THE DEVELOPMENT AND IMPLEMENTATION OF MASS SPECTROMETRY METHODS FOR THE CHARACTERIZATION OF PROTEINS AND METABOLITES

by

Matthew J. P. Rush

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

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN–MADISON

2018

Date of final oral examination: 05/29/2018

The dissertation is approved by the following members of the Final Oral Committee:
Joshua J. Coon, Professor, Chemistry and Biomolecular Chemistry
Lingjun Li, Professor, Chemistry and Pharmacy
Ying Ge, Associate Professor, Chemistry and Cell and Regenerative Biology
David J. Pagliarini, Associate Professor, Biochemistry

ACKNOWLEDGMENTS

I would like to begin by thanking Professor Joshua Coon, my research advisor, for his support and for fostering the environment I have worked in for the past five years. I came into graduate school unsure of what focus within analytical chemistry I wanted to take, and I can say that I undoubtedly landed in the right place. He has facilitated a research program where I was able to work on incredibly interesting projects and utilize state-of-the-art instrumentation, while being surrounded by motivated and knowledgeable coworkers. I would also like to extend my thanks to the team of people who worked with me during my time here. Every member of the lab that I overlapped with, present and past, have shaped my graduate school experience for the better. I think of my initial interactions with Chris Rose, who was the first person to train me on mass spectrometry and introduce me to the fundamental principles of their operation. I also remember my first days of GC-MS, when Arne Ulbrich showed me the ropes on the original GC Orbitrap system, including translating the German instructions into English for me. I would also like to acknowledge Derek Bailey and Nick Kwiecien who taught me to program and provided data analysis tools that enabled my research. I would also like to deeply thank Kyle Connors and Jacob Buboltz who I was fortunate to mentor during my time here. Their intelligence and commitment made my job a breeze, and working with them was one of my most rewarding experiences. I am confident they will go on to successful careers and I hope their scientific training benefits them along the way. Similarly, I would like to thank those who

mentored me, particularly Mike Westphall, Alex Hebert, and Jason Russell who were always there when I needed concepts explained or guidance on instrument issues or experimental designs. I would also like to thank those who went through the program with me, Elyse Freiberger, Evgenia Shishkova, and Emily Wilkerson. Having them around made the trying times much easier. I have to thank Paul Hutchins as well for tolerating sitting across the desk from me for the past four years and for all the lunch conversations that we used to have before we ran out of things to talk about. I know I say some pretty outlandish stuff sometimes and Paul always engaged with my crazy theories. Outside of the lab I was fortunate to have made some great friends in Madison. Sean Staudt, Geoff Eddinger, Amber Sledzik, and Jon Voss were always down to have a beer or a coke and grenadine when I needed to unwind for an evening. Tim Rhoads deserves a special acknowledgement for introducing me to the wonderful world of craft beer which I am certain will be a lifelong interest of mine. I would also like to thank Kayla Sailor who always believed in me and has been a source of encouragement and joy during my time in Wisconsin. And lastly, I have to give an enormous thank you to Nick Riley. He has been a second advisor to me. He has probably edited everything I have written over the last five years, and has been a cornerstone member of the lab. The early days of NETD were the most exciting times I have ever had in my working life and I am grateful I was able to share them with him.

I would also like to acknowledge some of the teachers I have had throughout my life who have shaped me into the person I am today and have helped guide my academic pursuits. Mrs. Kimsey sparked my early interest in computers. I remember in elementary school I would come in early to help make the announcement slides that would run on the hallway monitors. Little did I know I would continue to employ this skillset in graduate school as the resident A/V specialist for the lab. But in all honesty, this initial introduction to computers and her encouragement made a permanent impact on me and certainly influenced the person that I have become. Next I would like to thank Jenay Weekly. Mrs. Weekly had an incredible knack for making any subject fascinating. I remember the science fair, disassembling computers, and messing around with an old record player as just some of the ways that she made learning fun. I thank Madamme Waldo for serving as an exceptional example of how a teacher can be challenging and engaging. Thank you for tolerating me during one of my more rebellious streaks in high school. I had countless amazing professors during my time at Washburn University, but none had a greater influence on me than Dr. Stephen Angel. It was during his second semester General Chemistry class that I knew for certain I would pursue chemistry as a career. I would also like to thank Dr. Angel's math counterpart, Dr. Donna LaLonde, for being such a wonderful teacher and person. I thank her for her contagious and inspirational joy of learning during my linear algebra and computer programming courses. Dr. Sam Leung first introduced me to academic research and was an extremely thoughtful and positive advisor. I also would like to thank Kellis Bayless. It was through Kellis's Human Impact on the Environment course that I renewed my love of science and passion for the environment. I hope the gardens are looking good this year. I thank the many friends that I passed the hours with in Stoffer Science Hall, notably Corey Suelter, Christian Gomez, Mike Benfer, Taylor Easley, Diana Crain, Yangfei

Mao, and Wendy Logan. These individuals truly made me feel like part of the fabric of the Washurn community.

I have been lucky to have many great friends, but none better than Tom Rosen and Willie Christ. While we have not all lived in the same place for a long time, I ensure we don't lose touch for long by my all too regular phone calls. I'm sure I could right pages about times we have spent together over the past 25 years, but most of them should remain undocumented. But I'll just say that it's been nice growing up with you guys, you are my brothers. Hope the bagel business continues to flourish, and I'll be down to Austin as often as I can. Keep me posted if you find a spot for a chemist and I'll pack my bags immediately. I was extremely privileged to grow up with two sets of parents. Libby and Eric Rosen have been with me from the beginning and are people that I aspire to be like. I would also like to thank my Grandpa Allen and Grandma Evelyn for their constant support and love. My Grandma Donna and Uncle Kim have had enormous bearing on the person that I have become. Thank you both for everything. My sister Sarah has also been an incredible support to me. As we get older I realize more and more all of the subtle ways we are similar. And lastly, I must thank my parents Mark and Ann Rush who raised me to follow my dreams. Without their love and support I would never have accomplished this goal. Thank you for being the best parents and role models that I could ever ask for.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

4-HB	4-Hydroxybenzoic acid
4-HBz	4-Hydroxybenzaldehyde
4-HMA	4-hydroxymandelate
4-HPA	4-hydroxyphenylacetate
4-HPAA	4-hydroxyphenylacetaldehyde
4-HPL	4-hydroxyphenyllactate
A	Ampere
ACN	Acetonitrile
AGC	Automatic gain control
AI-ETD	Activated ion electron transfer dissociation
AI-NETD	Activated ion negative electron transfer dissociation
ALDH	Aldehyde dehydrogenase
ATP	Adenosine triphosphate
BME	2-mercaptoethanol
C	Celsius
C#	C sharp (a programming language)
CAA	Chloroacetamide
CAD	Collision-activated dissociation
CHI	Chalcone isomerases

CI	Chemical ionization
COMPASS	Coon OMSSA Proteomic Analysis Software Suite
CoQ	Coenzyme Q
Da	Dalton (the atomic mass unit)
DNA	Deoxyribonucleic acid
ECD	Electron-capture dissociation
EDD	Electron detachment dissociation
EET	Excitation energy transfer
EI	Electron ionization
ETD	Electron transfer dissociation
ETF	Electron transfer flavoprotein
EThcD	ETD with supplemental activation of all reaction products via HCD
ETnoD	Non-dissociative electron transfer (electron transfer without dissociation)
eV	Electronvolt
FC	Fold-change
FDR	False discovery rate
g	Gram
GC	Gas-chromatography
GO	Gene ontology
h	Hour

HCD	Higher-energy collisional dissociation
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HRF	High-resolution filtering
I.C.	Ion current
i.d.	Inner diameter
IE	Ionization energy
IR	Infrared
IRMPD	Infrared multiphoton dissociation
L	Liter
LC-MS/MS	Liquid chromatography coupled online with tandem mass spectrometry
LFQ	Label free quantitation
Lys-C	Endoproteinase lysine C
m	Meter
M	Molar concentration
m	Mass
m/z	Mass-to-charge ratio
M+H	Protonated molecular ion
MCNA	Molecule covariance network analysis
MDC	Multipurpose dissociation cell
min	Minute

mRNA	Messenger ribonucleic acid
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MS^1	Survey mass analysis
MSTFA	N-Methyl-N-(trimethylsilyl) trifluoroacetamide
MW	Molecular weight
MXP	Mitochondrial uncharacterized protein
NAD+	Oxidized nicotinamide adenine dinucleotide
nce	Normalized collision energy
NETD	Negative electron transfer dissociation
NETnoD	Non-dissociative negative electron transfer (negative electron transfer without dissociation)
NIST	National Institute of Standards in Technology
nLC	Nanoflow liquid chromatography
o.d.	Outer diameter
OD	Optical density
OMSSA	Open Mass Spectrometry Search Algorithm
OxPhos	Oxidative phosphorylation
pABA	Para-aminobenzoic acid
pABA-	Para-aminobenzoic acid depleted media

PC	Principal component
PCR	Polymerase chain reaction
pH	Potential of hydrogen
PMSF	Phenylmethane sulfonyl fluoride
PPHB	3-polyprenyl-4-hydroxybenzoate
ppm	Part per million
psi	Pounds per square inch
PSM	Peptide-spectrum match
PTM	Post-translational modification
p-value	Probability value
q-value	Unit-less dimension for ion motion derived from Mathieu equation
RC	Respiration competent
RD	Respiration-deficient
RDR	Respiration deficiency response
ROC	Receiver operating characteristic
rpm	Rotations per minute
s	Second
s.d.	Standard deviation
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
S/N	Signal-to-noise ratio

TCEP	Tris(2-carboxyethyl)phosphine
TFA	Trifluoroacetic acid
Th	Thomson (the unit of the mass-to-charge ratio)
TIC	Total-ion chromatogram
TMCS	Trimethylsilyl chloride
TMS	Trimethylsilyl
Tris	Tris(hydroxymethyl)aminomethane
Tyr	Tyrosine
UV	Utraviolet
UVPD	Ultraviolet photodissociation
V	Volt
w	Weight
W	Watt
WT	Wild-type
Y3K	Yeast three thousand
z	Charge
Δgene	Yeast knockout strain

ABSTRACT

The work described in this dissertation involves the development and application of mass spectrometry methods for the analysis of peptides, proteins, and metabolites. **Chapter 1** gives a brief overview of the history and fundamentals of bioanalytical mass spectrometry, the utility of electron transfer dissociation (ETD) for the analysis of peptides and proteins, and the use of high resolution gas-chromatography mass spectrometry (GC-MS) to characterize volatile small molecule metabolites from complex mixtures. **Chapter 2** explores the role of the reagent cation in negative electron transfer dissociation (NETD) for the analysis of peptide anions. Then **Chapter 3** demonstrates the use of activated-ion NETD (AI-NETD) to analyze the negative-mode proteome of yeast. This work proved to generate far greater depth of coverage than previous negative-mode experiments, owing largely to the introduction of infrared radiation concurrent to the NETD reaction. For this study, a high pH chromatography method was also developed to greatly improve the ionization of peptide anions using negative electrospray ionization. **Chapter 4** characterizes activated-ion ETD (AI-ETD) as a fragmentation method to interrogate proteins with intact disulfide bonds. This method greatly improved the sequence coverage and sequence ion generation compared to other commonly used fragmentation techniques for a set standard proteins. **Chapter 5** outlines the creation of a high resolution metabolite mass spectral library using a Q Exactive GC mass spectrometer. This library is then employed to identify metabolites from yeast and human cell cultures, significantly improving the identification confidence

over commercially available spectral libraries. Lastly, **Chapter 6** describes the multi-omic analysis of yeast strains with a single gene deletion, aiming to correlate genes of known function with genes unknown function.

Chapter 1

INTRODUCTION

Portions of this chapter are adapted from published manuscripts and are reprinted with permission from:

Rose CM, **Rush MJP**, Riley NM, Merrill AE, Kwiecien NW, Holden DD, Mullen C, Westphall MS, Coon JJ. *A calibration routine for efficient ETD in large-scale proteomics.* Journal of the American Society for Mass Spectrometry. **2015**, *26 (11)*, 1848-1857. doi: 10.1007/s13361-015-1183-1.

Kwiecien, NW, Bailey, DJ, **Rush, MJP**, Colse, JS, Ulbrich, A, Hebert, AS, Westphall, MS, Coon JJ. *High-Resolution Filtering for Improved Small Molecule Identification via GC/MS.* Analytical Chemistry. **2015**, *87 (16)*, 8328-8335. doi: 10.1021/acs.analchem.5b01503.

Background

The blueprint for life is written using a sequence of small organic molecules called nucleotides which string together to create deoxyribonucleic acid (DNA). Segments of DNA encode for a particular biological process (gene), such as the sequence of a protein. The nucleic acids of DNA are read in sets of three, with each set corresponding to an amino acid. The DNA molecule is transcribed by the enzyme RNA polymerase to create a messenger ribonucleic acid (mRNA). This mRNA molecule can then be used by the ribosome to create a protein comprised of the appropriate sequence of amino acids. This process is commonly referred to as the central dogma of molecular biology. Proteins are incredibly dynamic, performing a wide variety of cellular functions. Metabolism comprises the set of biochemical reactions within an organism which sustain life, where commonly the protein actors perform a biochemical transformation on a small molecule metabolite for a plethora of purposes, such as to produce energy in the case of the citric acid cycle. These classes of biomolecules comprise the majority of matter within a cell. The ability to identify them and measure their abundance is key to increasing our understanding of the complex interplay underlying all biochemical processes within an organism. Within the past decade enormous advances have been made to this end. The ability to globally measure the DNA and RNA within an organism leapt forward with the Human Genome Project and has become routine thanks to the invention of high-throughput sequencing technologies.[1–3] However, the complete measurement of the more dynamic proteome and metabolome of organisms is not yet possible to the same degree. The work described in this dissertation discusses mass spectrometry methods to advance the study of proteins and metabolites within biological samples.

Bioanalytical mass spectrometry

For decades mass spectrometry (MS) has been a cornerstone analytical technique for measuring chemical compounds. Mass spectrometers are instruments which are able to detect the mass to charge ratio (*m/z*) of compounds which are introduced into the system. Mass is not measured directly, as the mass spectrometer requires the use of charged particles (ions) for detection, because the detectors used measure the charge induced or current produced by the ions as it passes by or hits the surface of the detector. Charged particles can also be manipulated by electric or magnetic fields, allowing mass spectrometers to shuffle ions around, trap ions in particular segments of the instrument, mix populations of ions, perform gas phase ion chemistry, or select ions with particular *m/z* values and eliminate those ions that are not important. Because mass spectrometers rely upon analyzing charged gas phase particles, the production of gas phase ions has been an area of active research for decades. There are numerous methods for creating ions which fall into the categories of hard ionization or soft ionization. Hard ionization techniques are those which ionize the chemical compound while concurrently fragmenting it. An example of this is electron ionization which shoots a beam of electrons directly at the analyte, ionizing the molecule and generating fragments which can be used to structurally elucidate the molecule.[4,5] Soft ionization ionizes the molecule with little to no additional fragmentation. The two premier methods for soft ionization are matrix-assisted laser desorption ionization (MALDI), developed by Koichi Tanaka and coworkers, and electrospray ionization (ESI), invented by John Fenn.[6–8] These two methods, developed in the late 1980's, opened the way for the analysis of large biomolecules by mass spectrometry and the inventors shared the Nobel Prize in chemistry in 2002 for their contribution.

The work described in this dissertation utilizes three distinct mass analyzers. The quadrupole (Q) mass filter is a device comprising four parallel linear rods where a radio frequency (RF) voltage is applied to one set of opposing rods and a direct current (DC) offset voltage is applied to the other pair of rods. By manipulating the voltages applied to the sets of rods, regions of the *m/z* space can be selectively transmitted through the quadrupole device and allowed to reach the detector. Commonly a set of three quadrupoles are used in series, called a triple quadrupole mass spectrometer. In this arrangement, the first quadrupole is used as a mass filter to select a specific analyte, the second quadrupole operates as a collision cell where the analyte is fragmented, and the third quadrupole can be used to select a particular fragment ion or can scan the entire fragmentation spectrum.[9] The next mass analyzer is the 2-dimensional quadrupole ion trap, which operates similarly to a quadrupole mass filter, but can confine ions within the cell. This is accomplished by separating the quadrupole into three segments, one center section and two end electrodes. A RF is applied to trap the ions radially and an electric potential is applied to the end electrodes to confine the ions axially. This analyzer can then operate as a mass-selective filter or an ion trap, providing flexibility to the device and facilitating ion/ion reactions.[10] The last mass analyzer is the Orbitrap, which is an ion trap where an inner spindle-shaped electrode is surrounded by an outer electrode.[11,12] The ions are injected into the Orbitrap with enough energy that their inertia counterbalances the electrostatic attraction to the inner electrode. Ions then rotate around the center electrode radially forming helical packets. They also oscillate axially at rates dependent on their *m/z* ratio. In this way, ions separate out based on *m/z*, with each packet inducing an image current on the outer electrode. The image current for each packet is measured and the *m/z* value can be calculated from the frequency of the ion packet axial oscillation within the trap. The ability to continuously detect the image current for each ion packet over extended periods of time allows the Orbitrap mass analyzer to have substantially greater mass resolution capabilities than the quadrupole or linear ion trap mass analyzers. Finally, some modern instruments integrate two or three mass analyzers together within a single instrument, frequently referred to as hybrid mass spectrometers. These systems allows for the most flexibility in experimental design and leverages each mass analyzers primary advantages. The two types of hybrid instruments used in this work comprise quadrupole-Orbitrap or linear ion trap-quadrupole-Orbitrap geometries.[13,14]

Frequently, mass spectrometers are paired with a front-end chromatographic separation method, such as gas or liquid chromatography (GC and LC, respectively), to simplify complex mixtures of analytes.[15,16] Discussion of GC-MS application to discovery metabolomics will be covered in the section “Metabolomics”. The use of LC-MS has been the tool of choice for the analysis of complex protein mixtures, such as those from cell culture or tissue samples. The workflow most commonly used in modern-day proteomics is shown in Figure 1. The sample is enzymatically digested into peptides, which are separated by liquid chromatography using a reverse-phase packing material and eluting peptides are ionized by electrospray ionization.[17] Once in the mass spectrometer a survey MS scan (MS^1) is performed and the most abundant ions are selected for secondary MS (MS^2), where the peptide is fragmented and the analytes sequence can be identified by matching to an *in silico* peptide database.[18,19] The most commonly used fragmentation methods are collisional-based, where peptide ions are collided with gas molecules in the mass spectrometer, however alternative fragmentation methods have been developed and are often preferred for certain experiments.

Electron transfer dissociation

As proteomics has advanced to the point of being able to detect representative peptides from thousands of proteins in a single analysis, there remain significant challenges regarding complete coverage of each protein in the proteome and access to a protein's post-translational modification complement. Most proteins in the cell can be modified post-translationally, but these modifications can easily be lost when peptides are fragmented using collision-based methods. This limitation spawned research into alternative fragmentation methods that utilize electrons, photons, or metastable atoms.[20–24] Electron transfer dissociation (ETD) has become one of the most prominent alternative dissociation methods, owing to

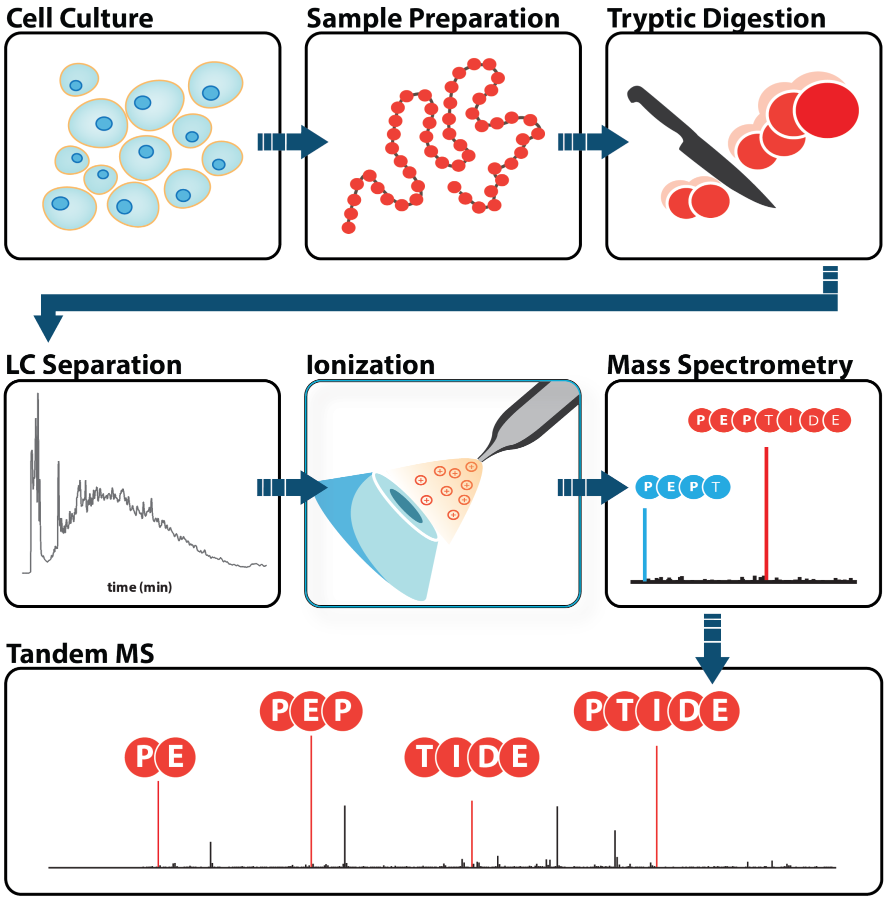


Figure 1.1: LC-MS proteomics workflow. An example LC-MS proteomics workflow is shown. Cells are lysed, and proteins are then enzymatically digested into peptides. Samples are then injected onto a reverse phase LC column where peptides are separated and eluting peptides are ionized by electrospray ionization. Survey MS scans are then performed on all eluting peptides and then individual peptides are selected for MS/MS analysis. The resulting raw data can then be searched using proteomics software, such as COMPASS or MaxQuant.

its compatibility to various MS instrument platforms and utility in fragmenting a wide variety of biomolecules.[25,26] ETD is an ion-ion reaction which uses a singly charged radical reagent anion to donate an electron to a peptide or protein cation, causing radical induced dissociation and generating *c*- and $z^{\bullet}$*-type* sequencing ions. This process has been shown to greatly increase the retention of labile PTMs, allowing for their localization on the protein.[27,28] ETD reactions are governed by many operational parameters including the precursor and reagent ion populations, precursor charge (z), ion-ion reaction vessel characteristics, and reaction duration.[29–35] Samples analyzed with traditional shotgun methodology comprise a diverse pool of precursor peptide cations so that optimal reaction conditions vary considerably from one scan to the next. For collisional activation methods this variation is accommodated by normalizing collision energy as a function of precursor mass and charge.[36,37] Ion/ion reactions, however, involve more parameters and normalization is not as straightforward. The ETD reaction follows pseudo-first order kinetics and can be modelled using an exponential decay function, **Equation 1**:

$$N_p(t) = N_p(0) \times e^{-k[R]t}$$

where $N_p(0)$ represents the initial precursor population, $N_p(t)$ is the amount of precursor remaining at time t, [R] is the average number density of ion cloud overlap, and k is the ion-ion reaction rate constant.[38–41] This rate constant is defined as (**Equation 2)**:

$$k = c(\,|\,\nu\,|\,) \times Z_p^{2} Z_r^{2} \times ([\mathrm{m_p} + \mathrm{m_r}]\ /\ \mathrm{m_p m_r})$$

where Z_p is precursor charge, m_p is precursor mass, Z_r is reagent charge, m_r is reagent mass, and the quantity $c(|\nu|)$ is a function of the distribution of the magnitude of differential velocities, $|\nu|$, of precursor and reagent ions in the overlapping ion clouds. Holding $N_p(0)$ and [R] constant, as a typical MS/MS experiment would, the ion-ion reaction is dependent upon *k* and *t* (**Equation 1**). Thus, optimal reaction duration (*t*) (*i.e.*, the shortest reaction time to achieve the maximal quality spectra), can be achieved by selecting the appropriate ETD rate constant, *k*. From **Equation 2**, *k* scales with the square of the precursor charge and the number of precursor and reagent ions. Note the contribution of reduced mass to *k* — for precursors spanning 1000 to 10,000 Da —is negligible relative to precursor charge and ion populations and, thus, is presumed constant. The upshot is that higher charge state precursors react more quickly, requiring reduced reaction times for ideal performance. Besides improving scan speed, shortened reactions limit the occurrence of secondary electron transfer events that erode signal and complicate spectra. Thus, for every selected precursor there is an ETD reaction time that will result in an optimal creation of product ions. **Figure 1.2** illustrates this concept by plotting the *c*- and $z^{\bullet}$-type product ion signal-to-noise (*S/N*) as a function of ion-ion reaction duration for triply protonated cations of angiotensin I (DRVYIHPFHL). In this case, the maximum product ion *S/N* is achieved using a 40 ms ion-ion reaction, which leaves ~12% of total MS/MS signal attributed to the intact precursor. Continuing the reaction beyond this duration consumes remaining precursor; however, product ion *S/N* is likewise reduced because of secondary electron transfer events.

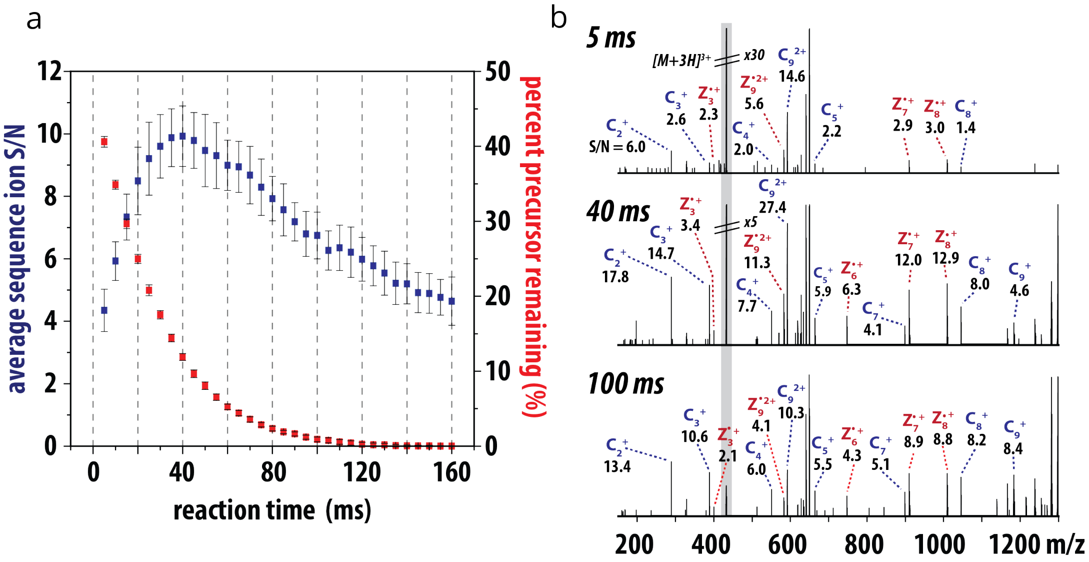


Figure 1.2: ETD reaction product ion generation. ETD reaction duration is a key variable for production of quality MS/MS spectra. Panel A plots the product ion *S/N* (blue) and rate of precursor (red, triply protonated angiotensin 1) consumption for reaction times ranging from 5 to 160 ms. From these data we conclude the maximal product ion *S/N* is achieved when between 10 and 15% of the precursor remains. Extension of the reaction beyond this point both degrades spectral quality and slows the instrument scan cycle. Single scan ETD MS/MS spectra from the 5, 40, and 100 ms reactions are shown in panel B. Note that while the precursor is nearly absent in the 100 ms scan, the overall product ion *S/N* is lower than the optimal 40 ms reaction.

Optimization of reaction time is one way of enhancing the effectiveness of ETD. Another is in improving the efficiency of product ion generation. This can be accomplished in multiple ways. One challenge with ETD is that some amount of electron transfer does not lead to dissociation (ETnoD). These products form due to non-covalent interactions that exist between the dissociated segments of a peptide or protein. By providing the electron transfer reaction with supplemental energy non-covalent interactions can be disrupted, allowing for the sequence-informative peptides to be detected. Several strategies have been developed to introduce this energy, two of which will be highlighted in this dissertation. The first is by increasing the reaction enthalpy of the electron transfer reaction.[42] Excess energy can redistribute into the peptide precursor, mitigating the extent of ETnoD reaction products. The second is the introduction of IR photons concurrent with the ETD reaction.[43] These photons impart energy into the peptide, heating the ions, and prevent the non-covalent interactions from forming.

Electron transfer dissociation continues to be utilized in new and inventive ways within the mass spectrometry community. ETD has emerged as an excellent fragmentation method for the profiling of post-translational modifications such as glycosylation and phosphorylation, the study of proton transfer and ion-ion reactions, and in structural characterization of native proteins and protein complexes.[44] Two areas for the application of ETD are explored within this dissertation: 1) the ion-ion reaction dynamics and fragmentation process of peptide anions using negative ETD (NETD) is used as a method for bottom-up proteomics (**Chapter 2**) and the development an LC-MS method leveraging activated ion

NETD (AI-NETD) for the description of the negative mode proteome of *Saccharomyces cerevisiae* (**Chapter 3**) and 2) the sequencing of proteins with intact disulfide bonds using AI-ETD (**Chapter 4**).

Metabolomics

Metabolites serve as a direct signature of biochemical activity in the system, and the metabolome is the collection of small molecules which are produced by the cells.[45] To fully understand the state of a cell, we would ideally be able to monitor the abundance of all molecules in the system. The ability to map the metabolome currently lags far behind genomics and proteomics for two primary reasons. The first is that the production of metabolites is non-template driven, which means that we cannot infer the entire compliment of expressed metabolites from an organism's genetic information, which is estimated to be in the tens of thousands for many complex organisms.[46,47] The second major hurdle is that metabolites contain enormous compositional heterogeneity. This makes the ability to assess a large swath of metabolites using a single analytical technique extremely difficult. This chemical diversity also means that metabolites cannot be identified en masse from mass spectral data using a single set of fragmentation rules without comparison to spectral libraries, which is possible for peptides or proteins.

At present, most metabolite measurements by mass spectrometry utilize a targeted approach, which allows for the analysis of a small number of metabolites of similar chemical makeup. However, recent advancements in separation, mass spectrometer instrumentation, and data analysis have facilitated a greater number of metabolites to be profiled using a single method. Lipids, a subset of metabolites generally consisting of a head group and nonpolar hydrocarbon tails, can be identified and quantified by the hundreds from biological samples.[48,49] This approach usually entails a reverse phase chromatographic separation with electrospray ionization and data dependent MS/MS. Because lipids fragment according to a set number of fragmentation rules, by analyzing standard lipids from each separate lipid class, sets of fragmentation rules can be generated which allows the identification of similar lipids without the need to analyze an authentic standards for each compound. However, the challenge of discovery analyses of other metabolite classes remains.

Gas chromatography is a technique which separates molecules based primarily on boiling point.[50] Because analytes are separated by boiling point, and not directly by polarity, both polar and non-polar molecules can be separated and analyzed. This makes it a premier tool for simplifying complex mixtures of volatile small molecule metabolites from biological samples, and as the analytes are already in the gas phase, makes it easily compatible with mass spectrometry.[51] To increase the breadth of molecules which can be analyzed by GC-MS, analytes can be derivatized to lower their boiling point. This facilitates the GC-MS analysis of molecules such as disaccharides and steroids, which would otherwise decompose before boiling at standard pressures. A typical GC-MS metabolomics workflow is shown in **Figure 1.3**. Here samples are extracted, dried, and resuspended in the derivatization solvent. The most common derivatization reaction is silylation which replaces acidic protons such as those found on alcohols, amines, or carboxylic acids with silyl groups in order to disrupt

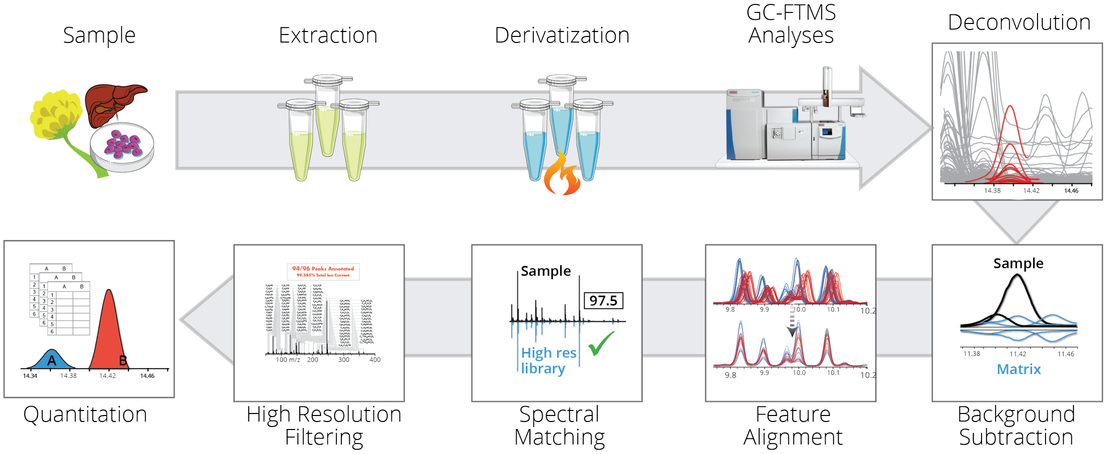


Figure 1.3: GC-MS metabolomics workflow. A sample GC-MS metabolomics is shown. First the sample is extracted to lyse the cells and extract metabolites. Then samples are dried and resuspended in a derivatization solvent and heated for 30 to 60 minutes. Next samples are injected onto a GC where metabolites are volatilized and separated. Eluting metabolites are then ionized, typically by electron ionization, and MS scans are performed on the resulting ions. The raw data files are then processed through a data analysis pipeline, which performed spectral deconvolution, background subtraction, feature grouping, spectral matching, and quantitation.

hydrogen bonding, which can lower the boiling point of molecules by more than 100° C.[52] The samples are then injected into a GC inlet where they are volatilized and then separated prior to introduction into the mass spectrometer. The two primary ionization modes for GC-MS are electron ionization and chemical ionization, described previously. Electron ionization is the most widely used and most metabolite spectral databases, which must be compared against for confident metabolite identification, are composed of spectra collected in EI mode using an electron energy of 70 eV.[53,54] However, because it is a hard ionization method, compounds are both ionized and fragmented in a single step. This is disadvantageous because the intact molecular ion is often absent from the spectrum, which confounds metabolite identification. For discovery GC-MS methods, the mass spectrometer can then be operated by collecting only MS scans, without the need to perform tandem MS/MS. Because all eluting molecules are fragmented and analyzed, the mass spectrometry data acts as a digital record of all volatile analytes in the samples, unlike targeted approaches or data dependent methods which only fragment selected species.

The GC-MS metabolomics work described in this dissertation uses a Q Exactive GC Orbitrap mass spectrometer. This instrument was initially conceived by the Coon lab and the prototype system was built here in collaboration with, and subsequently commercialized by, Thermo Fisher Scientific.[55–57] The commercial model was used to collect the GC-MS metabolomics analysis described here. The development of the high-resolution Orbitrap GC-MS system was motivated by the ability to increase compound identification and identification confidence in discovery metabolomics analyses. Because of the novelty of this

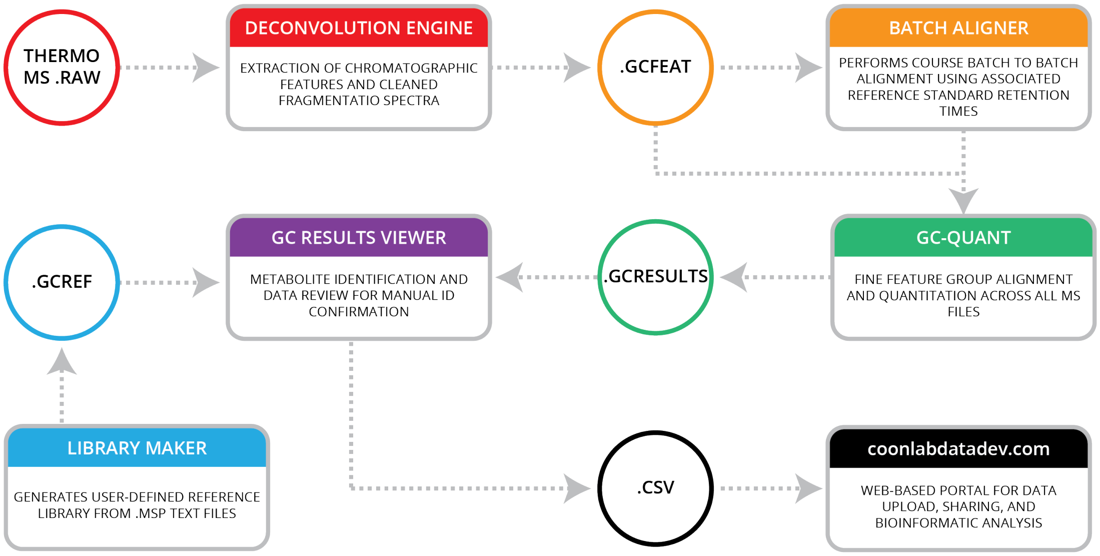


Figure 1.4: GC-MS metabolomics data analysis pipeline. A detailed workflow of the high resolution GC-MS metabolomics data analysis pipeline is shown. Batches of raw data files are deconvolved using the Deconvolution Engine program, and then optionally realigned using the Batch Aligner node. Next, feature groups are matched between separate feature files and quantitative values are extracted from each, using the GC-Quant software. Then results files for the experiment can be visualized with the GC Results Viewer application and metabolites can be searched against library databases and manually annotated. Last, resultant metabolite peak tables can be exported and uploaded to downstream data analysis software.

system, and the desire to design discovery metabolomics experiments of very large scope, an in-house data analysis pipeline was developed, outlined in **Figure 1.4**. In this pipeline raw data files are first deconvolved, which is the process of grouping ions based on the shape and retention time apex of the associated elution profile, using the Deconvolution Engine program. Next, if necessary, files from separate batches of data which are to be compared against one another are realigned to correct for retention time deviations due to instrument drift of the GC column, using the Batch Aligner node. Next, the aligned raw files are then compared to one another and features which show high spectral similarity and retention time are grouped together in the GC-Quant program. Then the files are imported into the GC Results Viewer program where the spectra are searched against user-generated spectral libraries. The Library Maker program can be used to tailor the spectral library to the user's specifications, such as eliminating non-derivatized molecules or restricting the search to only high or low resolution spectra. The GC Results Viewer program allows the user to manually assign compound identifications by visualizing the library and collected spectra, and by calculating the spectral similarity score and high-resolution filtering (HRF) score, which is described in detail below. Once the metabolite identifications have been assigned, the data can be exported as a .csv file and into uploaded to excel, to the coonlabdatadev.com website, or imported into any other downstream pipeline for further analysis.

One way in which high resolution GC-MS spectra can be leveraged to improve compound identification is by using high-resolution filtering. In this method accurate mass GC-MS data are deconvolved into component feature groups as shown in **Figure 1.5** panels

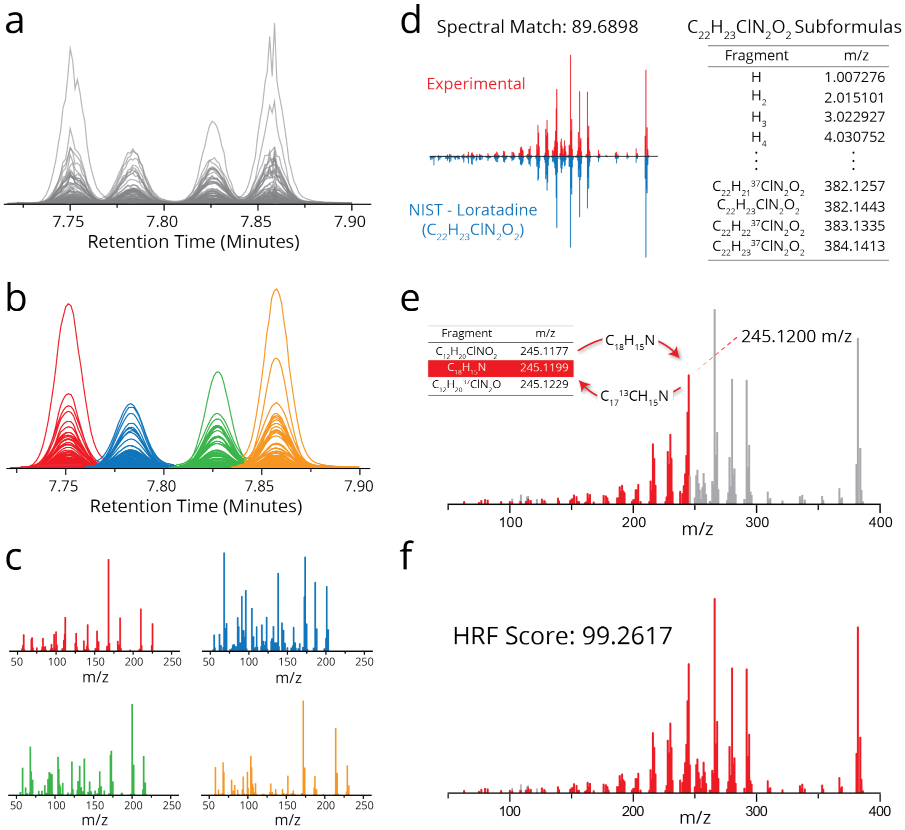


Figure 1.5: High-resolution filtering workflow. Overview of HRF workflow. (a) *m/z* peaks observed across consecutive scans are condensed into data features, which are smoothed and grouped based on elution apex (b). All features within a group are assumed to arise from a singular precursor. (c) Individual spectra are derived from feature groups, using average *m/z* and apex intensity, and are then subjected to spectral matching. (d) A high-scoring spectral match of an GC-Orbitrap mass spectrum of loratadine against the corresponding NIST reference spectrum. All sub-formulas from $C_{22}H_{23}ClN_2O_2$, the molecular formula of loratadine, are generated and sorted by exact formula mass, less the mass of an electron. (e) Sub-formulas are matched to *m/z* peaks in ascending order. For each matched fragment a variant containing appropriate heavy isotopes is created and placed into the list of sub-formulas in sorted-order. (f) For the experimental mass spectrum of loratadine 99.2617% of the measured signal can be annotated with sub-formulae of $C_{22}H_{23}ClN_2O_2$.

a, b, and c. **Figure 1.5** panels d-f illustrates the HRF strategy using an EI mass spectrum of loratadine collected from an authentic standard using the Q Exactive GC system. A unit resolution database search returns a reference spectrum of loratadine as a strong candidate match. To evaluate the quality of this putative identification, the chemical formula of loratadine ($C_{22}H_{23}ClN_2O_2$) is used to calculate all non-repeating combinations of atoms (*i.e.*, sub-formulas) which are then ordered by ascending exact mass, less an electron (**Figure 1.5** panel d). Starting with the smallest measured *m/z* peak, sub-formulas are matched based on exact mass. To accommodate isotopic clusters present in spectra, a variant containing an appropriate number of heavy isotopes is created for each matched fragment and placed back onto the list of sub-formulas. For example, once the highlighted *m/z* peak at 245.1200 is matched to $C_{18}H_{15}N$ a formula containing a substituted ^{13}C isotope ($C_{17}{}^{13}CH_{15}N$) is added to the list of candidate sub-formulas (**Figure 1.5** panel e). This strategy enables annotation of non-monoisotopic fragments without unduly increasing sub-formula search space. Once every *m/z* peak in the spectrum has been considered, the total percentage of measured ion current that has been annotated is returned in the form of a HRF score. In the example case of loratadine we find that 99.2617% of all measured ion current can be annotated using a sub-formula of its true parent precursor (**Figure 1.5** panel f). By enabling discrimination between candidate molecular precursors on the basis of both measured fragmentation profiles and accurate mass, this method effectively bridges the gap between high resolution spectral acquisition and unit resolution mass spectral libraries. **Chapters 5** and **6** of this dissertation outline advancements and applications of the described discovery metabolomics pipeline using the Q Exactive GC system. **Chapter 5** covers the development of a library of metabolite GC-EI-MS spectra which drastically improves compound identification confidence. **Chapter 6** describes the multi-omic analysis of a suite of single gene deletion *Saccharomyces cerevisiae* knockout strains in order to elucidate the function of unknown genes.

