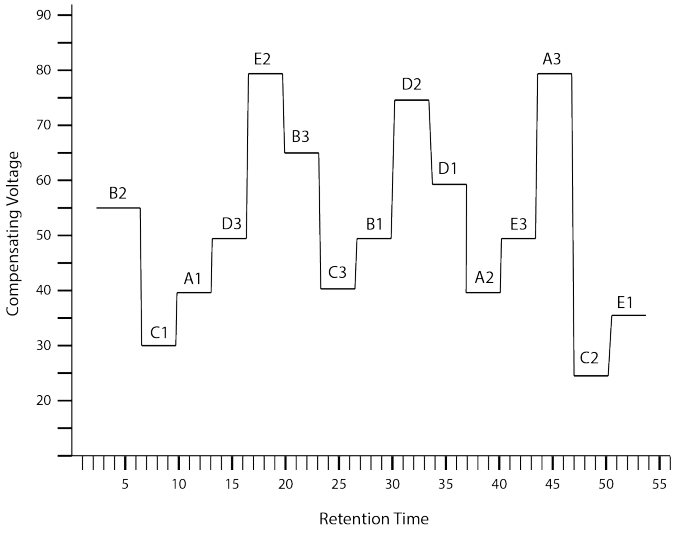


Figure 6.1: Simulated Scheduled PRM Experiment using CV and RT predictions. Table shows target peptides along with retention time (RT), mass-to-charge(m/z) and favorable compensating voltage (CV). Line plot depicts short retention time windows at which scans would be collected using the indicated compensating voltage, with each window labelled by the targeted peptide.

given their reduced proteomic complexity. Method design using this optimization strategy may also allow for shorter gradients. Chromatographic gradients as short as five minutes have already been utilized in combination with FAIMS successfully[86], as well as direct infusion with no liquid chromatography[87].

The rapid development of objective-specific human proteomic methods, whether based on specific tissues, diseases or individuals is key to advancing mass spectrometry-based proteomics as a tool for improving human health. Although the examples provided here are quite general, these method development pipelines are highly adaptable and could be tailored to any tissue or disease.

Proteomics by mass spectrometry is an incredibly powerful tool for monitoring, detecting, and understanding human diseases. MS-based proteomics has the capacity to rapidly, accurately and sensitively quantify the molecular machinery of the cell. A host of variables at the individual and cellular level affect protein abundance. To accurately capture physiologically relevant proteomic changes, research must isolate, account for and quantify these variables. As the findings and strategies described here are further advanced and improved, mass spectrometry will add increasing value to the field's understanding of the role of proteomics in characterizing human health and disease. It is the author's hope that the contributions from this research will help others parse and illuminate the incredibly complex and dynamic landscape that is the human proteome.

References

[1] E. M. McBride, R. J. Lawrence, K. McGee, P. M. Mach, P. S. Demond, M. W. Busch, J. W. Ramsay, E. K. Hussey, T. Glaros, and E. S. Dhummakupt, "Rapid liquid chromatography tandem mass spectrometry method for targeted quantitation of human performance metabolites in saliva," *J. Chromatogr. A*, vol. 1601, pp. 205–213, Sept. 2019.

[2] B. L. Schulz, J. Cooper-White, and C. K. Punyadeera, "Saliva proteome research: current status and future outlook," *Crit. Rev. Biotechnol.*, vol. 33, pp. 246–259, Sept. 2013.

[3] L. Azzi, G. Carcano, F. Gianfagna, P. Grossi, D. D. Gasperina, A. Genoni, M. Fasano, F. Sessa, L. Tettamanti, F. Carinci, V. Maurino, A. Rossi, A. Tagliabue, and A. Baj, "Saliva is a reliable tool to detect SARS-CoV-2," *J. Infect.*, vol. 81, pp. e45–e50, July 2020.

[4] J. Ceron, E. Lamy, S. Martinez-Subiela, P. Lopez-Jornet, F. Capela-Silva, P. Eckersall, and A. Tvarijonaviciute, "Use of saliva for diagnosis and monitoring the SARS-CoV-2: A general perspective," 2020.

[5] F. A. R. R. Hartenbach, É. Velasquez, F. C. S. Nogueira, G. B. Domont, E. Ferreira, and A. P. V. Colombo, "Proteomic analysis of whole saliva in chronic periodontitis," *J. Proteomics*, vol. 213, p. 103602, Feb. 2020.

[6] S. K. Al-Tarawneh, M. B. Border, C. F. Dibble, and S. Bencharit, "Defining salivary

biomarkers using mass spectrometry-based proteomics: a systematic review," *OMICS*, vol. 15, pp. 353–361, June 2011.

[7] N. Christodoulides, P. N. Floriano, C. S. Miller, J. L. Ebersole, S. Mohanty, P. Dharshan, M. Griffin, A. Lennart, K. L. M. Ballard, C. P. King, Jr, M. C. Langub, R. J. Kryscio, M. V. Thomas, and J. T. McDevitt, "Lab-on-a-chip methods for point-of-care measurements of salivary biomarkers of periodontitis," *Ann. N. Y. Acad. Sci.*, vol. 1098, pp. 411–428, Mar. 2007.

[8] P. Sivadasan, M. K. Gupta, G. J. Sathe, L. Balakrishnan, P. Palit, H. Gowda, A. Suresh, M. A. Kuriakose, and R. Sirdeshmukh, "Human salivary proteome–a resource of potential biomarkers for oral cancer," *J. Proteomics*, vol. 127, pp. 89–95, Sept. 2015.

[9] S. Hu, M. Arellano, P. Boontheung, J. Wang, H. Zhou, J. Jiang, D. Elashoff, R. Wei, J. A. Loo, and D. T. Wong, "Salivary proteomics for oral cancer biomarker discovery," *Clin. Cancer Res.*, vol. 14, pp. 6246–6252, Oct. 2008.

[10] H. Xiao, L. Zhang, H. Zhou, J. M. Lee, E. B. Garon, and D. T. W. Wong, "Proteomic analysis of human saliva from lung cancer patients using two-dimensional difference gel electrophoresis and mass spectrometry," *Mol. Cell. Proteomics*, vol. 11, p. M111.012112, Feb. 2012.

[11] A. K. Dey, GARBH-Ini Study Group*, B. Kumar, A. K. Singh, P. Ranjan, R. Thiruvengadam, B. K. Desiraju, P. Kshetrapal, N. Wadhwa, S. Bhatnagar, F. Rashid, D. Malakar,

D. M. Salunke, and T. K. Maiti, "Salivary proteome signatures in the early and middle stages of human pregnancy with term birth outcome," 2020.

[12] M. Bayani, M. Pourali, and M. Keivan, "Possible interaction between visfatin, periodontal infection, and other systemic diseases: A brief review of literature," *Eur. J. Dent.*, vol. 11, pp. 407–410, July 2017.

[13] B. Manconi, B. Liori, T. Cabras, F. Vincenzoni, F. Iavarone, L. Lorefice, E. Cocco, M. Castagnola, I. Messana, and A. Olianas, "Top-down proteomic profiling of human saliva in multiple sclerosis patients," *J. Proteomics*, vol. 187, pp. 212–222, Sept. 2018.

[14] H. Xiao, Y. Zhang, Y. Kim, S. Kim, J. J. Kim, K. M. Kim, J. Yoshizawa, L.-Y. Fan, C.-X. Cao, and D. T. W. Wong, "Differential proteomic analysis of human saliva using tandem mass tags quantification for gastric cancer detection," *Sci. Rep.*, vol. 6, p. 22165, Feb. 2016.

[15] N. Grassl, N. A. Kulak, G. Pichler, P. E. Geyer, J. Jung, S. Schubert, P. Sinitcyn, J. Cox, and M. Mann, "Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome," *Genome Med.*, vol. 8, p. 44, Apr. 2016.

[16] S.-G. Ra, S. Maeda, R. Higashino, T. Imai, and S. Miyakawa, "Metabolomics of salivary fatigue markers in soccer players after consecutive games," *Appl. Physiol. Nutr. Metab.*, vol. 39, pp. 1120–1126, Oct. 2014.