References

[1] F. Sanger, A. R. Coulson, B. G. Barrell, A. J. Smith, and B. A. Roe, "Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing," *Journal of Molecular Biology*, vol. 143, no. 2, pp. 161–178, 1980.

[2] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen,

H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, and C. International Human Genome Sequencing, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[3] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. C. E. Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G. D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver,

S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith, "Accurate whole human genome sequencing using reversible terminator chemistry," *Nature*, vol. 456, no. 7218, pp. 53–59, 2008.

[4] A. J. Dempster, "A New Method of Positive Ray Analysis," *Physical Review*, vol. 11, pp. 316–325, apr 1918.

[5] W. Bleakney, "A New Method of Positive Ray Analysis and Its Application to the Measurement of Ionization Potentials in Mercury Vapor," *Physical Review*, vol. 34, pp. 157–160, jul 1929.

[6] M. Karas, D. Bachmann, and F. Hillenkamp, "Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules," *Analytical Chemistry*, vol. 57, pp. 2935–2939, dec 1985.

[7] K. Tanaka, H. Waki, Y. Ido, S. Akita, Y. Yoshida, T. Yoshida, and T. Matsuo, "Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 2, pp. 151–153, aug 1988.

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

[9] R. A. Yost and C. G. Enke, "Selected ion fragmentation with a tandem quadrupole mass spectrometer," *Journal of the American Chemical Society*, vol. 100, pp. 2274–2275, mar 1978.

[10] J. C. Schwartz, M. W. Senko, and J. E. P. Syka, "A two-dimensional quadrupole ion trap mass spectrometer," *Journal of the American Society for Mass Spectrometry*, vol. 13, pp. 659–669, jun 2002.

[11] A. Makarov, "Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis," *Analytical Chemistry*, vol. 72, pp. 1156–1162, mar 2000.

[12] Q. Hu, R. J. Noll, H. Li, A. Makarov, M. Hardman, and R. Graham Cooks, "The Orbitrap: a new mass spectrometer," *Journal of Mass Spectrometry*, vol. 40, pp. 430–443, apr 2005.

[13] A. Michalski, E. Damoc, J.-P. Hauschild, O. Lange, A. Wieghaus, A. Makarov, N. Nagaraj, J. Cox, M. Mann, and S. Horning, "Mass spectrometry-based proteomics using

Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer.," *Molecular & cellular proteomics : MCP*, vol. 10, p. M111.011015, sep 2011.

[14] A. Michalski, E. Damoc, O. Lange, E. Denisov, D. Nolting, M. Müller, R. Viner, J. Schwartz, P. Remes, M. Belford, J.-J. Dunyach, J. Cox, S. Horning, M. Mann, and A. Makarov, "Ultra High Resolution Linear Ion Trap Orbitrap Mass Spectrometer (Orbitrap Elite) Facilitates Top Down LC MS/MS and Versatile Peptide Fragmentation Modes," *Molecular & Cellular Proteomics*, vol. 11, p. O111.013698, mar 2012.

[15] R. S. Gohlke, "Time-of-Flight Mass Spectrometry and Gas-Liquid Partition Chromatography," *Analytical Chemistry*, vol. 31, pp. 535–541, apr 1959.

[16] M. P. Washburn, D. Wolters, and J. R. Yates, "Large-scale analysis of the yeast proteome by multidimensional protein identification technology.," *Nature biotechnology*, vol. 19, pp. 242–7, mar 2001.

[17] 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.," *Molecular & cellular proteomics : MCP*, vol. 13, pp. 339–47, jan 2014.

[18] J. K. Eng, A. L. McCormack, and J. R. Yates, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *Journal of the American Society for Mass Spectrometry*, vol. 5, pp. 976–989, nov 1994.

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

[20] R. A. Zubarev, D. M. Horn, E. K. Fridriksson, N. L. Kelleher, N. A. Kruger, M. A. Lewis, B. K. Carpenter, and F. W. McLafferty, "Electron Capture Dissociation for Structural Characterization of Multiply Charged Protein Cations," *Analytical Chemistry*, vol. 72, no. 3, pp. 563–573, 2000.

[21] D. P. Little, J. P. Speir, M. W. Senko, P. B. O'Connor, and F. W. McLafferty, "Infrared Multiphoton Dissociation of Large Multiply Charged Ions for Biomolecule Sequencing," *Analytical Chemistry*, vol. 66, no. 18, pp. 2809–2815, 1994.

[22] J. A. Madsen, D. R. Boutz, and J. S. Brodbelt, "Ultrafast ultraviolet photodissociation at 193 nm and its applicability to proteomic workflows," *Journal of Proteome Research*, vol. 9, no. 8, pp. 4205–4214, 2010.

[23] S. L. Cook and G. P. Jackson, "Metastable atom-activated dissociation mass spectrometry of phosphorylated and sulfonated peptides in negative ion mode," *Journal of the American Society for Mass Spectrometry*, vol. 22, no. 6, pp. 1088–1099, 2011.

[24] A. Kalli, G. Grigorean, and K. Håkansson, "Electron Induced Dissociation of Singly Deprotonated Peptides," *Journal of The American Society for Mass Spectrometry*, vol. 22, pp. 2209–2221, dec 2011.

[25] J. E. P. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt, "Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry," *Proceedings of the National Academy of Sciences*, vol. 101, pp. 9528–9533, jun 2004.

[26] J. J. Coon, J. Shabanowitz, D. F. Hunt, and J. E. P. Syka, "Electron transfer dissociation of peptide anions," *Journal of the American Society for Mass Spectrometry*, vol. 16, pp. 880–882, jun 2005.

[27] A. Chi, C. Huttenhower, L. Y. Geer, J. J. Coon, J. E. P. Syka, D. L. Bai, J. Shabanowitz, D. J. Burke, O. G. Troyanskaya, and D. F. Hunt, "Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry," *Proceedings of the National Academy of Sciences*, vol. 104, pp. 2193–2198, feb 2007.

[28] N. M. Riley, A. S. Hebert, G. Dürnberger, F. Stanek, K. Mechtler, M. S. Westphall, and J. J. Coon, "Phosphoproteomics with Activated Ion Electron Transfer Dissociation," *Analytical Chemistry*, vol. 89, no. 12, pp. 6367–6376, 2017.

[29] V. Larraillet, R. Antoine, P. Dugourd, and J. Lemoine, "Activated-electron photodetachment dissociation for the structural characterization of protein polyanions," *Analytical Chemistry*, vol. 81, no. 20, pp. 8410–8416, 2009.

[30] S. A. McLuckey and M. Mentinova, "Ion/neutral, ion/electron, ion/photon, and ion/ion interactions in tandem mass spectrometry: Do we need them all? Are they enough?," *Journal of the American Society for Mass Spectrometry*, vol. 22, no. 1, pp. 3–12, 2011.

[31] W. J. Herron, D. E. Goeringer, and S. A. McLuckey, "Gas-Phase Electron Transfer Reactions from Multiply-Charged Anions to Rare Gas Cations," *Journal of the American Chemical Society*, vol. 117, pp. 11555–11562, nov 1995.

[32] Y. Gao, J. Yang, and M. Cancilla, "Top-Down Interrogation of Chemically Modified Oligonucleotides by Negative Electron Transfer and Collision Induced Dissociation," *Analytical chemistry*, 2013.

[33] T. Y. Huang and S. A. McLuckey, "Gas-phase ion/ion reactions of rubrene cations and multiply charged DNA and RNA anions," *International Journal of Mass Spectrometry*, vol. 304, no. 2-3, pp. 140–147, 2011.

[34] D. M. Horn, Y. Ge, and F. W. McLafferty, "Activated ion electron capture dissociation for mass spectral sequencing of larger (42 kDa) proteins," *Analytical Chemistry*, vol. 72, no. 20, pp. 4778–4784, 2000.

[35] Y. Xia, H. P. Gunawardena, D. E. Erickson, and S. A. McLuckey, "Effects of cation charge-site identity and position on electron-transfer dissociation of polypeptide cations," *Journal of the American Chemical Society*, vol. 129, pp. 12232–12243, oct 2007.

[36] Y. Zhang, S. B. Ficarro, S. Li, and J. A. Marto, "Optimized orbitrap HCD for quantitative analysis of phosphopeptides," *Journal of the American Society for Mass Spectrometry*, vol. 20, pp. 1425–1434, aug 2009.

[37] G. C. McAlister, D. H. Phanstiel, J. Brumbaugh, M. S. Westphall, and J. J. Coon, "Higher-energy Collision-activated Dissociation Without a Dedicated Collision Cell," *Molecular & Cellular Proteomics*, vol. 10, p. O111.009456, may 2011.

[38] S. A. McLuckey, J. L. Stephenson, and K. G. Asano, "Ion/Ion Proton-Transfer Kinetics: Implications for Analysis of Ions Derived from Electrospray of Protein Mixtures," *Analytical Chemistry*, vol. 70, no. 6, pp. 1198–1202, 1998.

[39] A. V. Tolmachev, H. R. Udseth, and R. D. Smith, "Modeling the ion density distribution in collisional cooling RF multipole ion guides," *International Journal of Mass Spectrometry*, vol. 222, pp. 155–174, jan 2003.

[40] J. L. Stephenson and S. A. McLuckey, "Ion/Ion Reactions in the Gas Phase: Proton Transfer Reactions Involving Multiply-Charged Proteins," *Journal of the American Chemical Society*, vol. 118, pp. 7390–7397, jan 1996.

[41] S. A. McLuckey and J. L. Stephenson, "Ion/ion chemistry of high-mass multiply charged ions," *Mass Spectrometry Reviews*, vol. 17, no. 6, pp. 369–407, 1998.

[42] H. P. Gunawardena, M. He, P. A. Chrisman, S. J. Pitteri, J. M. Hogan, B. D. M. Hodges, and S. A. McLuckey, "Electron Transfer versus Proton Transfer in Gas-Phase Ion/Ion Reactions of Polyprotonated Peptides," *Journal of the American Chemical Society*, vol. 127, pp. 12627–12639, sep 2005.

[43] A. R. Ledvina, G. C. McAlister, M. W. Gardner, S. I. Smith, J. A. Madsen, J. C. Schwartz, G. C. Stafford, J. E. P. Syka, J. S. Brodbelt, and J. J. Coon, "Infrared Photoactivation Reduces Peptide Folding and Hydrogen-Atom Migration following ETD Tandem Mass Spectrometry," *Angewandte Chemie International Edition*, vol. 48, pp. 8526–8528, oct 2009.

[44] N. M. Riley and J. J. Coon, "The Role of Electron Transfer Dissociation in Modern Proteomics," *Analytical Chemistry*, vol. 90, no. 1, pp. 40–64, 2018.

[45] G. J. Patti, O. Yanes, and G. Siuzdak, "Metabolomics: The apogee of the omics trilogy," *Molecular Cell Biology*, vol. 13, pp. 263–269, 2012.

[46] D. S. Wishart, C. Knox, A. C. Guo, R. Eisner, N. Young, B. Gautam, D. D. Hau, N. Psychogios, E. Dong, S. Bouatra, R. Mandal, I. Sinelnikov, J. Xia, L. Jia, J. A. Cruz, E. Lim, C. A. Sobsey, S. Shrivastava, P. Huang, P. Liu, L. Fang, J. Peng, R. Fradette, D. Cheng, D. Tzur, M. Clements, A. Lewis, A. De souza, A. Zuniga, M. Dawe, Y. Xiong, D. Clive, R. Greiner, A. Nazyrova, R. Shaykhutdinov, L. Li, H. J. Vogel, and I. Forsythei, "HMDB: A knowledgebase for the human metabolome," *Nucleic Acids Research*, vol. 37, no. SUPPL. 1, pp. 603–610, 2009.

[47] C. A. Smith, G. O. Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, and G. Siuzdak, "METLIN," *Therapeutic Drug Monitoring*, vol. 27, pp. 747–751, dec 2005.

[48] P. D. Hutchins, J. D. Russell, and J. J. Coon, "LipiDex: An Integrated Software Package for High-Confidence Lipid Identification," *Cell Systems*, pp. 1–5, 2018.

[49] T. Kind, K.-H. Liu, D. Y. Lee, B. DeFelice, J. K. Meissen, and O. Fiehn, "LipidBlast in silico tandem mass spectrometry database for lipid identification," *Nature Methods*, vol. 10, no. 8, pp. 755–758, 2013.

[50] James A.T. and Martin A.J.P., "GAS-LIQUID CHROMATOGRAPHY: A Technique for the Analysis and Identification of Volatile Materials," *Br. Med. Bull. 10(3): 170-176.*, vol. 10, no. 3, pp. 170–176, 1954.

[51] E. C. Horning and M. G. Horning, "Metabolic profiles: gas-phase methods for analysis of metabolites.," *Clinical chemistry*, vol. 17, pp. 802–9, aug 1971.

[52] M. M. Koek, R. H. Jellema, J. van der Greef, A. C. Tas, and T. Hankemeier, "Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives," *Metabolomics*, vol. 7, pp. 307–328, sep 2011.

[53] S. E. Stein and D. R. Scott, "Optimization and Testing of Mass-Spectral Library Search Algorithms for Compound Identification," *Journal of the American Society for Mass Spectrometry*, vol. 5, no. 9, pp. 859–866, 1994.

[54] S. E. Stein, "An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data," *Journal of the American Society for Mass Spectrometry*, vol. 10, no. 8, pp. 770–781, 1999.

[55] A. C. Peterson, G. C. McAlister, S. T. Quarmby, J. Griep-Raming, and J. J. Coon, "Development and characterization of a GC-enabled QLT-Orbitrap for high-resolution and high-mass accuracy GC/MS.," *Analytical chemistry*, vol. 82, pp. 8618–28, oct 2010.

[56] A. C. Peterson, J. P. Hauschild, S. T. Quarmby, D. Krumwiede, O. Lange, R. A. S. Lemke, F. Grosse-Coosmann, S. Horning, T. J. Donohue, M. S. Westphall, J. J. Coon, and J. Griep-Raming, "Development of a GC/quadrupole-orbitrap mass spectrometer, Part I: Design and characterization," *Analytical Chemistry*, vol. 86, no. 20, pp. 10036–10043, 2014.

[57] A. C. Peterson, A. J. Balloon, M. S. Westphall, and J. J. Coon, "Development of a GC/quadrupole-orbitrap mass spectrometer, part II: New approaches for discovery metabolomics," *Analytical Chemistry*, vol. 86, no. 20, pp. 10044–10051, 2014.

Chapter 2

SULFUR PENTAFLUORIDE IS A PREFERRED REAGENT CATION FOR NEGATIVE ELECTRON TRANSFER DISSOCIATION

This chapter has been published and is reprinted with permission from:

Rush MJP, Riley, NM, Westphall MS, Coon JJ. *Sulfur Pentafluoride is a Preferred Reagent Cation for Negative Electron Transfer Dissociation.*
Journal of the American Society for Mass Spectrometry. **2017**, *28 (7)*, 1324-1332. doi: 10.1007/s13361-017-1600-8.

Abstract

Negative mode proteome analysis offers access to unique portions of the proteome and several acidic post-translational modifications; however, traditional collision-based fragmentation methods fail to reliably provide sequence information for peptide anions. Negative electron transfer dissociation (NETD), on the other hand, can sequence precursor anions in a high-throughput manner. Similar to other ion–ion methods, NETD is most efficient with peptides of higher charge state because of the increased electrostatic interaction between reacting molecules. Here we demonstrate that NETD performance for lower charge state precursors can be improved by altering the reagent cation. Specifically, the recombination energy of the NETD reaction—largely dictated by the ionization energy (IE) of the reagent cation—can affect the extent of fragmentation. We compare the NETD reagent cations of $C_{16}H_{10}{}^{\bullet+}$ (IE = 7.9 eV) and $SF_5{}^{\bullet+}$ (IE = 9.6 eV) on a set of standard peptides, concluding that $SF_5{}^{\bullet+}$ yields greater sequence ion generation. Subsequent proteome-scale nLC-MS/MS experiments comparing $C_{16}H_{10}{}^{\bullet+}$ and $SF_5{}^{\bullet+}$ further supported this outcome: analyses using $SF_5{}^{\bullet+}$ yielded 4637 peptide spectral matches (PSMs) and 2900 unique peptides, whereas $C_{16}H_{10}{}^{\bullet+}$ produced 3563 PSMs and 2231 peptides. The substantive gain in identification power with $SF_5{}^{\bullet+}$ was largely driven by improved identification of doubly deprotonated precursors, indicating that increased NETD recombination energy can increase product ion yield for low charge density precursors. This work demonstrates that $SF_5{}^{\bullet+}$ is a viable, if not favorable, reagent cation for NETD, and provides improved fragmentation over the

commonly used fluoranthene reagent.

Introduction

Modern proteome characterization relies on liquid chromatography coupled with tandem mass spectrometry (nLC-MS/MS) to detect and quantify proteins from complex biological samples.[1–5] Despite continual advances in proteomic depth achieved in such experiments, these technologies do not detect all the proteins present in a sample and, in fact, typically only monitor a portion of those proteins that are detected. More specifically, one or two peptides resulting from the enzymatic digestion of a protein can map uniquely to the parent protein and allow for its unambiguous detection and quantification. New approaches that can offer increased diversity of peptides measured are therefore of considerable significance as they can reveal new proteins and offer access to portions of proteins that were previously not detectable. One factor that may limit the scope of the present technology is the unilateral use of positive electrospray ionization. Many proteins, and portions of most proteins, are acidic and thus are more easily ionized in the negative mode.[6–8] Negative electrospray ionization can generate multiply deprotonated peptide anions, but the commonly used collision-based dissociation methods are ineffective at producing sequence-informative fragmentation of negatively charged peptides.[9,10] These limitations have driven the development of alternative dissociation methods for peptide anions that utilize electrons, photons, and metastable atoms.[11–30]

Scott McLuckey, the recipient of the Award for Distinguished Contributions in Mass Spectrometry, whom we honor in this issue of JASMS, has been a pioneer in this field, especially in the development of gas-phase ion–ion chemistry[31,32] Indeed, this investigation stems from foundational work his group and others described. Most notably, in 1995 McLuckey and co-workers first introduced ion–ion reactions of positive reagent cations with anionic oligonucleotides, and have continued this innovating in this space.[33–35] Inspired by McLuckey's experiment, we developed ion–ion reaction chemistries to abstract electrons from multiply deprotonated peptide anions using singly charged reagent cations, termed negative electron transfer dissociation (NETD).[11] NETD has emerged as one of the most promising peptide anion dissociation methods and has been successfully utilized in nLC-MS/MS experiments to access the acidic proteome.[11,12,16,17,30]

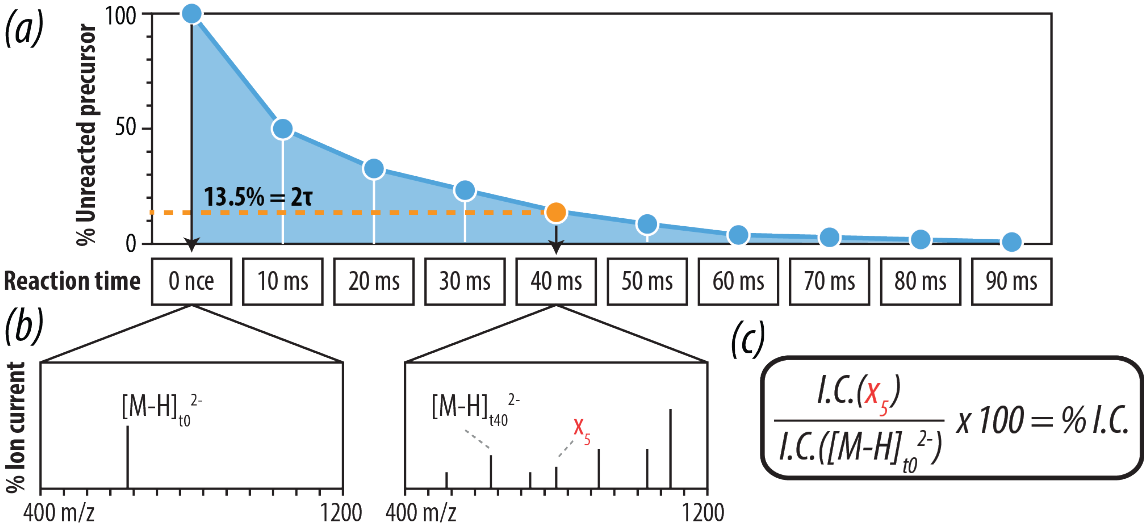
One challenge of using NETD in large scale proteomic experiments is its limited product ion yield for low charge density precursors. In NETD experiments, peptide anions are oxidized by positively charged reagent cations to initiate dissociation and production of $a^\bullet$ - and *x-type* product ions. Sometimes electron transfer from the anion to the cation occurs without concomitant dissociation (*i.e.*, non-dissociative negative electron transfer, or NETnoD).[36–39] NETnoD frequently occurs and is one of the primary causes for reduced product ion yields in NETD, especially in the case of low charge density precursors.[39] In these cases, once electron transfer has occurred, peptide backbone cleavage may be achieved; however, the resultant product ions can be held together by noncovalent bonding and detected as a charge-reduced product. To maximize the production of NETD product ions, a reduction of NETnoD species can be accomplished by supplying the charge-reduced product with more energy, either concurrent with, or post- electron transfer. The additional energy can disrupt the noncovalent interactions holding these ions together, yielding sequence informative products.[40,41] One such approach, termed activated-ion NETD (AI-NETD), has been implemented by concurrently irradiating the ions with infrared photons as they are interacting; however, this approach requires the addition of an IR laser to the system.[16,30] An alternate approach to increase the energy of the system during electron transfer events is to alter the reaction exothermicity, which is determined by the difference between the ionization energy of the reagent cation and the electron affinity of the peptide anion. For example, the ionization energy of fluoranthene, $C_{16}H_{10}^{\bullet}$, is 7.9 eV and the electron affinity of the carboxylate ion of a peptide is 3.4 eV, yielding a reaction enthalpy of 4.5 eV, known as the recombination energy. This energy surplus is redistributed into the peptide anion and drives fragmentation. Use of NETD reagent cations with higher ionization energies result in increased recombination energies and, potentially, an increase in NETD fragmentation efficiency.[42] The recombination energy of the reagent cation and its effect on NETD has been previously explored by Polfer and co-workers for use in determining phosphorylation sites on standard peptides[43], as well as by McLuckey et al. to investigate transition metal complexes and their interaction with peptide.[44] Polfer and colleagues compared two NETD reagents (fluoranthene and xenon) and determined that the increase in ionization energy of xenon led to considerable phosphate and side-chain neutral loss and, therefore, fluoranthene, having a lower ionization energy than xenon, should be used for sequencing phosphopeptides. Alternatively, McLuckey and co-workers showed that transition metal complexes yield electron transfer as well as metal insertion reactions, allowing further control of the cation–anion interaction. To date, only xenon, fluoranthene, and phenanthroline complexes of Fe, Cu, and Co have been investigated as NETD reagent cations for peptide analysis, and fluoranthene remains by far the most commonly used NETD reagent.

The primary motivation of this work is to investigate the hypothesis that increasing the ionization energy of the NETD reagent cation will increase the NETnoD conversion to product ions, thereby yielding greater peptide identification rates and protein sequence coverage in large-scale shotgun proteome analyses. To test this hypothesis, we used a set of synthetic peptides to compare the NETD fragmentation efficiency of sulfur pentafluoride cations ($SF_5^{\bullet+}$, IE = 9.6 eV) and fluoranthene cations ($C_{16}H_{10}^{\bullet+}$, IE = 7.9 eV) over a range of available precursor charge states (z = –2 to –6). Concluding that $SF_5^{\bullet+}$ cations provided increased sequence ion production for low charge state precursors, we then performed nanoflow liquid chromatography-tandem mass spectrometry (nLC-MS/MS) experiments employing high pH separations and optimized NETD reaction kinetics to compare $C_{16}H_{10}^{\bullet+}$ with $SF_5^{\bullet+}$ NETD reagent ions for analysis of a complex mixture of yeast peptides. From these data, we revealed that up to 40% more peptide spectral matches (PSMs) could be made when using reagent ions from sulfur pentafluoride compared with fluoranthene. The overall peptide spectral match and unique peptide identification numbers improved 30% when using $SF_5^{\bullet+}$ as the NETD reagent instead of fluoranthene. From the data, we conclude that the use of $SF_5^{\bullet+}$ offers a direct route to boosting the performance of NETD dissociation.

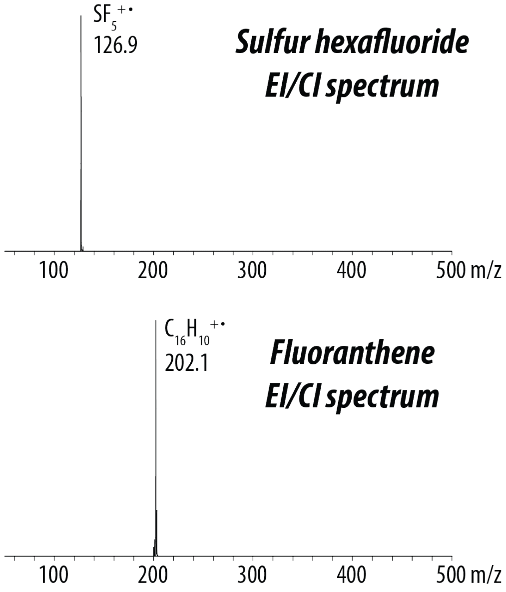
Materials and Methods

NETD on Standard Peptides. Synthetic peptides that have the sequences SVFAVNWISY-LASK, EEAQALEDLTGFK, and ELVNDDEDIDWVQTEK were obtained from New England Peptides (Gardner, MA, USA) and were individually suspended in 3:1 methanol/water with 5 mM piperidine to a concentration of 10 ppm. The peptides were infused into a LTQ Velos mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) modified to perform NETD. For each peptide precursor, a 0 nce CAD MS/MS scan was performed, followed by a series of NETD MS/MS scans with increasing reaction time, encompassing the 2τ time point (see **Supplementary Figure S2.1** for further experimental design diagram). This series was repeated at least 15 times. Once complete, the next precursor charge state was reacted until all accessible charge states were reacted. The precursor AGC target was set to 10,000 and the reagent AGC target was set to 1,000,000 for all MS/MS acquisitions. The q-value for the NETD reaction was kept at 0.4 for all experiments. This procedure was repeated using both $SF_5^{\bullet+}$ and $C_{16}H_{10}^{\bullet+}$ reagent cations. The solid phase NETD reagent fluoranthene was introduced to the system using the standard glass vials contained in the ETD module's reagent vial heater. The gaseous sulfur hexafluoride reagent was introduced to the system by connecting the high purity SF_6 gas cylinder (Concorde Specialty Gases, Eatontown, NJ, USA) to a precision regulator (Porter Instruments, Hatfield, PA, USA) with 6 feet of 1/8 inch outer diameter and 0.065 inch inner diameter copper tubing. Then a 100 μm inner diameter capillary tube was attached to the out port of the precision regulator and fed directly into the vacuum manifold of the ETD module and the pressure reading of the precision regulator was adjusted to generate appropriate reagent cation signal (~2 psi). Reagent cations were generated via the EI/CI source in the presence of nitrogen gas, and reagent signal was optimized by varying lens voltages using an automated calibration routine. The filament emission current was set to 70 μA for both reagents. The nitrogen gas pressure was also optimized for maximum signal for each reagent cation. For fluoranthene, the reagent vial temperature was set to 108° C. For both reagents, the ion source, transfer line, and restrictor temperature were held at 160° C. The fluoranthene reagent produced ion radicals $C_{16}H_{10}{}^{\bullet+}$, whereas the predominant cation for sulfur hexafluoride was the fluorine-loss species $SF_5{}^{\bullet+}$, shown in **Supplementary Figure S2.2**.

Yeast Sample Preparation. Tryptic yeast (*Saccharomyces cerevisiae*) peptides were prepared as previously described.[16] Briefly, cultured yeast cells were lysed by glass bead milling (Retsch GmnH, Germany), and proteins were reduced and alkylated using 5 mM dithiothreitol and 15 mM iodoacetamide, respectively. Trypsin digestion was performed during an overnight incubation at room temperature with a 1:50 (w/w) enzyme to protein ratio. A second trypsin addition was done the following morning at 1:100 (w/w) enzyme to protein ratio for 1 h, followed by desalting over a C18 SepPak (Waters Corporation, Milford, MA, USA).



Supplementary Figure S2.1: Illustration of mass spectrometry method for NETD reagent cation evaluation. The mass spectrometry method for reaction of standard peptides with both NETD reagents is shown. Each precursor of interest is reacted with 0 nce CAD followed by a series of NETD reactions with increasing reaction times encompassing the 2τ time point, (a). This sequence is repeated until at least 15 scans are completed for each reaction time. The reaction time closest to the 2τ time point, identified by the dotted orange line, is then analyzed to extract all ion currents, (b), which are normalized by taking the average ion current (I.C.) of an individual fragment ion and dividing by the average ion current of the 0 nce CAD scan, shown in (c). This allows for an accurate percent conversion from initial starting ion population to fragment ion population to be determined, labelled the % I.C.



Supplementary Figure S2.2: EI/CI spectrum for reagent cations. The EI/CI ionization spectrum for both SF_6 and fluoranthene are shown. Sulfur hexafluoride produces the fluorine loss product $SF_5^{\bullet+}$ as the primary ion, while fluoranthene produces the molecular ion $C_{16}H_{10}^{\bullet+}$.

High pH nLC-MS/MS. An ETD-enabled hybrid dual cell-quadrupole ion trap-Orbitrap mass spectrometer (Orbitrap Elite, Thermo Fisher Scientific) coupled to a nanoACQUITY UltraPerformance liquid chromatograph (Waters) was used for the nLC-MS/MS analyses. The mass spectrometer was modified to perform NETD as described previously.[41,45–47] Briefly, the higher energy collisional dissociation cell (HCD) was replaced with a multi-purpose dissociation cell (MDC) that can conduct ion–ion reactions, allowing for NETD to be performed within. Fluoranthene and sulfur hexafluoride reagents were introduced as is described above, with the exception of the SF_6 pressure being adjusted to 10 psi. The solvent compositions for liquid chromatography were mobile phase A (5 mM piperidine in water) and mobile phase B (5 mM piperidine in 85% ACN and 15% water). The reverse phase columns were prepared in-house using 75 μm i.d., 360 o.d. bare fused silica capillary tubing packed to a 30 cm length with 3.5 μm, 130 Å pore size, Ethylene Bridged Hybrid C18 particles (Waters). For each analysis, 1 μg of yeast digest was loaded onto the column equilibrated with 95% A at 400 nL/min. The gradient elution was performed at 400 nL/min increasing from 5% mobile phase B to 30% B over 70 min, followed by an increase to 70% B at 76 min and a wash at 70% B for 4 more min. Peptides were ionized in the negative mode using electrospray ionization with a spray voltage of –1.5 kV. The inlet capillary temperature was set to 300° C. Survey MS scans were analyzed in the Orbitrap mass analyzer with a resolving power of 60,000 at 400 *m/z* and an AGC precursor ion target value of 1,000,000 over a mass range of 300–1250 *m/z*. Data-dependent MS/MS events were triggered off of the 10 most intense peaks in the survey scan. Each MS/MS scan used a precursor AGC target of 100,000 ions and was analyzed in the Orbitrap with a resolving power of 15,000 at 400 *m/z*. Precursors were isolated at ±0.9 Th and an exclusion window of ±10 ppm was created around the monoisotopic peak of the precursor for 45 s. All nLC-MS/MS experiments were performed in duplicate.

Data Analysis. Peptide standard infusion data were searched using an in-house C# script, which extracted ion current intensities directly from raw data files. These values were normalized relative to the ion current of the precursor from a 0 nce CAD scan collected previous to the NETD reacted spectra (see **Supplementary Figure S2.1** for further experimental design diagram). The nLC-MS/MS raw data files were searched using the open mass spectrometry search algorithm (OMSSA), modified to allow for anionic peptide $a^{\bullet}$ - and *x-type* fragment ions to be searched. Search parameters included carbamidomethylation of cysteine as a fixed modification and oxidation of methionine as a variable modification.[17,48] A multi-isotope search was employed using three isotopes with a mass tolerance of ±125 ppm for the precursors and a monoisotopic mass tolerance of ±0.02 Da for product ions. Three missed cleavages were allowed for the trypsin digestion. The data processing was done through the COMPASS software suite designed for OMSSA searching. A UniProt database for *Saccharomyces cerevisiae* (downloaded September 29, 2014) was concatenated with reversed sequences and used to determine peptide spectral matches (PSMs). Scored spectra were filtered using a false discovery rate of 1% at the unique peptide level. False discovery rates for spectra from each set of duplicate nLC-MS/MS experiments were calculated for the combined set of spectra as opposed to separate calculations for each nLC-MS/MS run. Additionally, prior to OMSSA searching, each spectrum was preprocessed to remove the unreacted precursor ion (±3 Da) and neutral loss ions of the oxidized precursor by removing ions within a window of 55 to 5 Da below the oxidized precursor ion. The nLC-MS/MS experiments were also searched using an in-house C# script to extract ion intensities for expected sequencing ions as well as ions resulting from neutral losses from both oxidized precursor ions and $a^\bullet$ - and *x-type* product ions. Ion intensities from spectra in nLC-MS/MS experiments were normalized relative to the total ion current (TIC) of each individual MS/MS scan.

Results and Discussion

NETD of Standard Peptides with Alternative Reagent Cations. The primary metric in determining an effective reagent cation is the production of sequence informative fragment ions relative to all other product ions produced. In these experiments, reagent ions $C_{16}H_{10}^{\bullet+}$ and $SF_5^{\bullet+}$ were tested for their effectiveness as NETD reagent cations using an ETD-enabled dual-pressure linear ion trap mass spectrometer. These species were reacted with three standard peptides (sequences given in top left of **Figure 2.1**), each with a C-terminal lysine to mimic those yielded from protein digestion with trypsin, and all were synthesized without additional post-translational modifications. The peptides had lengths of 13, 14, and 16 amino acid residues, isoelectric points of 3.62, 4.00, and 8.31, and generated peptide anions having charge states ranging from $z = -2$ to -6.

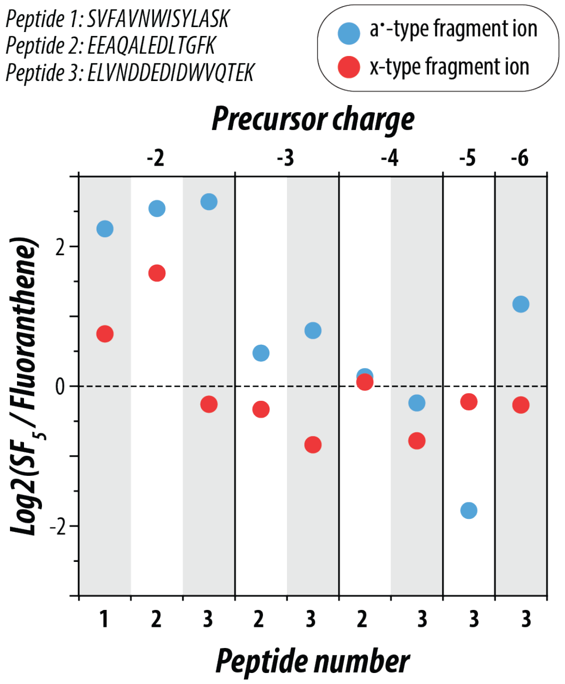
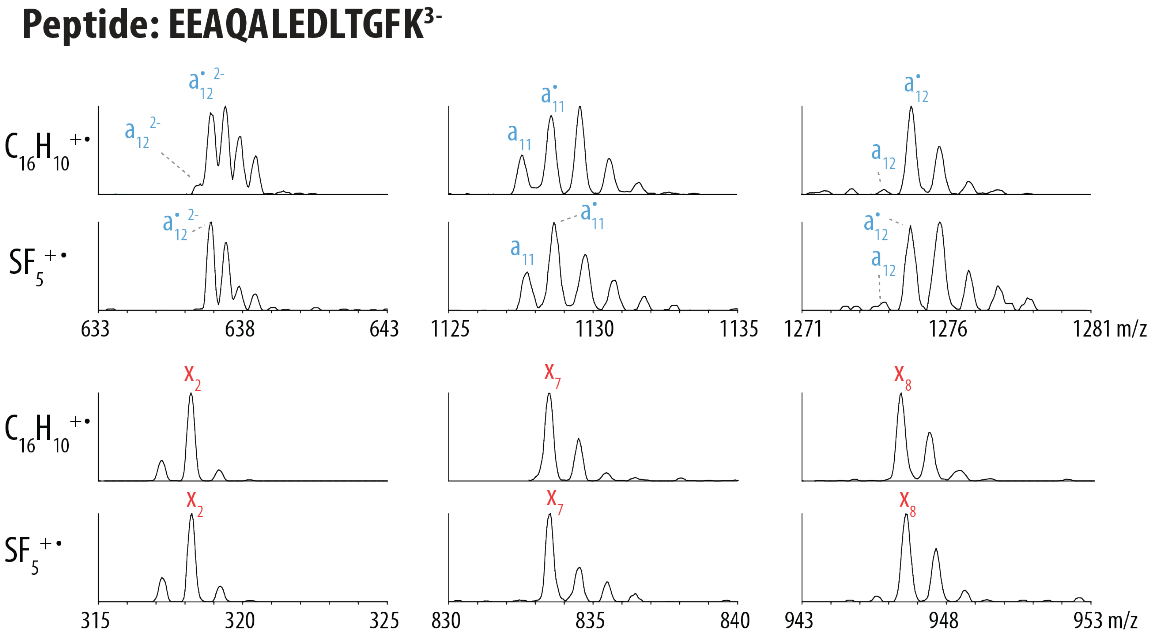


Figure 2.1: Product ion signal of standard peptides using $C_{16}H_{10}^{\bullet+}$ or $SF_5^{\bullet+}$. Comparison of product ion signal from NETD reactions using reagents $C_{16}H_{10}^{\bullet+}$ and $SF_5^{\bullet+}$ for all precursor charge states of three synthetic peptides (sequences shown top left). The intensities of both $a^{\bullet}$ - and *x-type* fragment ions were normalized to the total ion current of the unreacted precursor from a preceding scan as explained in the text. Following this normalization, the $\log_2$ fold change of the normalized fragment ion current intensity was calculated between $SF_5^{\bullet+}$ and $C_{16}H_{10}^{\bullet+}$. All precursor ion populations were reacted to the 2τ time point (*i.e.*, 13.5 ± 3% unreacted precursor ion current remaining in the MS/MS spectrum)

To provide a straightforward comparison of their product ion generation efficiencies, the extent of reaction for the two reagent cations was carefully controlled. To do so, we standardized several conditions that dictate the number of electron transfer events that occur per reaction, isolating the chemistry of the reagent cations as the main variable of the experiment. The rate of ion–ion reactions and, thus, the number of electron transfer events, are governed by a number of parameters, including ion population, reaction q-value, reaction cell architecture, and reaction time. Previous work has shown that to maximize peptide identifications, the ideal electron transfer extent occurs when the precursor ion population has been reduced by 86%.[49] Such reaction conditions minimize the amount of secondary electron transfer while still offering sufficient sequence informative fragment ion production. Using this as a model, we created a method, illustrated in **Supplementary Figure S2.1**, where a series of increasing reaction times were employed surrounding the optimal reaction extent (13.5% unreacted precursor remaining), keeping all other reaction parameters constant. **Figure 2.1** summarizes the production of sequence informative $a^\bullet$ - and *x-type* fragment ions produced when all accessible charge states of the three standard peptides were reacted with $SF_5^{\bullet+}$ or $C_{16}H_{10}^{\bullet+}$. Use of $SF_5^{\bullet+}$as the reagent cation more than quadrupled the $a^\bullet$*-type* ion signal relative to $C_{16}H_{10}^{\bullet+}$ for all doubly deprotonated precursors. An increase in *x-type* fragment ion production was observed for two out of three doubly deprotonated precursors.