[17] D. J. Michael, S. Daugherty, A. Santos, B. C. Ruby, and J. E. Kalns, "Fatigue biomarker index: An objective salivary measure of fatigue level," 2012.

[18] D. J. Michael, B. Valle, J. Cox, J. E. Kalns, and D. L. Fogt, "Salivary biomarkers of physical fatigue as markers of sleep deprivation," *J. Clin. Sleep Med.*, vol. 9, pp. 1325–1331, Dec. 2013.

[19] Y.-L. Xu, Y.-N. Gong, D. Xiao, C.-X. Zhao, X.-H. Gao, X.-H. Peng, A.-P. Xi, L.-H. He, L.-P. Lu, M. Ding, Y. Li, H. Jianjun, X.-H. Su, F.-L. Liu, J.-Z. Wang, Z.-J. Liu, and J.-Z. Zhang, "Discovery and identification of fatigue-related biomarkers in human saliva," *Eur. Rev. Med. Pharmacol. Sci.*, vol. 22, pp. 8519–8536, Dec. 2018.

[20] R. K. Marvin, M. B. Saepoo, S. Ye, D. B. White, R. Liu, K. Hensley, P. Rega, V. Kazan, D. R. Giovannucci, and D. Isailovic, "Salivary protein changes in response to acute stress in medical residents performing advanced clinical simulations: a pilot proteomics study," 2017.

[21] E. K. Adam, L. C. Hawkley, B. M. Kudielka, and J. T. Cacioppo, "Day-to-day dynamics of experience–cortisol associations in a population-based sample of older adults," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, pp. 17058–17063, Nov. 2006.

[22] M. Yamaguchi, M. Deguchi, J. Wakasugi, S. Ono, N. Takai, T. Higashi, and Y. Mizuno, "Hand-held monitor of sympathetic nervous system using salivary amylase activity

and its validation by driver fatigue assessment," *Biosens. Bioelectron.*, vol. 21, pp. 1007–1014, Jan. 2006.

[23] T. Wiewelhove, C. Raeder, T. Meyer, M. Kellmann, M. Pfeiffer, and A. Ferrauti, "Markers for routine assessment of fatigue and recovery in male and female team sport athletes during High-Intensity interval training," 2015.

[24] A. Hecksteden, S. Skorski, S. Schwindling, D. Hammes, M. Pfeiffer, M. Kellmann, A. Ferrauti, and T. Meyer, "Blood-Borne markers of fatigue in competitive athletes - results from simulated training camps," *PLoS One*, vol. 11, p. e0148810, Feb. 2016.

[25] A. Sarkar, S. Harty, S. M. Lehto, A. H. Moeller, T. G. Dinan, R. I. M. Dunbar, J. F. Cryan, and P. W. J. Burnet, "The microbiome in psychology and cognitive neuroscience," *Trends Cogn. Sci.*, vol. 22, pp. 611–636, July 2018.

[26] C. Xu, S. K. Lee, D. Zhang, and P. S. Frenette, "The gut microbiome regulates Psychological-Stress-Induced inflammation," *Immunity*, vol. 53, pp. 417–428.e4, Aug. 2020.

[27] K. Thomadaki, E. J. Helmerhorst, N. Tian, X. Sun, W. L. Siqueira, D. R. Walt, and F. G. Oppenheim, "Whole-saliva proteolysis and its impact on salivary diagnostics," *J. Dent. Res.*, vol. 90, pp. 1325–1330, Nov. 2011.

[28] Z. Rolfs, R. J. Millikin, and L. M. Smith, "An algorithm to improve the speed of semi

and Non-Specific enzyme searches in proteomics," *Curr. Bioinform.*, vol. 15, no. 9, pp. 1065–1074, 2020.

[29] Y.-H. Lin, R. V. Eguez, M. G. Torralba, H. Singh, P. Golusinski, W. Golusinski, M. Masternak, K. E. Nelson, M. Freire, and Y. Yu, "Self-Assembled STrap for global proteomics and salivary biomarker discovery," *J. Proteome Res.*, vol. 18, pp. 1907–1915, Apr. 2019.

[30] X. Zhang, P. Sadowski, and C. Punyadeera, "Evaluation of sample preparation methods for label-free quantitative profiling of salivary proteome," *J. Proteomics*, vol. 210, p. 103532, Jan. 2020.

[31] D. Vuckovic, E. Cudjoe, F. M. Musteata, and J. Pawliszyn, "Automated solid-phase microextraction and thin-film microextraction for high-throughput analysis of biological fluids and ligand-receptor binding studies," *Nat. Protoc.*, vol. 5, pp. 140–161, Jan. 2010.

[32] M. R. Bladergroen, R. J. E. Derks, S. Nicolardi, B. de Visser, S. van Berloo, Y. E. M. van der Burgt, and A. M. Deelder, "Standardized and automated solid-phase extraction procedures for high-throughput proteomics of body fluids," *J. Proteomics*, vol. 77, pp. 144–153, Dec. 2012.

[33] A.-B. Arul, M. Byambadorj, N.-Y. Han, J. M. Park, and H. Lee, "Development of an automated, high-throughput sample preparation protocol for proteomics analysis," 2015.

[34] T. Müller, M. Kalxdorf, R. Longuespée, D. N. Kazdal, A. Stenzinger, and J. Krijgsveld, "Automated sample preparation with SP3 for low-input clinical proteomics," *Mol. Syst. Biol.*, vol. 16, p. e9111, Jan. 2020.

[35] N. Matic, T. Lawson, G. Ritchie, A. Stefanovic, V. Leung, S. Champagne, M. G. Romney, and C. F. Lowe, "Automated molecular testing of saliva for SARS-CoV-2 detection," *Diagn. Microbiol. Infect. Dis.*, vol. 100, p. 115324, May 2021.

[36] A. W.-H. Chu, C. C.-Y. Yip, W.-M. Chan, A. C.-K. Ng, D. L.-S. Chan, R. H.-P. Siu, C. Y. T. Chung, J. P.-L. Ng, H. Kittur, G. L. Mosley, R. W.-S. Poon, R. Y.-T. Chiu, and K. K.-W. To, "Evaluation of an automated High-Throughput Liquid-Based RNA extraction platform on pooled nasopharyngeal or saliva specimens for SARS-CoV-2 RT-PCR," 2021.

[37] P. Theofilas, A. J. Ehrenberg, A. Nguy, J. M. Thackrey, S. Dunlop, M. B. Mejia, A. T. Alho, R. E. Paraizo Leite, R. D. Rodriguez, C. K. Suemoto, C. F. Nascimento, M. Chin, D. Medina-Cleghorn, A. M. Cuervo, M. Arkin, W. W. Seeley, B. L. Miller, R. Nitrini, C. A. Pasqualucci, W. J. Filho, U. Rueb, J. Neuhaus, H. Heinsen, and L. T. Grinberg, "Probing the correlation of neuronal loss, neurofibrillary tangles, and cell death markers across the alzheimer's disease braak stages: a quantitative study in humans," *Neurobiol. Aging*, vol. 61, pp. 1–12, Jan. 2018.

[38] M. Fricker, A. M. Tolkovsky, V. Borutaite, M. Coleman, and G. C. Brown, "Neuronal cell death," *Physiol. Rev.*, vol. 98, pp. 813–880, Apr. 2018.

[39] C. Lynch, "World alzheimer report 2019: Attitudes to dementia, a global survey," 2020.

[40] C. Vigo-Pelfrey, D. Lee, P. Keim, I. Lieberburg, and D. B. Schenk, "Characterization of beta-amyloid peptide from human cerebrospinal fluid," *J. Neurochem.*, vol. 61, pp. 1965–1968, Nov. 1993.

[41] A. E. Roher, J. D. Lowenson, S. Clarke, A. S. Woods, R. J. Cotter, E. Gowing, and M. J. Ball, "beta-amyloid-(1-42) is a major component of cerebrovascular amyloid deposits: implications for the pathology of alzheimer disease," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 90, pp. 10836–10840, Nov. 1993.