As charge density increased, the difference in fragment ion production was reduced. In the case of triply deprotonated precursors, $SF_5^{\bullet+}$ generated spectra with more $a^\bullet$*-type* fragment ions, whereas use of $C_{16}H_{10}{}^{\bullet+}$ cations produced more *x-type* fragment ions, but the magnitude of the difference is considerably less than for the lower charge state precursors. Generally, the $C_{16}H_{10}{}^{\bullet+}$ reagent cation produced spectra with marginally more *x-type* fragment ions for higher charge state species, but the product ion signal is largely comparable between the two reagent cations for $z \geqslant 3$ precursors. This suggests that the higher IE of $SF_5{}^{\bullet+}$ can benefit fragmentation of low charge density precursors (*i.e.*, $z = -2$) where the predominance of NETnoD can adversely affect dissociation product ion generation. Note, however, $SF_5{}^{\bullet+}$ reagent cations retain the good performance of $C_{16}H_{10}{}^{\bullet+}$ for more highly charged ions. **Supplementary Figure S2.3** also considers the distribution of even and odd electron fragment ions (odd electron species containing a radical electron and an additional hydrogen atom), showing that *a•-type* and *x-type* fragment ions are the predominant species formed upon NETD, and that there is no significant difference in the even and odd electron ratios between the two reagent cations.

nLC-MS/MS of Yeast Tryptic Digest Using $SF_5{}^{\bullet}$ and $C_{16}H_{10}{}^{\bullet}$. To expand the scope of our study, we compared the performance of $C_{16}H_{10}{}^{\bullet+}$ and $SF_5{}^{\bullet+}$ as reagent cations in nLC-MS/MS analyses of peptides derived following tryptic digestion of yeast proteins. Importantly, the majority of precursors sampled in negative mode nLC-MS/MS experiments are doubly deprotonated[16], suggesting that the use of $SF_5{}^{\bullet+}$ reagent cations could improve the depth of analysis in whole-proteome shotgun sequencing. Indeed, in 90-min nLC-MS/MS experiments, NETD with $SF_5{}^{\bullet+}$ generated 30% more PSM and unique peptide



Supplementary Figure S2.3: Isotope distribution of $a^{\bullet}$ and x ions. A selection of a and x fragment ions are shown from the NETD reaction of -3 charge state of the standard peptide EEAQALEDLTGFK, reacted using the reagent cation shown to the left. The distribution of a to $a^{\bullet}$ fragment ions are very similar, with $a^{\bullet}$ being the predominant species. The first isotope of the x fragment ions and the $x^{\bullet}$ ion are indistinguishable at this resolution, however the isotope ratio is similar between reagents, indicating that recombination energy has little effect on the production of even and odd electron x fragment ion species.

identifications than did NETD using $C_{16}H_{10}^{\bullet +}$ (**Figure 2.2**). **Figure 2.2b** displays the charge state distribution of peptide spectral matches (PSMs) following use of the two NETD reagents. We note a substantial increase in the identification of $z = -2$ peptides for analyses with $SF_5^{\bullet +}$ (1039 more $z = -2$ PSMs than those with $C_{16}H_{10}^{\bullet +}$), which comprised more than 70% of all PSMs identified in either dataset. This finding is consistent with the previous results using standard peptides, further showing that peptides with lower charge density benefit more from excess recombination energy, while higher charge density peptides are not as impacted.

Figure 2.3 presents representative spectra from the nLC-MS/MS experiments illustrating the improved fragmentation afforded by use of $SF_5^{\bullet +}$ reagent cations as compared to $C_{16}H_{10}^{\bullet +}$. For each spectrum, the doubly deprotonated precursor of the peptide ETAESYLGAK was reacted with either NETD reagent to a reaction extent of 37.9% and 38.4% for $C_{16}H_{10}^{\bullet +}$ and $SF_5^{\bullet +}$, respectively. Reaction extent is defined as the ion current of the unreacted precursor divided by the total ion current of the scan. While in the standard peptide infusion data, a 0 nce CAD scan was used to measure the total ion abundance for each peptide and fragment ion currents could be normalized to it, allowing direct calculation of the precursor to product ion conversion ratio, in the discovery nLC-MS experiments no 0 nce CAD scans were performed and, therefore, this normalization approach was not possible, and reaction extent was calculated instead. Comparing spectra with similar reaction extents ensures that the main contributor to the difference is the ion chemistries of the reagents themselves. The spectra show many similarities but with considerably more

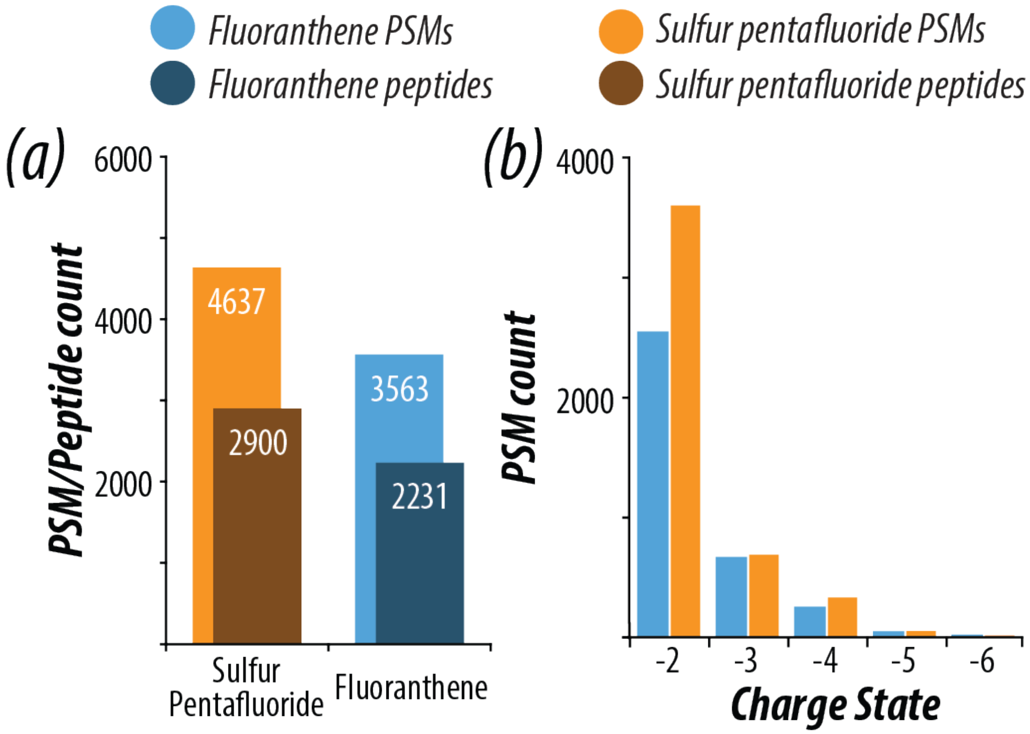


Figure 2.2: PSM and peptide counts and associated charge state distribution using different reagent cations. Summary of peptide spectral matches (PSM) and unique peptide identifications from nLC-MS/MS experiments using a yeast tryptic digest (a). Experiments were conducted using either $SF_5^{\bullet+}$ or $C_{16}H_{10}^{\bullet+}$ as reagent cations. The charge state distributions of successfully sequenced PSMs from these experiments are also shown (b), revealing a significant increase in $z = -2$ peptides identified with $SF_5^{\bullet+}$

product ions generated when $SF_5^{\bullet+}$ was used. Specifically, $SF_5^{\bullet+}$ produced two more *a*$^{\bullet}$*-type* and three more *x-type* fragment ions than the corresponding NETD spectrum produced when $C_{16}H_{10}^{\bullet+}$ was used as the reagent, and it also generated three CO_2 neutral losses from *x-type* fragment ions compared with only one when $C_{16}H_{10}^{\bullet+}$ cations were used. In all, use of $SF_5^{\bullet+}$ as the reagent cation produced a spectrum with 100% peptide sequence coverage for ETAESYLGAK, whereas the corresponding spectrum when $C_{16}H_{10}^{\bullet+}$ was used as the reagent yields only 55.5% coverage. Note, we define peptide sequence coverage as the ratio of the number of inter-residue positions broken to the total possible positions (residue length – 1) for a given sequence, expressed here as a percentage.

Figure 2.4a expands on the change in peptide sequence coverage between reagent cations $SF_5^{\bullet+}$and $C_{16}H_{10}^{\bullet+}$ by showing the composite difference for all peptides identified in the NETD experiments. Only peptides found in both data sets were considered, and of the 1427 peptides in common, 812 peptides yielded an increase in sequence coverage when $SF_5^{\bullet+}$ was used as the NETD reagent while only a fourth of that (n = 202) showed an increase when $C_{16}H_{10}^{\bullet+}$ was used. On average, $SF_5^{\bullet+}$ accounted for a nearly 10% improvement in peptide sequence coverage for all overlapping peptides.

Differences in peptide sequence coverage are even more pronounced when delineating across charge states. **Figure 2.4b** displays peptides categorized by charge state, where charge states of $z = -2, -3, -4$, and -5 showed an average sequence coverage difference of 9.0%, 4.7%, 2.8%, and –4.7%, respectively. Note, as indicated in the ΔSequence Coverage equation at the top of **Figure 2.4**, a positive value indicates higher sequence coverage

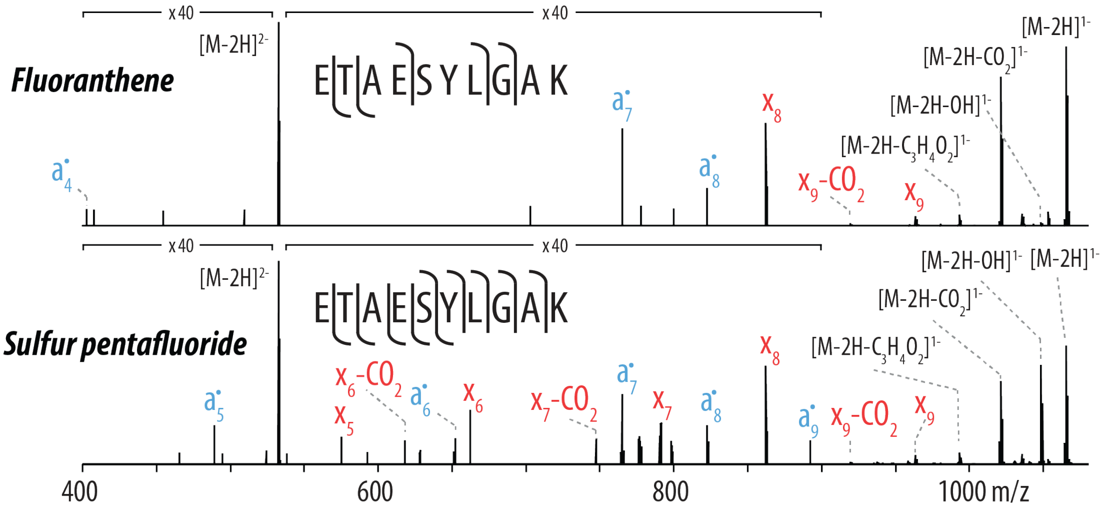


Figure 2.3: Example NETD spectra using $C_{16}H_{10}^{\bullet +}$ or $SF_5^{\bullet +}$ reagent cation. Single-scan spectra for the NETD fragmentation of the peptide ETAESYLGAK, $z = -2$. The unreacted precursor signal accounts for 37.9% and 38.4% of the total ion current in each MS/MS scan for $C_{16}H_{10}^{\bullet +}$ and $SF_5^{\bullet +}$, respectively, indicating both precursors were reacted to similar extents. Use of $SF_5^{\bullet +}$ as the NETD reagent provided greater sequence coverage than $C_{16}H_{10}^{\bullet +}$, yielding 11 sequence informative fragment ions compared with 5 for $C_{16}H_{10}^{\bullet +}$. The NETD reagent $SF_5^{\bullet +}$ produced three CO_2 product ion neutral loss species while the fluoranthene spectrum only contains a single CO_2 product ion neutral loss fragment. Both spectra were acquired in the nLC-MS/MS experiments and represent a single scan (*i.e.*, un-averaged)

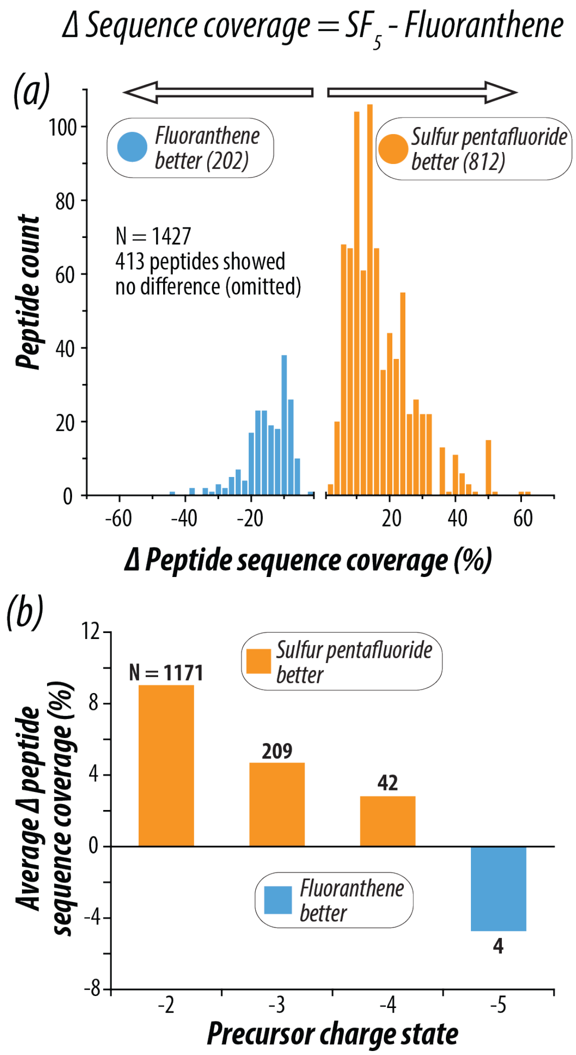
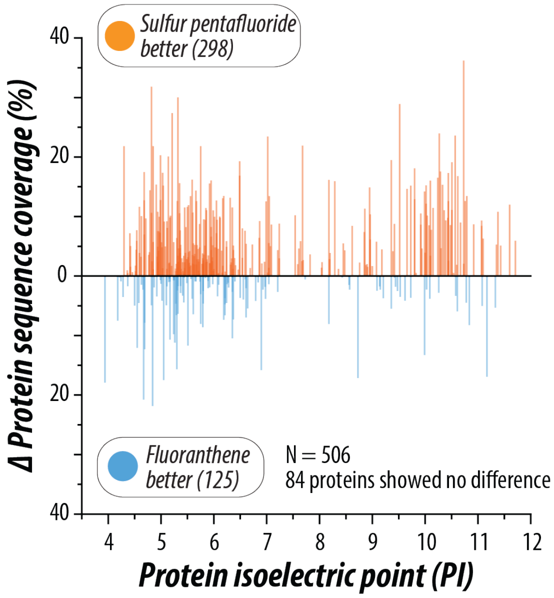


Figure 2.4: Comparison of peptide sequence coverage using $C_{16}H_{10}{}^{\bullet+}$ or $SF_5{}^{\bullet+}$ reagent cation. (a) Percent peptide sequence coverage was calculated for each of the 1427 peptides in common between the $SF_5{}^{\bullet+}$ and $C_{16}H_{10}{}^{\bullet+}$ analyses. The difference between peptide sequence coverage with $SF_5{}^{\bullet+}$ and $C_{16}H_{10}{}^{\bullet+}$ (Δ Peptide Sequence Coverage) was calculated for each peptide, and the distribution of the Δ Peptide Sequence Coveragevalues are shown. The orange distribution shows peptides with greater sequence coverage with $SF_5{}^{\bullet+}$ (n = 812), and the blue shows peptides with better sequence coverage with $C_{16}H_{10}{}^{\bullet+}$ (n = 202). Panel (b) shows the average Δ Peptide Sequence Coverage for all peptides in common between the two analyses as a function of peptide precursor charge state. The number of precursors averaged is shown in black text above the bar for each charge state. Concordant with previously shown data, lower charged precursors benefit most from the use of $SF_5{}^{\bullet+}$

with $SF_5^{\bullet+}$. Interestingly, $C_{16}H_{10}^{\bullet+}$ as the NETD reagent only performs better on $z = -5$ precursors, which account for 1.3% of the total precursors sampled in the experiments. In all other cases, peptide sequence coverage was improved by use of $SF_5^{\bullet+}$, especially for lower precursor charge states. **Supplementary Figure S2.4** considers the impact of reagent cation on total protein sequence coverage for all proteins in common between the two data sets, as the increased number of peptide identifications with $SF_5^{\bullet+}$ translates to high coverage of the proteins mapped in the experiments. Here, protein sequence coverage is defined as the number of amino acid residues comprising each identified peptide divided by all amino acid residues in the protein sequence. Use of $SF_5^{\bullet+}$ improved coverage of 298 proteins, $C_{16}H_{10}^{\bullet+}$ improved coverage of 125 proteins, and 84 showed no difference between the reagent cations, with an average improvement of 3.1% in favor of $SF_5^{\bullet+}$.

Comparison of Neutral Losses. Utilizing the higher ionization energy $SF_5^{\bullet+}$ reagent cation yields greater fragmentation in regards to sequence informative fragment ions. As the Polfer group noted with xenon, however, high ionization energies can drive the production of neutral losses from fragment ions. Using the same pool of doubly deprotonated peptides in common between the $C_{16}H_{10}^{\bullet+}$ and $SF_5^{\bullet+}$data sets discussed above, **Figure 2.5** examines how the oxidized (*i.e.*, charge-reduced) precursor ions and their associated neutral losses are affected by the use of the two reagent cations. **Figure 2.5a** displays the average percent ion current accounted for by the oxidized precursor as well as the signal from corresponding neutral losses of CO_2 and either NH_3 or OH from this charge-reduced species. Doubly



Supplementary Figure S2.4: Comparison of protein sequence coverage change using $C_{16}H_{10}^{\bullet +}$ or $SF_5^{\bullet +}$. Protein sequence coverage for all proteins common to both analyses are shown. A Δ Protein Sequence Coverage value was calculated for all 507 proteins seen in both analyses, and each individual protein is shown as a semi-transparent bar. The average sequence coverage for all common proteins was 3.1% higher for $SF_5^{\bullet +}$ (orange) compared to $C_{16}H_{10}^{\bullet +}$ (blue). Note the difference of peptide and protein sequence coverage discussed in the text.

charged peptide precursors reacted with $C_{16}H_{10}^{\bullet+}$ show on average 8.5% more CO_2 neutral loss while the $SF_5^{\bullet+}$ data yielded greater NH_3 and OH neutral loss. **Figure 2.5b** compares the abundance of amino acid-specific side chain neutral losses from the oxidized precursor [12]. The predominant differences are found in methionine, serine, and threonine residues. Interestingly, methionine and serine show a substantial increase in neutral loss when $C_{16}H_{10}^{\bullet+}$ is used, despite the lower recombination energy relative to $SF_5^{\bullet+}$. However, in total, very little difference is found in amino acid-specific side chain neutral loss when comparing the two NETD reagent cations.

Lastly, we investigated the $a^{\bullet}$ - and *x-type* product ions generated in these experiments, including how CO_2 neutral losses from product ions differed between the two reagent cations. **Figure 2.6a** provides a holistic look at the total number of $a^{\bullet}$ - and *x-type* fragment ions produced in the NETD datasets. As shown with our standard peptide data above, $SF_5^{\bullet+}$ yields more fragment ions species than $C_{16}H_{10}^{\bullet+}$, particularly more $a^{\bullet}$-*type* fragments, with an increase of 32%. **Figure 2.6b** shows the percentage of those sequence ions that also yield a CO_2 neutral loss peak. $SF_5^{\bullet+}$ cations generated 5.4% more *x-type* ion species with CO_2 neutral losses. To investigate how the occurrence of these neutral losses from sequencing ions impacts data analysis, we searched our yeast peptide database allowing for a variety of different fragment ion combinations. No combination of product ions and neutral losses yielded a greater number of PSMs when searching the $SF_5^{\bullet+}$ dataset than just $a^{\bullet}$ - and *x-type* fragment ions alone, which concurs with the same analysis performed on the fluoranthene dataset (not shown). We conclude that the amount of CO_2 neutral loss

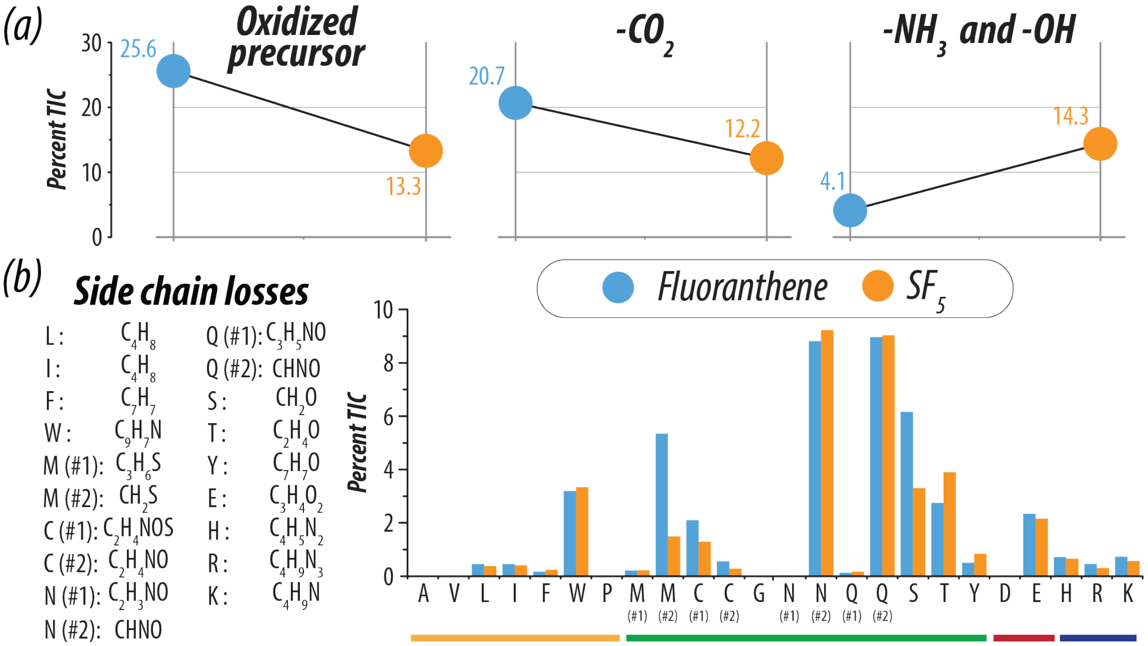


Figure 2.5: Distribution of neutral losses using $C_{16}H_{10}^{\bullet+}$ and $SF_5^{\bullet+}$ reagent cations. (a) The average percent of total ion current accounted for by the oxidized precursor ion and neutral losses of CO_2 and either NH_3 or OH from the oxidized precursor ion are shown for $z = -2$ PSMs from $C_{16}H_{10}^{\bullet+}$ (blue) and $SF_5^{\bullet+}$ (orange) nLC-MS/MS analyses. (b) The average percent of total ion current is shown for side chain neutral losses from the oxidized precursor ion for $z = -2$ PSMs. In order for a spectrum to be searched for a given side chain loss, the PSM sequence had to contain that residue. Note, amino acids are organized by their side chain properties: nonpolar (yellow), polar (green), acidic (red), and basic (blue)

is not significant enough to impact or adversely affect automated spectral annotation.

Conclusion

NETD provides direct access to analysis of proteomes in the negative mode, but improvements are still needed to make NETD amenable to the large proportion of low charge density precursors generated in whole-proteome analyses. In this experiment, we sought to improve peptide anion fragmentation using $SF_5^{\bullet+}$, a reagent cation with a higher IE of 9.6 eV, compared with $C_{16}H_{10}^{\bullet+}$, the most common NETD reagent cation with an IE of 7.9 eV. Using peptide fragmentation efficiency and unique peptide identifications as our primary metrics, we determined that $SF_5^{\bullet+}$ significantly improves shotgun proteomic analyses with NETD, especially considering the fragmentation of doubly deprotonated precursors. We contribute this gain in identification power to the increase in ionization energy for $SF_5^{\bullet+}$ compared with $C_{16}H_{10}^{\bullet+}$, which impacts the exothermicity of the ion–ion reaction and provides greater fragment ion yield and less non-dissociative negative electron transfer. The predominant gain in identification for low charge density precursors may also be useful when considering protein digestion using proteases other than trypsin, such as Lys-C, which often yields longer, less charge-dense peptides. Another benefit of using a gaseous reagent for NETD is the simplification of the ion source as no reagent vial heaters or heated transfer lines are required to volatilize and transfer the solid fluoranthene reagent. Although $SF_5^{\bullet+}$ appears to be a favorable choice for large-scale negative mode proteomics analyses, its use in other types of NETD applications may benefit as well. While previous

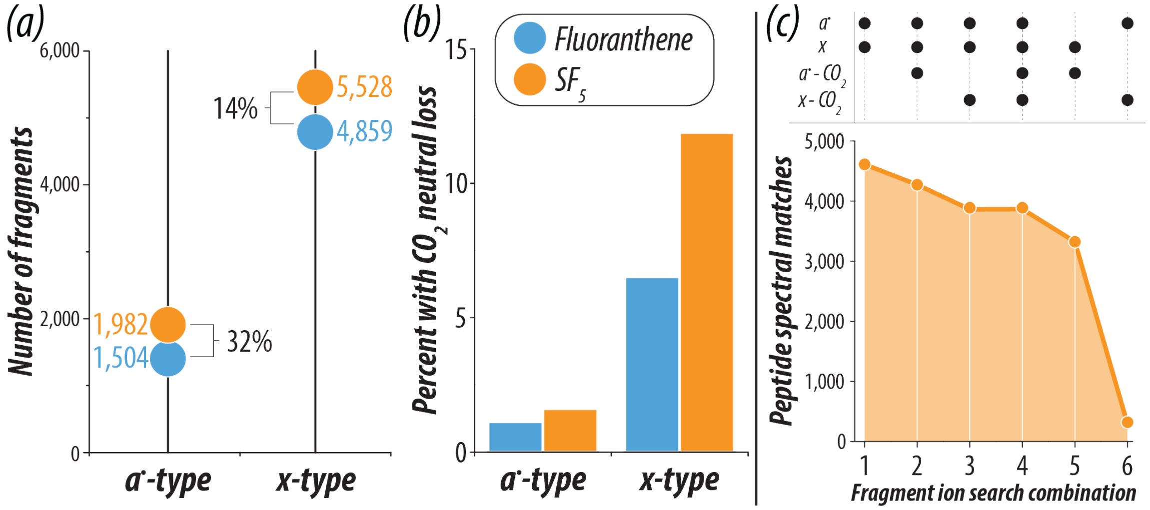


Figure 2.6: Abundance of $a^\bullet$ - and *x-type* fragment ions and fragment ion search settings. (a) The number of total $a^\bullet$ - and *x-type* fragment ions generated from the peptides common to both $C_{16}H_{10}{}^{\bullet+}$ (blue) and $SF_5{}^{\bullet+}$ (orange) analyses are shown. For both fragment ion types $SF_5{}^{\bullet+}$ produces the greater number of sequencing ions, and the percent gains over the number of ions from $C_{16}H_{10}{}^{\bullet+}$ analyses are shown in bold. Panel (b) compares the percentage of $a^\bullet$ - and *x-type* product ions from Panel (a) that have a corresponding neutral loss of CO_2. (c) Despite the small increase in CO_2neutral losses observed, incorporating CO_2 neutral losses from $a^\bullet$ - and *x-type* product ions as fragment ion types to query in a database search does not improve peptide identifications over using standard $a^\bullet$ - and *x-type* product ions only for the $SF_5{}^{\bullet+}$ data. The combination of product ion types used in the database searches are shown in black at the top, and the number of identified peptide spectral matches are shown in orange at the bottom. Similar results were obtained with the $C_{16}H_{10}{}^{\bullet+}$analyses (data not shown)

work has shown that lower recombination energy NETD reactions reduce the occurrence of labile PTM neutral loss [43], many other molecules might benefit from the more energetic reaction. In addition to peptides, NETD has been useful in the study of polynucleotides[35] and carbohydrates[50–53], and further improvements may be obtained through the utilization of $SF_5^{\bullet +}$ as the reagent cation.

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. Moghaddas 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. 582–7, May 2014.

[3] 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.," *Molecular & cellular proteomics : MCP*, vol. 13, pp. 339–47, Jan 2014.

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

[5] N. M. Riley, A. S. Hebert, and J. J. Coon, "Proteomics Moves into the Fast Lane," *Cell Systems*, vol. 2, no. 3, pp. 142–143, 2016.

[6] M. Yamashita and B. J. Fenn, "Negative ion production with the electrospray ion source," *J Phys.Chem.*, vol. 88, no. 20, pp. 4671–4675, 1984.

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

[8] R. F. Straub, R. D. Voyksner, and D. Voyksner, "Negative ion formation in electrospray mass spectrometry," *J Am Soc Mass Spectrom*, vol. 4, no. 7, pp. 578–587, 1993.

[9] C. S. Brinkworth, S. Dua, A. M. McAnoy, and J. H. Bowie, "Negative ion fragmentations of deprotonated peptides: backbone cleavages directed through both Asp and Glu.," *Rapid communications in mass spectrometry : RCM*, vol. 15, pp. 1965–73, Jan 2001.

[10] J. H. Bowie, C. S. Brinkworth, and S. Dua, "Collision-induced fragmentations of the (M-H)-parent anions of underivatized peptides: An aid to structure determination and some unusual negative ion cleavages," *Mass Spectrometry Reviews*, vol. 21, pp. 87–107, Jan 2002.

[11] J. J. Coon, J. Shabanowitz, D. F. Hunt, and J. E. P. Syka, "Electron transfer dissociation of peptide anions.," *Journal of the American Society for Mass Spectrometry*, vol. 16, pp. 880–2, Jun 2005.

[12] N. G. Rumachik, G. C. McAlister, J. D. Russell, D. J. Bailey, C. D. Wenger, and J. J. Coon, "Characterizing Peptide Neutral Losses Induced by Negative Electron-Transfer Dissociation (NETD)," *Journal of The American Society for Mass Spectrometry*, vol. 23, pp. 718–727, Apr 2012.

[13] J. W. Flora and D. C. Muddiman, "Selective, Sensitive, and Rapid Phosphopeptide Identification in Enzymatic Digests Using ESI-FTICR-MS with Infrared Multiphoton Dissociation," *Analytical Chemistry*, vol. 73, pp. 3305–3311, Jul 2001.

[14] J. A. Madsen, T. S. Kaoud, K. N. Dalby, and J. S. Brodbelt, "193-nm photodissociation of singly and multiply charged peptide anions for acidic proteome characterization," *PROTEOMICS*, vol. 11, pp. 1329–1334, Apr 2011.

[15] J. B. Shaw, D. a. Kaplan, and J. S. Brodbelt, "Activated ion negative electron transfer dissociation of multiply charged peptide anions.," *Analytical chemistry*, vol. 85, pp. 4721–8, May 2013.

[16] N. M. Riley, M. J. P. Rush, C. M. Rose, A. L. Richards, N. W. Kwiecien, D. J. Bailey, A. S. Hebert, M. S. Westphall, and J. J. Coon, "The Negative Mode Proteome with Activated Ion Negative Electron Transfer Dissociation," *Molecular & Cellular Proteomics*, vol. 14, pp. 2644–2660, 2015.

[17] G. C. McAlister, J. D. Russell, N. G. Rumachik, A. S. Hebert, J. E. P. Syka, L. Y. Geer, M. S. Westphall, D. J. Pagliarini, and J. J. Coon, "Analysis of the acidic proteome with negative electron-transfer dissociation mass spectrometry," *Analytical chemistry*, vol. 84, no. 6, pp. 2875–2882, 2012.

[18] S. L. Cook and G. P. Jackson, "Metastable atom-activated dissociation mass spectrometry of phosphorylated and sulfonated peptides in negative ion mode," *Journal of the American Society for Mass Spectrometry*, vol. 22, no. 6, pp. 1088–1099, 2011.

[19] S. A. Smith, C. L. Kalcic, L. Cui, and G. E. Reid, "Femtosecond laser-induced ionization/dissociation tandem mass spectrometry (fsLID-MS/MS) of deprotonated phosphopeptide anions," *Rapid Communications in Mass Spectrometry*, vol. 27, no. 24, pp. 2807–2817, 2013.

[20] J. Jai-nhuknan and C. J. Cassady, "Negative ion postsource decay time-of-flight mass spectrometry of peptides containing acidic amino acid residues," *Analytical Chemistry*, vol. 70, no. 24, pp. 5122–5128, 1998.

[21] V. Larraillet, R. Antoine, P. Dugourd, and J. Lemoine, "Activated-electron photodetachment dissociation for the structural characterization of protein polyanions," *Analytical Chemistry*, vol. 81, no. 20, pp. 8410–8416, 2009.

[22] V. Larraillet, A. Vorobyev, C. Brunet, J. Lemoine, Y. O. Tsybin, R. Antoine, and P. Dugourd, "Comparative dissociation of peptide polyanions by electron impact and photo-induced electron detachment," *Journal of the American Society for Mass Spectrometry*, vol. 21, no. 4, pp. 670–680, 2010.

[23] R. Antoine, L. Joly, T. Tabarin, M. Broyer, P. Dugourd, and J. Lemoine, "Photo-induced formation of radical anion peptides. Electron photo-detachment dissociation experiments [1]," 2007.

[24] K. Haselmann, B. Budnik, F. Kjeldsen, M. Nielsen, J. Olsen, and R. Zubarev, "Electronic excitation gives informative fragmentation of polypeptide cations and anions," *European Journal of Mass Spectrometry*, vol. 8, p. 117, 2002.

[25] F. Kjeldsen, O. A. Silivra, I. A. Ivonin, K. F. Haselmann, M. Gorshkov, and R. A. Zubarev, "Cα-C backbone fragmentation dominates in electron detachment dissociation of gas-phase polypeptide polyanions," *Chemistry - A European Journal*, vol. 11, no. 6, pp. 1803–1812, 2005.

[26] M. A. Halim, M. Girod, L. MacAleese, J. Lemoine, R. Antoine, and P. Dugourd, "213 nm Ultraviolet Photodissociation on Peptide Anions: Radical-Directed Fragmentation Patterns," *Journal of the American Society for Mass Spectrometry*, vol. 27, no. 3, pp. 474–486, 2016.

[27] S. M. Greer, J. R. Cannon, and J. S. Brodbelt, "Improvement of Shotgun Proteomics in the Negative Mode by Carbamylation of Peptides and Ultraviolet Photodissociation Mass Spectrometry.," *Analytical chemistry*, Nov 2014.

[28] M. R. Robinson, K. L. Moore, and J. S. Brodbelt, "Direct identification of tyrosine sulfation by using ultraviolet photodissociation mass spectrometry," *Journal of the American Society for Mass Spectrometry*, vol. 25, no. 8, pp. 1461–1471, 2014.

[29] J. A. Madsen, H. Xu, M. R. Robinson, A. P. Horton, J. B. Shaw, D. K. Giles, T. S. Kaoud, K. N. Dalby, M. S. Trent, and J. S. Brodbelt, "High-throughput Database Search and Large-scale Negative Polarity Liquid Chromatography–Tandem Mass Spectrometry with Ultraviolet Photodissociation for Complex Proteomic Samples," *Molecular & Cellular Proteomics*, vol. 12, pp. 2604–2614, Sep 2013.

[30] N. M. Riley, M. Bern, M. S. Westphall, and J. J. Coon, "A full-featured search algorithm for negative electron transfer dissociation," *Journal of Proteome Research*, vol. 15, no. 8, pp. 2768–2776, 2016.

[31] B. M. Prentice and S. A. McLuckey, "Gas-phase ion/ion reactions of peptides and proteins: acid/base, redox, and covalent chemistries.," *Chemical communications (Cambridge, England)*, vol. 49, pp. 947–65, 2013.

[32] S. A. McLuckey and M. Mentinova, "Ion/neutral, ion/electron, ion/photon, and ion/ion interactions in tandem mass spectrometry: Do we need them all? Are they enough?," *Journal of the American Society for Mass Spectrometry*, vol. 22, no. 1, pp. 3–12, 2011.

[33] W. J. Herron, D. E. Goeringer, and S. A. McLuckey, "Gas-Phase Electron Transfer Reactions from Multiply-Charged Anions to Rare Gas Cations," *Journal of the American Chemical Society*, vol. 117, pp. 11555–11562, Nov 1995.

[34] Y. Gao, J. Yang, and M. Cancilla, "Top-Down Interrogation of Chemically Modified Oligonucleotides by Negative Electron Transfer and Collision Induced Dissociation," *Analytical chemistry*, 2013.

[35] T. Y. Huang and S. A. McLuckey, "Gas-phase ion/ion reactions of rubrene cations and multiply charged DNA and RNA anions," *International Journal of Mass Spectrometry*, vol. 304, no. 2-3, pp. 140–147, 2011.

[36] D. M. Horn, Y. Ge, and F. W. McLafferty, "Activated ion electron capture dissociation for mass spectral sequencing of larger (42 kDa) proteins," *Analytical Chemistry*, vol. 72, no. 20, pp. 4778–4784, 2000.

[37] Y. Xia, H. P. Gunawardena, D. E. Erickson, and S. A. McLuckey, "Effects of cation charge-site identity and position on electron-transfer dissociation of polypeptide cations," *Journal of the American Chemical Society*, vol. 129, pp. 12232–12243, Oct 2007.

[38] J. B. Shaw, J. a. Madsen, H. Xu, and J. S. Brodbelt, "Systematic comparison of ultraviolet photodissociation and electron transfer dissociation for peptide anion characterization.," *Journal of the American Society for Mass Spectrometry*, vol. 23, pp. 1707–15, Oct 2012.

[39] D. L. Swaney, G. C. McAlister, M. Wirtala, J. C. Schwartz, J. E. Syka, and J. J. Coon, "Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors," *Anal Chem*, vol. 79, no. 2, pp. 477–485, 2007.

[40] H. B. Hamidane, D. Chiappe, R. Hartmer, A. Vorobyev, M. Moniatte, and Y. O. Tsybin, "Electron Capture and Transfer Dissociation: Peptide Structure Analysis at Different Ion Internal Energy Levels," *Journal of the American Society for Mass Spectrometry*, vol. 20, no. 4, pp. 567–575, 2009.

[41] A. R. Ledvina, C. M. Rose, G. C. McAlister, J. E. P. Syka, M. S. Westphall, J. Griep-Raming, J. C. Schwartz, and J. J. Coon, "Activated ion ETD performed in a modified collision cell on a hybrid QLT-Oribtrap mass spectrometer.," *Journal of the American Society for Mass Spectrometry*, vol. 24, pp. 1623–33, Nov 2013.

[42] H. P. Gunawardena, M. He, P. A. Chrisman, S. J. Pitteri, J. M. Hogan, B. D. M. Hodges, and S. A. McLuckey, "Electron transfer versus proton transfer in gas-phase ion/ion reactions of polyprotonated peptides," *Journal of the American Chemical Society*, vol. 127, no. 36, pp. 12627–12639, 2005.

[43] M. Huzarska, I. Ugalde, D. A. Kaplan, R. Hartmer, M. L. Easterling, and N. C. Polfer, "Negative electron transfer dissociation of deprotonated phosphopeptide anions: choice of radical cation reagent and competition between electron and proton transfer," *Analytical Chemistry*, vol. 82, pp. 2873–8, Apr 2010.

[44] D. M. Crizer, Y. Xia, and S. A. McLuckey, "Transition Metal Complex Cations as Reagents for Gas-Phase Transformation of Multiply Deprotonated Polypeptides," *Journal of the American Society for Mass Spectrometry*, vol. 20, no. 9, pp. 1718–1722, 2009.

[45] A. R. Ledvina, N. A. Beauchene, G. C. McAlister, J. E. P. Syka, J. C. Schwartz, J. Griep-Raming, M. S. Westphall, and J. J. Coon, "Activated-ion electron transfer dissociation improves the ability of electron transfer dissociation to identify peptides in a complex mixture," *Analytical Chemistry*, vol. 82, no. 24, pp. 10068–10074, 2010.

[46] C. M. Rose, J. D. Russell, A. R. Ledvina, G. C. McAlister, M. S. Westphall, J. Griep-Raming, J. C. Schwartz, J. J. Coon, and J. E. P. Syka, "Multipurpose dissociation cell

for enhanced ETD of intact protein species.," *Journal of the American Society for Mass Spectrometry*, vol. 24, pp. 816–27, Jun 2013.

[47] N. M. Riley, M. S. Westphall, and J. J. Coon, "Activated Ion Electron Transfer Dissociation for Improved Fragmentation of Intact Proteins," *Analytical Chemistry*, vol. 87, no. 14, pp. 7109–7116, 2015.

[48] L. Y. Geer, S. P. Markey, J. a. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant, "Open mass spectrometry search algorithm.," *Journal of proteome research*, vol. 3, no. 5, pp. 958–64, 2004.

[49] C. M. Rose, M. J. P. Rush, N. M. Riley, A. E. Merrill, N. W. Kwiecien, D. D. Holden, C. Mullen, M. S. Westphall, and J. J. Coon, "A Calibration Routine for Efficient ETD in Large-Scale Proteomics," *Journal of the American Society for Mass Spectrometry*, vol. 26, no. 11, pp. 1848–1857, 2015.

[50] J. J. Wolff, F. E. Leach, T. N. Laremore, D. A. Kaplan, M. L. Easterling, R. J. Linhardt, and I. J. Amster, "Negative Electron Transfer Dissociation of Glycosaminoglycans," *Analytical Chemistry*, vol. 82, pp. 3460–3466, May 2010.

[51] F. E. Leach, J. J. Wolff, Z. Xiao, M. Ly, T. N. Laremore, S. Arungundram, K. Al-Mafraji, A. Venot, G.-J. Boons, R. J. Linhardt, and I. J. Amster, "Negative Electron Transfer Dissociation Fourier Transform Mass Spectrometry of Glycosaminoglycan Carbohydrates," *European Journal of Mass Spectrometry*, vol. 17, pp. 167–176, Apr 2011.

[52] Y. Huang, X. Yu, Y. Mao, C. E. Costello, J. Zaia, and C. Lin, "De Novo Sequencing of Heparan Sulfate Oligosaccharides by Electron-Activated Dissociation," *Analytical Chemistry*, vol. 85, pp. 11979–11986, Dec 2013.