[42] D. P. Hanger, J. C. Betts, T. L. Loviny, W. P. Blackstock, and B. H. Anderton, "New phosphorylation sites identified in hyperphosphorylated tau (paired helical filament-tau) from alzheimer's disease brain using nanoelectrospray mass spectrometry," *J. Neurochem.*, vol. 71, pp. 2465–2476, Dec. 1998.

[43] À. Bayés, M. O. Collins, C. M. Galtrey, C. Simonnet, M. Roy, M. D. R. Croning, G. Gou, L. N. van de Lagemaat, D. Milward, I. R. Whittle, C. Smith, J. S. Choudhary, and S. G. N. Grant, "Human post-mortem synapse proteome integrity screening for proteomic studies of postsynaptic complexes," *Mol. Brain*, vol. 7, p. 88, Nov. 2014.

[44] E. B. Dammer, D. M. Duong, I. Diner, M. Gearing, Y. Feng, J. J. Lah, A. I. Levey, and N. T. Seyfried, "Neuron enriched nuclear proteome isolated from human brain," 2013.

[45] S. Davis, C. Scott, O. Ansorge, and R. Fischer, "Development of a sensitive, scalable method for spatial, Cell-Type-Resolved proteomics of the human brain," *J. Proteome Res.*, vol. 18, pp. 1787–1795, Apr. 2019.

[46] Y. M. Gozal, D. M. Duong, M. Gearing, D. Cheng, J. J. Hanfelt, C. Funderburk, J. Peng, J. J. Lah, and A. I. Levey, "Proteomics analysis reveals novel components in the detergent-insoluble subproteome in alzheimer's disease," *J. Proteome Res.*, vol. 8, pp. 5069–5079, Nov. 2009.

[47] J. McKetney, R. M. Runde, A. S. Hebert, S. Salamat, S. Roy, and J. J. Coon, "Proteomic atlas of the human brain in alzheimer's disease," *J. Proteome Res.*, vol. 18, pp. 1380–1391, Mar. 2019.

[48] B. C. Carlyle, R. R. Kitchen, J. E. Kanyo, E. Z. Voss, M. Pletikos, A. M. M. Sousa, T. T. Lam, M. B. Gerstein, N. Sestan, and A. C. Nairn, "A multiregional proteomic survey of the postnatal human brain," *Nat. Neurosci.*, vol. 20, pp. 1787–1795, Dec. 2017.

[49] N. L. Komarova and C. J. Thalhauser, "High degree of heterogeneity in alzheimer's disease progression patterns," *PLoS Comput. Biol.*, vol. 7, p. e1002251, Nov. 2011.

[50] G. Devi and P. Scheltens, "Heterogeneity of alzheimer's disease: consequence for drug trials?," *Alzheimers. Res. Ther.*, vol. 10, p. 122, Dec. 2018.

[51] W. M. Van der Flier, "Clinical heterogeneity in familial alzheimer's disease," *Lancet Neurol.*, vol. 15, pp. 1296–1298, Dec. 2016.

[52] G. Bartzokis, D. Sultzer, P. H. Lu, K. H. Nuechterlein, J. Mintz, and J. L. Cummings, "Heterogeneous age-related breakdown of white matter structural integrity: implications for cortical "disconnection" in aging and alzheimer's disease," *Neurobiol. Aging*, vol. 25, pp. 843–851, Aug. 2004.

[53] U. Işıldak, M. Somel, J. M. Thornton, and H. M. Dönertaş, "Temporal changes in the gene expression heterogeneity during brain development and aging," *Sci. Rep.*, vol. 10, p. 4080, Mar. 2020.

[54] U. Distler, S. Schumann, H.-G. Kesseler, R. Pielot, K.-H. Smalla, M. Sielaff, M. J. Schmeisser, and S. Tenzer, "Proteomic analysis of brain region and Sex-Specific synaptic protein expression in the adult mouse brain," *Cells*, vol. 9, Jan. 2020.

[55] S. Billington, L. Salphati, C. E. C. A. Hop, X. Chu, R. Evers, D. Burdette, C. Rowbottom, Y. Lai, G. Xiao, W. G. Humphreys, T. B. Nguyen, B. Prasad, and J. D. Unadkat, "Interindividual and regional variability in drug transporter abundance at the human Blood-Brain barrier measured by quantitative targeted proteomics," *Clin. Pharmacol. Ther.*, vol. 106, pp. 228–237, July 2019.

[56] D. Zhang, X. Dong, X. Liu, L. Ye, S. Li, R. Zhu, Y. Ye, and Y. Jiang, "Proteomic analysis of brain regions reveals brain regional differences and the involvement of multiple keratins in chronic alcohol neurotoxicity," *Alcohol Alcohol*, vol. 55, pp. 147–156, Mar. 2020.

[57] L. C. Graham, M. J. Naldrett, S. G. Kohama, C. Smith, D. J. Lamont, B. W. McColl, T. H. Gillingwater, P. Skehel, H. F. Urbanski, and T. M. Wishart, "Regional molecular mapping of primate synapses during normal healthy aging," *Cell Rep.*, vol. 27, pp. 1018–1026.e4, Apr. 2019.

[58] S. Y. Jung, J. M. Choi, M. W. C. Rousseaux, A. Malovannaya, J. J. Kim, J. Kutzera, Y. Wang, Y. Huang, W. Zhu, S. Maity, H. Y. Zoghbi, and J. Qin, "An anatomically resolved mouse brain proteome reveals parkinson disease-relevant pathways," *Mol. Cell. Proteomics*, vol. 16, pp. 581–593, Apr. 2017.

[59] J. Rasmussen, J. Mahler, N. Beschorner, S. A. Kaeser, L. M. Häsler, F. Baumann, S. Nyström, E. Portelius, K. Blennow, T. Lashley, N. C. Fox, D. Sepulveda-Falla, M. Glatzel, A. L. Oblak, B. Ghetti, K. P. R. Nilsson, P. Hammarström, M. Staufenbiel, L. C. Walker, and M. Jucker, "Amyloid polymorphisms constitute distinct clouds of conformational variants in different etiological subtypes of alzheimer's disease," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, pp. 13018–13023, Dec. 2017.

[60] G. Di Fede, M. Catania, E. Maderna, R. Ghidoni, L. Benussi, E. Tonoli, G. Giaccone, F. Moda, A. Paterlini, I. Campagnani, S. Sorrentino, L. Colombo, A. Kubis, E. Bistaffa, B. Ghetti, and F. Tagliavini, "Molecular subtypes of alzheimer's disease," *Sci. Rep.*, vol. 8, p. 3269, Feb. 2018.

[61] E. C. B. Johnson, E. Kathleen Carter, E. B. Dammer, D. M. Duong, E. S. Gerasimov, Y. Liu, J. Liu, R. Betarbet, L. Ping, L. Yin, G. E. Serrano, T. G. Beach, J. Peng, P. L.

De Jager, C. Gaiteri, D. A. Bennett, M. Gearing, T. S. Wingo, A. P. Wingo, J. J. Lah, A. I. Levey, and N. T. Seyfried, "Large-Scale deep Multi-Layer analysis of alzheimer's disease brain reveals strong proteomic Disease-Related changes not observed at the RNA level."

[62] C. D. Whelan, N. Mattsson, M. W. Nagle, S. Vijayaraghavan, C. Hyde, S. Janelidze, E. Stomrud, J. Lee, L. Fitz, T. A. Samad, G. Ramaswamy, R. A. Margolin, A. Malarstig, and O. Hansson, "Multiplex proteomics identifies novel CSF and plasma biomarkers of early alzheimer's disease," *Acta Neuropathol Commun*, vol. 7, p. 169, Nov. 2019.

[63] E. C. B. Johnson, E. B. Dammer, D. M. Duong, L. Ping, M. Zhou, L. Yin, L. A. Higginbotham, A. Guajardo, B. White, J. C. Troncoso, M. Thambisetty, T. J. Montine, E. B. Lee, J. Q. Trojanowski, T. G. Beach, E. M. Reiman, V. Haroutunian, M. Wang, E. Schadt, B. Zhang, D. W. Dickson, N. Ertekin-Taner, T. E. Golde, V. A. Petyuk, P. L. De Jager, D. A. Bennett, T. S. Wingo, S. Rangaraju, I. Hajjar, J. M. Shulman, J. J. Lah, A. I. Levey, and N. T. Seyfried, "Large-scale proteomic analysis of alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation," *Nat. Med.*, vol. 26, pp. 769–780, May 2020.

[64] O. Preische, S. A. Schultz, A. Apel, J. Kuhle, S. A. Kaeser, C. Barro, S. Gräber, E. Kuder-Buletta, C. LaFougere, C. Laske, J. Vöglein, J. Levin, C. L. Masters, R. Martins, P. R. Schofield, M. N. Rossor, N. R. Graff-Radford, S. Salloway, B. Ghetti, J. M. Ringman, J. M. Noble, J. Chhatwal, A. M. Goate, T. L. S. Benzinger, J. C. Morris, R. J. Bateman, G. Wang,

A. M. Fagan, E. M. McDade, B. A. Gordon, M. Jucker, and Dominantly Inherited Alzheimer Network, "Serum neurofilament dynamics predicts neurodegeneration and clinical progression in presymptomatic alzheimer's disease," *Nat. Med.*, vol. 25, pp. 277–283, Feb. 2019.

[65] Z. Guo, C. Shao, Y. Zhang, W. Qiu, W. Li, W. Zhu, Q. Yang, Y. Huang, L. Pan, Y. Dong, H. Sun, X. Xiao, W. Sun, C. Ma, and L. Zhang, "A global multiregional proteomic map of the human cerebral cortex."

[66] S. Srivast, D. Biswas, S. Shenoy, C. Chetanya, A. Athithyan, M. Lachén-Montes, S. Ghosh, K. Ausín, M. Zelaya, J. Fernández-Irigoyen, A. Manna, S. Roy, A. Barpanda, G. Ball, E. Santamaría, and A. Talukdar, "A high-resolution brain proteome map uncovers the inter-hemispheric laterality & inter-regional protein expression changes."

[67] O. E. Curran, Z. Qiu, C. Smith, and S. G. N. Grant, "A single-synapse resolution survey of PSD95-positive synapses in twenty human brain regions," *Eur. J. Neurosci.*, June 2020.

[68] M. J. Hawrylycz, E. S. Lein, A. L. Guillozet-Bongaarts, E. H. Shen, L. Ng, J. A. Miller, L. N. van de Lagemaat, K. A. Smith, A. Ebbert, Z. L. Riley, C. Abajian, C. F. Beckmann, A. Bernard, D. Bertagnolli, A. F. Boe, P. M. Cartagena, M. M. Chakravarty, M. Chapin, J. Chong, R. A. Dalley, B. David Daly, C. Dang, S. Datta, N. Dee, T. A. Dolbeare, V. Faber, D. Feng, D. R. Fowler, J. Goldy, B. W. Gregor, Z. Haradon, D. R. Haynor, J. G. Hohmann, S. Horvath, R. E. Howard, A. Jeromin, J. M. Jochim, M. Kinnunen, C. Lau, E. T. Lazarz,

C. Lee, T. A. Lemon, L. Li, Y. Li, J. A. Morris, C. C. Overly, P. D. Parker, S. E. Parry, M. Reding, J. J. Royall, J. Schulkin, P. A. Sequeira, C. R. Slaughterbeck, S. C. Smith, A. J. Sodt, S. M. Sunkin, B. E. Swanson, M. P. Vawter, D. Williams, P. Wohnoutka, H. R. Zielke, D. H. Geschwind, P. R. Hof, S. M. Smith, C. Koch, S. G. N. Grant, and A. R. Jones, "An anatomically comprehensive atlas of the adult human brain transcriptome," *Nature*, vol. 489, pp. 391–399, Sept. 2012.

[69] C. F. Mendonça, M. Kuras, F. C. S. Nogueira, I. Plá, T. Hortobágyi, L. Csiba, M. Palkovits, É. Renner, P. Döme, G. Marko-Varga, G. B. Domont, and M. Rezeli, "Proteomic signatures of brain regions affected by tau pathology in early and late stages of alzheimer's disease," *Neurobiol. Dis.*, vol. 130, p. 104509, Oct. 2019.

[70] J. Xu, S. Patassini, N. Rustogi, I. Riba-Garcia, B. D. Hale, A. M. Phillips, H. Waldvogel, R. Haines, P. Bradbury, A. Stevens, R. L. M. Faull, A. W. Dowsey, G. J. S. Cooper, and R. D. Unwin, "Regional protein expression in human alzheimer's brain correlates with disease severity," *Commun Biol*, vol. 2, p. 43, Feb. 2019.

[71] M. Pirmoradian, H. Budamgunta, K. Chingin, B. Zhang, J. Astorga-Wells, and R. A. Zubarev, "Rapid and deep human proteome analysis by single-dimension shotgun proteomics," *Mol. Cell. Proteomics*, vol. 12, pp. 3330–3338, Nov. 2013.

[72] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "The one hour yeast proteome," *Mol. Cell. Proteomics*, vol. 13, pp. 339–347, Jan. 2014.

[73] A. S. Hebert, S. Prasad, M. W. Belford, D. J. Bailey, G. C. McAlister, S. E. Abbatiello, R. Huguet, E. R. Wouters, J.-J. Dunyach, D. R. Brademan, M. S. Westphall, and J. J. Coon, "Comprehensive Single-Shot proteomics with FAIMS on a hybrid orbitrap mass spectrometer," *Anal. Chem.*, vol. 90, pp. 9529–9537, Aug. 2018.

[74] S. J. Valentine, A. E. Counterman, and D. E. Clemmer, "A database of 660 peptide ion cross sections: use of intrinsic size parameters for bona fide predictions of cross sections," *J. Am. Soc. Mass Spectrom.*, vol. 10, pp. 1188–1211, Nov. 1999.

[75] S. J. Valentine, A. E. Counterman, C. S. Hoaglund-Hyzer, and D. E. Clemmer, "Intrinsic amino acid size parameters from a series of 113 Lysine-Terminated tryptic digest peptide ions," 1999.

[76] B. Wang, S. Valentine, M. Plasencia, S. Raghuraman, and X. Zhang, "Artificial neural networks for the prediction of peptide drift time in ion mobility mass spectrometry," *BMC Bioinformatics*, vol. 11, p. 182, Apr. 2010.

[77] A. R. Shah, K. Agarwal, E. S. Baker, M. Singhal, A. M. Mayampurath, Y. M. Ibrahim, L. J. Kangas, M. E. Monroe, R. Zhao, M. E. Belov, G. A. Anderson, and R. D. Smith, "Machine learning based prediction for peptide drift times in ion mobility spectrometry," *Bioinformatics*, vol. 26, pp. 1601–1607, July 2010.

[78] S. Tiwary, R. Levy, P. Gutenbrunner, F. Salinas Soto, K. K. Palaniappan, L. Deming, M. Berndl, A. Brant, P. Cimermancic, and J. Cox, "High-quality MS/MS spectrum

prediction for data-dependent and data-independent acquisition data analysis," *Nat. Methods*, vol. 16, pp. 519–525, June 2019.

[79] S. Gessulat, T. Schmidt, D. P. Zolg, P. Samaras, K. Schnatbaum, J. Zerweck, T. Knaute, J. Rechenberger, B. Delanghe, A. Huhmer, U. Reimer, H.-C. Ehrlich, S. Aiche, B. Kuster, and M. Wilhelm, "Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning," *Nat. Methods*, vol. 16, pp. 509–518, June 2019.

[80] C. Ma, Y. Ren, J. Yang, Z. Ren, H. Yang, and S. Liu, "Improved peptide retention time prediction in liquid chromatography through deep learning," *Anal. Chem.*, vol. 90, pp. 10881–10888, Sept. 2018.

[81] V. A. Fusaro, D. R. Mani, J. P. Mesirov, and S. A. Carr, "Prediction of high-responding peptides for targeted protein assays by mass spectrometry," *Nat. Biotechnol.*, vol. 27, pp. 190–198, Feb. 2009.