[53] H. Hu, Y. Huang, X. Yu, X. Yongmei, J. Liu, C. Zong, G.-J. Boons, C. Lin, Y. Xia, and J. Zaia, "A Computational Framework for Heparan Sulfate Sequencing Using High-resolution Tandem Mass Spectra," *Molecular and cellular Proteomics*, vol. 13, no. 9, pp. 2490–502, 2014.

Chapter 3

THE NEGATIVE MODE PROTEOME WITH ACTIVATED ION NEGATIVE ELECTRON TRANSFER DISSOCIATION (AI-NETD)

This chapter has been published and is reprinted with permission from:

Riley NM, **Rush MJP**, Rose CM, Richards AL, Kwiecien NW, Bailey DJ, Hebert AS, Westphall MS, Coon JJ. *The Negative Mode Proteome with Activated Ion Negative Electron Transfer Dissociation (AI-NETD).* Molecular & Cellular Proteomics **2015**, *14 (10)*, 2644-2660. doi: 10.1074/mcp.M115.049726.

Abstract

The field of proteomics almost uniformly relies on peptide cation analysis, leading to an underrepresentation of acidic portions of proteomes, including relevant acidic posttranslational modifications. Despite the many benefits negative mode proteomics can offer, peptide anion analysis remains in its infancy due mainly to challenges with high-pH reversed-phase separations and a lack of robust fragmentation methods suitable for peptide anion characterization. Here, we report the first implementation of activated ion negative electron transfer dissociation (AI-NETD) on the chromatographic timescale, generating 7,601 unique peptide identifications from *Saccharomyces cerevisiae* in single-shot nLC-MS/MS analyses of tryptic peptides — a greater than 5-fold increase over previous results with NETD alone. These improvements translate to identification of 1,106 proteins, making this work the first negative mode study to identify more than 1,000 proteins in any system. We then compare the performance of AI-NETD for analysis of peptides generated by five proteases (trypsin, LysC, GluC, chymotrypsin, and AspN) for negative mode analyses, identifying as many as 5,356 peptides (1,045 proteins) with LysC and 4,213 peptides (857 proteins) with GluC in yeast — characterizing 1,359 proteins in total. Finally, we present the first deep-sequencing approach for negative mode proteomics, leveraging offline low-pH reversed-phase fractionation prior to online high-pH separations and peptide fragmentation with AI-NETD. With this platform, we identified 3,467 proteins in yeast with trypsin alone and characterized a total of 3,730 proteins using multiple proteases, or nearly 83% of the expressed yeast

proteome. This work represents the most extensive negative mode proteomics study to date, establishing AI-NETD as a robust tool for large-scale peptide anion characterization and making the negative mode approach a more viable platform for future proteomic studies.

Introduction

Global protein analysis continues to enjoy substantial technological leaps forward in its ability to characterize protein expression in a variety of organisms with both speed and sensitivity.[1–4] Nevertheless, the impressive advances in protein sequence technology over past decades have rigidly adhered to positive electrospray ionization for MS[1] analysis, limiting the scope of peptides and posttranslational modifications that can be analyzed. Widely utilized acidic mobile phases both permit stable and reproducible reversed-phase separations and also provide optimal conditions for ionization and detection of peptides and proteins that readily accept positive charge via protonation (i.e. basic species). Acidic peptides and proteins, however, favor deprotonation, making positive electrospray regimes ill-suited for their characterization. Moreover, important classes of posttranslational modifications (PTMs), such as phosphorylation, sulfation, and glycosylation, can impart acidic properties to the peptides and proteins they modify, often producing entire classes of biomolecules that preferentially ionize as anions.[5–10]

Electrospray ionization operated in the negative mode can generate multiply deprotonated species;[11,12] however, canonical collisional activation methods produce MS/MS spectra riddled with neutral losses and internal fragments that are difficult, if not impossible, to interpret.[13–16] Alternatively, a number of emerging fragmentation techniques, including electron-based dissociation methods and photodissociation approaches, can generate sequence informative MS/MS spectra from peptide anions.[17–23] Both negative electron transfer dissociation (NETD) and ultraviolet photodissociation (UVPD) have been employed in large-scale proteomic studies, enabling sequencing of thousands of unique peptides in a single experiment.[24,25] In NETD, the negative mode analog of electron transfer dissociation (ETD),[26] peptide anions are oxidized with reagent cations, causing the radical peptide anions to undergo electron rearrangement steps that often lead to cleavage of the C-Cα backbone bond, producing $a^{\bullet}$- and x-type product ions.[18,24,27,28] Sometimes, however, an electron is abstracted from the precursor anion and backbone cleavage is achieved, but the product ions are held together by intramolecular noncovalent interactions as long-lived charge-reduced species that do not separate. The probability of this phenomenon, called nondissociative negative electron transfer, is directly related to precursor anion charge density; as charge density decreases, i.e. the precursor mass-to-charge ratio (*m/z*) increases, so does the magnitude of nondissociative negative electron transfer, limiting the amount of sequence information derived from the NETD MS/MS event.[29–31]

Many approaches have been explored to mitigate the effects of nondissociative electron transfer in ETD reactions of peptide cations, including collisional activation of all product ions,[32,33] activation of nondissociative electron transfer products,[34,35] elevated bath gas temperatures,[36] and infrared photon bombardment concurrent to the ETD reaction.[37] The last of these approaches, termed activated ion ETD (AI-ETD), has shown substantial promise for proteomics applications.[38,39] The concomitant IR photoirradiation disrupts the secondary gas-phase structure responsible for nondissociative electron transfer, increasing the efficiency of sequence-informative product ion generation; furthermore, the introduction of additional energy to the reaction during AI-ETD occurs only during the ion–ion reaction time, keeping the instrument cycle time as short as possible and reducing problematic hydrogen-atom rearrangements that can occur prior to fragment ion separation with other supplemental activation techniques.[34,37] Indeed, activated ion NETD (AI-NETD), which uses simultaneous IR irradiation during the NETD reaction, has been reported to improve peptide anion fragmentation for a handful of standard peptides,[30] but a large-scale implementation of AI-NETD for negative mode shotgun proteomics has yet to be demonstrated.

We recently described the development of a multipurpose dissociation cell (MDC) specifically for improved ion–ion reactions on an ETD-enabled dual cell quadrupole ion trap-Orbitrap hybrid mass spectrometer, the same platform on which we reported the successful implementation of NETD for high-throughput peptide analyses.[40] The MDC, equipped with a higher operating rf frequency and longer axial dimensions, allows faster ETD reaction times and larger precursor ion populations for improved fragment ion signal-to-noise. The placement of MDC is ideal for implementing AI-NETD on an NETD-capable mass spectrometer because a continuous wave CO_2 laser can be easily introduced concentrically to the trapping volume of the MDC.[39,41,42]

Here, we present the first description of NETD in the MDC reaction vessel, in addition to the first implementation of AI-NETD for large-scale peptide anion analysis. Analyzing complex mixtures of peptides from *Saccharomyces cerevisiae* whole cell lysates, we show that AI-NETD, in combination with judiciously chosen high-pH chromatographic conditions, enables the most robust analysis of peptide anions to date, identifying over 1,100 proteins in single-shot experiments. Furthermore, we extend our studies to investigate the performance of AI-NETD with five distinct proteases (trypsin, LysC, GluC, chymotrypsin, and AspN), providing the first large-scale investigation into the performance of multiple enzymes for peptide anion characterization. Using both single-shot analyses and deep sequencing via offline low-pH fractionation for each protease, we assert that entire proteomes can be investigated in the negative mode with AI-NETD, providing a new platform to thoroughly explore biologically relevant hypotheses, e.g. acidic PTM networks, which were previously inaccessible with canonical positive mode approaches.

Materials and Methods

Mass Spectrometry Instrumentation. The multipurpose dissociation cell (MDC) replaces the preexisting HCD cell in the ETD-enabled dual cell quadrupole linear ion trap-Orbitrap hybrid mass spectrometer system (**Figure 3.1c**)[43,44] (Thermo Fisher Scientific, San Jose, CA), retaining its basic geometry but requiring additional electronics to supply higher trapping rf voltages for faster reaction times, axial rf voltages for charge-sign independent trapping, and independently controlled DC (direct current) biases to its four sections. Offsets used for previous cation analysis with the MDC were inverted to accommodate injection and trapping of precursor anions rather than cations. Consistent with our earlier NETD work,[24,28] positive reagent ions were generated without hardware modification through optimized electron ionization/chemical ionization (EI/CI) source conditions, and reagent transmission through ion-transfer optics was tuned via automated optimization routines. Radical fluoranthene reagent cations were formed in a CI ion volume in the presence of nitrogen gas and were then accumulated in the back two sections of the MDC for NETD reactions. Following the NETD reaction via charge-sign independent trapping, the reaction was quenched by setting the center sections of the MDC to a positive DC offset (10 V), retaining anionic product ions and ejecting remaining cationic reagents. Product ions were then transferred to the C-trap for subsequent Orbitrap mass analysis using an extraction gradient analogous to that used for HCD scans. AI-NETD was performed by irradiating the trapping volume of the MDC during the entirety of the NETD reaction with a Firestar T-100 Synrad 100-W CO_2 continuous wave laser (Mukilteo, WA). The laser was introduced into the cell via an excavated ion passage in the reagent ion transfer multipole and a ZnSe window that was installed concentric to the MDC. Using instrument firmware and modification to instrument code in conjunction with a gated laser controller, laser power output (in Watts) from the laser (10.6 μm) was modulated remotely through voltage inputs to the controller and was triggered to fire only during the NETD reaction as it was being conducted in the MDC. The nitrogen pressure in the MDC was lowered to a ΔN_2 pressure of ~0.1×10^{-10} Torr, as measured by the Penning gauge in the Orbitrap chamber, to prevent collisional cooling that negates the additional energy supplied by the infrared laser. Lowered nitrogen pressure also increased transmission of the fluoranthene reagent cation, so pressure in the MDC was held at similar levels for NETD analyses (ΔN_2 of ~0.3×10^{-10} Torr). For more detailed descriptions of the MDC and affixed laser, see previous work.[39–41]

Sample Preparation. All protein lysates were derived from *S. cerevisiae* strain BY4741, as described previously.[45] Briefly, cells were grown to an optimal density at 600 nm (OD600) of ~0.6 and pelleted by centrifugation. A pellet was resuspended in lysis buffer (50 mM Tris, pH 8; 8 M urea; 75 mM sodium chloride; 10 mM sodium butyrate:protease and phosphatase inhibitor tablet (Roche Diagnostics, Indianapolis, IN)), and yeast cells were lysed by glass bead milling (Retsch GmbH, Haan, Germany). Two ml of acid-washed glass beads were combined with 2 ml of resuspended yeast cells in a 5 ml stainless steel container and shaken eight times at 30 Hz for 4 min with a 1 min rest in between. Lysate protein concentration was measured via BCA (Thermo Pierce, Rockford, IL), and yeast proteins were reduced through incubation in 5 mM dithiothreitol for 45 min at 58 °C. Free cysteines were alkylated in 15 mM iodoacetamide in the dark for 30 min. The alkylation was stopped with 5 mM DTT. For trypsin digestion, a 1 mg protein aliquot was digested overnight at room temperature in 1.5 M urea with trypsin (Promega, Madison, WI) added at a 1:50 (w/w) enzyme to protein ratio. A second trypsin addition was performed in the morning at a 1:100 (w/w) enzyme to protein ratio for 1 h. The digestion was quenched by the addition of TFA. For LysC digestion, a 1 mg protein aliquot was digested overnight at room temperature in 4 M urea with endo LysC (Wako Chemicals, Richmond, VA) at a 1:50 enzyme:protein ratio. Following overnight digestion, a second 1:50 aliquot of LysC was added and the digestion was allowed to proceed for 1 h. For GluC digestion, a 1 mg protein aliquot was digested overnight with

25 μg GluC (Roche) at room temperature in 0.5 M urea. Following overnight digestion, the sample was incubated with an additional 25 μg GluC for 1 h. For chymotrypsin digestion, a 1 mg protein aliquot was digested overnight with 12.5 μg of chymotrypsin (Promega) in 1 M urea. Following overnight digestion, an additional 12.5 μg of chymotrypsin was added to the sample, and the digestion was allowed to proceed for 1 h. For digestion with AspN, a 1 mg protein aliquot was incubated with 6 μg AspN (Roche) at room temperature overnight. Each digest was quenched by the addition of TFA and desalted over tC18 Sep-Pak cartridges (Waters, Milford, MA).

High-pH nLC-MS/MS. Five total reversed-phase solvent systems were tested for online high-pH separations. Ammonium hydroxide, ammonium formate, and piperidine were purchased from Sigma Aldrich (St. Louis, MO) and HPLC-grade water and acetonitrile were purchased from Fisher Scientific (Waltham, MA). The two ammonium formate solvent systems consisted of mobile phase A (5 mM ammonium formate in water) and mobile phase B (5 mM ammonium formate in 85% acetonitrile), basified to either pH 10 or pH 11.5 with ammonium hydroxide. Two different piperidine solvents systems used a mobile phase A of water and mobile phase B of 85% acetonitrile, 15% water, with either 5 mM or 10 mM piperidine in both A and B. Another 5 mM piperidine solvent system had the same mobile phase B, but mobile phase A consisted of 95% water, 5% DMSO with 5 mM piperidine. Reversed-phase columns were packed in-house using 75 μm inner diameter, 360 μm outer diameter bare fused silica capillary. A nanoelectrospray tip was laser pulled

(Sutter Instrument Company, Novato, CA) and packed with 3.5 μm diameter, 130 Å pore size ethylene bridged hybrid C18 particles (Waters) to a length of 30–35 cm. The column was installed on a nanoACQUITY UltraPerformance LC (Waters) using a stainless steel ultra-high pressure union formatted for 360 μm outer diameter columns (IDEX). One microgram of sample was loaded onto the column in 95% A for 10 min at 400 nl min^{-1}. Gradient elution was performed at 400 nl min^{-1}, and gradients increased linearly from 5 to 30% B over 70 min, followed by an increase to 70% B at 76 min and a wash at 70% B for 4 min. The column was then re-equilibrated at 5% B for 10 min.

Eluting peptide anions were converted to gas-phase ions by electrospray ionization at -1.5 kV with respect to ground, and the inlet capillary temperature was held at 300 °C. Survey scans of peptide precursors were collected over the 300–1,250 Th range with an automatic gain control (AGC) target value of 1,000,000, followed by data-dependent NETD MS/MS scans of the 10 most intense peaks (maximum injection times of 200 ms for both full and tandem MS scans). Precursors with charge states equal to one or unassigned were rejected. NETD reactions were performed in either the mass-analyzing quadrupole linear ion trap (A-QLT) or the MDC (50,000 or 100,000 AGC target values, respectively), followed by mass analysis in either the A-QLT or Orbitrap, as indicated in the text. The radical cation of fluoranthene was the reagent used for all NETD analyses. For NETD reactions conducted in the A-QLT, the reagent AGC target value was set to 1,000,000 and the default reaction time was set to 100 ms for $z = -2$ precursors, with reaction time scaling enabled as discussed in the text. For all MDC analyses, reagent accumulation times were set to 20 ms, and reaction times were as follows: 40 ms for z = -2, 30 ms for z = -3, 20 ms for z = -4, 15 ms for z = -5, and 10 ms for z = -6 and higher charge states. During AI-NETD reactions, the external CO_2 continuous wave laser was triggered using instrument firmware and modification to instrument code in conjunction with a gated laser controller. The laser irradiated the trapping volume of the MDC during the entirety of the NETD reaction at 70% total output. Precursors were isolated using a ±0.9 Th isolation window, and an exclusion window of ±10 ppm was constructed around the monoisotopic peak of each selected precursor for 45 s. Resolving powers of 60,000 and 15,000 at 400 *m/z* were used for survey scans and MS/MS scans in the Orbitrap, respectively.

Low-pH nLC-MS/MS. For comparison to positive mode methods, single-shot HCD and ETD analyses were collected for tryptic yeast peptides. Reversed-phase columns were prepared as described above. Mobile phase A was 0.2% formic acid in water with 5% DMSO, and mobile phase B was 0.2% formic acid in acetonitrile. One microgram of sample was loaded onto the column in 95% A for 10 min at 400 nl min^{-1}. Gradient elution was performed at 400 nl min^{-1}, with the gradient increased linearly from 5 to 25% B over 70 min, followed by an increase to 70% B at 76 min and a wash at 70% B for the 4 min. Electrospray voltage was set to 2 kV with respect to ground, and the inlet capillary was held at 275 °C. Survey scans of peptide precursors were collected over the 300–1,250 Th range with an AGC target value of 1,000,000 and 60,000 resolution, followed by data-dependent HCD MS/MS scans of the 15 most-intense peaks or ETD MS/MS scans of the 10 most-intense peaks (maximum injection times of 75 ms and 200 ms for full and tandem MS scans, respectively). Precursors with charge states equal to 1 or unassigned were rejected. HCD and ETD MS/MS events were both performed in the MDC with an AGC target value of 100,000, followed by mass analysis in the Orbitrap at 15,000 resolution. Normalized collision energies of 30 were used for HCD events. The radical anion of fluoranthene was the reagent used for ETD reactions. Reagent accumulation times were set to 20 ms, and reaction times were analogous to NETD reaction times described above. Precursor isolation and dynamic exclusion were the same as above.

Low-pH Prefractionation. In addition to single-shot nLC-MS/MS runs, deep-sequencing analyses were performed on digests from each enzyme, leveraging low-pH RPLC offline fractionation for an orthogonal degree of separation prior to online high-pH chromatography. Peptides were fractionated on a Phenomenex (Torrance, CA) Gemini 5 μm, 110 Å pore size C18 column (250 × 4.6 mm) with 0.1% TFA in pure water and 80% acetonitrile (mobile phases A and B, respectively). The separation gradient had a flow rate of 0.8 ml min^{-1} starting at 5% B for 4 min. From 4 to 8 min, B was increased to 12% and then to 45% at 49 min. At 51 min, a 5 min wash of 100% B started, followed by 20 min of re-equilibration in 5% B. Fractions were collected every minute from 4 min to 54 min for a total of 50 fractions, which were then combined into 10 total fractions in concatenated fashion. Each set of fractions was run in triplicate.

Data Analysis. Tandem mass spectra were searched with the Open Mass Spectrometry Search Algorithm, which was previously modified to accommodate anionic peptide fragments and NETD spectra, which contain $a^{\bullet}$- and x-type product ions.[24,46] Prior to the Open Mass Spectrometry Search Algorithm search, spectra were "cleaned" such that charge-reduced product ions and neutral losses within the window 55 Da below and 5 Da above the charge-reduced peaks were removed in addition to a ±3 Da window around the unreacted precursor.[47,48] A multi-isotope search using three isotopes with a mass tolerance of ±125 ppm was used for precursors, and a monoisotopic mass tolerance of ±0.30 Da or ±0.02 Da was used for product ions ($a^{\bullet}$- and x-type) in the ion trap or Orbitrap, respectively. Oxidation of methionine was specified as a variable modification, while carbamidomethylation of cysteine was a set as a fixed modification. For all enzymes, three missed cleavages were allowed with the following specificity: trypsin, full with P-rule; LysC, full with P-rule; GluC, full DE; chymotrypsin, full with P-rule; and AspN, full. Data processing was performed using in-house software (COMPASS) designed for Open Mass Spectrometry Search Algorithm search outputs.[49] Peptide spectral matches (PSMs) were made against the UniProt yeast database downloaded on September 29, 2014 (6,726 entries), which was concatenated with a reversed sequence version of the forward database. Peptides were filtered to a 1% false discovery rate using both e-value and precursor mass accuracy. When pooling spectra from multiple nLC-MS/MS analyses, the false discovery rate was calculated for the aggregate set of data rather than calculating a separate false discovery rate for each run prior to combining results. Information pertinent to fragmentation evaluation for NETD and

AI-NETD was extracted from MS/MS scans using a C# script developed in-house. Protein isoelectric points (pI) were calculated using ExPASy (http://www.expasy.org/). The same pipeline was used for positive mode comparisons, searching *b*- and *y*-type product ions for HCD and *c*- and $z^{\bullet}$-type product ions for ETD, with product ion search tolerances of ±0.02 Da. ETD spectra were cleaned (above). Large-scale positive mode data[50] were downloaded from Chorus (ID# 183) and analyzed with COMPASS. Raw files were searched as previously described.

Results

Negative Electron Transfer Dissociation in the Multipurpose Dissociation Cell. In our previous work using NETD for large-scale analyses of peptide anions,[24] we conducted all NETD reactions in the high-pressure trap of the dual-cell mass-analyzing quadrupole linear ion trap (A-QLT) and subsequent mass analysis was performed in the low-pressure trap of the A-QLT (**Figure 3.1c**). At the time, the sensitivity and speed of the A-QLT for analysis of tandem mass spectra offered more benefit than the higher resolution/accurate mass provided by Orbitrap mass analysis, especially considering the low precursor anion flux observed in those experiments; however, Orbitrap mass analysis did provide more confident spectral identification and was used in ensuing studies.[28] Following these investigations, we described the multipurpose dissociation cell (MDC) that demonstrated improved ETD performance (faster reaction times and higher product ion signal-to-noise) for precursor cations.[40] We hypothesized that the MDC could offer similar benefits for NETD analysis, granting access to the superior spectral quality of higher resolution/accurate mass product ion mass analysis in the Orbitrap without sacrificing the scan speed achieved with the A-QLT.

As described above, we modified the MDC to perform charge-sign-independent trapping of precursor anions and reagent cations, permitting NETD of multiply deprotonated peptides with radical fluoranthene cations (**Figure 3.1a**).[51] Using a 5 mM piperidine solvent system (*vide infra*), we compared performance of the A-QLT and MDC for NETD analyses using 90-min shotgun nLC-MS/MS experiments on a complex mixture of *S. cerevisiae* peptides, performed in triplicate. Reflecting our previous studies for NETD reactions in the A-QLT, product ion mass analysis in the A-QLT outperformed product ion mass analysis in the Orbitrap (3,530 vs 3,134 peptides). Thus, all experiments using the A-QLT for NETD reactions also used the A-QLT for product ion mass analysis. All NETD reactions in the MDC, however, used the Orbitrap for product ion mass analysis, as dictated by instrument geometry. To compare directly to our previous work, we first used an NETD reaction time of 100 ms for all precursor charge states for A-QLT analyses but used dynamic reaction times for MDC reactions, as optimal reaction times scale with precursor charge.[52] Because the MDC allows faster reaction times, only 40 ms were needed to achieve high-quality spectra for doubly deprotonated peptides compared with the 100 ms required for the reaction in the A-QLT. NETD experiments using the MDC as a reaction vessel afforded slightly more MS/MS scans on average than those using the A-QLT (12,494 vs 12,374, respectively) in addition to providing higher MS/MS success rates, i.e. percentage of tandem mass

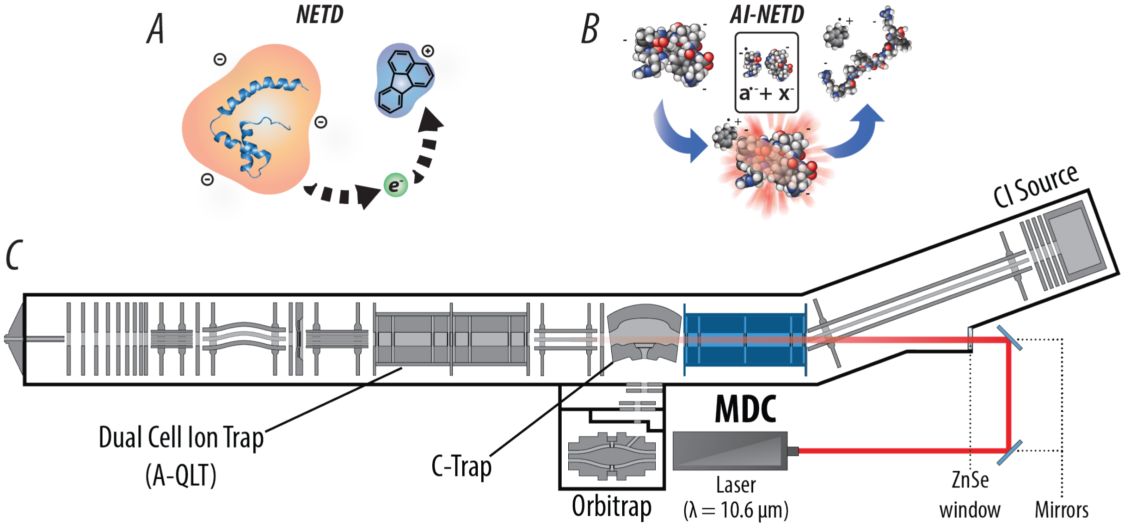


Figure 3.1: Modified linear ion trap-Orbitrap hybrid MS system for performing AI-NETD. a) NETD, which involves oxidation of precursor peptide anions by a reagent cation, can now be performed in the mass-analyzing quadrupole linear ion trap (A-QLT) and multipurpose dissociation cell (MDC). **b)** Modifying the MDC to perform NETD opens access to AI-NETD, in which anionic peptide precursors are concurrently irradiated with IR photons during the ion–ion reaction. This additional photoactivation disrupts peptide secondary gas-phase structure, increasing the efficiency of the NETD reaction. **c)** Instrument schematic of the hybrid linear ion trap-Orbitrap mass spectrometer is modified with the MDC in place of the traditional HCD collision cell, in addition to an excavated beam path, ZnSe window, and affixed 10.6 μm CO_2 laser for concentric irradiation of the MDC.

spectra that map back to sequence (29% vs 27%). With these advantages, NETD single-shot analyses in the MDC identified 3,805 unique peptides to the A-QLT's 3,411 unique peptides, both of which outmatched our previous results. Furthermore, we conducted a third set of experiments, this time enabling dynamic reaction times for NETD conducted in the A-QLT (scaled with precursor charge). With these conditions, NETD in the A-QLT averaged 12,582 MS/MS scans per run, illustrating the increase in scan speed afforded by scaled reaction times; however, the MS/MS success rates in these experiments also averaged 27%, matching that produced with a static reaction time in the A-QLT. Despite the increase in the number of MS/MS scans with dynamic reaction times enabled, the MDC still outperformed the A-QLT in the number of unique peptides identified (3,805 vs 3,530). These results illustrate the advantages the MDC provides for shotgun nLC-MS/MS peptide anion analyses, making high-quality, higher resolution/accurate mass NETD tandem mass spectra accessible for routine experiments.

Activated Ion NETD for Large-Scale Sequencing of Peptide Anions. Beyond the advantages the MDC provides for NETD alone, the ability to conduct NETD in this reaction vessel provides straightforward access to AI-NETD. The instrument geometry enables simple alignment of an external infrared laser that can be introduced concentric to the trapping volume of the MDC,[39–42] facilitating concurrent photoactivation during NETD for improved fragmentation efficiency (**Figure 3.1b**). Following laser alignment, we conducted another set of triplicate 90-min nLC-MS/MS experiments with yeast peptides, this time

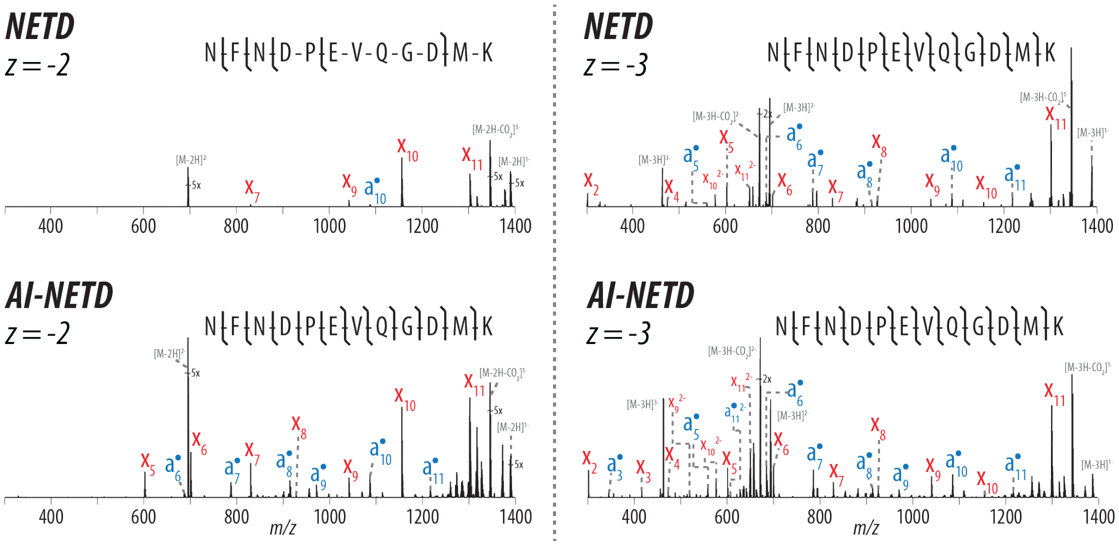


Figure 3.2: NETD- and AI-NETD-MS/MS spectra for the same peptide at z = -2 and -3. AI-NETD improves precursor-to-product ion conversion for enhanced peptide dissociation for precursors in lower charge states, increasing peptide sequence coverage in this case from 45% to 100%. NETD and AI-NETD both perform well on the same peptide at a higher charge (100% sequence coverage with both), although AI-NETD still provides a greater total number of sequencing ions. NETD and AI-NETD spectra are on the same scale for each precursor charge state.

comparing NETD in the MDC to AI-NETD in the MDC. The improvement in peptide fragmentation was immediately apparent. **Figure 3.2** provides an example of a peptide, NFNDPEVQGDMK, successfully identified at z = -2 and z = -3 in both NETD and AI-NETD analyses. The doubly deprotonated species of this 12-residue peptide has a moderate charge density (*m/z* 695.29), and fragmentation with NETD is somewhat limited (top left panel), providing only 45% peptide sequence coverage (as defined by number of bonds broken divided by total number of bonds); however, AI-NETD provides extensive fragmentation of the peptide, permitting straightforward annotation of fragment ions that provide 100% peptide sequence coverage (bottom left panel). NETD for the more charge-dense triply deprotonated species (*m/z* 463.19) provides more comprehensive fragmentation than with its doubly deprotonated counterpart, as expected (top right panel). AI-NETD maintains its high level of performance for the z = -3, too — again enabling 100% sequence coverage and providing more sequencing ions than NETD alone (bottom right panel).

The overall performance of AI-NETD for the large-scale analyses was just as compelling. The MS/MS success rate for AI-NETD experiments averaged nearly 53%, a drastic improvement over the 29% success rate of NETD alone. To demonstrate how AI-NETD could provide such a boost, we chose 409 unique peptides that were identified using both NETD and AI-NETD with z = -2 and compared the extent of fragmentation achieved with the two fragmentation techniques by extracting the product ions detected in each MS/MS spectrum. **Figure 3.3** displays fragment maps for these spectra, providing a separate column for $a^{\bullet}$-type and x-type fragments. Each column is further divided into subcolumns, the number

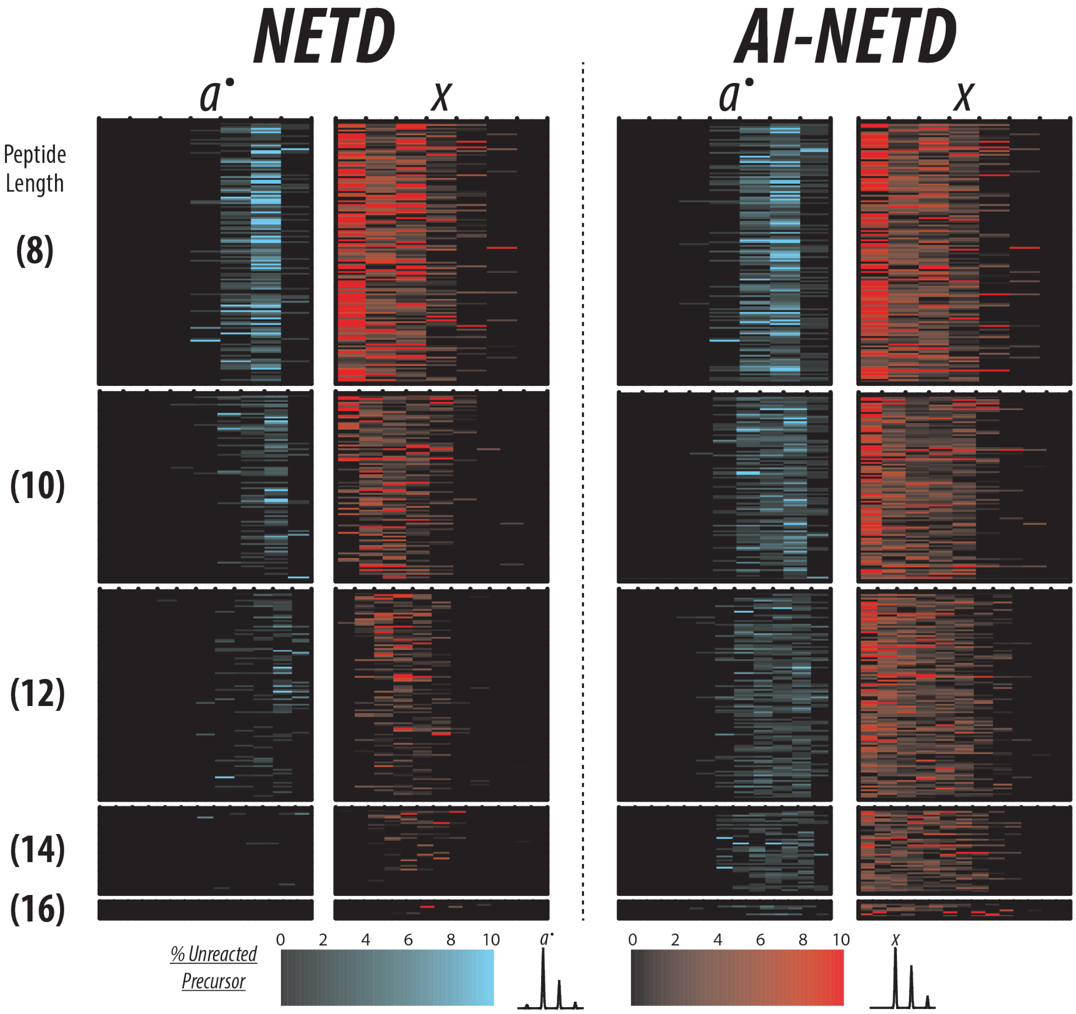


Figure 3.3: A fragment map of peptides identified with both NETD and AI-NETD. Here, each row is a unique peptide so that the same row across all four columns represents the same peptide sequence. Each subcolumn corresponds a peptide backbone bond so that a peptide with eight residues has seven backbone bonds and thus seven subcolumns for both $a^\bullet$- and x-type product ions. The numbers in parenthesis to the left show peptide length in number of residues, and all peptides shown here are z = -2, meaning precursor charge density decreases from top to bottom. With NETD, $a^\bullet$- and x-type fragments decrease in number and intensity as precursor charge density decreases (i.e. as peptide length increases). AI-NETD maintains superior fragment ion generation even with decreasing precursor charge density, greatly increasing peptide dissociation and sequence coverage compared with NETD.

of which corresponds to the number of backbone bonds in the peptide. For example, a peptide that is eight residues long has seven columns representing the seven backbone bonds. The furthest most left subcolumn for the $a^\bullet$-type fragments represents fragment $a^\bullet{}_1$, while the furthest most right subcolumn shows $a^\bullet{}_7$. Conversely, the furthest most left subcolumn for x-type fragments shows the x_7 fragment, whereas the right most subcolumn shows x_1. The color scale for the two fragment ion types, shown at the bottom of the figure, indicates the intensity of the fragment. To permit comparisons among multiple spectra, the intensity of all fragments in a given spectrum was normalized to the intensity of the unreacted precursor detected in that spectrum; thus, the intensities of the fragment ions are reported as percentages of this intensity. Additionally, the peptides shown in this figure are first grouped by length, which is depicted by the number in parentheses on the far left side, and the peptides with the same length are organized by *m/z* values in ascending fashion. With this organization, the charge density of the peptide precursors decreases from top to bottom, making the charge density dependence trends of NETD apparent. As peptide length increases, the extent of fragmentation achieved with NETD, alone, decreases noticeably, corresponding to expected trends for decreasing charge density. By comparison, AI-NETD provides extensive fragmentation for nearly all peptides shown, even as charge densities of the precursors decrease, increasing the number and intensity of both $a^\bullet$-type and x-type fragments. By mitigating the detrimental charge density dependence of NETD, AI-NETD clearly offers advantages for peptide anion fragmentation, explaining the increase in MS/MS success rate between NETD and AI-NETD experiments.

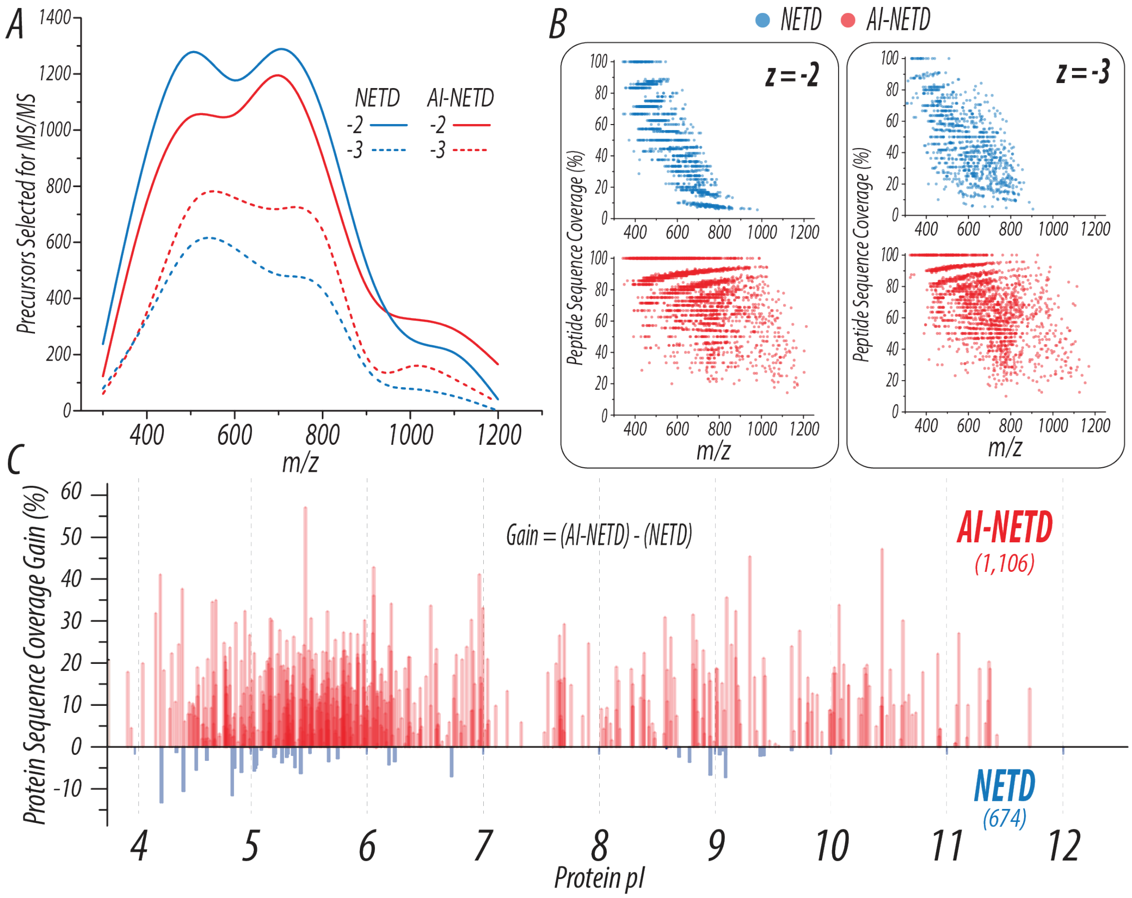


Figure 3.4: AI-NETD outperforms NETD on a global scale. a) The distribution of precursors selected for NETD or AI-NETD in a given run show roughly equivalent populations of peptides to fragment. **b)** Using the PSMs generated from the runs shown in (a), peptide sequence coverage is plotted as a function of precursor *m/z*. AI-NETD increases the number of peptides identified, extends the *m/z* range that can generate successful PSMs, and provides overall higher peptide sequence coverage at given *m/z* values. **c)** The combination of these improvements with AI-NETD at the peptide level translates to benefits at the protein level as well. AI-NETD characterized 1,106 proteins to NETD's 674. Beyond this, for proteins detected with both methods, AI-NETD overwhelmingly enhanced protein sequence coverage, by more than 50% in some cases.

Importantly, these improvements afforded by AI-NETD translate not only to higher quality MS/MS spectra with more robust fragmentation but also to remarkable improvements in peptide and protein identifications. Compared with the 3,805 peptides reported above for NETD analysis using the MDC, AI-NETD identified 7,601 unique peptides, essentially doubling the number of peptide identifications achievable in the same amount of analysis time. **Figure 3.4** summarizes the benefits of AI-NETD. The distribution of precursors across *m/z* space was similar for both NETD and AI-NETD runs (**Figure 3.4a**), but the distribution of peptides successfully sequenced is noticeably more extensive with AI-NETD (**Figure 3.4b**). Here, the peptide sequence coverage achieved for each doubly and triply deprotonated peptide identified with NETD or AI-NETD is plotted as a function of its *m/z* value. Successful identification of doubly deprotonated peptides with NETD dropped abruptly beyond 800 Th, and the sequence coverage achieved for these peptides decreased with higher *m/z* values. AI-NETD, on the other hand, successfully identified peptides across the entire *m/z* range and provided higher peptide sequence coverage, even maintaining 100% sequence coverage for doubly deprotonated peptides up to 1,000 Th. NETD did perform more favorably for triply deprotonated precursors than for doubly deprotonated ones, but AI-NETD remained distinctly superior for this population of peptides as well.