[82] F. Meier, N. D. Köhler, A.-D. Brunner, J.-M. H. Wanka, E. Voytik, M. T. Strauss, F. J. Theis, and M. Mann, "Deep learning the collisional cross sections of the peptide universe from a million experimental values," *Nat. Commun.*, vol. 12, p. 1185, Feb. 2021.

[83] A. R. Ahmad Izaham, C.-S. Ang, S. Nie, L. E. Bird, N. A. Williamson, and N. E. Scott, "What are we missing by using hydrophilic enrichment? improving bacterial

glycoproteome coverage using total proteome and FAIMS analyses," *J. Proteome Res.*, vol. 20, pp. 599–612, Jan. 2021.

[84] M. A. Baird and A. A. Shvartsburg, "Localization of Post-Translational modifications in peptide mixtures via High-Resolution differential ion mobility separations followed by electron transfer dissociation," *J. Am. Soc. Mass Spectrom.*, vol. 27, pp. 2064–2070, Dec. 2016.

[85] L. K. Muehlbauer, A. S. Hebert, M. S. Westphall, E. Shishkova, and J. J. Coon, "Global phosphoproteome analysis using High-Field asymmetric waveform ion mobility spectrometry on a hybrid orbitrap mass spectrometer," *Anal. Chem.*, vol. 92, pp. 15959–15967, Dec. 2020.

[86] D. B. Bekker-Jensen, A. Martínez-Val, S. Steigerwald, P. Rüther, K. L. Fort, T. N. Arrey, A. Harder, A. Makarov, and J. V. Olsen, "A compact Quadrupole-Orbitrap mass spectrometer with FAIMS interface improves proteome coverage in short LC gradients," *Mol. Cell. Proteomics*, vol. 19, pp. 716–729, Apr. 2020.

[87] J. Meyer, N. M. Niemi, D. J. Pagliarini, and J. J. Coon, "Peptide identification and quantification by gas phase fractionation enables proteomics without liquid chromatography."

Chapter 7

GENERAL PUBLIC CHAPTER

JM wrote this chapter in collaboration with the Wisconsin Initative for Science Literacy (WISL).

Introduction

All human beings begin life as a single cell. The cell eventually divides into two cells, then four cells, all with nearly identical DNA. Over time these cells begin to diversify, some become skin cells while others become lung cells, organs with two very different appearances and functions. Yet these cells maintain their DNA similarity with only small differences that arise from errors in DNA replication as cells divide. Understandably some mistakes are made when copying 6 billion bases to replicate the DNA. If each of these bases were a word, it would be the equivalent of writing out Leo Tolstoy's War and Peace more than 10 times over. So how does the body develop the diversity to form these different cell types, when all cells read from the same DNA script? It would be like Mercedes Benz making every automobile they offer from a single blueprint. This differentiation occurs in part through the control of proteins, the molecular machinery of the cell, which are encoded in the genes of the DNA.

Where do proteins come from?

Proteins are made up of long chains of amino acids chemically bound together. Some people might be familiar with amino acids as a dietary supplement. Protein expression describes the generation of these proteins from the corresponding code in the DNA. Each cell must translate the nucleic acid sequence that make up the genes within DNA into the sequence of amino acids that make up proteins, similar to translating a book to another language.

Although DNA is made up of sequences of four possible letters, or bases, Adenine (A), Guanine (G), Thiamine (T) and Cytosine (C), proteins are made up of a sequence of 20 possible amino acids. Three letter segments of nucleic acid encode a single amino acid letter to allow for this increase in vocabulary (**Figure 7.1**). Segments of DNA are used as a template to first generate an intermediate nucleic acid called RNA, before being translated into protein sequences (**Figure 7.1**).

Many factors influence the quantity of proteins in a cell or tissue. Specific protein enzymes read the DNA, translate new proteins and degrade old proteins, controlling the protein life cycle. The accessibility of a DNA segment for reading, the speed of translation, and the speed of degradation all impact a protein's abundance (**Figure 7.1**). Some proteins travel to certain parts of the cell or exit the cell entirely. The different presence, levels and location of proteins in a cell differentiates the appearance and function of myocardial cells in the heart from neurons in the brain. By measuring the changing abundances of these proteins, researchers are able to better understand how cells function and whether they are sick or healthy. As a single cell divides, the resulting differential protein expression guides cells to form our lungs, hearts and brains. These proteins help form our "molecular self" and in organs like our brain they help form our memories, emotions and personalities. Similar to DNA, proteins make up who we are.

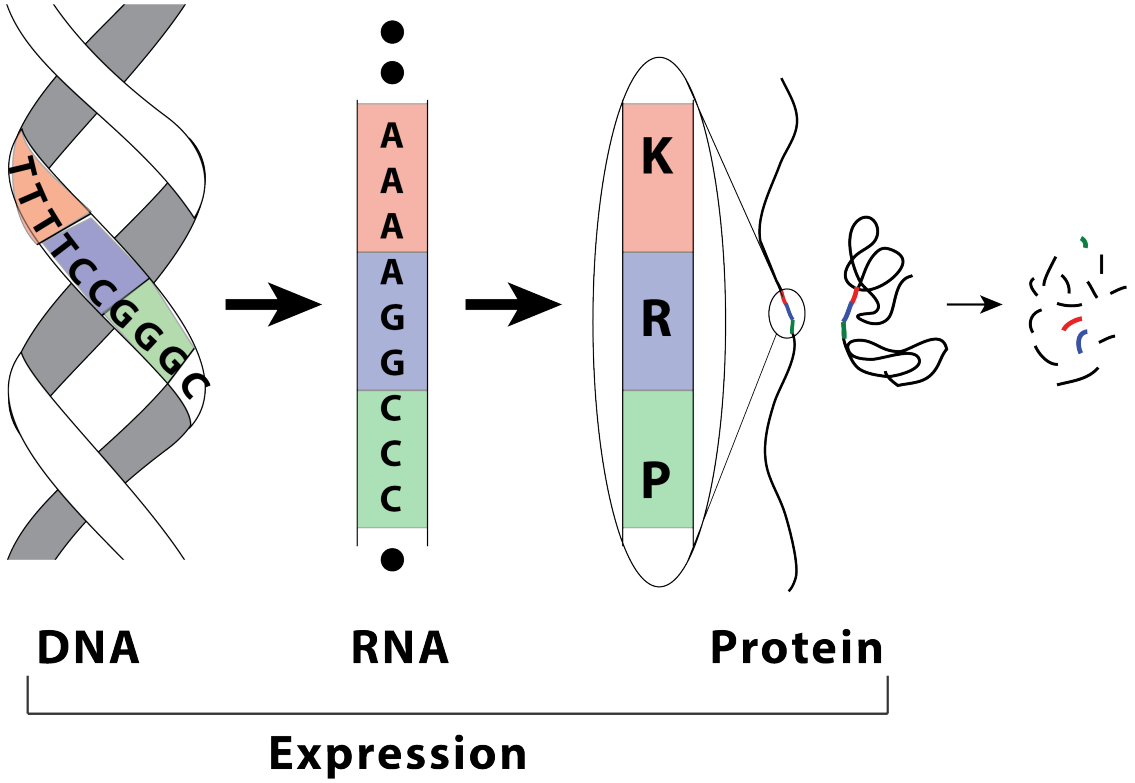


Figure 7.1: Translation, the process of generating proteins from DNA. Gene sequences within the DNA are first converted to sequences of RNA, another nucleic acid with a highly similar composition. This RNA is then used as a template for translation into the amino acid sequence of proteins. Protein strands fold into specific shapes that allow proteins to perform their functions. When proteins are no longer needed or start to malfunction, they are degraded back to their amino acid components by degradation machinery.