Beyond the substantial improvements in fragmentation at the peptide level, AI-NETD also performed advantageously at the protein level. In triplicate single-shot 90 min analyses, AI-NETD identified 1,106 proteins in yeast, making it the first technique to achieve identification of more than 1,000 proteins using the negative mode approach. Comparatively, NETD identified 674 yeast proteins. Panel **c** of **Figure 3.4** illustrates the gain in protein sequence coverage afforded by AI-NETD for proteins identified using both methods. Here, protein sequence coverage represents number of total amino acids explained from peptide identifications divided by the total number of amino acids. The gain in sequence coverage is defined as sequence coverage with NETD subtracted from sequence coverage with AI-NETD for a given protein; thus, a negative value means that higher sequence coverage was seen with NETD. AI-NETD provided a gain in sequence coverage for the overwhelming majority of proteins, many of which were acidic in nature (as indicated by the x-axis). In all, the enhancement in fragmentation that AI-NETD afforded for peptide anions unequivocally translates to more robust protein characterization, making it a premier tool for negative mode proteomics.

Selection of Robust High-pH Solvent System. Concurrent to our investigations into NETD and AI-NETD, we also explored how to increase precursor anion flux with different high-pH solvent systems. We previously employed ammonium formate solvents for high-pH separations, but charge state distributions favoring lowly charged precursors, degradation of silica-based packing materials, and low precursor flux were significant challenges with this system.[24,28] Although others have used similar systems with success, we also took note of several studies utilizing piperidine buffers for high-pH chromatography.[20,25,53,54] With this knowledge, we evaluated five different solvent systems using either piperidine or ammonium formate buffers over triplicate 90-min nLC-MS/MS analysis of

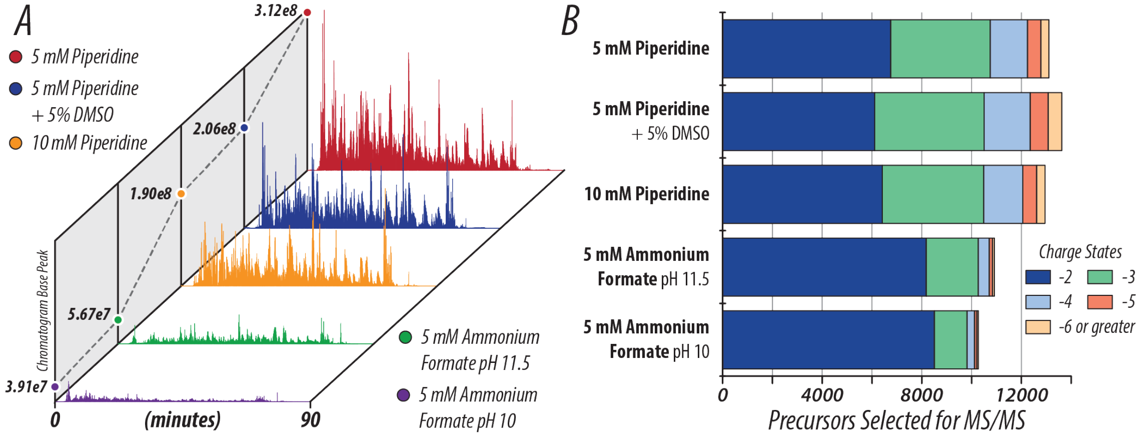
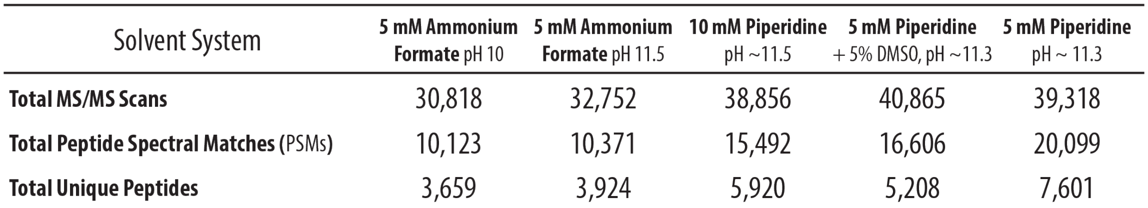


Figure 3.5: A comparison of high-pH solvent systems used. a) Base peak chromatograms show that choice of buffer additives, rather than pH alone, generate increased ionization with negative ESI. Piperidine solvents generate nearly an order of magnitude higher precursor ion signal than ammonium formate solvents. **b)** Beyond increasing the number of precursors selected for MS/MS, piperidine solvents shift the charge state distributions of precursor anions to be more highly charged (more negative) than ammonium formate solvents. Notably, DMSO in the 5 mM piperidine solvent system expanded the distribution of charge states to more highly charged precursors, rather than collapsing it to lower charges as reported in positive electrospray.

yeast peptides with AI-NETD fragmentation. A 10 mM piperidine system and 5 mM ammonium formate, pH 10, were first compared. We observed increased peptide anion signal with piperidine solvents, so we prepared ammonium formate buffers that were basified to pH 11.5 with ammonium hydroxide to match the pH of the 10 mM piperidine system. In the same vein, we also prepared 5 mM piperidine solvents, reducing the pH of buffer A slightly to ~11.3. Additionally, leveraging the recent descriptions of boosts provided by addition of DMSO to buffers for positive mode analyses, we assessed a 5 mM piperidine solvent system that contained 5% DMSO in buffer A.

Panel **a** of **Figure 3.5** displays base peak chromatograms from an experiment with each of the five solvent systems, highlighting the increases in signal, i.e. anion flux, observed with the piperidine buffers. The 5- to 10-fold improvements in base peak signal with piperidine solvents are reflected in the boosts observed in the number of tandem mass spectra, peptide spectral matches, and unique peptide identifications summarized in **Table I** (**Figure 3.6**). Panel **b** of **Figure 3.5** shows that, beyond increasing the number of MS/MS scans acquired, piperidine solvents shift the charge state distributions precursors to more highly charged (more negative) species, a phenomenon also reported for peptide standards and simple peptide mixtures.[20,53]. Although increasing the pH of ammonium formate solvents to 11.5 did show a small expansion in precursor charge state distributions, this difference was minimal compared with the pH 10 ammonium formate solvents and failed to match that observed with piperidine. Interestingly, the 5 mM piperidine buffers with 5% DMSO provided neither the highest base peak signal nor a gain in peptide identifications, which



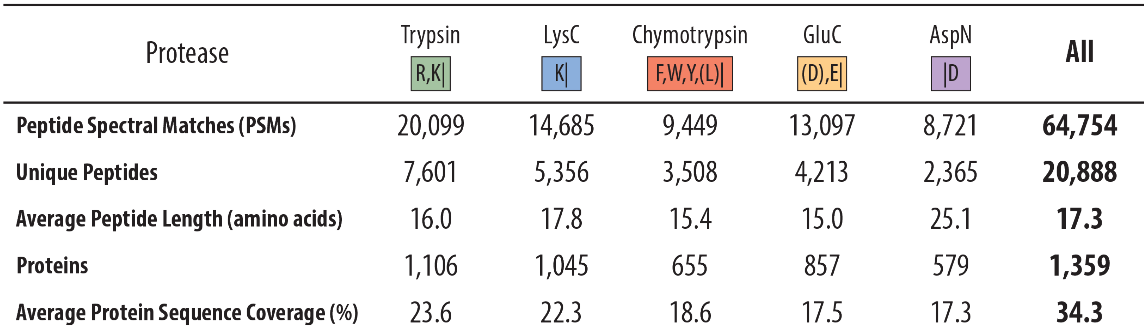
Solvent System	**5 mM Ammonium Formate** pH 10	**5 mM Ammonium Formate** pH 11.5	**10 mM Piperidine** pH ~11.5	**5 mM Piperidine** + 5% DMSO, pH ~11.3	**5 mM Piperidine** pH ~ 11.3
Total MS/MS Scans	30,818	32,752	38,856	40,865	39,318
Total Peptide Spectral Matches (PSMs**)**	10,123	10,371	15,492	16,606	20,099
Total Unique Peptides	3,659	3,924	5,920	5,208	7,601

Figure 3.6: Table I. Identifications with various solvent systems. Summary of AI-NETD experiments with five different high-pH solvent systems.

are both advantages reported in positive mode analyses.[45,55,56]. As shown in **Table I** (**Figure 3.6**) and **Figure 3.5b**, DMSO in the 5 mM piperidine solvent did permit the acquisition of the most MS/MS scans and displayed the widest distribution of precursor charge states; however, peptide identifications were lower with the DMSO additive than both the other two piperidine solvents. Regardless, high-pH solvent systems that utilize piperidine performed consistently better than ammonium formate solvent systems, maximizing the number of unique peptides identified with AI-NETD.

Multiple Proteases for Single-Shot Negative Mode Proteomics. Even with the robust analyses afforded by the combination of AI-NETD with prudently chosen piperidine high-pH solvents, use of just one protease limits the portion of the proteome that is accessible in a given experiment. The value of multiple proteases for canonical positive mode shotgun proteomics has been shown by us and others;[50,57–59] however, negative mode studies to date have largely failed to capitalize on the advantages offered by use of multiple enzymes for protein digestion. Instead, most studies rely mainly on trypsin for enzymatic digestion, although GluC has also been used. In this study, we extended our success with AI-NETD for tryptic peptides to single-shot analyses of peptides derived from four other proteases (LysC, GluC, chymotrypsin, and AspN) with the goals of exploring the amenability of different proteases to negative mode experiments and increasing the proteomic depth that can be achieved with negative mode analyses.

Table II (**Figure 3.7**) summarizes the results from triplicate 90-min AI-NETD nLC-



Protease	Trypsin R,K|	LysC K|	Chymotrypsin F,W,Y,(L)|	GluC (D),E|	AspN |D	**All**
Peptide Spectral Matches (PSMs)	20,099	14,685	9,449	13,097	8,721	**64,754**
Unique Peptides	7,601	5,356	3,508	4,213	2,365	**20,888**
Average Peptide Length (amino acids)	16.0	17.8	15.4	15.0	25.1	**17.3**
Proteins	1,106	1,045	655	857	579	**1,359**
Average Protein Sequence Coverage (%)	23.6	22.3	18.6	17.5	17.3	**34.3**

Figure 3.7: Table II. Identifications with various proteases. AI-NETD performance for peptides generated from five different proteases in single-shot experiments. Proteolytic specificity for each enzyme is shown.

MS/MS experiments that analyzed five different complex yeast peptide mixtures, each from one of the five proteases investigated. Immediately evident is the superior performance of trypsin and LysC, both of which enabled the identification of more than 1,000 proteins. Similar performance of these two enzymes is unsurprising considering their related proteolytic specificity C-terminal to lysine (trypsin and LysC) and arginine (trypsin) residues. GluC, which cleaves C-terminal to glutamic acid (and at slower rates, aspartic acid[60]), enabled the identification of more than 4,200 unique peptides, a greater than 7-fold increase over our previous NETD results with the enzyme.[24]. These peptides mapped back to 857 proteins, which also outmatches the best results achieved to date for any protease in negative mode approaches.[25] Chymotrypsin and AspN performed considerably well, too, illustrating the flexibility AI-NETD can offer for peptide anion characterization. Toward our goal of increasing proteomic depth, we batched the results from these five proteases together, providing nearly 21,000 unique peptide identifications. This combination of results provided valuable depth at the protein level, bolstering the number of proteins identified by ~23% and providing a jump in average protein sequence coverage from 23.6% to 34.4% over analysis with trypsin alone.

Beyond the beneficial information derived from the multiple protease approach for proteome characterization, we were also curious how AI-NETD performed for fragmenting families of peptides that were chemically distinct due to their proteolytic origins. We constructed fragment maps (*vide supra*) for AI-NETD fragmentation of peptides from all five enzymes and used results from trypsin as a point of reference. The trends in fragmentation between peptides from trypsin, LysC, chymotrypsin, and AspN were similar (data not shown); the fragmentation for GluC peptides, however, seemed to be more extensive than that seen with trypsin (**Figure 3.8a**). Here, we looked at peptides of relatively moderate and long length (12 and 24 residues, respectively), considering two precursor charge states for each. All peptides of the given length and charge are represented for peptides from both enzymes. This includes 443, 98, 94, and 65 tryptic peptides and 323, 146, 43, and 34 GluC peptides for length 12, z = -2; length 12, z = -3; length 24, z = -3; and length 24, z = -4, respectively. AI-NETD appears to produce fragment ions in greater number and intensity for peptides derived from GluC compared with trypsin, which is especially notable for z = -3 peptides that are 12 amino acids (AA) long. Juxtaposed to tryptic peptides with basic C termini, GluC peptides have a C-terminal acidic residue (D/E), meaning at least one negative charge is fixed at the C terminus. This could explain the improved electron-driven fragmentation seen for GluC peptides, especially for shorter peptides where proximity of backbone bonds to this C-terminal negative charge is greater. Broadening this idea beyond the subset of peptides investigated in panel **a**, we calculated peptide sequence coverage values for all identified tryptic and GluC peptides, z = -2 through -4. Density plots in **Figure 3.8b** show the frequency of peptides characterized with a given sequence coverage, serving as a surrogate for extent of fragmentation. AI-NETD fragmentation shows a similar trend for z = -2 peptides from trypsin and GluC, favoring higher sequencing coverage; however, AI-NETD with GluC peptides maintains high peptide sequence coverage for z = -3 and -4 peptides while the sequence coverage for tryptic peptides is much more evenly distributed, mirroring the more extensive fragmentation seen with the subset of GluC peptides in panel **a**.

Although it appears AI-NETD fragments GluC peptides more extensively, results with GluC do not rival those achieved with trypsin. Intrigued by this, we performed an *in silico* digest of the yeast proteome for both proteases, allowing up to two missed cleavages with a minimum peptide length of six amino acids and a maximum length of 75 amino acids. The larger histogram in **Figure 3.8c** shows that the distribution of possible tryptic peptides favors shorter peptides (average length of 20.6 residues) while GluC produces fewer total peptides and does not favor short peptides as drastically (average length of 27.8 residues). The inset in **Figure 3.8c** displays a histogram of peptide lengths for peptides actually identified in AI-NETD analyses, providing the complementary experimental measurement to the theoretical data derived from the in silico digest. Interestingly, despite the notable difference in distributions of peptide length from the *in silico* digestion, the experimental distributions of peptide length are very similar, with average peptide lengths of 16 and 15 residues for trypsin and GluC, respectively. We surmise that lower peptide identifications with GluC compared with trypsin is a function of the population of peptides generated by GluC rather than the fragmentation achieved for these peptides. Although GluC peptides may fragment better, a smaller portion of peptides derived from GluC are in the ideal range of peptide length for AI-NETD, which appears to be 10–25 residues. This translates to a higher percentage of fragmentation events (i.e. MS/MS scans) occurring on larger, more difficult-to-sequence peptides for GluC. In fact, **Figure 3.8d** supports this, showing

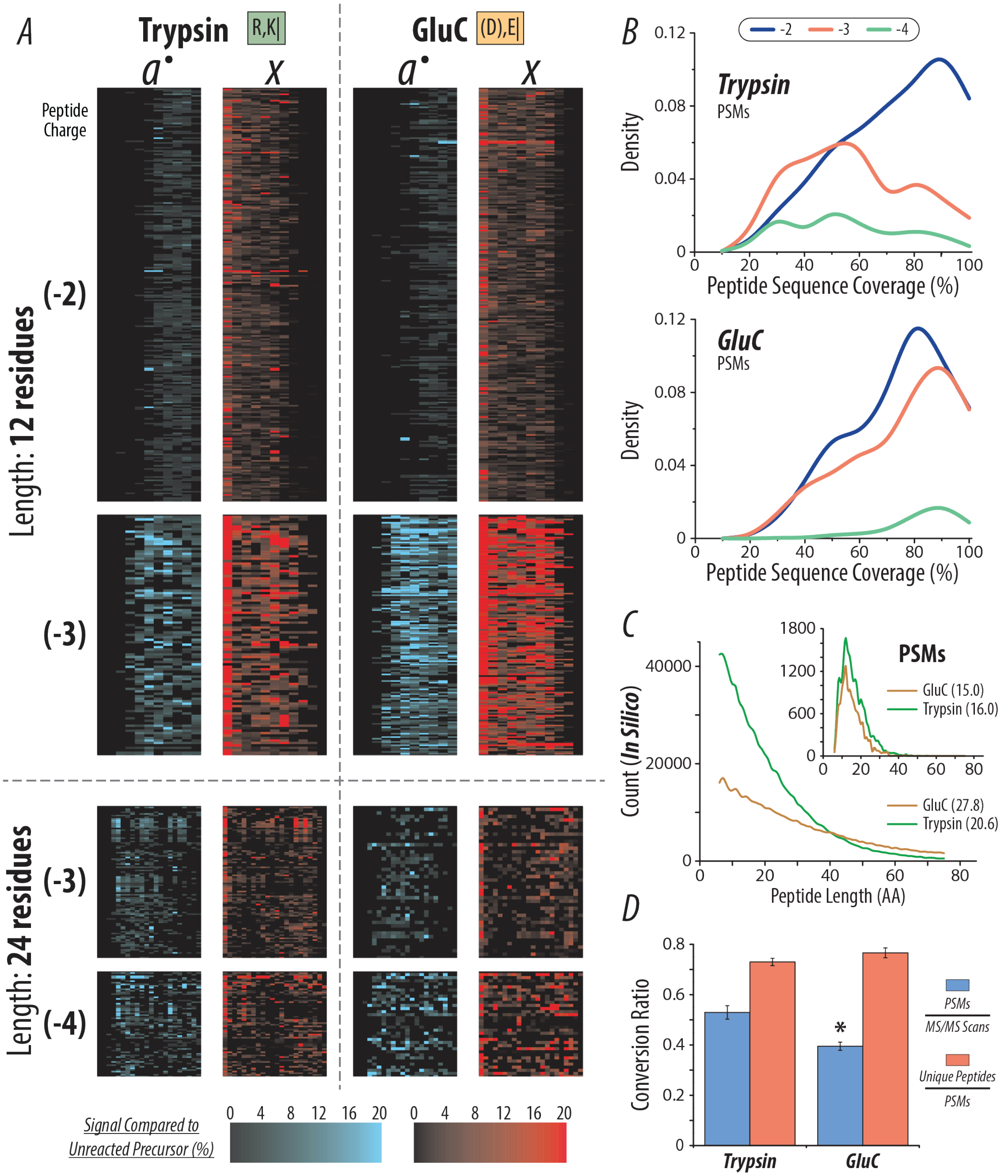


Figure 3.8: Comparison of single-shot AI-NETD for peptides produced by either trypsin or GluC. a) AI-NETD fragment map for peptides (12 and 24 amino acids in length) derived from both trypsin and GluC digestions. The numbers in parenthesis to the left indicate peptide charge. **b)** Density plots for peptide sequence coverage for PSMs from trypsin and GluC.

Figure 3.8: **c)** The larger histogram shows the distribution of peptide lengths from an *in silico* digest for trypsin and GluC. The inset displays the distribution of lengths of peptides identified in the trypsin and GluC experiments. The numbers in parenthesis in the respective legends show the average peptide length for each protease. **d)** The MS/MS success rate (blue) for GluC peptides is significantly lower than tryptic peptides ($p < .01$, indicated by *), while the ratio of unique peptides to total PSMs detected (red) is not statistically different ($p < .05$).

that the ratio of unique peptides to total PSMs is the same for the two enzymes, but the MS/MS success rate is significantly lower ($p < .01$) for GluC. Additionally, our entire workflow was originally optimized for tryptic peptides, including sample preparation and chromatographic conditions, which could also contribute to this discrepancy. These observations may explain why, even with superior fragmentation and higher peptide sequence coverage for the peptides we do sequence, fewer peptides are ultimately identified using GluC as a protease rather than trypsin. Even so, this does not make AI-NETD analyses of GluC peptides any less valuable; rather, these results suggest that combinations of proteases can be used not only to enhance proteomic depth but also to access the advantages AI-NETD can provide for peptides with distinct chemical properties.

Deep Sequencing in the Negative Mode with Offline Low-pH Fractionation and Multiple Proteases. Encouraged by these results, we sought to improve upon the proteomic depth we could achieve with purely negative mode techniques. All previous large-scale peptide anion analyses have used online one-dimensional high-pH chromatography for single-shot experiments, similar to the approaches we have described thus far. Common practice in traditional proteomic experiments is to increase the achievable proteomic depth by decreasing sample complexity via online and/or offline fractionation;[61,62] we used the same logic to fractionate peptide mixtures from trypsin, LysC, GluC, chymotrypsin, and AspN digestions prior to nLC-MS/MS analysis with AI-NETD. In positive mode approaches, offline high-pH reversed-phase fractionation provides an orthogonal mode of separation to the online acidic reversed-phase chromatographic conditions used for nLC-MS/MS analysis.[63] It holds that the two should remain orthogonal even if the order in which they are performed is inverted. Thus, we employed a simple low-pH reversed-phase fractionation system to separate a complex mixture of peptides into 50 fractions, which were then concatenated into 10 total fractions for subsequent negative mode nLC-MS/MS analysis. This fractionation was done for a digestion from each protease.

The offline fractionation approach extraordinarily improved peptide and protein identification for all five proteases with AI-NETD (**Figure 3.9**). Where single-shot AI-NETD experiments with tryptic peptides produced 7,601 unique peptide identifications and 1,106 proteins, analysis of the same mixture of tryptic peptides split equally into 10 fractions enabled the identification of 36,713 unique peptides and 3,467 proteins. **Figure 3.9a** shows the results achieved from analysis of 10 fractions for each of the five proteases. Here, the area of the circle represents the total number of proteins identified. The circles along the diagonal present the results for the proteases individually. This figure also displays the average percentage of protein sequence coverage observed with each enzyme with a color gradient. Further, panel **a** presents pairwise comparisons that show the combination of

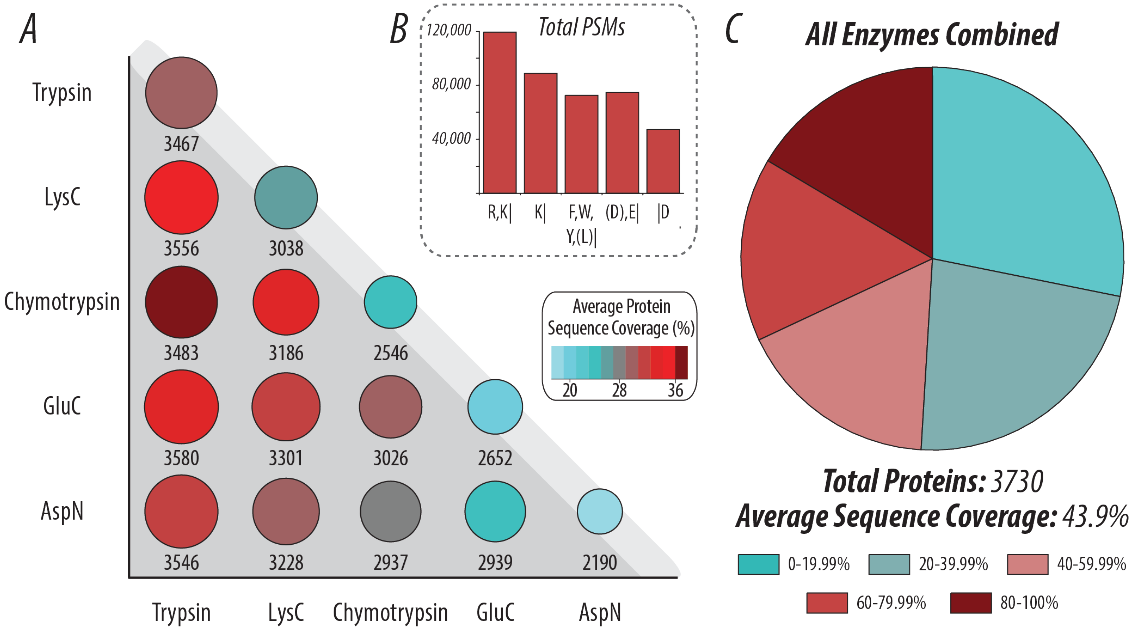


Figure 3.9: Proteome coverage in the negative mode with various proteases using AI-NETD and low-pH fractionation. a) Pairwise comparisons of the number of proteins (area of circle) and average protein sequence coverage (color) when using different proteases, illustrating the degree of orthogonality of each protease with the others. The circles along the top edge (light gray background) show each enzyme by itself. **b)** The total number of PSMs for each protease. **c)** When combining all PSMs from the five enzymes, AI-NETD characterizes over 80% of the yeast proteome (3,730 proteins). The pie chart here shows what proportions of these proteins had sequence coverage in the given range.

results from two proteases batched together. This analysis allows an evaluation of how combinations of different proteases affect the number and coverage of proteins detected, illustrating the degree of orthogonality of each protease with the others. Intriguingly, the combination of trypsin and GluC proteases for this large-scale, deep-sequencing approach provided the largest number of proteins identified in these pairwise comparisons with an average protein sequence coverage of 35.14%. The combination of trypsin and chymotrypsin, however, provided the greatest average sequence coverage (36.5%) even though it did not appreciably increase the number of proteins identified with trypsin alone. This analysis provides insight into what combinations of proteases, whether they cleave at basic, acidic, or hydrophobic residues, may be the most beneficial as negative mode approaches continue to advance. Panel **b** shows the number of total PSMs collected for each enzyme, providing some concept of the complexity of peptide mixtures produced from each. When integrating the identifications from all five enzymes into a batched analysis of proteins identified, AI-NETD facilitated the characterization of 3,730 proteins with an average sequence coverage of 43.9%. This represents nearly 83% of the expressed yeast proteome (estimated to be ~4,500 proteins[64]), demonstrating that comprehensive proteome analysis, which has been previously confined to positive mode analyses, can be achieved in the negative mode via peptide fragmentation with AI-NETD. **Figure 3.9c** shows the proportions of these 3,730 proteins with a given protein sequence coverage; while approximately half of the proteins had sequence coverage under 40%, an appreciable fraction of these protein identifications had excellent sequence coverage (80–100%).

Comparing Positive and Negative Mode Data. Although the data presented above stand alone as a demonstration of what can be considered realistic and achievable in negative mode proteomics, a natural and valuable extension lies in the comparison of these large-scale peptide anion analyses to large-scale positive mode experiments. First, we performed triplicate single-shot analyses of yeast tryptic peptides using HCD and ETD fragmentation for comparison of positive mode data to AI-NETD. Both HCD and ETD were performed in the MDC to provide the most direct comparison possible, thus keeping both reaction cell and mass analyzer (i.e. the Orbitrap) consistent with the negative mode experiments. **Figure 3.10a** displays the number of peptides identified with each method and shows the overlap in peptides between the three fragmentation types. Surprisingly, AI-NETD performed as well as, if not better, than ETD in the number of peptides identified (7,601 vs 7,414). Only one-third of the total peptides sequenced by ETD and NETD were identified in both methods, highlighting the complementarity of the positive and negative mode approaches for electron-driven dissociation techniques. Furthermore, although HCD produced more peptide identifications than both ETD and AI-NETD, the overlap in peptides was higher for the two positive mode techniques — 83% of ETD peptides were also identified with HCD while 68.5% of AI-NETD peptides were also seen in the HCD data. This provides clear evidence that negative mode analysis with AI-NETD is fully capable of affording greater orthogonality to positive mode collision-based peptide identification than offered by positive mode ETD, maximizing the number of peptides than can be identified in a sample (although we note that more vigorous and extensive comparisons are need to fully explore this issue).

To further our comparison to positive mode methods, we examined the overlap in peptides and proteins characterized in our deep-sequencing experiments to those identified in an extensive published proteomic analysis of yeast peptides that used offline fractionation and multiple proteases, two key components of our deep proteome sequencing with AI-NETD.[50] Additionally this study utilized decision tree logic to tailor fragmentation (either CID or ETD) to each peptide, maximizing the chances of an MS/MS scan being successfully mapped to sequence.[65] All of these components make this data set one of the most robust available characterizations of the yeast proteome, with the added benefit of the ability to compare multiple proteases in positive and negative modes. **Figure 3.10b** shows the distribution of percentage protein sequence coverages achieved for *n* proteins in the positive mode when using trypsin and various combinations of peptides from other proteases [*trypsin(+)* and *proteaseX(+)*, where proteaseX is LysC, GluC, AspN, or ArgC]. Also included in that plot is the combination of tryptic peptides from the positive mode data and the AI-NETD data set [*trypsin(+)* and *trypsin(-)*] (gray background). Compellingly, using one protease with both the positive and negative modes outperforms all other positive mode combinations of trypsin with a different protease. Not only does the *trypsin(+)* and *trypsin(-)* combination identify more proteins than any *trypsin(+)* and *proteaseX(+)* combination, but it also provides the greatest median (29.17%) and average (34.02%) percentage sequence coverage — indicating that negative mode analyses for tryptic peptides can provide more orthogonality to positive mode experiments than the use of different proteases. For the

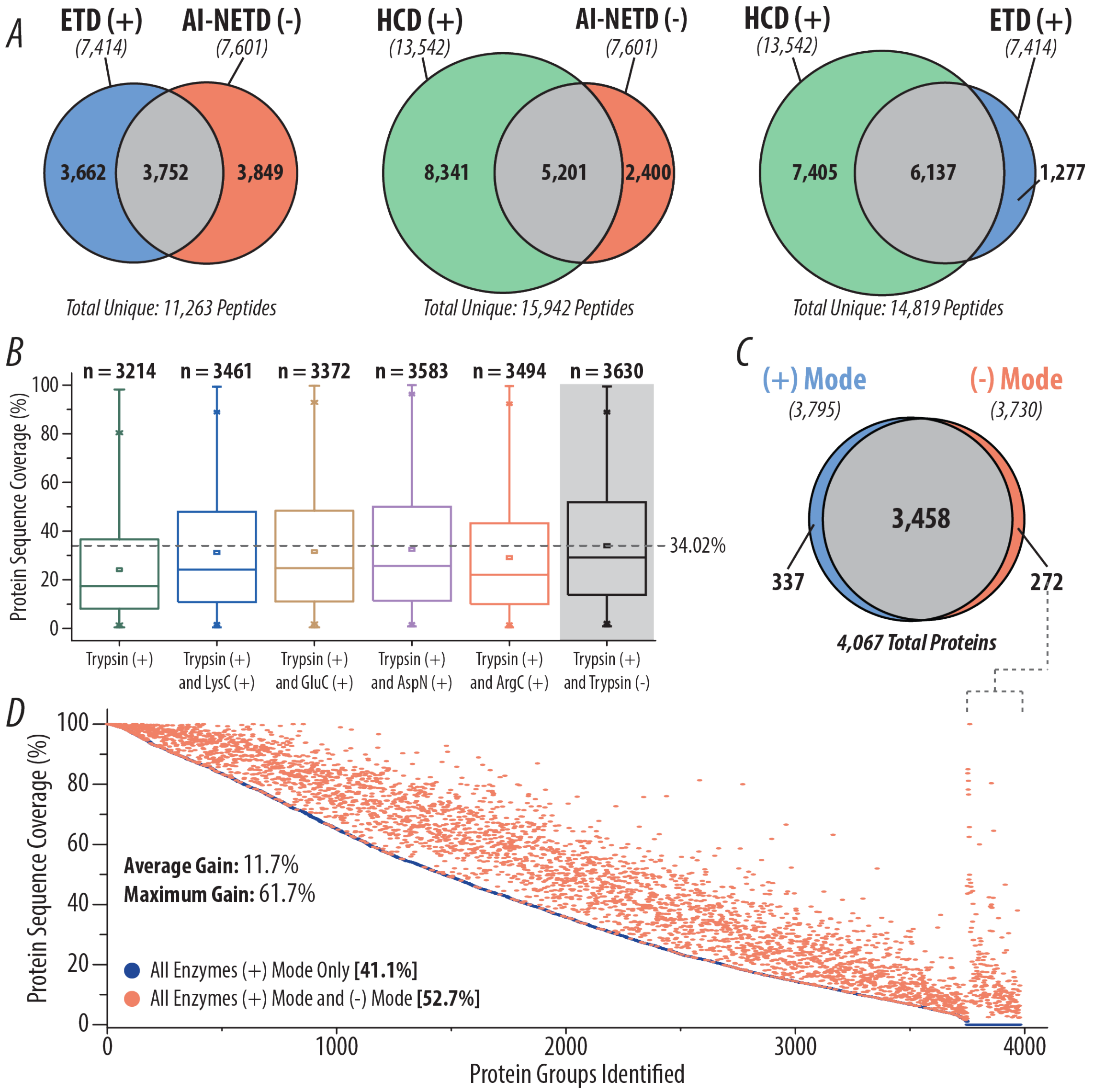


Figure 3.10: Comparison of AI-NETD to positive mode analyses. a) Overlap in yeast tryptic peptides identified in HCD, ETD, and AI-NETD single-shot experiments. The number of unique peptides identified with each fragmentation type are indicated in italics below the appropriate label. **b)** Distribution of protein sequence coverages achieved with combinations of trypsin with different proteases in positive mode analyses with CID and ETD fragmentation compared with positive mode data combined with negative mode data with AI-NETD (gray background) using only trypsin. The dotted line shows the highest average sequence coverage. **c)** Overlap in yeast proteins identified using positive mode (CID and ETD) and negative mode (AI-NETD) analyses. For both sets of proteins, peptides from five different proteases were batched together.

Figure 3.10: d) Proteins detected in the positive mode data from **(c)** are rank ordered by percentage sequence coverage and plotted in blue, with an average sequence coverage of 41.1%. The protein sequence coverage achieved for a given protein when combining the positive and negative mode data is plotted in red, highlighting the gain in sequence coverage afforded by the addition of negative mode analyses. To the far right, sequence coverages are shown for the 272 proteins that were not detected in positive mode analyses but were characterized with AI-NETD.

other three proteases in common between the two studies (LysC, GluC, and AspN), the addition of negative mode analyses with AI-NETD increases "positive mode only" protein sequence coverage with LysC from 23.72% to 30.97%, with GluC from 19.57% to 24.10%, and with AspN from 20.98% to 23.34%.

Finally, we examined the overlap of all proteins from all five proteases characterized in the positive mode data to all proteins from all five proteases in the AI-NETD data set, which is depicted in **Figure 3.10c**. The degree of overlap in proteins identified is noticeably large (~85% of all proteins identified were seen in both data sets), which is not wholly unsurprising as >90% of the expressed yeast proteome is represented. What this overlap fails to display, however, is the difference in the populations of peptides sequenced. Of the 106,861 unique peptides identified between the two experiments, only 19,697 of them (~18%) were detected in both data sets. To visualize how this impacts protein sequence coverage, **Figure 3.10d** displays all proteins identified in the positive mode data, rank ordered by percentage protein sequence coverage (blue circles). For each protein there is a corresponding red circle that shows the sequence coverage achieved when including the negative mode data with the positive mode data set; thus, the magnitude of the difference along the y-axis between a red and blue circle for any given protein shows the gain in percentage protein sequence coverage provided by the inclusion of the AI-NETD data. The average percentage sequence coverage for the positive mode data is 41.1%, which is boosted by nearly 12 percentage points to 52.7% by the addition of the negative mode analyses. This combination of positive and negative mode data puts the average sequence coverage above 50%, a mark that neither data set reached on their own. The highest gain in percentage sequence coverage was 61.7%, starting at 18.29% sequence coverage with positive mode methods alone and going to 80% sequence coverage with combined positive and negative mode analyses. This 174 residue protein, calcineurin subunit B - which is a calcium-dependent, calmodulin stimulated protein phosphatase that confers calcium sensitivity - is a notably acidic protein (pI ~ 4.36) with 36 negatively charged residues (Asp and Glu) and only 21 basic residues (Lys and Arg). Many proteins that showed comparable gains in sequence coverage shared similar characteristics in isoelectric point and residue frequencies, illustrating the ability of AI-NETD to add coverage to the acidic portions of the proteome that may be missed by positive mode methods. The far right of the graph in **Figure 3.10d** shows sequence coverage for 272 proteins that were only characterized by the inclusion of the AI-NETD data. Interestingly, many of these proteins are membrane proteins, including mitochondrial membrane, Golgi apparatus membrane, and other transport proteins; this characterization of membrane proteins via negative mode methods may point to an interesting avenue to explore in future investigations. In all, these comparisons demonstrate that large-scale negative mode analyses with AI-NETD are a valuable complement to positive mode methods, perhaps even offering more orthogonality than ETD or the use of multiple proteases.

Discussion

Negative mode approaches for the characterization of peptide anions offer a valuable dimension to proteomic analyses,[21,24,25,66] especially as biologically relevant posttranslational modifications and other analytes that pose challenges to canonical positive mode techniques continue to emerge.[7,53,67–72] Platforms for shotgun analysis of complex mixtures of peptide anions have been introduced, most notably using UVPD and NETD; however, these approaches have yet to provide considerable proteomic depth (fewer than ~800 total proteins identified in a given experiment), restricting the degree to which peptide anion characterization can benefit the proteomic community.

With a new implementation of AI-NETD, we have introduced a robust negative mode platform for the characterization of over 1,100 proteins in a eukaryotic system (yeast) using 90-min single-shot analyses. Moreover, we demonstrate that AI-NETD is compatible with a diverse array of commonly used proteases, increasing proteomic depth (>1,350 proteins) and degree of protein characterization, i.e. sequence coverage, achievable in single-shot experiments. The ability to utilize several proteases makes AI-NETD an especially viable technique for analysis of PTMs, where access to chemically distinct peptides or a combinatorial pattern of sequences may greatly increase confidence in identification and localization. This approach may also prove beneficial for more extensive characterization of proteins that are challenging to current techniques, e.g. proteins with dominantly acidic sequences or those with both highly hydrophobic and hydrophilic regions.

Integral to the improvements observed with our AI-NETD work was the selection of a high-pH solvent system that increased anion flux and provided optimal distribution of precursor charge states. Previous studies have shown the benefits of piperidine buffers for anion analysis, but this work is the first to demonstrate the consequential effects it can have on large-scale shotgun proteomic experiments. Based on the benefits we and others observed with DMSO[45,55,56] we expected the addition of DMSO to our piperidine solvents would provide an additional boon to our negative mode work. This was not the case, however, as the piperidine solvents with DMSO included provided the fewest numbers of unique peptide identifications among the three piperidine systems tested. It has been suggested that the benefits of DMSO in positive mode comes from the charge state coalescence for peptide cations, making signal more concentrated for a fewer number of charge states and reducing redundant sampling of the same peptide with different *m/z* values. In the negative mode, DMSO appears to have the opposite effect, expanding the charge state distribution of peptide precursors sampled for MS/MS events (**Figure 3.5b**). The base peak intensity of the chromatogram from 5 mM piperidine solvent with 5% DMSO was slightly lower than 5 mM piperidine alone (**Figure 3.5a**), also juxtaposing the trends observed in positive mode. Noticeable, though, is that the greatest number of MS/MS scans was taken with the piperidine/DMSO buffers than with any other system. These results suggest that DMSO is further spreading signal among many charge states for peptide precursor anions, rather than collapsing it as in positive mode. Thus, it is possible that DMSO does not merely condense a signal into lower charge states, as was suggested based on positive mode studies but that it makes the precursor charge “more negative” or “less positive” than before. Surely validation of this hypothesis requires further exploration beyond the scope of this work, but this observation highlights an unexpected outcome that emphasizes the value negative mode proteomics can have as a complementary tool to positive mode techniques.

Additionally, challenges with column longevity and hindered performance of LC pumps have been reported for high-pH reversed-phase separations for negative mode proteomics.[7,24,73] We previously struggled with both precolumn and analytical column degradation due to the instability of silica in basic conditions. In this work, we eliminated the precolumns that used silica frits and employed a polymer-silica hybrid reversed-phase packing material that is stable at both acidic and basic conditions.[74] With these modifications, our columns often lasted a week or more of constant runs, showing impressive longevity compared with our previous column setup that lasted roughly a day. LC maintenance was required periodically, including changing of silica capillaries lines prior to the analytical column, but this still permitted straightforward and consistent data collection for our single-shot and deep-sequencing experiments.

In summary, we demonstrated that AI-NETD, in conjunction with robust high-pH separations, multiple proteases, and offline low-pH prefractionation, can be leveraged to characterize the large majority (>80%) of the yeast proteome, matching the average sequence coverage observed in similar positive mode experiments using multiple fragmentation types (~45%).[50] Moreover, AI-NETD analyses provided a substantial improvement in protein sequence coverage (an average of ~12% but as much as ~62%) over what could be achieved with positive mode methods alone, even those utilizing a multiple protease approach. That being said, we do not see negative mode proteomics replacing any of the wide array of positive mode approaches, much less competing with the acquisition rate and proteomic depth they can achieve;[45,75,76] instead, we foresee negative mode proteomics continuing to advance in sensitivity and speed — especially as robust fragmentation techniques like AI-NETD are implemented on the newest generations of instruments — serving as a powerful complement to traditional positive mode methods. Most importantly, this work demonstrates that large-scale analysis of peptide anions can be used to characterize nearly an entire proteome, enabling a much more thorough investigation of previously intractable portions of the proteome and critical PTMs that will greatly benefit from analysis in the negative mode.

References

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

[2] 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. 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. 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, no. 7502, pp. 575–581, 2014.

[3] M. Wilhelm, J. Schlegl, H. Hahne, A. Moghaddas 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, no. 7502, pp. 582–587, 2014.

[4] A. L. Richards, A. E. Merrill, and J. J. Coon, "Proteome sequencing goes deep.," *Current opinion in chemical biology*, vol. 24, pp. 11–7, mar 2015.

[5] K. Engholm-Keller and M. R. Larsen, "Technologies and challenges in large-scale phosphoproteomics.," *Proteomics*, vol. 13, pp. 910–31, mar 2013.

[6] D. S. Dalpathado and H. Desaire, "Glycopeptide analysis by mass spectrometry," *Analyst*, vol. 133, no. 6, pp. 731–738, 2008.

[7] M. R. Robinson, K. L. Moore, and J. S. Brodbelt, "Direct identification of tyrosine sulfation by using ultraviolet photodissociation mass spectrometry.," *Journal of the American Society for Mass Spectrometry*, vol. 25, pp. 1461–71, aug 2014.

[8] N. P. Ewing and C. J. Cassady, "Dissociation of multiply charged negative ions for hirudin (54-65), fibrinopeptide B, and insulin A (oxidized)," *J Am Soc Mass Spectrom*, vol. 12, no. 1, pp. 105–116, 2001.

[9] J. Kiraga, P. Mackiewicz, D. Mackiewicz, M. Kowalczuk, P. Biecek, N. Polak, K. Smolarczyk, M. R. Dudek, and S. Cebrat, "The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms," *BMC Genomics*, vol. 8, p. 163, 2007.

[10] N. Leymarie and J. Zaia, "Effective use of mass spectrometry for glycan and glycopeptide structural analysis.," *Analytical chemistry*, vol. 84, pp. 3040–8, apr 2012.

[11] M. Yamashita and J. B. Fenn, "Negative ion production with the electrospray ion source," *The Journal of Physical Chemistry*, vol. 88, pp. 4671–4675, sep 1984.

[12] R. F. Straub, R. D. Voyksner, and D. Voyksner, "Negative ion formation in electrospray mass spectrometry," *J Am Soc Mass Spectrom*, vol. 4, no. 7, pp. 578–587, 1993.

[13] J. H. Bowie, C. S. Brinkworth, and S. Dua, "Collision-induced fragmentations of the (M-H)- parent anions of underivatized peptides: an aid to structure determination and some unusual negative ion cleavages.," *Mass spectrometry reviews*, vol. 21, pp. 87–107, jan 2002.

[14] C. S. Brinkworth, S. Dua, A. M. McAnoy, and J. H. Bowie, "Negative ion fragmentations of deprotonated peptides: backbone cleavages directed through both Asp and Glu.," *Rapid communications in mass spectrometry : RCM*, vol. 15, pp. 1965–73, jan 2001.

[15] S. T. Steinborner and J. H. Bowie, "A comparison of the positive- and negative-ion mass spectra of bio-active peptides from the dorsal secretion of the Australian red tree frog, Litoria rubella.," *Rapid communications in mass spectrometry : RCM*, vol. 10, pp. 1243–7, jan 1996.

[16] S. T. Steinborner and J. H. Bowie, "The Negative Ion Mass Spectra of [MH]- Ions Derived From Caeridin and Dynastin Peptides. Internal Backbone Cleavages Directed Through Asp and Asn Residues," *Rapid Communications in Mass Spectrometry*, pp. 253–258, feb 1997.

[17] B. A. Budnik, K. F. Haselmann, and R. A. Zubarev, "Electron detachment dissociation of peptide di-anions: an electron–hole recombination phenomenon," *Chemical Physics Letters*, vol. 342, pp. 299–302, jul 2001.

[18] J. J. Coon, J. Shabanowitz, D. F. Hunt, and J. E. P. Syka, "Electron transfer dissociation of peptide anions.," *Journal of the American Society for Mass Spectrometry*, vol. 16, pp. 880–2, jun 2005.

[19] A. Kalli, G. Grigorean, and K. Håkansson, "Electron induced dissociation of singly deprotonated peptides.," *Journal of the American Society for Mass Spectrometry*, vol. 22, pp. 2209–21, dec 2011.

[20] J. A. Madsen, T. S. Kaoud, K. N. Dalby, and J. S. Brodbelt, "193-nm photodissociation of singly and multiply charged peptide anions for acidic proteome characterization.," *Proteomics*, vol. 11, pp. 1329–34, apr 2011.

[21] H. J. Yoo, N. Wang, S. Zhuang, H. Song, and K. Håkansson, "Negative-ion electron capture dissociation: radical-driven fragmentation of charge-increased gaseous peptide anions," *Journal of the American Chemical Society*, vol. 133, pp. 16790–3, oct 2011.

[22] R. Antoine, L. Joly, T. Tabarin, M. Broyer, P. Dugourd, and J. Lemoine, "Photo-induced formation of radical anion peptides. Electron photodetachment dissociation experiments," *Rapid Commun Mass Spectrom*, vol. 21, no. 2, pp. 265–268, 2007.

[23] F. Kjeldsen, O. A. Silivra, I. A. Ivonin, K. F. Haselmann, M. Gorshkov, and R. A. Zubarev, "C alpha-C backbone fragmentation dominates in electron detachment dissociation of gas-phase polypeptide polyanions," *Chemistry*, vol. 11, no. 6, pp. 1803–1812, 2005.

[24] G. C. McAlister, J. D. Russell, N. G. Rumachik, A. S. Hebert, J. E. P. Syka, L. Y. Geer, M. S. Westphall, D. J. Pagliarini, and J. J. Coon, "Analysis of the acidic proteome with negative electron-transfer dissociation mass spectrometry.," *Analytical chemistry*, vol. 84, pp. 2875–82, mar 2012.

[25] J. A. Madsen, H. Xu, M. R. Robinson, A. P. Horton, J. B. Shaw, D. K. Giles, T. S. Kaoud, K. N. Dalby, M. S. Trent, and J. S. Brodbelt, "High-throughput database search and large-scale negative polarity liquid chromatography-tandem mass spectrometry with ultraviolet photodissociation for complex proteomic samples," *Mol Cell Proteomics*, vol. 12, no. 9, pp. 2604–2614, 2013.

[26] J. E. P. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt, "Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 9528–33, jun 2004.

[27] M. Huzarska, I. Ugalde, D. A. Kaplan, R. Hartmer, M. L. Easterling, and N. C. Polfer, "Negative electron transfer dissociation of deprotonated phosphopeptide anions: choice of radical cation reagent and competition between electron and proton transfer.," *Analytical chemistry*, vol. 82, pp. 2873–8, apr 2010.

[28] N. G. Rumachik, G. C. McAlister, J. D. Russell, D. J. Bailey, C. D. Wenger, and J. J. Coon, "Characterizing peptide neutral losses induced by negative electron-transfer dissociation (NETD).," *Journal of the American Society for Mass Spectrometry*, vol. 23, pp. 718–27, apr 2012.

[29] D. M. Good, M. Wirtala, G. C. McAlister, and J. J. Coon, "Performance characteristics of electron transfer dissociation mass spectrometry.," *Molecular & cellular proteomics : MCP*, vol. 6, pp. 1942–51, nov 2007.

[30] J. B. Shaw, D. A. Kaplan, and J. S. Brodbelt, "Activated ion negative electron transfer dissociation of multiply charged peptide anions," *Analytical chemistry*, vol. 85, pp. 4721–8, may 2013.

[31] J. B. Shaw, J. A. Madsen, H. Xu, and J. S. Brodbelt, "Systematic comparison of ultraviolet photodissociation and electron transfer dissociation for peptide anion characterization.," *Journal of the American Society for Mass Spectrometry*, vol. 23, pp. 1707–15, oct 2012.

[32] Y. Xia, H. Han, and S. A. McLuckey, "Activation of intact electron-transfer products of polypeptides and proteins in cation transmission mode ion/ion reactions.," *Analytical chemistry*, vol. 80, pp. 1111–7, feb 2008.

[33] C. K. Frese, A. F. M. Altelaar, H. van den Toorn, D. Nolting, J. Griep-Raming, A. J. R. Heck, and S. Mohammed, "Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry.," *Analytical chemistry*, vol. 84, pp. 9668–73, nov 2012.

[34] D. L. Swaney, G. C. McAlister, M. Wirtala, J. C. Schwartz, J. E. P. Syka, and J. J. Coon, "Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors.," *Analytical chemistry*, vol. 79, pp. 477–85, jan 2007.

[35] J. J. Coon, "Collisions or electrons? Protein sequence analysis in the 21st century.," *Analytical chemistry*, vol. 81, pp. 3208–15, may 2009.

[36] S. J. Pitteri, P. A. Chrisman, and S. A. McLuckey, "Electron-transfer ion/ion reactions of doubly protonated peptides: effect of elevated bath gas temperature.," *Analytical chemistry*, vol. 77, pp. 5662–9, sep 2005.

[37] A. R. Ledvina, G. C. McAlister, M. W. Gardner, S. I. Smith, J. a. Madsen, J. C. Schwartz, G. C. Stafford, J. E. P. Syka, J. S. Brodbelt, and J. J. Coon, "Infrared photoactivation reduces peptide folding and hydrogen-atom migration following ETD tandem mass spectrometry.," *Angewandte Chemie (International ed. in English)*, vol. 48, pp. 8526–8, jan 2009.

[38] A. R. Ledvina, N. a. Beauchene, G. C. McAlister, J. E. P. Syka, J. C. Schwartz, J. Griep-Raming, M. S. Westphall, and J. J. Coon, "Activated-ion electron transfer dissociation improves the ability of electron transfer dissociation to identify peptides in a complex mixture.," *Analytical chemistry*, vol. 82, pp. 10068–74, dec 2010.

[39] A. R. Ledvina, C. M. Rose, G. C. McAlister, J. E. P. Syka, M. S. Westphall, J. Griep-Raming, J. C. Schwartz, and J. J. Coon, "Activated ion ETD performed in a modified collision cell on a hybrid QLT-Oribtrap mass spectrometer.," *Journal of the American Society for Mass Spectrometry*, vol. 24, pp. 1623–33, nov 2013.

[40] C. M. Rose, J. D. Russell, A. R. Ledvina, G. C. McAlister, M. S. Westphall, J. Griep-Raming, J. C. Schwartz, J. J. Coon, and J. E. P. Syka, "Multipurpose dissociation cell for enhanced ETD of intact protein species.," *Journal of the American Society for Mass Spectrometry*, vol. 24, pp. 816–27, jun 2013.

[41] N. M. Riley, M. S. Westphall, and J. J. Coon, "Activated Ion Electron Transfer Dissociation for Improved Fragmentation of Intact Proteins.," *Analytical chemistry*, vol. 87, pp. 7109–7116, jun 2015.

[42] Y. Zhao, N. M. Riley, L. Sun, A. S. Hebert, X. Yan, M. S. Westphall, M. J. P. Rush, G. Zhu, M. M. Champion, F. M. Medie, P. A. D. Champion, J. J. Coon, and N. J. Dovichi, "Coupling Capillary Zone Electrophoresis with Electron Transfer Dissociation and Activated Ion Electron Transfer Dissociation for Top-Down Proteomics.," *Analytical chemistry*, vol. 87, pp. 5422–5429, may 2015.

[43] A. Michalski, E. Damoc, O. Lange, E. Denisov, D. Nolting, M. Müller, R. Viner, J. Schwartz, P. Remes, M. Belford, J.-J. Dunyach, J. Cox, S. Horning, M. Mann, and A. Makarov, "Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes.," *Molecular & cellular proteomics : MCP*, vol. 11, p. O111.013698, mar 2012.

[44] G. C. McAlister, W. T. Berggren, J. Griep-Raming, S. Horning, A. Makarov, D. Phanstiel, G. Stafford, D. L. Swaney, J. E. P. Syka, V. Zabrouskov, and J. J. Coon, "A proteomics grade electron transfer dissociation-enabled hybrid linear ion trap-orbitrap mass spectrometer.," *Journal of proteome research*, vol. 7, pp. 3127–36, aug 2008.

[45] 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.," *Molecular & cellular proteomics : MCP*, vol. 13, pp. 339–47, jan 2014.

[46] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant, "Open mass spectrometry search algorithm.," *Journal of proteome research*, vol. 3, pp. 958–64, jan 2004.

[47] D. M. Good, C. D. Wenger, G. C. McAlister, D. L. Bai, D. F. Hunt, and J. J. Coon, "Post-acquisition ETD spectral processing for increased peptide identifications.," *Journal of the American Society for Mass Spectrometry*, vol. 20, pp. 1435–40, aug 2009.

[48] D. M. Good, C. D. Wenger, and J. J. Coon, "The effect of interfering ions on search algorithm performance for electron-transfer dissociation data," *Proteomics*, vol. 10, no. 1, pp. 164–167, 2010.

[49] C. D. Wenger, D. H. Phanstiel, M. V. Lee, D. J. Bailey, and J. J. Coon, "COMPASS: a suite of pre- and post-search proteomics software tools for OMSSA.," *Proteomics*, vol. 11, pp. 1064–74, mar 2011.

[50] D. L. Swaney, C. D. Wenger, and J. J. Coon, "Value of using multiple proteases for large-scale mass spectrometry-based proteomics.," *Journal of proteome research*, vol. 9, pp. 1323–9, mar 2010.

[51] J. J. Coon, J. E. Syka, J. C. Schwartz, J. Shabanowitz, and D. F. Hunt, "Anion dependence in the partitioning between proton and electron transfer in ion/ion reactions," *International Journal of Mass Spectrometry*, vol. 236, pp. 33–42, aug 2004.

[52] C. M. Rose, M. J. P. Rush, N. M. Riley, A. E. Merrill, N. W. Kwiecien, D. D. Holden, C. Mullen, M. S. Westphall, and J. J. Coon, “A Calibration Routine for Efficient ETD in Large-Scale Proteomics.,” *Journal of the American Society for Mass Spectrometry*, vol. 26, pp. 1848–1857, jun 2015.

[53] J. W. Flora and D. C. Muddiman, “Selective, sensitive, and rapid phosphopeptide identification in enzymatic digests using ESI-FTICR-MS with infrared multiphoton dissociation,” *Analytical Chemistry*, vol. 73, no. 14, pp. 3305–3311, 2001.

[54] B. Ganisl, M. Taucher, C. Riml, and K. Breuker, “Charge as you like! Efficient manipulation of negative ion net charge in electrospray ionization of proteins and nucleic acids,” *Eur J Mass Spectrom (Chichester, Eng)*, vol. 17, no. 4, pp. 333–343, 2011.

[55] H. Hahne, F. Pachl, B. Ruprecht, S. K. Maier, S. Klaeger, D. Helm, G. Medard, M. Wilm, S. Lemeer, and B. Kuster, “DMSO enhances electrospray response, boosting sensitivity of proteomic experiments,” *Nat Methods*, vol. 10, no. 10, pp. 989–991, 2013.

[56] J. G. Meyer and E. A Komives, “Charge state coalescence during electrospray ionization improves peptide identification by tandem mass spectrometry.,” *Journal of the American Society for Mass Spectrometry*, vol. 23, pp. 1390–9, aug 2012.

[57] G. Choudhary, S. L. Wu, P. Shieh, and W. S. Hancock, “Multiple enzymatic digestion for enhanced sequence coverage of proteins in complex proteomic mixtures using capillary LC with ion trap MS/MS,” *J Proteome Res*, vol. 2, no. 1, pp. 59–67, 2003.

[58] R. G. Biringer, H. Amato, M. G. Harrington, A. N. Fonteh, J. N. Riggins, and A. F. Huhmer, “Enhanced sequence coverage of proteins in human cerebrospinal fluid using multiple enzymatic digestion and linear ion trap LC-MS/MS,” *Brief Funct Genomic Proteomic*, vol. 5, no. 2, pp. 144–153, 2006.

[59] A. Leitner, R. Reischl, T. Walzthoeni, F. Herzog, S. Bohn, F. Förster, and R. Aebersold, “Expanding the chemical cross-linking toolbox by the use of multiple proteases and enrichment by size exclusion chromatography.,” *Molecular & cellular proteomics : MCP*, vol. 11, p. M111.014126, mar 2012.

[60] T. Jakoby, B. H. van den Berg, and A. Tholey, “Quantitative protease cleavage site profiling using tandem-mass-tag labeling and LC-MALDI-TOF/TOF MS/MS analysis.,” *Journal of proteome research*, vol. 11, pp. 1812–20, mar 2012.

[61] M. P. Washburn, D. Wolters, and J. R. Yates 3rd, “Large-scale analysis of the yeast proteome by multidimensional protein identification technology,” *Nat Biotechnol*, vol. 19, no. 3, pp. 242–247, 2001.

[62] Y. Zhang, B. R. Fonslow, B. Shan, M. C. Baek, and J. R. Yates 3rd, “Protein analysis by shotgun/bottom-up proteomics,” *Chem Rev*, vol. 113, no. 4, pp. 2343–2394, 2013.

[63] F. Yang, Y. Shen, D. G. Camp, and R. D. Smith, "High-pH reversed-phase chromatography with fraction concatenation for 2D proteomic analysis.," *Expert review of proteomics*, vol. 9, pp. 129–34, apr 2012.

[64] S. Ghaemmaghami, W.-K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman, "Global analysis of protein expression in yeast.," *Nature*, vol. 425, pp. 737–41, oct 2003.

[65] D. L. Swaney, G. C. McAlister, and J. J. Coon, "Decision tree-driven tandem mass spectrometry for shotgun proteomics.," *Nature methods*, vol. 5, pp. 959–64, nov 2008.

[66] F. Kjeldsen, O. B. Horning, S. S. Jensen, A. M. Giessing, and O. N. Jensen, "Towards liquid chromatography time-scale peptide sequencing and characterization of post-translational modifications in the negative-ion mode using electron detachment dissociation tandem mass spectrometry," *J Am Soc Mass Spectrom*, vol. 19, no. 8, pp. 1156–1162, 2008.

[67] K. E. Hersberger and K. Hakansson, "Characterization of O-sulfopeptides by negative ion mode tandem mass spectrometry: superior performance of negative ion electron capture dissociation," *Anal Chem*, vol. 84, no. 15, pp. 6370–6377, 2012.

[68] Y. Huang, X. Yu, Y. Mao, C. E. Costello, J. Zaia, and C. Lin, "De novo sequencing of heparan sulfate oligosaccharides by electron-activated dissociation," *Anal Chem*, vol. 85, no. 24, pp. 11979–11986, 2013.

[69] F. E. Leach 3rd, J. J. Wolff, Z. Xiao, M. Ly, T. N. Laremore, S. Arungundram, K. Al-Mafraji, A. Venot, G. J. Boons, R. J. Linhardt, and I. J. Amster, "Negative electron transfer dissociation Fourier transform mass spectrometry of glycosaminoglycan carbohydrates," *Eur J Mass Spectrom (Chichester, Eng)*, vol. 17, no. 2, pp. 167–176, 2011.

[70] J. A. Madsen, B. J. Ko, H. Xu, J. A. Iwashkiw, S. A. Robotham, J. B. Shaw, M. F. Feldman, and J. S. Brodbelt, "Concurrent automated sequencing of the glycan and peptide portions of O-linked glycopeptide anions by ultraviolet photodissociation mass spectrometry.," *Analytical chemistry*, vol. 85, pp. 9253–61, oct 2013.

[71] C. C. Nwosu, J. S. Strum, H. J. An, and C. B. Lebrilla, "Enhanced detection and identification of glycopeptides in negative ion mode mass spectrometry.," *Analytical chemistry*, vol. 82, pp. 9654–62, dec 2010.

[72] J. J. Wolff, F. E. Leach 3rd, T. N. Laremore, D. A. Kaplan, M. L. Easterling, R. J. Linhardt, and I. J. Amster, "Negative electron transfer dissociation of glycosaminoglycans," *Anal Chem*, vol. 82, no. 9, pp. 3460–3466, 2010.

[73] S. M. Greer, J. R. Cannon, and J. S. Brodbelt, “Improvement of Shotgun Proteomics in the Negative Mode by Carbamylation of Peptides and Ultraviolet Photodissociation Mass Spectrometry.,” *Analytical chemistry*, vol. 86, pp. 12285–12290, nov 2014.

[74] K. D. Wyndham, J. E. O’Gara, T. H. Walter, K. H. Glose, N. L. Lawrence, B. A. Alden, G. S. Izzo, C. J. Hudalla, and P. C. Iraneta, “Characterization and evaluation of C18 HPLC stationary phases based on ethyl-bridged hybrid organic/inorganic particles.,” *Analytical chemistry*, vol. 75, pp. 6781–8, dec 2003.

[75] C. D. Kelstrup, R. R. Jersie-Christensen, T. S. Batth, T. N. Arrey, A. Kuehn, M. Kellmann, and J. V. Olsen, “Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field Orbitrap mass spectrometer.,” *Journal of proteome research*, vol. 13, pp. 6187–95, dec 2014.

[76] R. A. Scheltema, J.-P. Hauschild, O. Lange, D. Hornburg, E. Denisov, E. Damoc, A. Kuehn, A. Makarov, and M. Mann, “The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer.,” *Molecular & cellular proteomics : MCP*, vol. 13, pp. 3698–708, dec 2014.

Chapter 4

TOP-DOWN CHARACTERIZATION OF PROTEINS WITH INTACT DISULFIDE BONDS USING ACTIVATED-ION ELECTRON TRANSFER DISSOCIATION

This chapter has been submitted for review:

Rush MJR, Riley, NM, Westphall, Coon JJ. *Top-Down Characterization of Proteins with Intact Disulfide Bonds Using Activated-Ion Electron Transfer Dissociation.* Analytical Chemistry. **2018**.

Abstract

Here we report the fragmentation of disulfide linked intact proteins using activated-ion electron transfer dissociation (AI-ETD) for top-down protein characterization. This fragmentation method is then compared to the alternative methods of HCD, ETD, and EThcD. We analyzed multiple precursor charge states of the protein standards bovine insulin, α-lactalbumin, lysozyme, β-lactoglobulin, and trypsin inhibitor. In all cases we found that AI-ETD provides a boost in protein sequence coverage information and the generation of fragment ions from within regions enclosed by disulfide bonds. AI-ETD shows the largest improvement over the other techniques when analyzing highly disulfide linked and low charge density precursors ions. This substantial improvement is attributed to the concurrent irradiation of the gas phase ions while the electron-transfer reaction is taking place, mitigating non-dissociative electron transfer, helping unfold the gas phase protein during the electron transfer event, and preventing disulfide bond reformation. We also show that AI-ETD is able to yield comparable sequence coverage information when disulfide bonds are left intact relative to proteins that have been reduced and alkylated. This work demonstrates that AI-ETD is an effective fragmentation method for the analysis of proteins with intact disulfide bonds, dramatically enhancing sequence ion generation and total sequence coverage compared to HCD and ETD.

Introduction

Top-down mass spectrometry allows researchers to interrogate proteins and protein modifications without the need for protein digestion or derivatization.[1,2] The potential benefits to avoid these steps are myriad and include investigation of genetic variants, alternative splicing, and site occupancy of post-translational modifications – information that is often lost upon enzymatic digestion.[3–5] One limitation to the top down approach is that intact protein cations do not dissociate as completely or readily as peptides during tandem mass spectrometry (MS/MS). MS/MS methods such as collision-activated dissociation (CAD) and infrared multiple photon dissociation (IRMPD) often selectively cleave the most labile bonds, limiting sequence coverage and PTM localization.[6–8] Offering more extensive dissociation and the ability to preserve most PTMs, the electron based dissociation methods, *i.e.* electron-capture dissociation (ECD) and electron-transfer dissociation (ETD), have become particularly important for top-down mass spectrometry.[9–11]

Accessing information about PTMs is indeed one of the most attractive strengths of the top-down approach. Disulfide bonds are among the most common PTMs and have fundamental roles in protein stabilization, structure, and function.[12–14] That said, they are challenging to study as they necessitate extensive fragmentation of interlinked peptide backbone sequences. To improve fragmentation, most top-down MS/MS methodologies reduce and alkylate disulfide bonds prior to analysis. Some methods use online disulfide bond reduction just prior to electrospray ionization in both shotgun and top-down proteomic regimes to improve precursor ion fragmentation, eliminating the reduction and alkylation step in peptide or protein sample preparation.[15–18] Other studies have left disulfide bonds intact but performed enzymatic digestion so that disulfide bridged peptides can be detected. That approach, however, suffers the same disadvantages of all shotgun approaches, namely that combinatorial patterns of modification cannot be discerned.[19–25] Producing sequence informative fragment ions from disulfide-bridge peptides or proteins with intact disulfide bonds remains challenging, as collision-based methods, the most prevalent fragmentation type used in proteomics experiments, do not efficiently cleave disulfide bonds, limiting sequence coverage within the region contained by disulfide bridges.

Twenty years ago McLafferty and co-workers demonstrated that disulfide bonds can be cleaved in the gas phase by ECD.[26–28] Since that time many dissociation methods have been examined for their potential application to disulfide bond characterization – e.g., electron transfer dissociation (ETD)24, electron detachment dissociation (EDD)[29], ultraviolet photodissociation (UVPD)[30,31], infrared multiple photon dissociation (IRMPD)[29], metal-cationization[32–34], excitation energy transfer (EET)[35], electron transfer and higher-energy collision dissociation (EThcD)[19], and radical induced dissociation.[36] Recently, Loo and colleagues described that pre-activation of ribonuclease A with UV and IR photons followed by ECD improved fragmentation over ECD alone, even allowing cleavage of multiple disulfide bonds.[37] In this same work, fragmentation of porcine insulin cleaved all disulfide bonds, yielding 73% sequence coverage, an improvement over previous studies.[26,32,37,38] Another important finding was that lengthening the time between UV and IR laser pulses allowed close proximity disulfide bonds to reform, revealing that the disulfide bond reformation of radical thiols was between 10 and 100 ms.[37] Despite these advancements, no one method can fragment disulfide intact proteins nearly as well as the same protein with disulfide bonds reduced and alkylated prior to analysis.

Activated-ion ETD (AI-ETD) greatly improves the sequence coverage of intact protein standards as compared to HCD, ETD, and EThcD[39–41] and improves analysis of post-translationally modified intact proteins, namely phosphoproteins.[42] Here we explore the utility of AI-ETD for interrogating proteins with intact disulfide bonds. Specifically, we compared the fragmentation of five protein standards (bovine insulin, bovine β-lactoglobulin, soybean trypsin inhibitor, α-lactalbumin, and chicken egg lysozyme) with molecular weights ranging from 5.7 to 20 kDa that have two to four intact disulfide bonds which enclose varying degrees of the protein backbone. Having dissociated these species using HCD, ETD, EThcD, and AI-ETD, we conclude that AI-ETD yields greater sequence coverage, fragment ion generation, and disulfide bond cleavage for all precursor charge states studied, with the greatest benefit arising from low charge density precursors. We hypothesize that the condensed state of the proteins with intact disulfide bonds greatly enhances the abundance of non-dissociative electron transfer and prevents electron transfer from occurring in the interior region of the gas phase protein. The benefit of concurrent supplemental infrared irradiation in AI-ETD is two-fold: 1) non-dissociative electron transfer product ions are converted to sequence informative product ions and 2) protein cations are unfolded to allow electron transfer to occur in the interior regions of the protein cation and prevent disulfide bond reformation. These two phenomenon work in tandem to allow for a significant improvement in fragment ion yield and therefore protein structure elucidation. Furthermore, AI-ETD removes the necessity for high charge density precursors that is a hallmark of electron-driven dissociation methods, making it amenable to a wide range of proteins – especially disulfide-bonded proteins that trend toward low charge density precursor ions.

Materials and Methods

Materials and Sample Preparation. The proteins bovine insulin, bovine β-lactoglobulin, and soybean trypsin inhibitor were purchased from Sigma-Aldrich (St. Louis, MO, USA) and α-lactalbumin and chicken egg lysozyme were obtained from Protea Biosciences (Morgantown, WV, USA). Formic acid ampules and acetonitrile were obtained from Thermo Fisher Scientific (Rockford, IL, USA). Solutions were prepared with Milli-Q water (Millipore Corporation, Billerica MA). Samples were prepared for infusion by suspending each protein in 49.9:49.9:0.2 acetonitrile/water/formic acid to a final concentration of 10 pmol per μL. For comparison to the disulfide intact protein, lysozyme was also reduced and alkylated. Lysozyme was suspended in buffer (8 M urea, 50 mM Tris, pH 8) and incubated with 5 mM diothiothreiol for 45 minutes at 58° C, then alkylated with 15 mM iodoacetamide for 45 minutes at room temperature in the dark. The sample was then desalted with a C2 SepPak (Waters, Milford, MA), evaporated, and resuspended in 49.9:49.9:0.2 acetonitrile/water/formic acid to a final concentration of 10 pmol per μL.

ESI-MS/MS Analysis. Each protein standard was infused via syringe pump at a flow rate of 5 µL per minute and electrosprayed with a spray voltage of 4 to 5 eV and inlet capillary temperature of 275° C. All mass spectrometry experiments were performed on a Fusion Lumos mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) modified with a Firestar T-100 Synrad 60-W CO_2 continuous wave laser (Mukilteo, WA) for performing AI-ETD, as previously described.[43] Survey scans using intact protein mode for each protein were performed at 240,000 resolution and averaged over 100 scans. MS/MS experiments were also performed using intact protein mode at 240,000 resolution with a precursor AGC target of 800,000 and averaged over 400 scans. For each protein, three precursor charge states spanning the protein envelope were selected for analysis. For HCD, normalized collision energies of 15, 20, and 25 were used to find the optimal energy for fragmentation. For ETD, EThcD, and AI-ETD, the reagent anion AGC target was set to 300,000 and the ETD reaction time was varied to optimize fragment ion generation and sequence coverage, from 20 to 38 milliseconds. Normalized collision energies of 8, 10, 12, and 15 were used for EThcD and laser powers of 18, 24, 30, and 36 Watts were used for AI-ETD to determine optimal fragmentation.

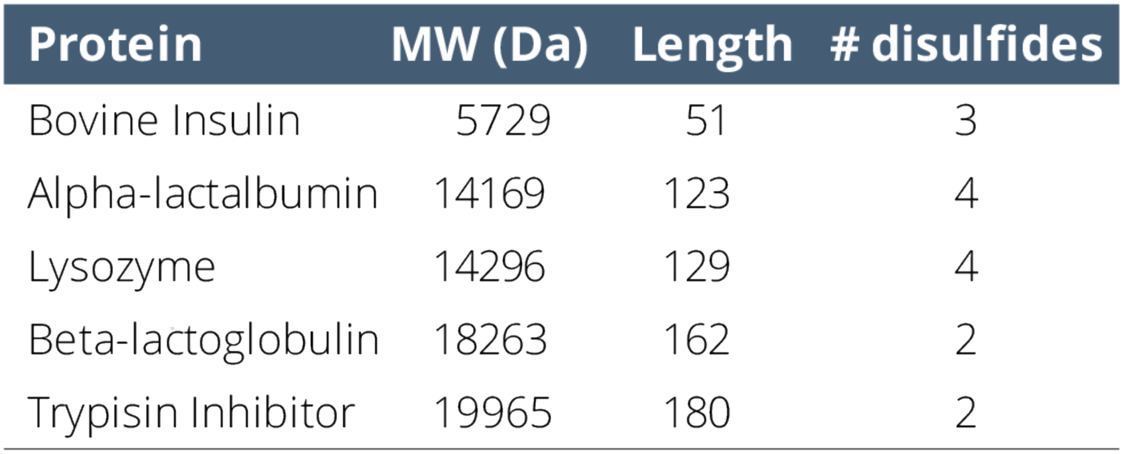
Data Analysis. Raw MS/MS spectra were deconvoluted using the Xtract algorithm (Thermo Fisher Scientific). The spectra were then compared against all possible b, y, c, and $z^{\bullet}$*-type* fragment ions which could be formed from that protein. Modifications were allowed for fragment ions containing a cysteine involved in a disulfide bond to consider all possible cleavage positions of the disulfide bond (S-S and C-S cleavage) and for hydrogen rearrangement products. Cleavage of all disulfide bonds were allowed but only one peptide backbone bond cleavage was considered (no internal fragment ions were searched). Fragment ions were matched within a mass tolerance of 10 part per million.

Results and Discussion

Dissociate proteins with intact disulfide bonds. AI-ETD improves the fragmentation of intact proteins when compared to ETD.[39,40] This boost is realized by the absorption of infrared photons by the protein cations – a process that ultimately induces gas-phase protein unfolding and concomitantly a boost in ETD efficiency. To investigate the potential of AI-ETD for dissociation of disulfide linked protein cations, we selected lysozyme, a protein that contains four disulfide bonds **Table I** (**Figure 4.1**), **Figure 4.2** panel A) enclosing 94% of the protein backbone. We analyzed the +12 charge state precursor of lysozyme using ETD and AI-ETD both with the disulfide bonds intact and cleaved (*i.e.*, reduced and alkylated prior to analysis). The precursor ion charge distributions for the lysozyme cations are shown in **Figure 4.2** panel B. Panel C of **Figure 4.2** presents a sequence coverage map afforded by each fragmentation method. When the disulfide bonds are reduced and cysteines are alkylated, ETD and AI-ETD cleave 58% and 82% of the backbone residues, respectively. Analysis of the protein with intact disulfides, however, presents a much greater challenge. While both methods have reduced sequence coverage, ETD provides only 23% coverage while AI-ETD achieves a much higher 58% coverage. The coverage map shown in panel

C of **Figure 4.2** illuminates the underlying cause for this discrepancy. Here the coverage map is divided into five regions (0, 1, 2, 3, and 4) where the region number corresponds to the number of disulfide bonds that must be cleaved in addition to a protein backbone bond in order to produce a sequence-informative product ion. For example, any observed fragment resulting from dissociation of the backbone between residues 31 and 63 (region 2) can only be formed if three bonds are broken – one protein backbone and two disulfides. Note that ETD does not generate any fragments where more than two dissociations are required – *i.e.*, one backbone and one disulfide. This indicates a clear relationship between the number of disulfide bonds enclosing a region and the amount of sequence informative fragment ions formed. Infrared photoactivation of protein cations during ETD (AI-ETD) disrupts non-covalent interactions and helps to reduce non-dissociative electron transfer (ETnoD).[44,45] These data demonstrate that the concurrent photoactivation used in AI-ETD can open the precursor gas-phase structure and expose the interior of a disulfide linked protein cation so that multiple dissociative electron transfer events can occur. In fact, for lysozyme we detect many fragments that result from three bond cleavages (one backbone and two disulfides) and several that result from four (one backbone and three disulfides).

Comparison of AI-ETD and other methods for dissociation of proteins entirely enclosed by disulfide bonds. Having established the efficacy of AI-ETD to dissociate multiple disulfide linkages, we next sought to characterize performance for various charge states of the same protein (*i.e.*, $z = +9$, $z = +11$, and $z = +13$ precursor ions of lysozyme) and benchmark



Protein	MW (Da)	Length	# disulfides
Bovine Insulin	5729	51	3
Alpha-lactalbumin	14169	123	4
Lysozyme	14296	129	4
Beta-lactoglobulin	18263	162	2
Trypisin Inhibitor	19965	180	2

Figure 4.1: Table I. Summary of standard proteins. List of protein standards chosen and relevant physical characteristics.

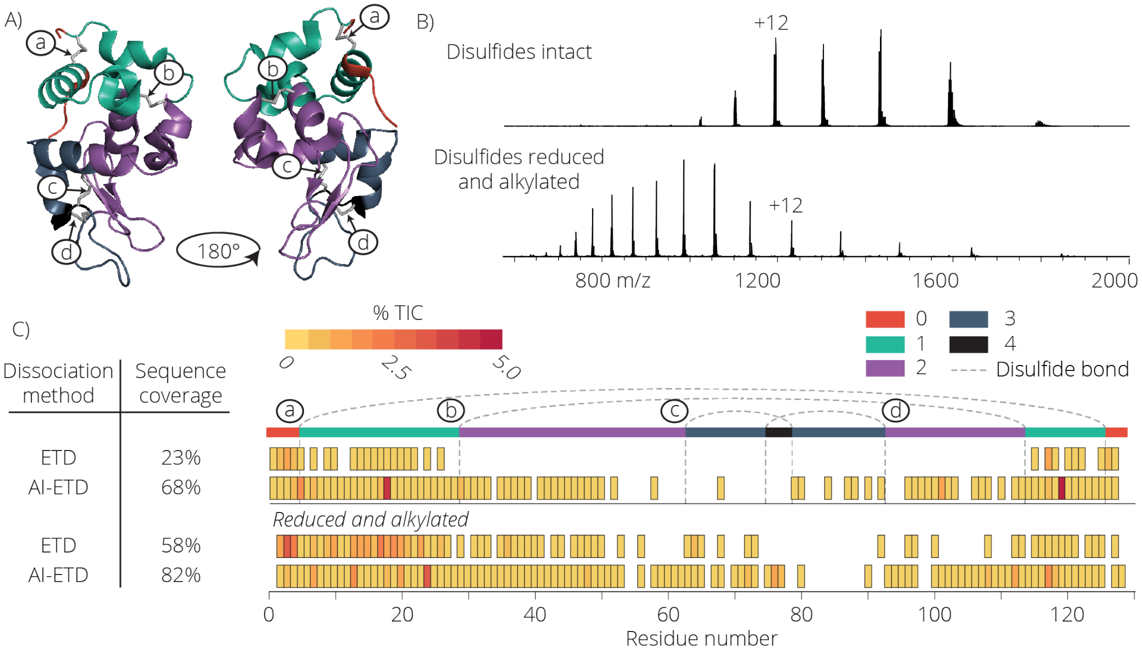


Figure 4.2: Disulfide intact or reduced and alkylated AI-ETD and ETD fragmentation sequence coverage map of lysozyme. Panel A shows the 3 dimensional structure of lysozyme with all 4 disulfide bonds intact in its native state. The regions are color coded based on the number of disulfide bond breakages required to generate a fragment ion for that region. The four disulfide bonds are labeled a,b,c, and d to illustrate where they occur along the protein backbone. The electrospray ionization MS^1 spectra for lysozyme with disulfide bonds intact and disulfide bonds reduced and alkylated is shown (B). Panel C compares the sequence coverage of lysozyme precursor charge state +12 using ETD and AI-ETD with disulfide bonds intact or reduced and alkylated. Sequence coverage of the highly disulfide linked region is hindered greatly using ETD, while AI-ETD shows pronounced coverage of this portion of the protein.

this performance to other common dissociation methods including beam-type collisional activation (HCD), ETD, and electron transfer dissociation with HCD as supplemental collisional activation (EThcD). **Figure 4.3** summarizes the results for lysozyme. **Figure 4.3** panel A compares the MS/MS spectra using each dissociation method for the z = +11 charge state. The percentage of the total product ion signal contained in sequence-informative fragment ion channels is 10%, 14%, 25%, and 49% for HCD, ETD, EThcD, and AI-ETD, respectively. From these data we conclude AI-ETD induces much more extensive fragmentation than any of the other tested methods. To see how these fragment ions facilitate sequence analysis, we generated sequence coverage maps for each dissociation method for all three precursor ion charge states (**Figure 4.3**, panels B and C). As with the example above, AI-ETD generates substantially more sequence informative fragment ions, especially for those ions requiring disulfide bond cleavages. Again AI-ETD allows for the observation of fragment ions that result from cleavage of up to five bonds (*i.e.*, one backbone and four disulfide linkages). For the z = +9 charge state precursor, only AI-ETD produced any fragmentation within the region enclosed by two or more disulfide bonds (show in purple, blue, and black along the top of the plot). As the charge density of the precursor increased, the sequence coverage for HCD, ETD, and EThcD is improved, while AI-ETD stays constant. This is consistent with previous work showing that higher charge states improve fragmentation for ETD while AI-ETD is more or less indifferent to precursor charge density. We next conducted a similar study but with a different protein – α-lactalbumin (14.2 kDa). This protein is very similar to lysozyme in terms of mass, length, and disulfide linkage, yet differing substantially in amino acid composition. The sequence coverage for the z = +10 charge state precursor of this protein reveals AI-ETD yields 64% sequence coverage while HCD, ETD, and EThcD show 6%, 12%, and 26%, respectively (**Supplementary Figure S4.1**).

Comparison of AI-ETD and other methods for dissociation of proteins partially enclosed by disulfide bonds. To contrast the highly disulfide linked proteins lysozyme and α-lactalbumin, we next analyzed β-lactoglobulin which has two disulfide bonds enclosing 59% of the protein. **Figure 4.4** presents sequence coverage maps following dissociation of the +10, +12, and +14 precursor charge states of β-lactoglobulin with either HCD, ETD, EThcD, or AI-ETD. To isolate the effect of the disulfide bond, we also calculate sequence coverage percentages between residues 1-65 (not enclosed by a disulfide bond) and residues 66-160 (enclosed by one or two disulfide bonds). Not surprisingly the region containing no disulfide bonds has high sequence coverage regardless of precursor charge state or dissociation method. The region enclosed by disulfide bonds, however, is much less accessible. AI-ETD cleaves at least 50% of the bonds in this region while ETD and HCD produce very few fragment ions. In addition, the structurally similar protein trypsin inhibitor was examined. This protein has two disulfide bonds enclosing 31% of the protein backbone. AI-ETD successfully sequenced 60% of the protein for all charge states examined (**Supplementary Figure S4.2**). These data demonstrate that the benefit of AI-ETD is most pronounced in regions that are enclosed by disulfide bonds.