Proteins in the brain

In the brain, specialized proteins and protein structures allow the transmission of thoughts, storage of memories, control of movement, and the processing of sights, tastes and smells. Proteins extruded from the cell build scaffolding that guides the growth and spread of the tiny tendrils of neurons as the brain develops and matures[1]. This guide functions like a roadmap for growing cells, allowing the formation of the diverse structures of the brain. Protein receptors serve as docks for signaling molecules in the brain. These receptors allow brain cells to communicate feelings of happiness or anxiety, like telephone lines connecting buildings. Scientists have shown that proteins play a key role in storing our memories, with alterations in proteins leading to loss of memories in mice[1–3]. As you are reading this, protein shuttles transport the signal that tells your hands to move and allows your brain to process words. When researchers constructed a compendium of human proteins from 44 tissues, more than 30 proteins were found exclusively in the brain[4]. All of these molecules work cooperatively in your brain to construct your personality, and allow you to move, think and feel every day. Your brain is an incredibly precise instrument that is tuned and operated based on changes in protein activity, location, and abundance. Due to this precise operation, changes in protein characteristics can lead to disastrous and deadly diseases in the brain, such as Alzheimer's disease (AD).

Proteins in Alzheimer's disease (AD)

The two primary proteins in AD, tau protein and amyloid beta protein, damage the brain by clumping together in a process called aggregation (**Figure 7.2**). As AD progresses, sticky, thread-like tendrils of tau protein spread within neurons, eventually leading to cell death, through a largely unknown mechanism. Outside of the cell, amyloid protein aggregates to form large sticky clumps to which other proteins and amyloid molecules adhere. Many times, these clumps grow large enough to impede communication between neurons, like sticking a piece of gum on a circuit board. Interneuron communication plays an important role in transmission of thoughts and emotions as well as promoting growth of the neurons. Besieged by growing clusters from both inside and outside, many neurons will cease to function and die as AD progresses, inhibiting cognitive abilities. The death of these neurons has disastrous consequences as mature neurons no longer divide, meaning this cerebral deforestation is largely irreversible. The brain attempts to fight back against this aggregate onslaught, marshalling immune cells to remove the amyloid aggregates and damaged or dying neurons, like bulldozers removing debris of a demolished building. These cells are often overwhelmed by the sheer scope of growing protein clusters[5,6]. The challenging nature of this battlefield is reflected in the fact that no drug currently exists to substantially perturb or reverse the death of these brain cells, or neurodegeneration, associated with Alzheimer's. Understanding how different cell populations change and work to remove these disease aggregates plays a crucial role in advancing treatments for

Alzheimer's disease. By measuring the protein levels in the nervous system, we can identify AD-related changes in these cells.

Although our analysis occurs at a molecular level, that should not undercut the devastating cost of Alzheimer's on afflicted individuals and their families. In the US alone, an individual is diagnosed with Alzheimer's almost every 65 seconds[7]. As patients develop these protein clusters, they begin to struggle to identify faces and remember people. They find it more difficult to think of words and to speak. Many times, afflicted individuals will eventually struggle to dress themselves and perform basic functions, like climbing stairs[8]. Yet other patients, with similar genetic backgrounds or family history, show less severe symptoms or symptoms that advance more slowly[7,9–11]. The diverse effects and progression of AD makes developing treatment challenging[12], but also provides hope that AD can be managed by drugs or other therapies. If the field can identify the protein expression differences between rapid and slow declining groups, we can target the affected systems with drugs to slow the disease. We investigate the protein differences and similarities caused by disease using an analytical technique called mass spectrometry.

Quantifying protein levels with mass spectrometry

Mass spectrometry identifies and quantifies proteins by relying on the predictable behavior of charged molecules, or ions, in an electric field. If we think back to our introductory physics course, or our younger siblings' first relationship, opposites attract. This means that a positively charged molecule in an electric field will move towards the negative pole

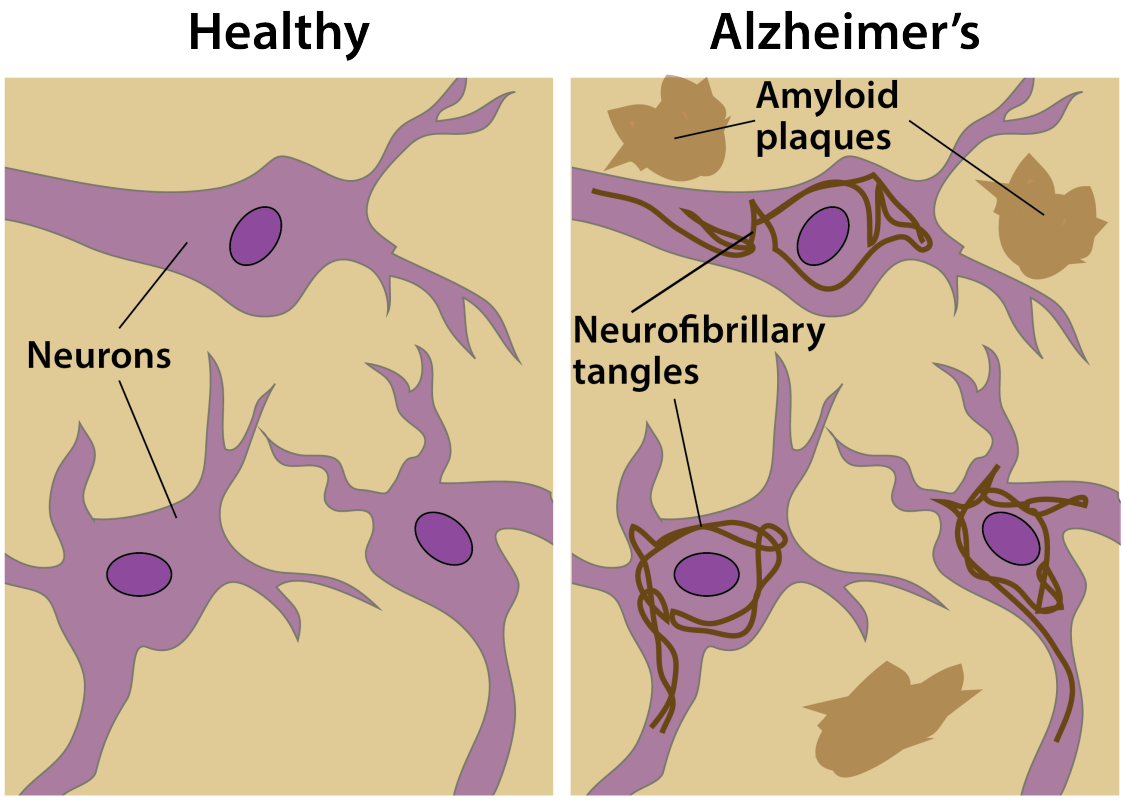


Figure 7.2: Aggregate formation in Alzheimer's disease. Tau protein clusters together inside of cells to form strand-like neurofibrillary tangles, while amyloid protein forms amorphous clumps outside the neurons, the primary cell of the brain.

(**Figure 7.3A**). The speed of the molecule will depend on two factors: the mass of the molecule and its charge. If we imagine the molecule as a car (**Figure 7.3B**), then the charge would be the number of wheels. If the charge is higher, then the force pushing the car increases, but if the car is heavier (greater mass) it takes more energy to speed it up and slow it down (**Figure 7.3C,D**). Due to this relationship, the ratio of mass to charge can be determined by applying an alternating electric field of known strength and measuring how ions accelerate and decelerate. Given that charge is a whole number, ions can often be identified using deductive reasoning for very simple mixtures with compounds of known mass.

We prepare samples by first breaking protein sequences into smaller sequences called peptides at specific amino acids (**Figure 7.4**). We then spray the peptides as tiny, charged droplets into the mass spectrometer (**Figure 7.4**). As the droplets fly through the air, they lose liquid until only the charged peptide remains (**Figure 7.4**). Inside of the instrument we determine the different peptides' mass-to-charge ratio. Once we have determined the mass-to-charge value we match that against a database of all human proteins and their component peptides in order to identify the peptide. The mass spectrometer uses an electric field to guide the peptides to strike a detector, which generates a signal proportional to the number of peptides colliding with it. We compare the relative abundances using this signal level. We determine protein relative abundance from their component peptide abundances.