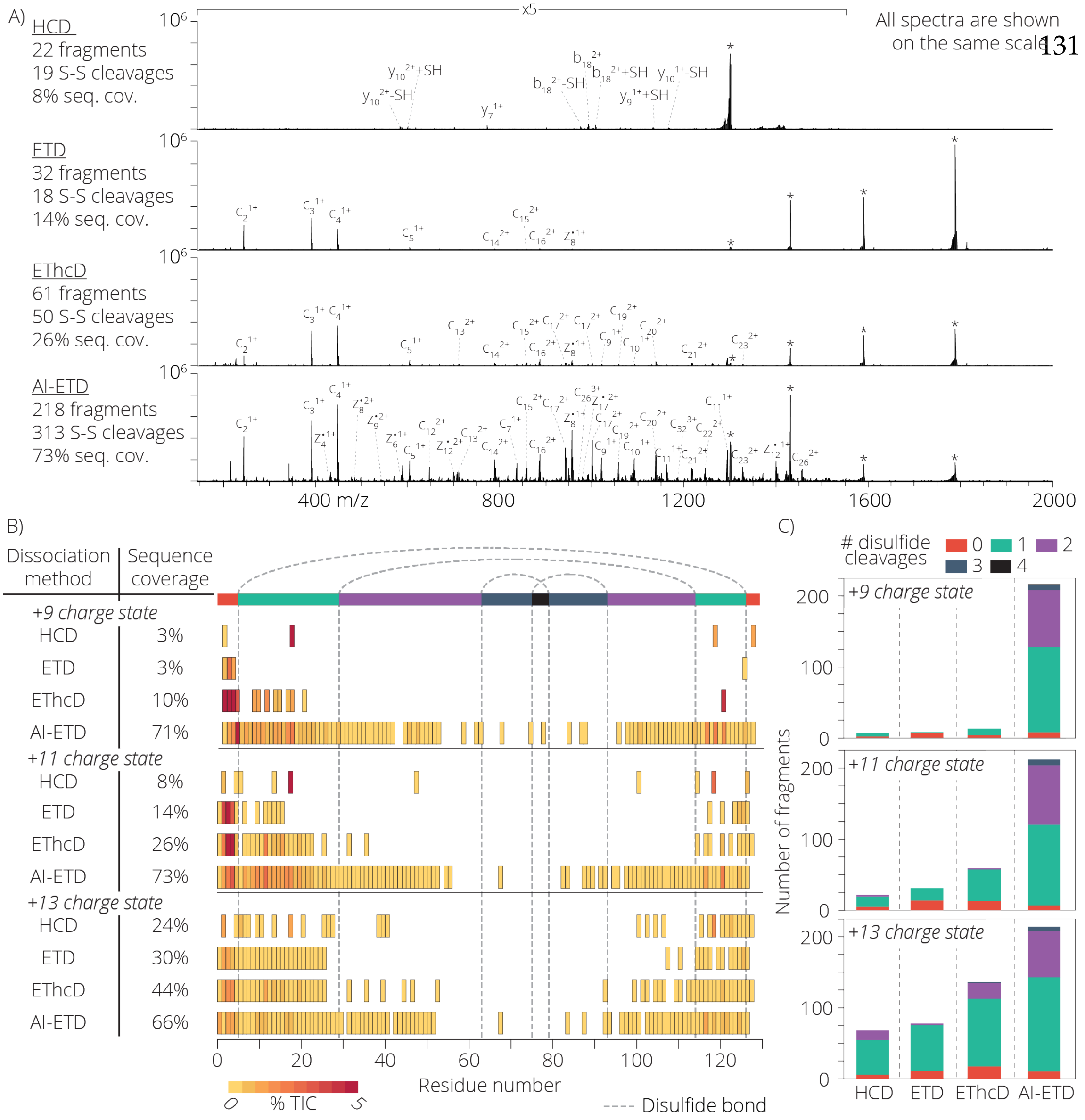
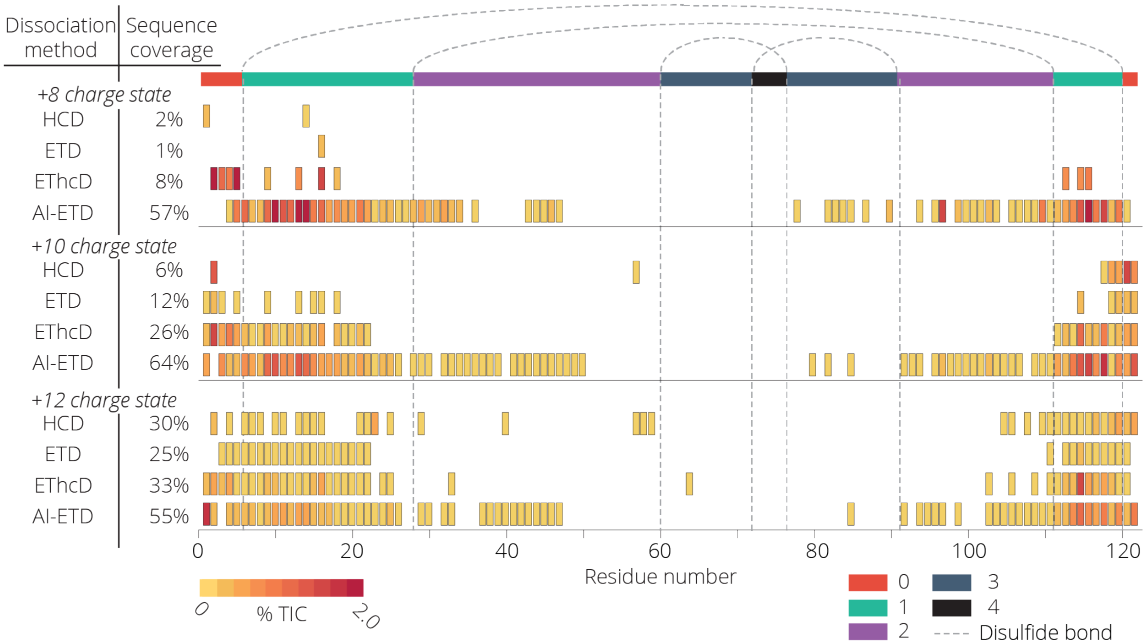


Figure 4.3: Lysozyme annotated spectra and sequence coverage map using HCD, ETD, EThcD, or AI-ETD fragmentation. Panel A shows a comparison of HCD, ETD, EThcD, and AI-ETD fragmentation of lysozyme with precursor charge state +11 with all 4 disulfide bonds intact. All spectra are 400 scan averages and are shown on the same scale. The scan range of 150 – 1550 *m/z* are shown at 5 times magnification. Peaks annotated with an asterisk (*) show the unreacted precursor and charge reduced precursor ions. A summary of the number of detected fragment ions, number of disulfide cleavages (either S-S or C-S) amongst identified fragments, and percentage of inter-residue bond cleavages (referred to as sequence coverage) is shown to the right for each dissociation method. Panel B illustrates the sequence coverage achieved for each dissociation method tested for the precursor charge states +9, +11, and +13. Notably, AI-ETD is able to generate fragment ions within the region of the protein requiring 2 or more disulfide bond cleavages far more often than the other dissociation methods. Panel C illustrates the number of identified fragment ions and the amount of disulfide bond cleavages necessary to form that ion.



Supplementary Figure S4.1: α-lactalbumin sequence coverage map using HCD, ETD, EThcD, or AI-ETD fragmentation. The sequence coverage map for α-lactalbumin precursor charge states $z = +8$, $z = +10$, and $z = +12$ is shown. The regions are codified by the number of disulfide bonds surrounding that portion of the protein backbone. The percent sequence coverage for each precursor charge state and dissociation method is shown in the table to the left.

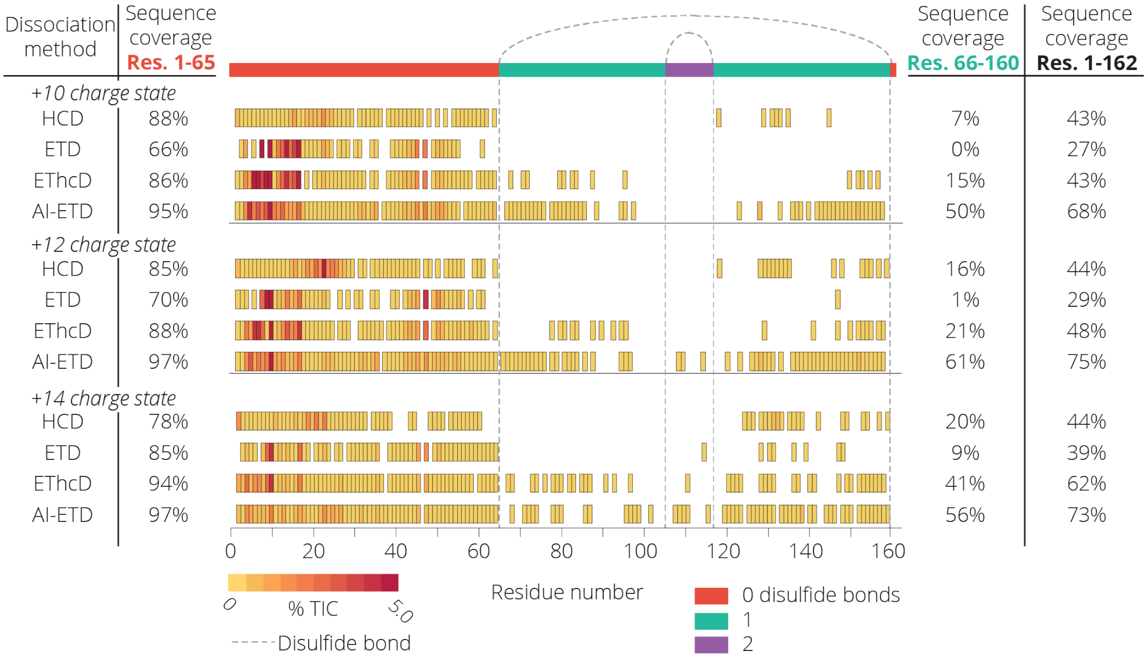
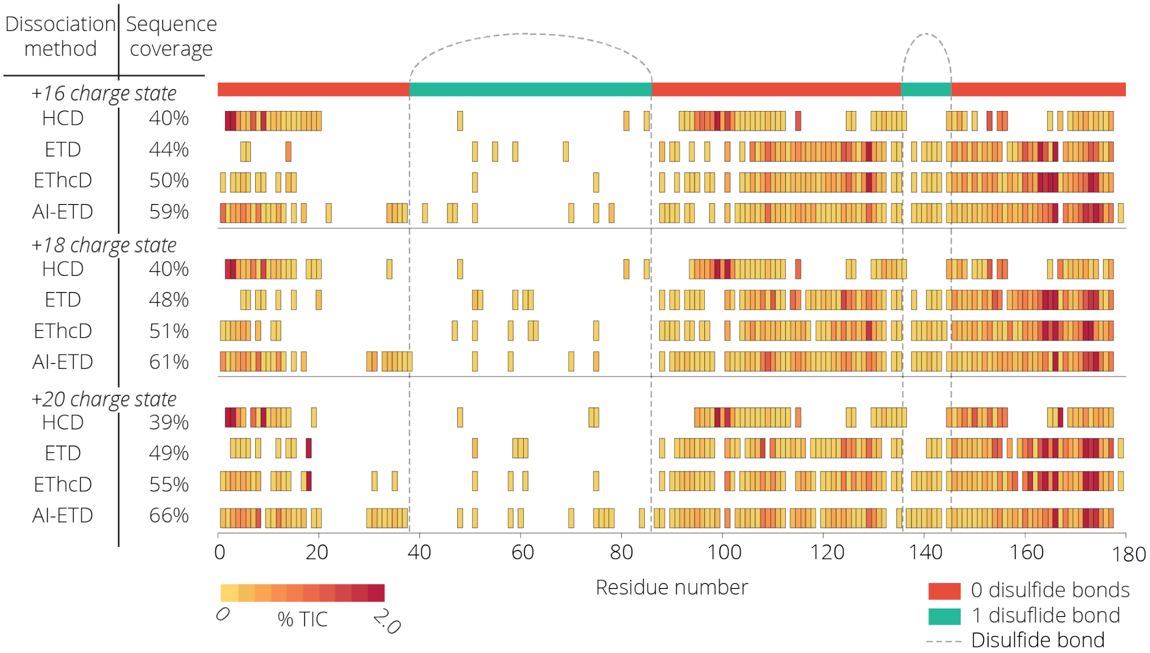


Figure 4.4: β-lactoglobulin sequence coverage map using HCD, ETD, EThcD, or AI-ETD fragmentation. The sequence coverage cleavage map for the +10, +12, and +14 charge state precursors of β-lactoglobulin are shown. The left sequence coverage values represent the region of the protein which does not contain any disulfide bonds and the left values show the sequence coverage which contains 1 or 2 disulfide bonds. All fragmentation methods perform well in the open region, while AI-ETD shows significant increase in coverage for the region enclosed by disulfide bonds.



Supplementary Figure S4.2: Trypsin inhibitor sequence coverage map using HCD, ETD, EThcD, or AI-ETD fragmentation. The sequence coverage map for trypsin inhibitor precursor charge states $z = +16$, $z = +18$, and $z = +20$ is shown. The regions are codified by the number of disulfide bonds surrounding that portion of the protein backbone. The percent sequence coverage for each precursor charge state and dissociation method is shown in the table to the left.

Comparison of AI-ETD and other methods for dissociation of a protein with interpeptide disulfide bonds. Lastly we investigated the fragmentation of bovine insulin with intact disulfides. Insulin comprises of two separate peptide chains linked by two disulfide bonds (**Figure 4.5** panel A). There is an additional intrapeptide disulfide bond on the A-chain. **Figure 4.5** panel A highlights the different types of fragment ions which can be formed, classifying each fragment ion by the number of disulfide bond cleavages necessary for the formation of the fragment. The results of the MS/MS fragmentation of insulin with HCD, ETD, EThcD, and AI-ETD for the $z = +5$ precursor charge state of insulin is shown in **Figure 4.5** panels B and C. Both EThcD and AI-ETD yield near complete sequence coverage of the protein; however, the intensity of fragment ions in AI-ETD make up a significantly larger percent of the total ion current. Fragment ion generation within the region of the protein enclosed by disulfide bonds is noticeably low with ETD. While ETD preferentially cleaves disulfide bonds, the highly compact and charge dense characteristics of bovine insulin likely causes substantial ETnoD product ion formation.

Conclusions

We demonstrate that AI-ETD is an extremely effective fragmentation method for five protein standards which contain intact disulfide linkages. The results for all proteins investigated are summarized in **Figure 4.6** where we show that the total sequence coverage for each protein, the number of fragments ions generated, and the number of total disulfide bonds broken across all fragment ions for all precursor ion charge states examined is optimal when

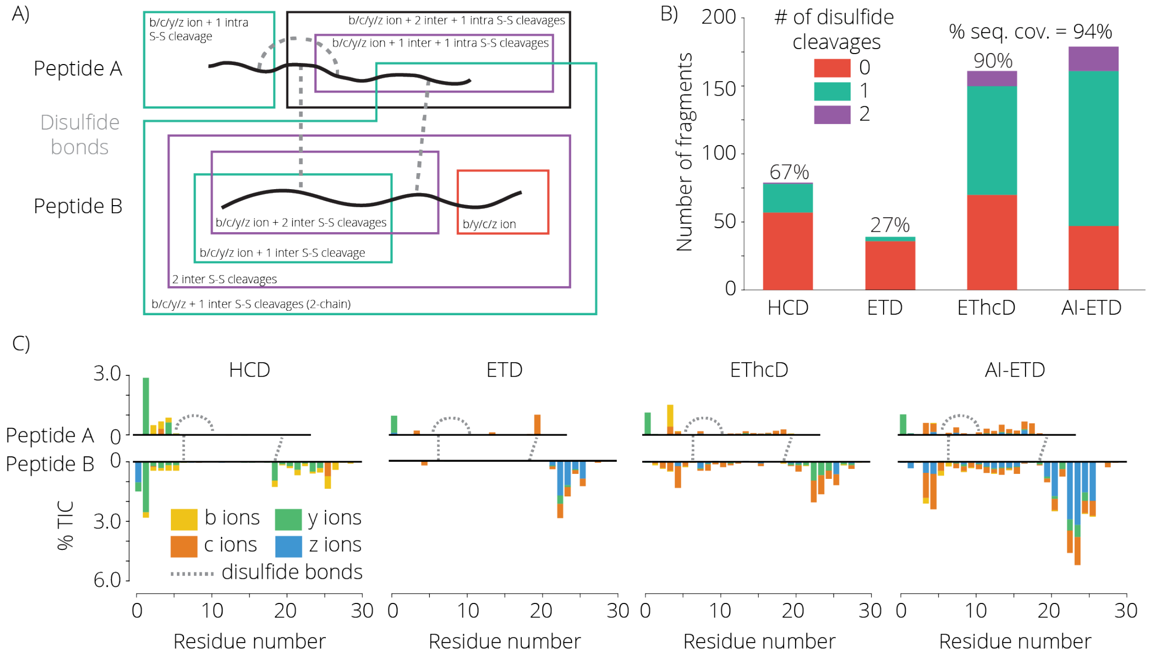


Figure 4.5: Bovine insulin disulfide linkage, fragment ion count, and fragmentation maps. A cartoon representation of bovine insulin with intact disulfide bonds is shown (A). Representative fragment ions are illustrated and are color coded based on the number of disulfide bond cleavages necessary to generate that fragment. Unlike the other proteins analyzed in this study, insulin consists of 2 separate disulfide linked peptides, generating a complex assortment of fragment ion possibilities. Panel B shows the number of fragment ions observed using each dissociation method and the number of disulfide cleavages necessary for each ion. The relative signal intensity and position of each ion is shown in panel C.

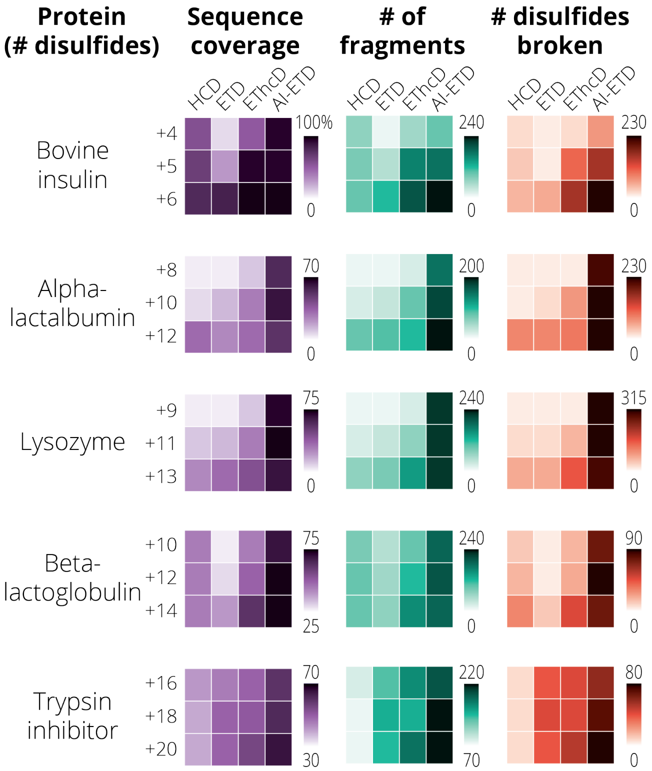


Figure 4.6: Summary of protein sequence coverage, fragment ion generation, and disulfide cleavage for HCD, ETD, EThcD, and AI-ETD fragmentation methods. A summary of the sequence coverage, number of identified fragments, and number of disulfide bonds cleaved for each protein and precursor charge state analyzed in this study. In all cases tested AI-ETD yields the greatest sequence coverage, number of fragment ions formed, and number of disulfide bond cleavages within the fragment ions.

AI-ETD as the fragmentation method. Note, when calculating the number of total disulfide bonds broken, a fragment requiring two disulfide cleavages would be counted twice. These results build upon previous observations that 1) ETD preferentially fragments disulfide bonds but can suffer from ETnoD and 2) AI-ETD improves fragmentation of intact proteins by reducing the amount of ETnoD product ion formation. Concurrent irradiation allows for the unfolding of the gas phase protein while the electron transfer reaction is occurring, exposing regions of the protein inaccessible to the ETD reagent otherwise and preventing disulfide bond reformation. The method is particularly effective with proteins which are highly compact in the gas phase, such as lysozyme and insulin, where a majority of the protein backbone is enclosed by disulfide linkages. Furthermore, AI-ETD shows effective fragmentation across the precursor ion charge state envelope, allowing the interrogation of low charge density precursor ions which generally offer poor fragmentation by ETD alone.

The ability to effectively fragment highly disulfide linked intact proteins with AI-ETD will likely advance efforts towards the structural characterization of many types of proteins such as intact antibodies, toxins, native proteins, and protein complexes. Additionally, the substantial sequence coverage afforded by AI-ETD may allow for decreased scan averaging, a significant bottle-neck in top-down experiments[46], which could enable chromatographic-timescale analyses of more complex mixtures of intact proteins. In all, AI-ETD is a superior fragmentation technique for proteins with intact disulfide bonds and will be a valuable tool for disulfide bond analysis in a variety of applications.

References

[1] T. K. Toby, L. Fornelli, and N. L. Kelleher, "Progress in Top-Down Proteomics and the Analysis of Proteoforms," *Annual Review of Analytical Chemistry*, vol. 9, no. 1, pp. 499–519, 2016.

[2] B. Chen, K. A. Brown, Z. Lin, and Y. Ge, "Top-Down Proteomics: Ready for Prime Time?," *Analytical Chemistry*, vol. 90, no. 1, pp. 110–127, 2018.

[3] N. Siuti and N. L. Kelleher, "Decoding protein modifications using top-down mass spectrometry," *Nature Methods*, vol. 4, no. 10, pp. 817–821, 2007.

[4] L. M. Smith and N. L. Kelleher, "Proteoform: A single term describing protein complexity," *Nature Methods*, vol. 10, no. 3, pp. 186–187, 2013.

[5] A. D. Catherman, O. S. Skinner, and N. L. Kelleher, "Top Down proteomics: Facts and perspectives," *Biochemical and Biophysical Research Communications*, vol. 445, no. 4, pp. 683–693, 2014.

[6] D. P. Little, J. P. Speir, M. W. Senko, P. B. O'Connor, and F. W. McLafferty, "Infrared Multiphoton Dissociation of Large Multiply Charged Ions for Biomolecule Sequencing," *Analytical Chemistry*, vol. 66, no. 18, pp. 2809–2815, 1994.

[7] S. A. Raspopov, A. El-Faramawy, B. A. Thomson, and K. W. M. Siu, "Infrared multiphoton dissociation in quadrupole time-of-flight mass spectrometry: Top-down characterization of proteins," *Analytical Chemistry*, vol. 78, no. 13, pp. 4572–4577, 2006.

[8] D. R. Ahlf, P. D. Compton, J. C. Tran, B. P. Early, P. M. Thomas, and N. L. Kelleher, "Evaluation of the compact high-field orbitrap for top-down proteomics of human cells," *Journal of Proteome Research*, vol. 11, no. 8, pp. 4308–4314, 2012.

[9] R. Zubarev, N. L. Kelleher, and F. W. McLafferty, "Electron capture dissociation of multiply charged protein cations. A …," *J. Am. Chem. Soc*, vol. 120, no. 16, pp. 3265–3266, 1998.

[10] J. E. P. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt, "Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry," in *Proceedings of the Nationalal Academy of Sciences*, vol. 101, pp. 9528 – 9533, 2004.

[11] N. M. Riley and J. J. Coon, "The Role of Electron Transfer Dissociation in Modern Proteomics," *Analytical Chemistry*, vol. 90, no. 1, pp. 40–64, 2018.

[12] J. M. Thornton, "Disulfide bridges in globular proteins," *J. Mol. Biol.*, vol. 151, pp. 261–287, 1981.

[13] M. Matsumura, G. Signor, and B. W. Matthews, “Substantial increase of protein stability by multiple disulphide bonds.,” *Nature*, vol. 342, no. 6247, pp. 291–293, 1989.

[14] W. J. Wedemeyer, E. Welker, M. Narayan, and H. A. Scheraga, “Disulfide bonds and protein folding,” *Biochemistry*, vol. 39, no. 15, pp. 4207–4216, 2000.

[15] Y. Zhang, H. D. Dewald, and H. Chen, “Online mass spectrometric analysis of proteins/peptides following electrolytic cleavage of disulfide bonds,” *Journal of Proteome Research*, vol. 10, no. 3, pp. 1293–1304, 2011.

[16] Y. Zhang, W. Cui, H. Zhang, H. D. Dewald, and H. Chen, “Electrochemistry-assisted top-down characterization of disulfide-containing proteins,” *Analytical Chemistry*, vol. 84, no. 8, pp. 3838–3842, 2012.

[17] S. Nicolardi, M. Giera, P. Kooijman, A. Kraj, J. P. Chervet, A. M. Deelder, and Y. E. M. Van Der Burgt, “On-line electrochemical reduction of disulfide bonds: Improved FTICR-CID and -ETD coverage of oxytocin and hepcidin,” *Journal of the American Society for Mass Spectrometry*, vol. 24, no. 12, pp. 1980–1987, 2013.

[18] D. S. Zhao, Z. R. Gregorich, and Y. Ge, “High throughput screening of disulfide-containing proteins in a complex mixture,” *Proteomics*, vol. 13, no. 22, pp. 3256–3260, 2013.

[19] F. Liu, B. van Breukelen, and A. J. R. Heck, “Facilitating Protein Disulfide Mapping by a Combination of Pepsin Digestion, Electron Transfer Higher Energy Dissociation (EThcD), and a Dedicated Search Algorithm SlinkS,” *Molecular & Cellular Proteomics*, vol. 13, no. 10, pp. 2776–2786, 2014.

[20] W. Ni, M. Lin, P. Salinas, P. Savickas, S. L. Wu, and B. L. Karger, “Complete mapping of a cystine knot and nested disulfides of recombinant human arylsulfatase a by multi-enzyme digestion and LC-MS analysis using CID and ETD,” *Journal of the American Society for Mass Spectrometry*, vol. 24, no. 1, pp. 125–133, 2013.

[21] S. L. Wu, H. Jiang, W. S. Hancock, and B. L. Karger, “Identification of the unpaired cysteine status and complete mapping of the 17 disulfides of recombinant tissue plasminogen activator using LC-MS with electron transfer dissociation/collision induced dissociation,” *Analytical Chemistry*, vol. 82, no. 12, pp. 5296–5303, 2010.

[22] Y. Wang, Q. Lu, S. L. Wu, B. L. Karger, and W. S. Hancock, “Characterization and comparison of disulfide linkages and scrambling patterns in therapeutic monoclonal antibodies: Using LC-MS with electron transfer dissociation,” *Analytical Chemistry*, vol. 83, no. 8, pp. 3133–3140, 2011.

[23] S. Lu, S. B. Fan, B. Yang, Y. X. Li, J. M. Meng, L. Wu, P. Li, K. Zhang, M. J. Zhang, Y. Fu, J. Luo, R. X. Sun, S. M. He, and M. Q. Dong, "Mapping native disulfide bonds at a proteome scale," *Nature Methods*, vol. 12, no. 4, pp. 329–331, 2015.

[24] S.-l. Wu, H. Jiang, Q. Lu, S. Dai, W. S. Hancock, and B. L. Karger, "Mass Spectrometric Determination of Disulfide Linkages in Recombinant Therapeutic Proteins Using Online LC-MS with Electron-Transfer Dissociation," *Analytical Chemistry*, vol. 81, pp. 112–122, jan 2009.

[25] D. F. Clark, E. P. Go, and H. Desaire, "Simple approach to assign disulfide connectivity using extracted ion chromatograms of electron transfer dissociation spectra," *Analytical Chemistry*, vol. 85, no. 2, pp. 1192–1199, 2013.

[26] R. A. Zubarev, N. A. Kruger, E. K. Fridriksson, M. A. Lewis, D. M. Horn, B. K. Carpenter, and F. W. McLafferty, "Electron Capture Dissociation of Gaseous Multiply-charged Proteins is Favored at Disulfide Bonds and Other Sites of High Hydrogen Atom Affinity," *J. Am. Chem. Soc.*, vol. 121, no. 8, pp. 2857–2862, 1999.

[27] B. Ganisl and K. Breuker, "Does Electron Capture Dissociation Cleave Protein Disulfide Bonds?," *ChemistryOpen*, vol. 1, no. 6, pp. 260–268, 2012.

[28] Y. Ge, M. ElNaggar, S. K. Sze, H. B. Oh, T. P. Begley, F. W. McLafferty, H. Boshoff, and C. E. Barry, "Top Down Characterization of Larger Proteins (45 kDa) by Electron Capture Dissociation Mass Spectrometry," *Journal of the American Chemical Society*, vol. 14, no. 3, pp. 253–261, 2003.

[29] A. Kalli and K. Håkansson, "Preferential cleavage of S{single bond}S and C{single bond}S bonds in electron detachment dissociation and infrared multiphoton dissociation of disulfide-linked peptide anions," *International Journal of Mass Spectrometry*, vol. 263, no. 1, pp. 71–81, 2007.

[30] Y. M. E. Fung, F. Kjeldsen, O. A. Silivra, T. W. D. Chan, and R. A. Zubarev, "Facile disulfide bond cleavage in gaseous peptide and protein cations by ultraviolet photodissociation at 157 nm," *Angewandte Chemie - International Edition*, vol. 44, no. 39, pp. 6399–6403, 2005.

[31] A. Agarwal, J. K. Diedrich, and R. R. Julian, "Direct elucidation of disulfide bond partners using ultraviolet photodissociation mass spectrometry," *Analytical Chemistry*, vol. 83, no. 17, pp. 6455–6458, 2011.

[32] M. Mentinova and S. A. McLuckey, "Cleavage of multiple disulfide bonds in insulin via gold cationization and collision-induced dissociation," *International Journal of Mass Spectrometry*, vol. 308, no. 1, pp. 133–136, 2011.

[33] H. Lioe, M. Duan, and R. A. J. O'Hair, "Can metal ions be used as gas-phase disulfide bond cleavage reagents? A survey of coinage metal complexes of model peptides containing an intermolecular disulfide bond," *Rapid Communications in Mass Spectrometry*, vol. 21, pp. 2727–2733, 2007.

[34] H. P. Gunawardena, R. A. O'Hair, and S. A. McLuckey, "Selective disulfide bond cleavage in gold(I) cationized polypeptide ions formed via gas-phase ion/ion cation switching," *Journal of Proteome Research*, vol. 5, no. 9, pp. 2087–2092, 2006.

[35] N. G. Hendricks, N. M. Lareau, S. M. Stow, J. A. McLean, and R. R. Julian, "Bond-specific dissociation following excitation energy transfer for distance constraint determination in the gas phase," *Journal of the American Chemical Society*, vol. 136, no. 38, pp. 13363–13370, 2014.

[36] M. Lee, Y. Lee, M. Kang, H. Park, Y. Seong, B. J. Sung, B. Moon, and H. B. Oh, "Disulfide bond cleavage in TEMPO-free radical initiated peptide sequencing mass spectrometry.," *Journal of mass spectrometry : JMS*, vol. 46, no. 8, pp. 830–839, 2011.

[37] P. Wongkongkathep, H. Li, X. Zhang, R. R. Ogorzalek Loo, R. R. Julian, and J. A. Loo, "Enhancing protein disulfide bond cleavage by UV excitation and electron capture dissociation for top-down mass spectrometry," *International Journal of Mass Spectrometry*, vol. 390, no. 2015, pp. 137–145, 2015.

[38] J. Liu, H. P. Gunawardena, T. Y. Huang, and S. A. McLuckey, "Charge-dependent dissociation of insulin cations via ion/ion electron transfer," *International Journal of Mass Spectrometry*, vol. 276, no. 2-3, pp. 160–170, 2008.

[39] N. M. Riley, M. S. Westphall, and J. J. Coon, "Activated Ion Electron Transfer Dissociation for Improved Fragmentation of Intact Proteins," *Analytical Chemistry*, vol. 87, no. 14, pp. 7109–7116, 2015.

[40] N. M. Riley, M. S. Westphall, and J. J. Coon, "Activated Ion-Electron Transfer Dissociation Enables Comprehensive Top-Down Protein Fragmentation," *Journal of Proteome Research*, vol. 16, no. 7, pp. 2653–2659, 2017.

[41] N. M. Riley, M. S. Westphall, and J. J. Coon, "Sequencing Larger Intact Proteins (30-70 kDa) with Activated Ion Electron Transfer Dissociation," *Journal of The American Society for Mass Spectrometry*, no. 2018, pp. 140–149, 2017.

[42] N. M. Riley, A. S. Hebert, G. Dürnberger, F. Stanek, K. Mechtler, M. S. Westphall, and J. J. Coon, "Phosphoproteomics with Activated Ion Electron Transfer Dissociation," *Analytical Chemistry*, vol. 89, no. 12, pp. 6367–6376, 2017.

[43] N. M. Riley, M. S. Westphall, A. S. Hebert, and J. J. Coon, “Implementation of Activated Ion Electron Transfer Dissociation on a Quadrupole-Orbitrap-Linear Ion Trap Hybrid Mass Spectrometer,” *Analytical Chemistry*, vol. 89, no. 12, pp. 6358–6366, 2017.

[44] S. J. Pitteri, P. A. Chrisman, and S. A. McLuckey, “Electron-transfer ion/ion reactions of doubly protonated peptides: Effect of elevated bath gas temperature,” *Analytical Chemistry*, vol. 77, pp. 5662–5669, sep 2005.

[45] S. J. Pitteri, P. A. Chrisman, J. M. Hogan, and S. A. McLuckey, “Electron transfer ion/ion reactions in a three-dimensional quadrupole ion trap: Reactions of doubly and triply protonated peptides with SO 2•-,” *Analytical Chemistry*, vol. 77, no. 6, pp. 1831–1839, 2005.

[46] N. M. Riley, C. Mullen, C. R. Weisbrod, S. Sharma, M. W. Senko, V. Zabrouskov, M. S. Westphall, J. E. Syka, and J. J. Coon, “Enhanced Dissociation of Intact Proteins with High Capacity Electron Transfer Dissociation,” *Journal of the American Society for Mass Spectrometry*, vol. 27, no. 3, pp. 520–531, 2016.

Chapter 5

A HIGH-RESOLUTION ORBITRAP GC-MS MASS SPECTRAL LIBRARY FOR DISCOVERY METABOLOMICS

Portions of this chapter are part of a manuscript in preparation:

Rush MJP, Connors, KJ, Kwiecien NW, Trujilla EA, Russell JD, Westphall MS, Coon JJ. *A High-Resolution Orbitrap GC-MS Mass Spectral Library for Discovery Metabolomics.* **2018**.

Abstract

Gas chromatography mass spectrometry has been a primary analytical tool for targeted metabolomics measurements for decades. More recently, the development of GC-compatible high-resolution accurate mass analyzers has popularized global metabolome profiling using discovery-based analytical approaches. Yet, confident compound identification in discovery experiments still poses significant challenges. In part, this can be attributed to the reliance on spectral libraries generated from low-resolution instruments which do not fully leverage accurate mass measurements at high resolution for compound identification. We endeavored to generate a curated metabolite specific spectral database using a Q Exactive GC Orbitrap mass spectrometer. We show the utility of this new high-resolution metabolite library by profiling a complex human cell lysate, resulting in increased confidence in metabolite identifications relative to a unit-resolution library.

Introduction

The use of gas-chromatography mass spectrometry has been an integral tool in the targeted identification and quantitation of small molecule metabolites for decades. Recent advancements in sample preparation, chromatographic separation, and mass analyzers has allowed for the analysis of hundreds of metabolites from whole cell lysates, termed metabolomics. As the downstream product of biochemical functions, the unbiased detection of metabolites from these samples, termed metabolomics, provides researchers with a snapshot of the metabolic state of the system.[1] Therefore, the goal of metabolomics technological development is to improve the breadth of coverage and identification confidence of metabolites detected using the analytical method. The use of mass spectral libraries is the most common way of interpreting the spectra collected by GC-MS experiments.[2] Software can be used to compare the collected spectra with the library entries to return scores which are proportional to the similarity of the collected spectra and the corresponding library match. In order for the compounds to match perfectly, they need to be collected under standardized conditions. The use of electron ionization (EI) at 70 eV is the most common ionization method for GC-MS library generation. Additionally, the Kovat's retention index (a standardized chromatographic retention time) is also compared to confirm a metabolites identity.[3] Therefore, spectral libraries will ideally comprise the electron impact fragmentation spectra, a retention index value, and additional information about the chemicals composition, structure, and other meta-data about the compound and instrument method used to acquire the spectral library entry. Despite the steps that have been taken to standardize the generation of GC-MS library spectra, spectral matching is imperfect as the conditions under which the library spectra are collected are not identical to those that the analyst uses when conducting their experiment, such as the type of mass spectrometer and the conditions of the EI source. Currently, many publically available mass spectral libraries and spectral repositories exist for use in compound identification of GC-MS experimental data.[4–8] The National Institutes for Standards in Technology (NIST) offers a library comprising over 250,000 EI spectra collected over many decades.[9] While this library is incredibly vast, all spectra are presented at unit-resolution. Therefore, modern instrumentation with much greater mass accuracy are unable to leverage this additional information when searching against these unit resolution databases. Furthermore, a significant portion of this library is composed of compounds which are non-biological, requiring the analyst to disregard numerous false matches when seeking to identify the metabolite of interest. For these reasons, we endeavored to build a metabolite spectral library using the high resolution Q Exactive GC platform, hand-selecting compounds most relevant to metabolism, health, and human disease1.[10–12]

Materials and Methods

Metabolite Standard Sample Preparation and GC-MS Analysis. Metabolite standards (obtained from Sigma Aldrich, Thermo Fisher Scientific, IROA Technologies, Santa Cruz Biotechnology, Toronto Research Chemicals, or ACROS Organics) were prepared by silylation with or without preceding methoxyamination, to a final concentration of 10 – 100 ppm. For methoxyamination, dried samples were suspended in 35 µL of 20 mg/mL methoxyamine hydrochloride (Sigma Aldrich) solution in pyridine (Sigma Aldrich) and heated at 37° C for 90 minutes, then 65 µL MSTFA with 1% TMCS (Restek) was added and the solution was heated to 60° C for 30 minutes. For silylation with MSTFA or MTBSTFA dried samples were suspended in 50 µL of a 1:1 mixture of pyridine and MSTFA with 1% TMCS or MTBSTFA with 1% t-BDCMS (Restek) and then heated to 60° C for 30 minutes. Derivatized samples were then analyzed using a GC-MS instrument comprising a Trace 1310 GC coupled to a Q Exactive Orbitrap mass spectrometer (Thermo Fisher Scientific). A linear temperature gradient ranging from 50° C to 320° C was employed, spanning a total runtime of 30 minutes. Analytes were injected onto a 30 meter TraceGOLD TG-5SILMS column (Thermo Fisher Scientific) using a 1:10 split ratio at an injector temperature of 275° C and ionized using electron ionization at 70 eV. The mass spectrometer was operated in full scan mode (50-650 *m/z*) using a resolution of 60,000 (m/Δm) relative to 200 *m/z*. For calculation of retention indices, after every ten standard injections, a C7 to C40 alkanes (Sigma Aldrich) sample, suspended to 10 ppm in hexane, was analyzed using the same GC-MS method.

Spectral Entry Data Analysis. Raw data files from each derivatized metabolite standard were deconvolved using the TraceFinder 3.3 Deconvolution Plugin (Thermo Fisher Scientific). Spectra were then exported from TraceFinder as text files in the .MSP format.

Information about the GC-MS method used, the Kovat's retention index, and other pertinent metadata were then appended to each .MSP file.