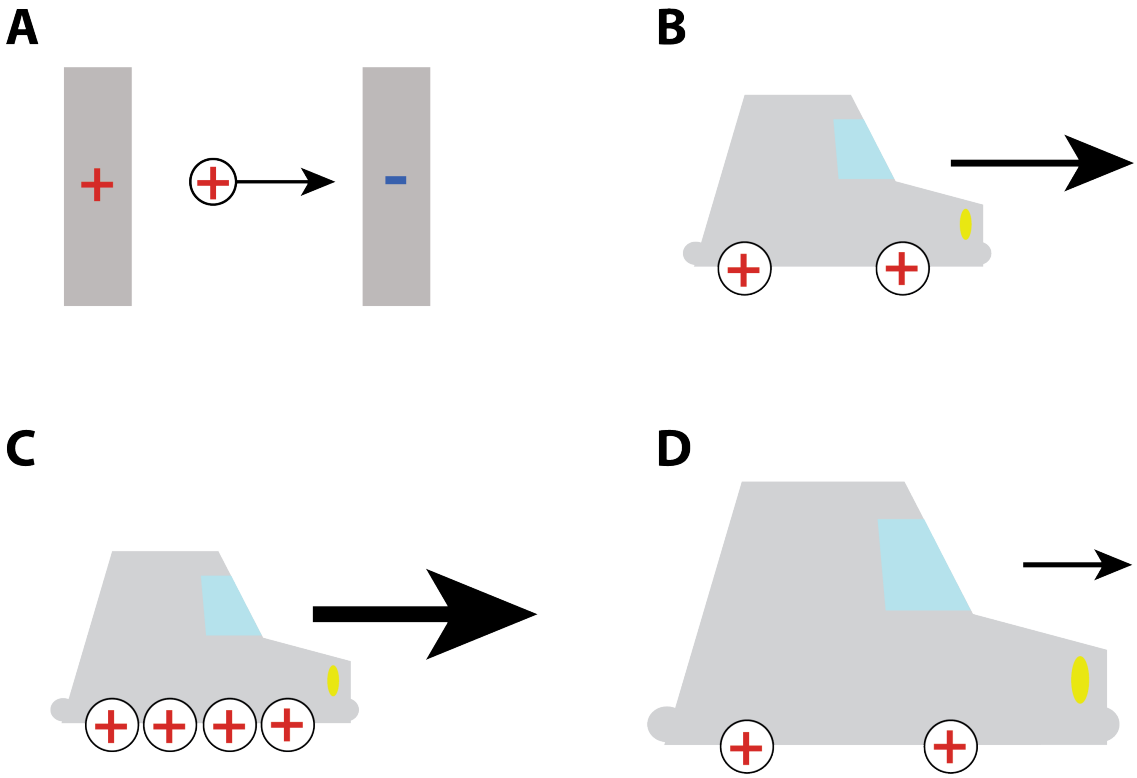


Figure 7.3: Predictable movement of charged molecules. (A) Positively charged ion moves towards the negative pole in an electric field. (B) If we imagine our charged molecule or ion as a car, wheels(charge) and weight(mass) of the car determine how fast it can accelerate (C) If more there is more charge but the same weight, the car can accelerate much more quickly (D) but a much larger car with the same number of wheels accelerates much more slowly.

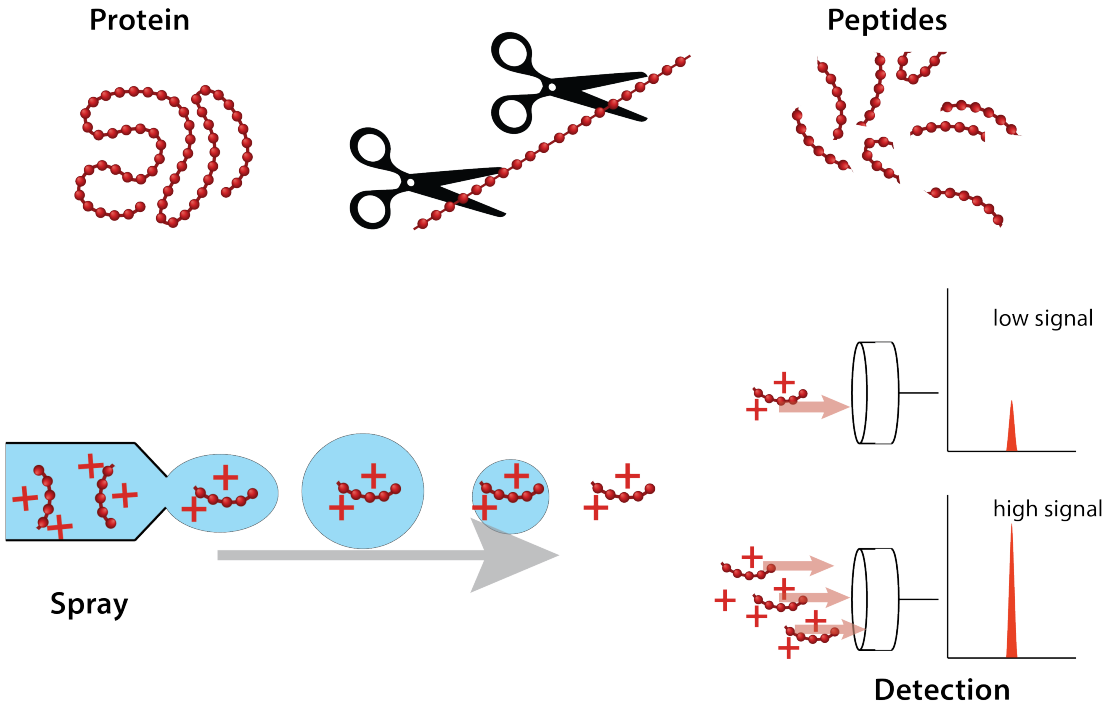


Figure 7.4: Detecting protein amounts. Proteins are first cut into shorter sequences called peptides. These peptides are then sprayed in charged droplets which dry as they fly through the air, eventually losing all of their liquid and leaving behind the charged peptide. Inside of the mass spectrometer, these peptides strike a detector and generate an electrical signal proportional to the number of peptide ions.

Regional brain protein analysis

A diverse set of structures and cell populations make up the different regions of the brain similar to neighborhoods in a city. The cerebellum, which controls aspects of movement and language, contains a high density of cells, like the busy downtown. While the caudate nucleus, which controls aspects of learning, consists of a long, stranded structure like a coastal peninsula (**Figure 7.5A**). Both of these structures contain differing neuron populations, which facilitate their different functions. The caudate nucleus houses a high proportion of spiny projection neurons, while the cerebellum contains many more Purkinje cells, a highly branched type of neuron. Differences like this exist for many of the structures and areas of the brain. Given these differences, we hypothesized that the damage caused by the protein aggregates described above may lead to unique responses for the different areas of the brain. We also theorized that differential regional effects in brain proteins could contribute to the wide diversity of clinical symptoms in AD. We identified overlap between proteins specific to certain brain regions and those affected by disease, suggesting that regional disease effects do exist. We also identified region-specific proteins that were altered by the normal aging process, a process closely intertwined with the development of Alzheimer's.

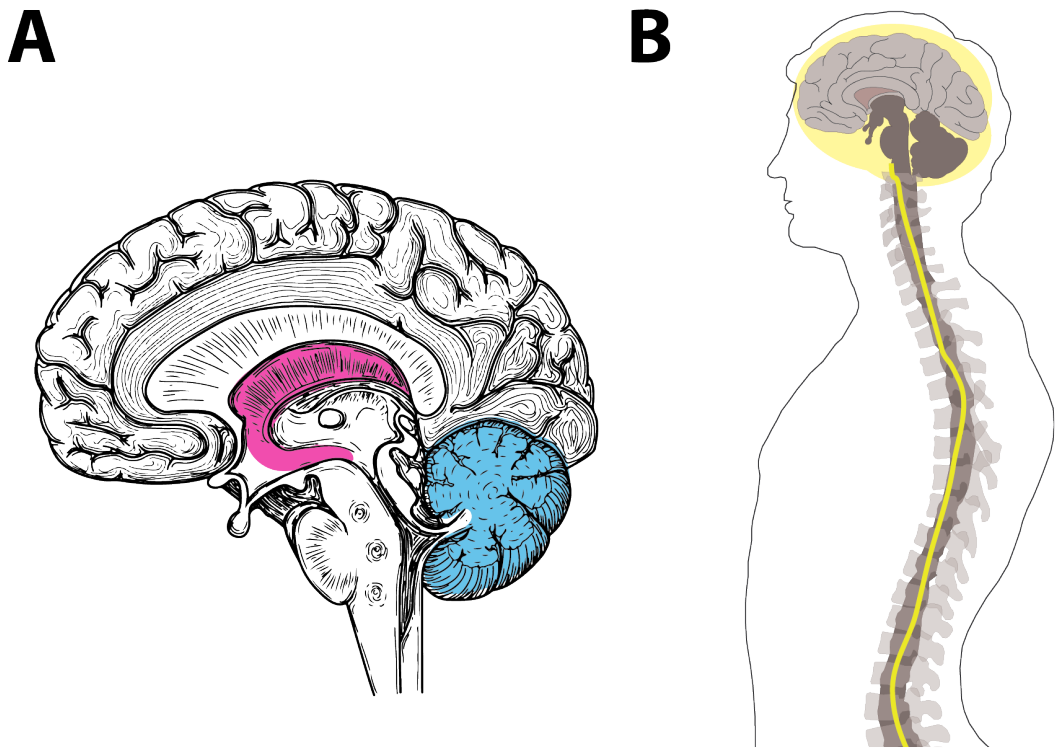


Figure 7.5: Diverse structures of the human nervous system. (A) Different regions of the brain take different shapes, with the cerebellum, pictured in blue, roughly circular and highly dense, and the caudate nucleus, pictured in pink, more strand-like. (B) Cerebrospinal fluid, indicated in yellow, surrounds the brain and flows up and down the spinal cord, delivering nutrients and removing waste.

Analysis using cerebrospinal fluid (CSF)

Our investigation of brain region-specific effects in Alzheimer's relied entirely on post-mortem tissue. This type of analysis provides only a snapshot of the final stage of the disease, forcing investigators to piece together processes from the destruction left behind, like detectives at a crime scene. Developing effective treatments for AD requires the ability to detect, identify, and diagnose the disease in its earliest stages, based on both clinical symptoms and molecular indicators. This requires analysis tools and strategies with the capacity to measure protein changes from living participants as disease develops and progresses. We sought to meet this need with an analysis of cerebrospinal fluid, which can be collected from living participants, as they develop symptoms or age normally. Cerebrospinal fluid (CSF) flows along the spinal cord and surrounds the brain, physically protecting the brain, supplying nutrients and removing molecular waste (**Figure 7.5B**). By studying the CSF, we can better understand the function of the brain as a whole in both healthy and diseased states. We performed a pilot analysis of CSF from twenty individuals, half of whom were diagnosed with AD, in order to test our capacity to detect protein differences. We identified more than 80 proteins to be associated with disease. Starting from vials of CSF, we prepared the samples and collected and analyzed the data in only five days. Rapid analyses of this type allow researchers to expand the number of samples analyzed in a study or clinical trial, leading to more representative results. This analysis strategy provides excellent scalability, allowing the theoretical analysis of more than 100

samples in less than two weeks.

Conclusions and future directions

The brain serves as a vault in which we keep our most treasured pieces of ourselves, our memories, our experiences, our emotions. Although all human brains perform the same basic functions, each is incredibly unique and completely irreplaceable. The brain inspires wonder at the impressive feat of bioengineering that allows it to function, and fear surrounding our own molecular mortality. While it is one of the most heavily studied organs, there is still so much we don't know about the brain. This mystery becomes a challenge when studying diseases that affect the brain such as Alzheimer's. Our findings provide general information about the cellular effects of AD, as well as additional foundational tools that could help future researchers develop precise and personalized treatments.

The distinctiveness of each human brain leads to unique effects of neurodegenerative disease regarding both symptoms and progression. Currently, Alzheimer's works in a devastating and insidious manner, moving across the brain, leaving damaged and dead neurons in its wake. The unpredictability of its progression only adds to the emotional trauma of patients and their loved ones, making it difficult to know what symptoms will develop and how rapidly they will worsen. This uncertainty also presents a challenge to researchers and doctors, as they attempt to develop treatments that prevent and reverse the advancing neurodegeneration. As different brain regions are affected by disease, they cause different symptoms, while the widespread nature of aggregates drives the severity

of these symptoms. Our streamlined analysis of cerebrospinal fluid in AD allows for the construction of a disease timeline, mapping specific protein changes to the different stages of disease and symptoms in the clinic. This analysis allows for more than 100 samples to be analyzed in less than two weeks. When combining this timeline with a brain region-specific atlas, AD progression can be tracked by both severity and location. In the short-term, this information will help neurologists prescribe preventive measures using better predictions regarding disease developments. In the long-term, as researchers develop compounds and drugs to inhibit or alter specific processes in Alzheimer's, these drugs can be strategically targeted to patients that would benefit most. Protein resources such as the atlas and timeline can be combined with high-throughput genome sequencing to create detailed and personalized molecular snapshots for individual patients. Diagnosis of AD-associated dementia can then be paired with highly specific drug and lifestyle regiments to allow for improved outcomes. Information about affected regions, molecular timelines, genetics, and clinical symptoms can be used to build a map of the neurodegeneration landscape, allowing medical practitioners to equip brain cells to succeed against Alzheimer's with tactical precision.

References

[1] C. S. Barros, S. J. Franco, and U. Müller, "Extracellular matrix: functions in the nervous system," *Cold Spring Harb. Perspect. Biol.*, vol. 3, p. a005108, Jan. 2011.

[2] X. Cao, H. Wang, B. Mei, S. An, L. Yin, L. P. Wang, and J. Z. Tsien, "Inducible and selective erasure of memories in the mouse brain via chemical-genetic manipulation," *Neuron*, vol. 60, pp. 353–366, Oct. 2008.

[3] S.-H. Lee, J.-H. Choi, N. Lee, H.-R. Lee, J.-I. Kim, N.-K. Yu, S.-L. Choi, S.-H. Lee, H. Kim, and B.-K. Kaang, "Synaptic protein degradation underlies destabilization of retrieved fear memory," 2008.

[4] M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, H. Wernerus, L. Björling, and F. Ponten, "Towards a knowledge-based human protein atlas," *Nat. Biotechnol.*, vol. 28, pp. 1248–1250, Dec. 2010.

[5] S. E. Hickman, E. K. Allison, and J. El Khoury, "Microglial dysfunction and defective -amyloid clearance pathways in aging alzheimer's disease mice," 2008.

[6] L. Rajendran and R. C. Paolicelli, "Microglia-Mediated synapse loss in alzheimer's disease," *J. Neurosci.*, vol. 38, pp. 2911–2919, Mar. 2018.

[7] C. Lynch, "World alzheimer report 2019: Attitudes to dementia, a global survey," 2020.

[8] D. A. Wolk and B. C. Dickerson, "Clinical features and diagnosis of alzheimer disease," *UpToDate, Waltham, MA*, 2016.

[9] D. Ferreira, L.-O. Wahlund, and E. Westman, "The heterogeneity within alzheimer's disease," *Aging*, vol. 10, pp. 3058–3060, Nov. 2018.

[10] N. L. Komarova and C. J. Thalhauser, "High degree of heterogeneity in alzheimer's disease progression patterns," *PLoS Comput. Biol.*, vol. 7, p. e1002251, Nov. 2011.

[11] W. M. Van der Flier, "Clinical heterogeneity in familial alzheimer's disease," *Lancet Neurol.*, vol. 15, pp. 1296–1298, Dec. 2016.

[12] G. Devi and P. Scheltens, "Heterogeneity of alzheimer's disease: consequence for drug trials?," *Alzheimers. Res. Ther.*, vol. 10, p. 122, Dec. 2018.

COLOPHON

This document was typesetted with LaTeX 2ε using Overleaf. It is based on the University of Wisconsin dissertation template created by William C. Benton (available at `https://github.com/willb/wi-thesis-template`).