Human Cell Sample Preparation and Analysis. Human HAP1 cells (Horizon) were grown in IMDM media (Thermo Fisher Scientific) to 1.5 x 10^5 cells per cm^2 on 10-cm plates. Media was aspirated and cells were washed with phosphate-buffered saline (3 x 5 mL washes). Metabolism was quenched with addition of liquid N_2 (2-5 mL) onto the plates. Cells were then extracted using 1 mL of MeOH:H_2O (80:20, v/v) and transferred into a 1.5 mL microcentrifuge tube. Then a 300 µL aliquot of this extract was transferred to an autosampler vial and dried by vacuum centrifuge (1 hour). The dried metabolites were derivatized using the MSTFA silylation method described previously. Samples were analyzed using a Q Exactive GC mass spectrometer. A temperature gradient ranging from 100° C to 320° C was employed spanning a total runtime of 25 minutes. Analytes were injected onto a 30 m TraceGOLD TG-5SILMS column (Thermo Scientific) using a 1:10 split at a temperature of 275° C and ionized using electron ionization (EI). The mass spectrometer was operated in full scan mode using a resolution of 30,000 (m/Δm) relative to 200 *m/z*. The resulting GC–MS data were processed using an in-house-developed software suite (https://github.com/coongroup/Y3K-Software), described previously.[13]

Results

Contents of the High Resolution Orbitrap GC-MS Library. The High Resolution Orbitrap GC-MS Spectral Library currently contains 625 spectral entries. **Table I** (**Figure 5.1**) outlines the metabolite classification of all spectra in the library (classifications obtained from the Human Metabolome Database, HMDB).[8] These spectra comprise 512 unique metabolites which were hand-selected as being those most commonly found in GC-MS metabolomics experiments, particularly focused on mammalian derived metabolites and those of biomedical importance. There are 517 TMS derivatized spectral entries, 51 TBDMS entries, 30 methoxyaminated entries, and 58 underivatized entries. To aid integration of the library into existing pipelines and in elucidating biological function from identified species, each entry includes an extensive list of metadata. This includes a broad list of common synonyms, metabolite structural information, chemical formula, compound classifiers, and compound identifiers from many commonly used chemical databases.[8,14–17] Additionally, the spectral information is stored in the NIST .MSP text file format for seamless integration into NIST MS Search, AMDIS, and TraceFinder software programs. Spectral entries also include semi-nonpolar retention index values to aid in compound identification as well as all relevant instrument method parameters. An example spectra, cholesterol 1TMS, from the library is shown in **Figure 5.2**. Due to the high mass accuracy and resolution of the spectra, the fragments can be automatically annotated with confidence. In this example, the program Thermo Xcalibur Qual Browser was used to annotate selected fragments with the constraints that at most 4 nitrogen, 10 oxygen, 30 carbon, 60 hydrogen, and 8 silicon atoms may be used. The fragments were all correctly assigned, including the molecular ion, and have a mass error within ±1 part per million. The ability to confidently and reliably annotate the collected spectra can aid in compound identification, and affords the use of high resolution filtering, a scoring metric which takes high resolution spectra and attempts to annotate the fragments using only subsets of the chemical formula of the library match.[18]

Comparison of High Resolution Orbitrap GC-MS Library to NIST Library. Human cell metabolite extracts were analyzed and metabolites were identified using the high resolution Orbitrap library and the NIST EI-MS library, restricted to only silylated compounds. The forward dot product score for each identified compound and each library entry were calculated using Equation 1:

$$\text{Forward score} = \{\Sigma[(m/z)^{\mathrm{n}}(\text{Intensity})^{\mathrm{m}}]_{\mathrm{exp}} * [(m/z)^{\mathrm{n}}(\text{Intensity})^{\mathrm{m}}]_{\mathrm{lib}}\}^{2} \backslash$$

$$\{\Sigma[(m/z)^{\mathrm{n}}(\text{Intensity})^{\mathrm{m}}]_{\mathrm{exp}}{}^{2} * \Sigma[(m/z)^{\mathrm{n}}(\text{Intensity})^{\mathrm{m}}])_{\mathrm{lib}}{}^{2}\} * 100$$

where *m/z* and intensity represent the mass to charge ratio and intensity pair of a fragment, m and n representing the weight factors for intensity and *m/z*, respectively. Additionally, the subscripts “exp” and “lib” denote whether the spectra was experimentally derived or from the library. The dot products were calculated using weight factors, m and n, of 0.53 and 1.3 respectively. These values were confirmed to maximize the score differential of correct identifications and incorrect identifications (**Figure 5.3** panel A) and

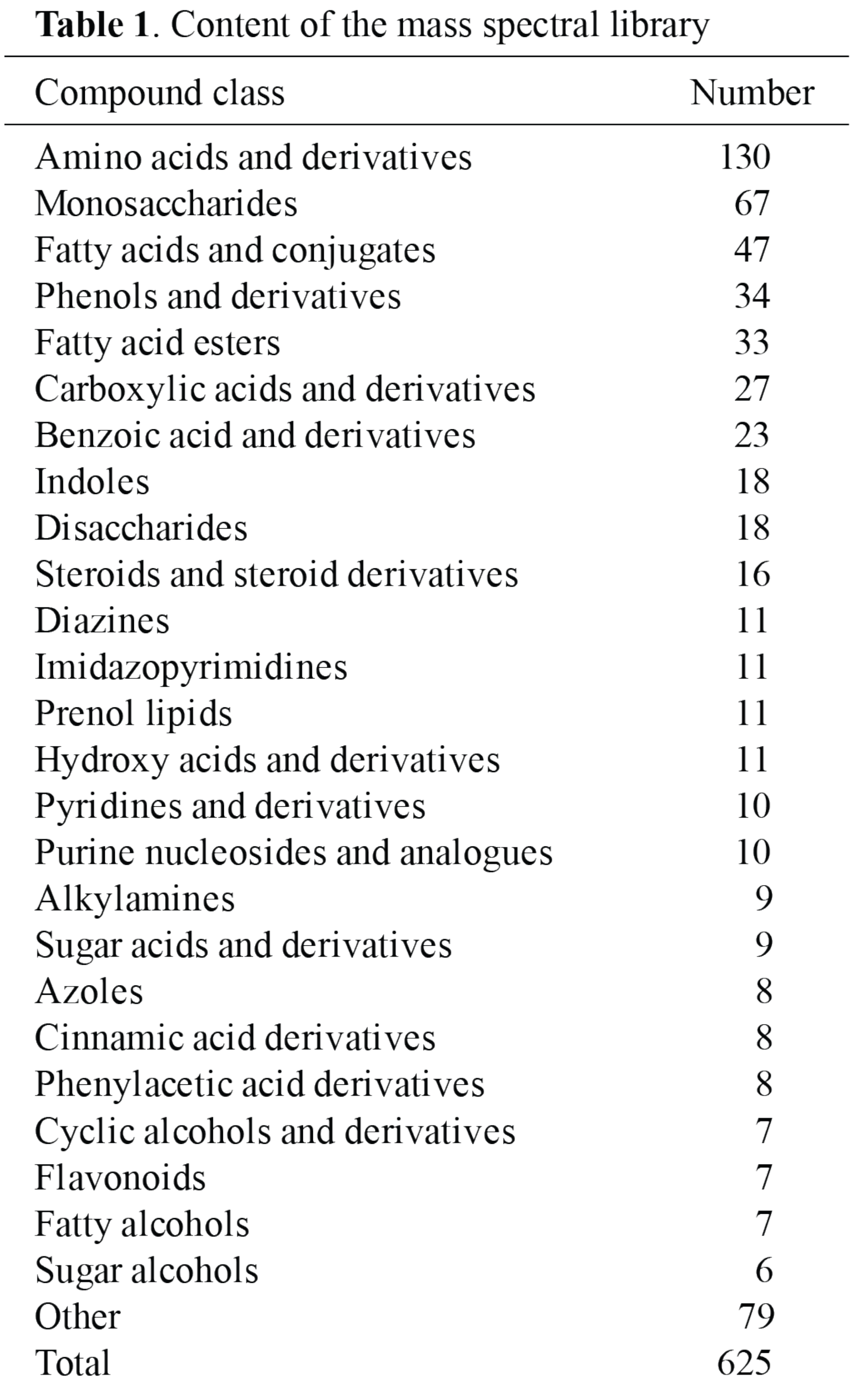


Table 1. Content of the mass spectral library

Compound class	Number
Amino acids and derivatives	130
Monosaccharides	67
Fatty acids and conjugates	47
Phenols and derivatives	34
Fatty acid esters	33
Carboxylic acids and derivatives	27
Benzoic acid and derivatives	23
Indoles	18
Disaccharides	18
Steroids and steroid derivatives	16
Diazines	11
Imidazopyrimidines	11
Prenol lipids	11
Hydroxy acids and derivatives	11
Pyridines and derivatives	10
Purine nucleosides and analogues	10
Alkylamines	9
Sugar acids and derivatives	9
Azoles	8
Cinnamic acid derivatives	8
Phenylacetic acid derivatives	8
Cyclic alcohols and derivatives	7
Flavonoids	7
Fatty alcohols	7
Sugar alcohols	6
Other	79
Total	625

Figure 5.1: Table I. Summary of metabolite classes. List of all metabolite spectra in the library grouped by chemical class

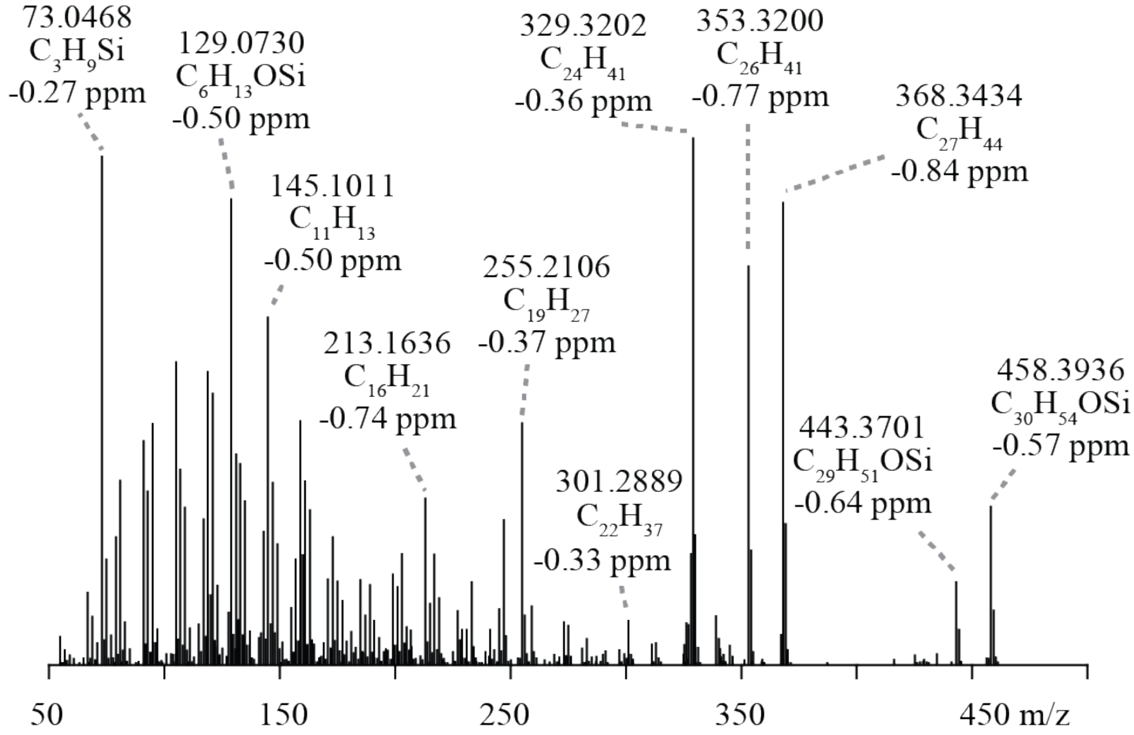


Figure 5.2: High resolution spectra of cholestrol 1 TMS (annotated). The EI spectra for cholesterol 1 TMS is shown. Above selected peaks are the *m/z* value, the automatically assigned chemical composition, and the mass error of the composition assignment, in parts per million.

align with weight factors used in previous studies.[19] Omitting monosaccharides, which generate spectra that are nearly indistinguishable, 71 unique spectral features were manually identified and contained in both the Orbitrap spectral library and in the NIST silylated subset library. Of these identified features, all but one was returned as the best scoring match when searched against the Orbitrap library, while 15 of these compounds were not correctly matched to the top scoring library hit when using the NIST library (**Figure 5.3** panel B). The high resolution orbitrap library also provided a significant boost in library match discrimination. The average forward dot product score between the 71 returned matches and the second best scoring library entry was 42 points (out of a 100 point scale), allowing the analyst to more easily rule out false hits. While this same set using the NIST library gave an average score difference of 9 points (**Figure 5.3** panel C). This benefit can likely be ascribed to the specificity of the high resolution library, containing only those species that are likely to be found in metabolomics experiments. While conversely, the breadth of the NIST library is beneficial for identifying less common or unexpected metabolites found in the sample. **Figure 5.3**, panels D and E, illustrate this difference in library matching. The deconvolved spectrum manually identified as uracil 2TMS is correctly returned as the top scoring library match using the Orbitrap library and the second match is that of thymine 2TMS, yielding a forward dot product score differential of 46.1 points and easily allowing the analyst to correctly assign the metabolite identity (**Figure 5.3** panel D). While that same deconvolved spectrum, when searched against the NIST library returns 4,6-dihydroxypyrimidine 2TMS, a structurally similar yet non-metabolic compound, as the top scoring entry, and the correct identification is returned as the second match (**Figure 5.3** panel E). Additionally, these two compounds nearly coelute, with retention index values within 10 units, further challenging the analyst when attempting to identify this feature. A particular spectral difference when comparing the high resolution library spectra of uracil 2TMS to that of the NIST uracil 2TMS spectra is that the low *m/z* ions (less than 100 *m/z*), represent a smaller percentage of the total ion current in the high resolution spectra (**Figure 5.3** D and E), resulting in a decrease in forward dot product score. We have noticed that, commonly, ions below 100 *m/z* tend to be less intense for the Orbitrap collected spectra as compared to unit-resolution library spectra, likely collected using quadrupole mass analyzers.

Conclusion

The high resolution Orbitrap GC-MS metabolite spectral library combined with the improved scoring formula offers exceptional improvement in metabolite identification using the same throughput techniques previously established in GC-MS metabolomics. Contributing to this advancement was the high resolution Orbitrap GC-MS metabolite spectral library, which offers substantial coverage of GC-amenable metabolites involved in metabolism. The standard spectra generated on this system, unsurprisingly, have higher spectral similarity to features analyzed using this same system, as compared to spectral libraries that were compiled on different GC-MS systems. We note that one reason for this difference in dot product scoring is that the Orbitrap-collected spectra have less intense low mass ions. Overall, the

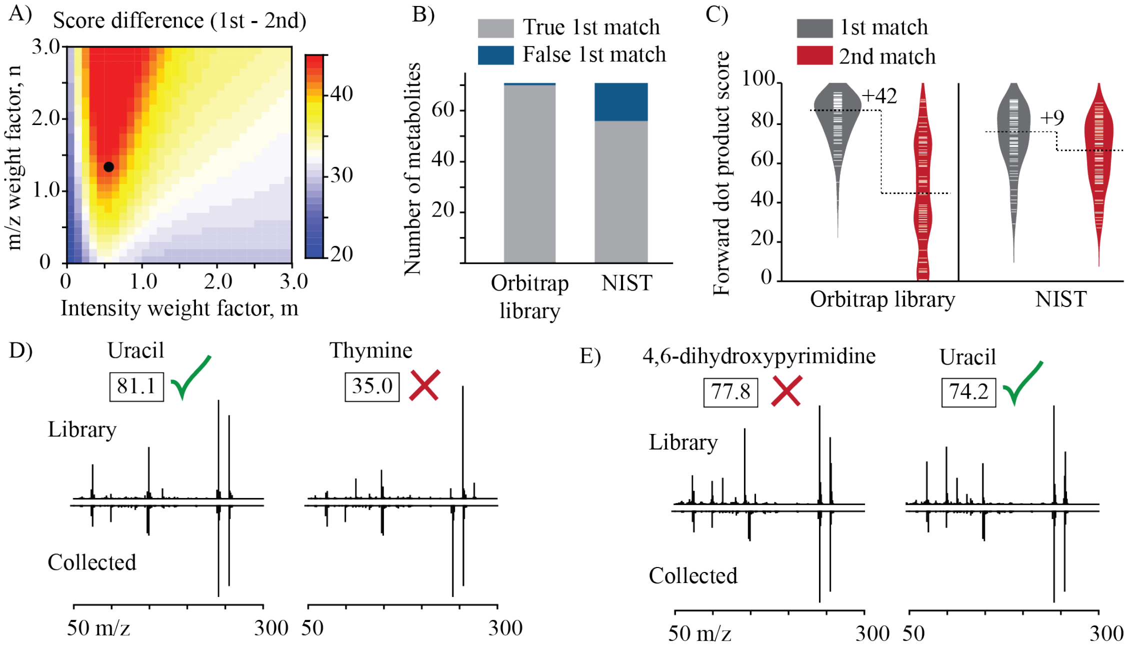


Figure 5.3: Comparison of Orbitrap and NIST libraries. Panel A shows a heat map of the average forward dot product score difference between the first and second library match returned from a set of 71 manually verified metabolites identified from human cells. The chosen, optimal, weight factors for the dot product calculation was found to be 0.53 and 1.3 for the m and n weight factors respectively (denoted as a black dot on the plot). Panel B shows the number of metabolites of the 71 confirmed species that were returned as the top scoring match using either the Orbitrap library or the NIST silylated subset library. Panel C shows the distribution of scores between the first and second returned library hit using either library, and the average score difference between the top two matches. Panel D and E compare the top two returned library entries for the metabolite uracil 2TMS using the Orbitrap library (D) and NIST library (E).

spectral matching of feature groups collected on the Q Exactive GC system is significantly improved when matching to the high resolution library, and metabolite identifications can be made with more confidence when compared to using a unit-resolution library, such as NIST.

References

[1] G. J. Patti, O. Yanes, and G. Siuzdak, “Metabolomics: The apogee of the omics trilogy,” *Molecular Cell Biology*, vol. 13, pp. 263–269, 2012.

[2] W. B. Dunn, A. Erban, R. J. M. Weber, D. J. Creek, M. Brown, R. Breitling, T. Hankemeier, R. Goodacre, S. Neumann, J. Kopka, and M. R. Viant, “Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics,” *Metabolomics*, vol. 9, no. SUPPL.1, pp. 44–66, 2013.

[3] E. Kováts, “Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone,” *Helvetica Chimica Acta*, vol. 41, no. 7, pp. 1915–1932, 1958.

[4] “NIST Mass Spectral Library,” 2017.

[5] “Wiley Registry of Mass Spectral Data,” 2010.

[6] J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmüller, P. Dörmann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A. R. Fernie, and D. Steinhauser, “GMD@CSB.DB: The Golm metabolome database,” *Bioinformatics*, vol. 21, no. 8, pp. 1635–1638, 2005.

[7] T. Kind, G. Wohlgemuth, D. Y. Lee, Y. Lu, M. Palazoglu, S. Shahbaz, and O. Fiehn, “FiehnLib: Mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry,” *Analytical Chemistry*, vol. 81, no. 24, pp. 10038–10048, 2009.

[8] D. S. Wishart, C. Knox, A. C. Guo, R. Eisner, N. Young, B. Gautam, D. D. Hau, N. Psychogios, E. Dong, S. Bouatra, R. Mandal, I. Sinelnikov, J. Xia, L. Jia, J. A. Cruz, E. Lim, C. A. Sobsey, S. Shrivastava, P. Huang, P. Liu, L. Fang, J. Peng, R. Fradette, D. Cheng, D. Tzur, M. Clements, A. Lewis, A. De souza, A. Zuniga, M. Dawe, Y. Xiong, D. Clive, R. Greiner, A. Nazyrova, R. Shaykhutdinov, L. Li, H. J. Vogel, and I. Forsythei, “HMDB: A knowledgebase for the human metabolome,” *Nucleic Acids Research*, vol. 37, no. SUPPL. 1, pp. 603–610, 2009.

[9] S. E. Stein, "An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data," *Journal of the American Society for Mass Spectrometry*, vol. 10, no. 8, pp. 770–781, 1999.

[10] A. C. Peterson, G. C. McAlister, S. T. Quarmby, J. Griep-Raming, and J. J. Coon, "Development and characterization of a GC-enabled QLT-Orbitrap for high-resolution and high-mass accuracy GC/MS.," *Analytical chemistry*, vol. 82, pp. 8618–28, oct 2010.

[11] A. C. Peterson, J. P. Hauschild, S. T. Quarmby, D. Krumwiede, O. Lange, R. A. S. Lemke, F. Grosse-Coosmann, S. Horning, T. J. Donohue, M. S. Westphall, J. J. Coon, and J. Griep-Raming, "Development of a GC/quadrupole-orbitrap mass spectrometer, Part I: Design and characterization," *Analytical Chemistry*, vol. 86, no. 20, pp. 10036–10043, 2014.

[12] A. C. Peterson, A. J. Balloon, M. S. Westphall, and J. J. Coon, "Development of a GC/quadrupole-orbitrap mass spectrometer, part II: New approaches for discovery metabolomics," *Analytical Chemistry*, vol. 86, no. 20, pp. 10044–10051, 2014.

[13] J. A. Stefely, N. W. Kwiecien, E. C. Freiberger, A. L. Richards, A. Jochem, M. J. 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," *Nature Biotechnology*, vol. 34, no. 11, pp. 1191–1197, 2016.

[14] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: Back to metabolism in KEGG," *Nucleic Acids Research*, vol. 42, no. D1, pp. 199–205, 2014.

[15] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, and C. Steinbeck, "ChEBI in 2016: Improved services and an expanding collection of metabolites," *Nucleic Acids Research*, vol. 44, no. D1, pp. D1214–D1219, 2016.

[16] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. H. Bryant, "PubChem substance and compound databases," *Nucleic Acids Research*, vol. 44, no. D1, pp. D1202–D1213, 2016.

[17] D. Weininger, "SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[18] N. W. Kwiecien, D. J. Bailey, M. J. P. Rush, J. S. Cole, A. Ulbrich, A. S. Hebert, M. S. Westphall, and J. J. Coon, "High-Resolution Filtering for Improved Small Molecule Identification via GC/MS," *Analytical Chemistry*, vol. 87, no. 16, pp. 8328–8335, 2015.

[19] S. Kim, I. Koo, X. Wei, and X. Zhang, "A method of finding optimal weight factors for compound identification in gas chromatography-mass spectrometry," *Bioinformatics*, vol. 28, no. 8, pp. 1158–1163, 2012.

Chapter 6

MITOCHONDRIAL PROTEIN FUNCTIONS ELUCIDATED BY MULTI-OMIC MASS SPECTROMETRY PROFILING

This chapter has been published and is reprinted with permission from:

Stefely JA*, Kwiecien NW*, Freiberger EC, Richards AL, Jochem A, **Rush MJP**, Ulbrich A, Robinson KP, Hutchins PD, Veling MT, Guo X, Kemmerer ZA, Connors KJ, Trujillo EA, Sokol J, Marx H, Westphall MS, Hebert AS, Pagliarini DJ, Coon JJ. *Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling.* Nature Biotechnology. **2016**, doi:10.1038/nbt.3683

* Authors contributed equally

Contributions

The following paper describes the mulit-omic analysis of a set of 174 yeast strains with a single gene deletion which encodes for a protein related to mitochondrial biology. Each strain was grown in triplicate under two growth conditions, a fermentation condition and a respiration condition. This cohort was selected to contain 39 strains with genes of unknown function, with one overarching goal being the association of the genes of unknown function to a gene of known function to elucidate its role in mitochondrial biology. This was to be accomplished by measuring as many biomolecules as possible which would, taken together, map the phenotype of this strain. To achieve this goal, the team endeavored to perform three discovery-based mass spectrometry methods to analyze separately the proteome, lipidome, and metabolome of extracts from each strain.

My contribution to this study began by developing the sample preparation and instrument method for the discovery metabolomics assay. Our lab had not ever performed this type of experiment before, and the instrument that we planned to use, the Q Exactive GC, had not yet been commercially released. The sample preparation portion of the development required optimizing the concentration of metabolites to maximize the number of detected metabolites while not overloading the GC column or saturating the detector. The derivatization method I chose was silylation, which is near-universally used for discovery metabolomics assays. This derivatization method broadly reacts with polar metabolites by reacting with acidic protons on amines, alcohols, carboxylic acids, phosphates, and thiols, making these molecules more volatile, and therefore, GC-amenable. The next step in the method development process was creating the instrument method. One facet of GC-MS that is often not recognized is the importance and diversity of the sample injection and introduction onto the column. The two most common modes of injection are split and splitless injection modes. In splitless injections, all of the withdrawn sample volume is introduced to the GC column. In split injections, a majority of the sample is syphoned to waste, with only a fractional amount input onto the head of the GC column. There are four primary reasons I chose to use a split injection for this method. First, the split method yielded much narrower peak shape than the splitless method. This is because the sample is loaded onto the column more quickly, and in a narrower band than the splitless mode. This is extremely beneficial in samples which are very complex, as it provides greater resolution of peaks that would otherwise not be resolvable with wider elution profiles. Secondly, splitless mode produced a much more symmetric peak shape. In splitless mode, early eluting peaks front because the sample injection process takes longer, and an analyte beings to elute before being fully injected onto the column. This is related to the next benefit, which is that the solvent front in split injections are dramatically reduced. In a typical splitless injection, the solvent front extends much further into the gradient and can mask analytes which elute near the beginning the gradient. Lastly, while splitless injections provided greater sensitivity by increasing metabolite signal, we were not sample limited, and were able to concentrate the sample in previous sample preparation steps to increase metabolite signal to a sufficient level in the split mode. One last instrument method decision was to collect the data at resolution 30,000 (m/Δm) relative to 200 *m/z*. This resolution allowed for sufficient scans to be collected, approximately 14 scans per second, which provided enough data points to effectively deconvolve the 2 to 3 second wide GC peaks.

Once the method was developed, I began analyzing the approximately 1300 samples. This includes wild-type yeast controls, solvent blanks, and instrument suitability controls to ensure the data quality was comparable across the analysis. In order to mitigate system error caused by instrument deviation and chromatographic drift, the samples were run consecutively over the span of 40 days. Each day 33 yeast extracts were prepared and analyzed, comprising three replicates of eleven separate yeast strains, with one set of three being a wild-type control. The samples were randomized and three solvent blanks were analyzed at the beginning of the set, in the middle, and at the end. Each analysis required 35 minutes of instrument time from one injection to the next, totaling 21 hours of analysis time each day. In the 3 hour down time between days, the instrument required calibration, any necessary maintenance, and each sample had to be examined to ensure no injections were missed or contained anomalous features.

Concurrent with the sample preparation and instrumentation method development, Nick Kwiecien began developing the data analysis pipeline. This pipeline in its current state is outlined in Chapter 1. However, at the time, we were not sure what the correct data analysis steps would be. Nick had already written a deconvolution algorithm to extract and group GC-MS data. We then discussed what other components would be critical, such as the need to group like features between files (what ended up becoming GC Quant) and the need to visually inspect feature groups, background feature groups, and then search these groups against library spectra (what became GC Results Viewer). My role was to be the primary beta-tester of this software package and to suggest modifications to allow for greater usability.

The most laborious aspect to the metabolomics analysis was annotating the measured metabolites. Each feature group was manually interpreted as being present in the background sample or not. This entailed examining 1000 feature groups per growth condition and comparing them to background samples. Next, those features which were not in the background (411 in total) were searched against the 250,000 spectra in the NIST mass spectral library, with each feature group returning dozens of potential matches. However as the GC-MS system used was brand new the fragmentation patterns of public mass spectral libraries were not identical to those collected on this mass spectrometer. And therefore to ensure the correct identifications were made dozens of authentic standards were acquired and analyzed to compare to the data to confirm metabolite identifications, initiating the development of our in-house spectral library, described in Chapter 5. A major finding of this manuscript is the description of the Hfd1 (and the human homolog ALDH3A1) protein's role in coenzyme Q (CoQ) biosynthesis. The effected pathway is the conversion of tyrosine to 4-hydroxybenzoic acid, which is an intermediate in the generation of CoQ. Initially the only metabolite in this pathway which was able to be confidently identified in the samples was tyrosine. This is due to it being present at abundances orders of magnitude greater than the other metabolites in this pathway. Fortunately as the assay was performed in a discovery fashion and GC-EI-MS experiments are data-independent I was able to acquire authentic standards and attempt to find the rest of these metabolites in the dataset. Using this approach p-coumaric acid, para-aminobenzoic acid, 4-hydroxybenzoic acid, and 4-hydroxybenzaldehyde were all able to be detected, confidently identified, and quantified within the data. This further validated the proposed mechanism of Hfd1 as the responsible enzyme for the conversion of 4-hydroxybenzaldehyde to 4-hydroxybenzoic acid. In total 350,000 metabolite measurements were made in this experiment making it one of the largest metabolic profiling of a yeast knockout library to date.

The resulting publication is reprinted here in its entirety to illustrate the scope and impact of the research. Facilitated by Dr. Josh Coon and Dr. Dave Pagliarini and lead by Jon Stefely and Nick Kwiecien, this manuscript describes not only important biochemical findings related to mitochondrial biology but also outlines a method of multi-omic analysis which can be used to potentially uncover the function of genes and proteins in any system. I was very fortunate to work on this project and view this as a perfect example of how academic research can effectively work to answer challenging questions. Where a group of researchers can come together and accomplish something far greater than any one person's knowledge, time, and skillset would allow. Seeing how dedicated Jon and Nick were to this work inspired me in my other research projects and I am a much better scientist because of this experience.

Abstract

Mitochondrial dysfunction is associated with many human diseases, including cancer and neurodegeneration, that are often linked to proteins and pathways that are not well-characterized. To begin defining the functions of such poorly characterized proteins, we used mass spectrometry to map the proteomes, lipidomes and metabolomes of 174 yeast strains, each lacking a single gene related to mitochondrial biology. 144 of these genes have human homologs, 60 of which are associated with disease and 39 of which are uncharacterized. We present a multi-omic data analysis and visualization tool that we use to find covariance networks that can predict molecular functions, correlations between profiles of related gene deletions, gene-specific perturbations that reflect protein functions, and a global respiration deficiency response. Using this multi-omic approach, we link seven proteins including Hfd1p and its human homolog ALDH3A1 to mitochondrial coenzyme Q (CoQ) biosynthesis, an essential pathway disrupted in many human diseases. This Resource should provide broad molecular insights into mitochondrial protein functions.

Introduction

High resolution mass spectrometry (MS) has become the primary analysis tool for many classes of biomolecules, including proteins, metabolites, and lipids. Major advancements in MS technology—particularly in the rate and depth of analysis—have enabled dozens of proteomes, metabolomes, and lipidomes to be analyzed in a single day[1–3]. Studies of bacteria demonstrated that parallel measurement of multiple molecule classes can synergistically enhance the biological insight afforded[4,5]. Recently, proteomics has been integrated with transcriptomics and genomics in mice[6,7]. However, large-scale, comprehensive (i.e., proteome-wide), multi-omic data acquisition, integration, and visualization tools remain underdeveloped, often lagging behind genomics in terms of coverage, speed, and broad accessibility for end users. Given the interdependence of proteins, lipids, and metabolites, we reasoned that coordinated analysis across all three biomolecule classes could afford new insight into eukaryotic biology. In particular, we hypothesized that this multi-omic profiling strategy, when coupled with genetic and environmental perturbations, could enable functional predictions for uncharacterized proteins.

We applied this strategy to study mitochondria, dynamic organelles whose dysfunction is associated with over 150 human diseases including cancer, diabetes, Parkinson's, and numerous genetic disorders[8–10]. While the yeast and mammalian mitochondrial proteomes were recently defined[11–13], functional annotation of these proteins lags behind[14], impeding biomedical research on the many diseases impacted by mitochondrial metabolism. Of the

~1,200 mammalian mitochondrial proteins, nearly 300 are "mitochondrial uncharacterized (x) proteins" (MXPs)[15,16] that have no well-established biochemical function within mitochondria. Here, toward defining functions for MXPs, we performed over 3,000 MS experiments in parallel to analyze the proteomes, metabolomes, and lipidomes of 174 single-gene deletion ("Δ*gene*") Saccharomyces cerevisiae yeast strains in biological triplicate across two metabolic conditions, fermentation and respiration **(Fig. 6.1a)**. To facilitate development of biological hypotheses based on the resultant "yeast-three-thousand (Y3K)" dataset **(Fig. 6.1b)**, we also developed a multi-omic data visualization approach (highlighted in **Fig. 6.1c** and online at http://y3kproject.org/). Our data establish many new connections between MXPs and proteins with well-established functions by virtue of gene-specific phenotypes or shared global biomolecular changes that result from the loss of each protein's expression. We leveraged a subset of these connections to address the incomplete mitochondrial pathway that generates ubiquinone (coenzyme Q, CoQ), an essential lipid required for oxidative phosphorylation (OxPhos) and linked to diseases ranging from severe infantile multisystemic disease to isolated myopathy and aging[17,18].

Results

Multi-omic mass spectrometry profiling. The 174 Δ*gene* yeast strains we analyzed covered 124 characterized genes that were selected to span a broad range of pathways to assist functional mapping, and 50 uncharacterized genes that encode MXPs **(Fig. 6.1a and Supplementary Fig. S6.1a)**. In selecting these targets, we prioritized genes with human homologs (144/174 genes) and those associated with disease (60/144 genes) based on primary literature analysis and online database gene annotation (e.g., omim.org). Inclusion of characterized genes, some of which could be considered as only partially characterized, also provided the ability to connect them to previously unrecognized functions. Each strain was grown in biological triplicate under two contrasting growth conditions, a standard fermentation culture condition and a carefully optimized respiration culture condition that stimulates mitochondrial function **(Fig. 6.1a, Supplementary Fig. S6.1b–e, and Supplementary Note 1)**—yielding six separate cultures per yeast strain.

Altogether we grew more than 1,050 yeast cultures (including WT cultures), each of which was analyzed using three separate high-resolution MS-based proteomic, metabolomic, and lipidomic techniques. These 3,000+ MS experiments yielded quantitation of 4,040 proteins, 411 metabolites, and 53 lipids (averaging 3,180 proteins, 252 metabolites, and 53 lipids per culture)—over 3.5 million biomolecule measurements in total **(Fig. 6.1a and Supplementary Fig. S6.2a,b)**. Key to our approach was streamlining procedures for proteome extraction and preparation to under two hours of hands-on time **(Supplementary Fig. S6.2c)**. Use of label-free quantitation negated the need for a chemical tagging step and further increased throughput. We observed a wide dynamic range across all profiled omes, with some molecule abundances spanning more than three orders of magnitude **(Supplementary Fig. S6.2d)**. Additionally, we observed remarkable reproducibility between replicate cultures, with a median coefficient of variation of 12.7% considering all profiled biomolecules, and high overlap of molecules quantified across cultures **(Supplementary**

Fig. S6.2e–g).

A high-level view of the Y3K dataset shows significant perturbations across all three omes, with more pronounced perturbations in respiration **(Fig. 6.1b and Supplementary Fig. S6.3a)**. Hierarchical clustering revealed groups of functionally related molecules (along the y-axis) and groups of functionally related Δ*gene* strains (along the x-axis). Protein clusters show significant gene ontology (GO) term enrichments for diverse processes and include both characterized and uncharacterized proteins **(Supplementary Fig. S6.3b)**. For example, the uncharacterized proteins Esbp6p and Ypr010c-a cluster with proteins involved in mitochondrial ATP synthesis and electron transport chain function, respectively **(Supplementary Fig. S6.4)**. Here, we leverage analyses from three different vantage points, each of which can be recapitulated with our online data visualization suite, exploiting unique biological perspectives afforded by a multi-omic dataset of diverse genetic perturbations **(Fig. 6.1c)**.

Identification of gene-specific phenotypes. First, we systematically surveyed the Y3K dataset for significant molecule perturbations unique to just one or two of the strains in the study **(Fig. S6.2a)**. This unbiased search revealed 714 Δ*gene*-specific phenotypes **(Fig. 6.2a and Supplementary Note 2)**, which can reveal functional relationships. For example, the electron transfer flavoprotein (ETF) subunit Aim45p was uniquely decreased in just two Δ*gene* strains: the Δ*aim45* strain, and the Δ*cir1* strain, which lacks the second ETF heterodimer subunit **(Fig. 6.2b)**. Numerous additional Δ*gene*-specific phenotypes were

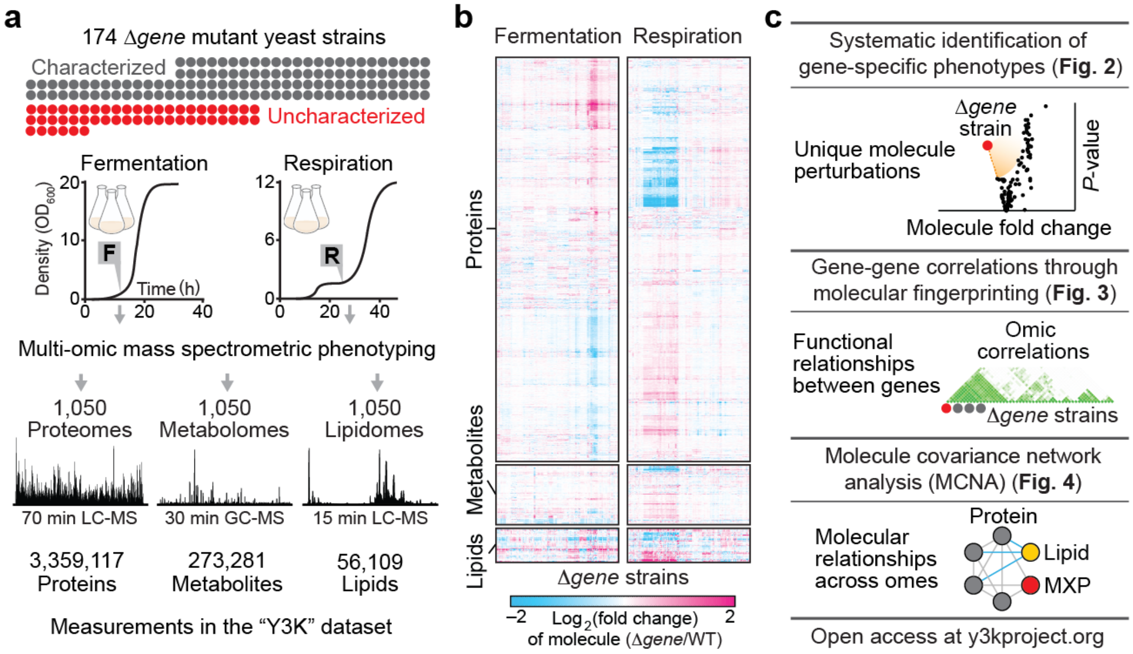
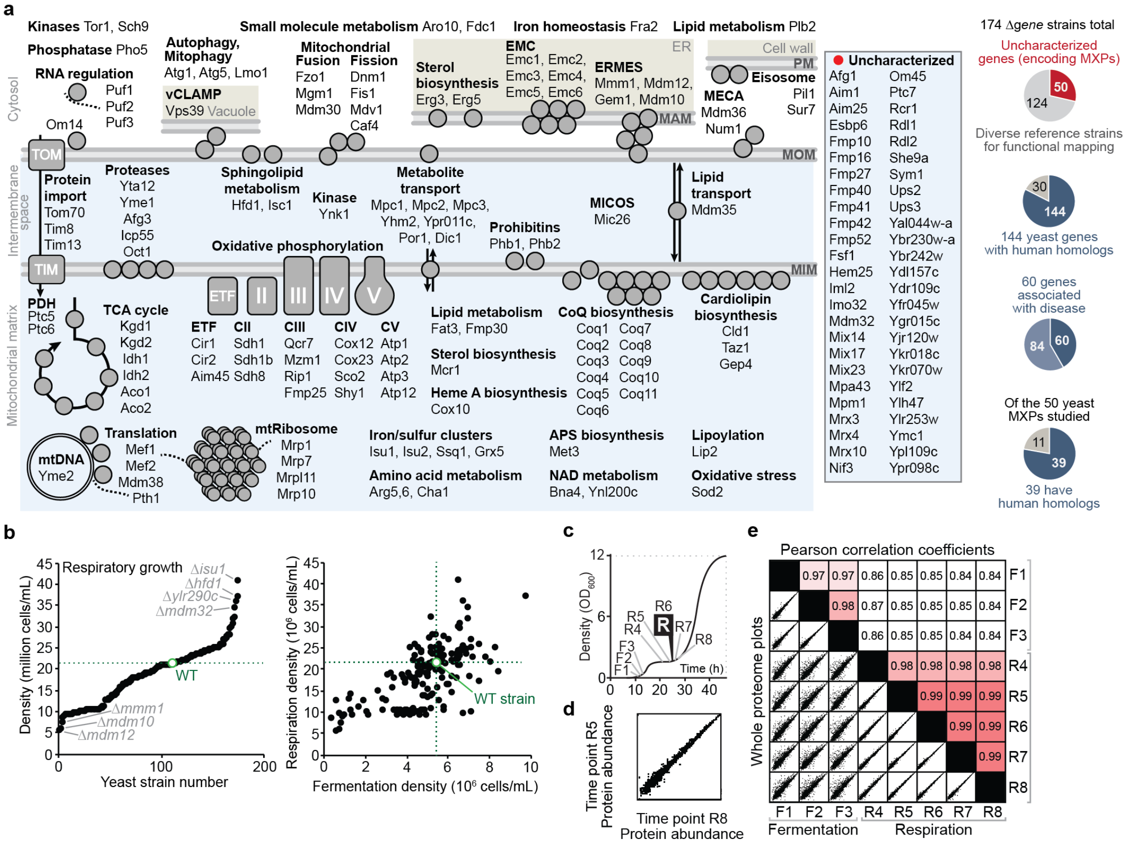


Figure 6.1: Multi-omic mass spectrometry profiling and data visualization. Multi-omic mass spectrometry profiling and data visualization. Overviews of **(a)** the experimental design and high resolution quantitative MS analysis, **(b)** the Y3K dataset, shown as hierarchical clusters of Δ*gene* strains and significantly perturbed molecules (relative abundances compared to WT as quantified by MS, mean, n = 3; $P < 0.05$, two-sided Student's t-test), and **(c)** the multi-omic data analysis and visualization tools developed here.



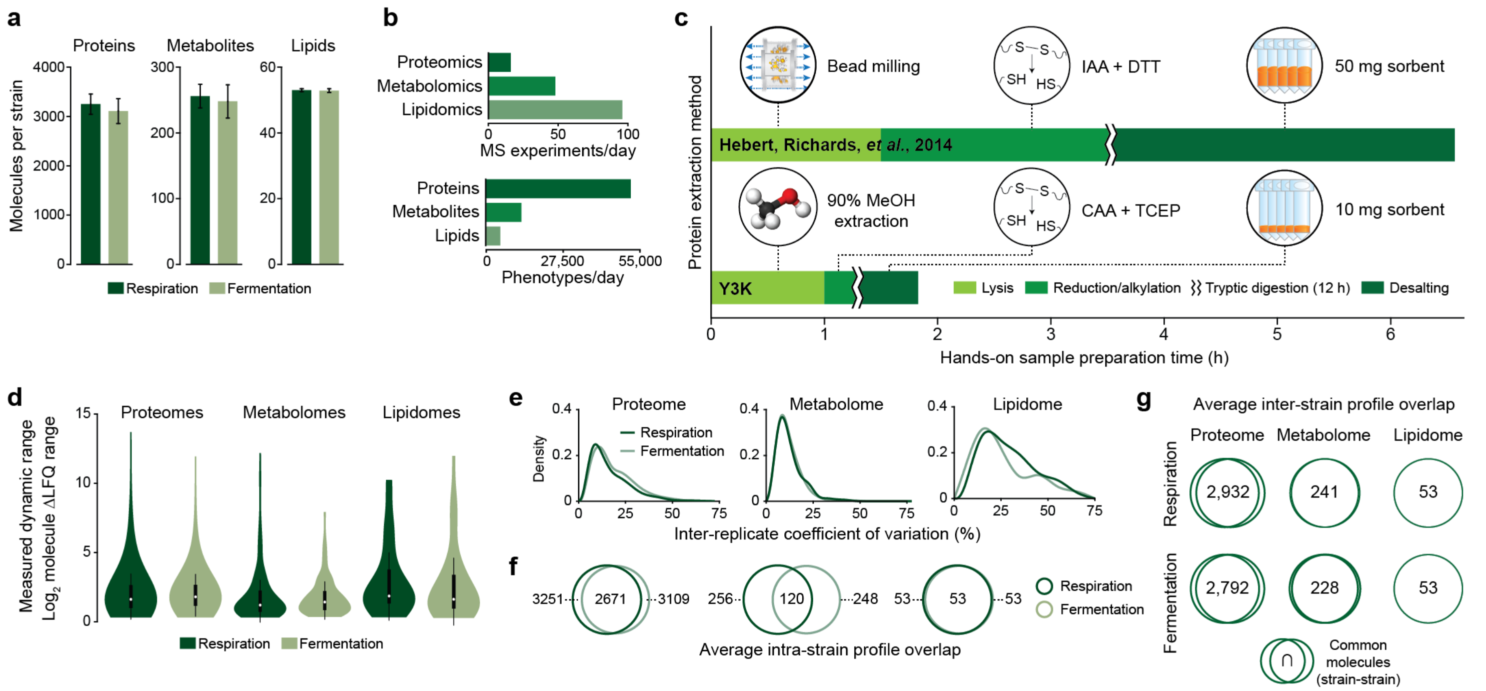
	F1	F2	F3	R4	R5	R6	R7	R8
F1		0.97	0.97	0.86	0.85	0.85	0.84	0.84
F2			0.98	0.87	0.85	0.85	0.85	0.84
F3				0.86	0.85	0.85	0.84	0.84
R4					0.98	0.98	0.98	0.98
R5						0.99	0.99	0.99
R6							0.99	0.99
R7								0.99
R8								

Supplementary Figure S6.1: Δ*Gene* target strain characteristics and respiration culture optimization. (a) Proteins encoded by the individual genes knocked out of the 174 yeast strains investigated in this study, shown in the context of biological pathways. APS, adenosine-5′-phosphosulfate; CII–CV, oxidative phosphorylation complexes II–V; ER, endoplasmic reticulum; EMC, ER membrane complex; ERMES, ER-mitochondria encounter structure; ETF, electron transfer flavoprotein complex; MAM, mitochondria-associated membrane; MECA, mitochondria-ER-cortex anchor; MICOS, mitochondrial contact site and cristae organizing system; MIM, mitochondrial inner membrane; MOM, mitochondrial outer membrane; mtDNA, mitochondrial DNA; mtRibosome, mitochondrial ribosome; NAD, nicotinamide adenine dinucleotide; PDH, pyruvate dehydrogenase; TCA, tricarboxylic acid cycle; vCLAMP, vacuole and mitochondria patch. The pie charts show the total number of characterized and uncharacterized genes profiled (top); the total number of profiled genes that have human homologs (upper middle); of these genes with human homologs, the number of profiled genes that are also associated with disease (lower middle); and of the uncharacterized genes profiled, the number of genes that have human homologs (bottom).

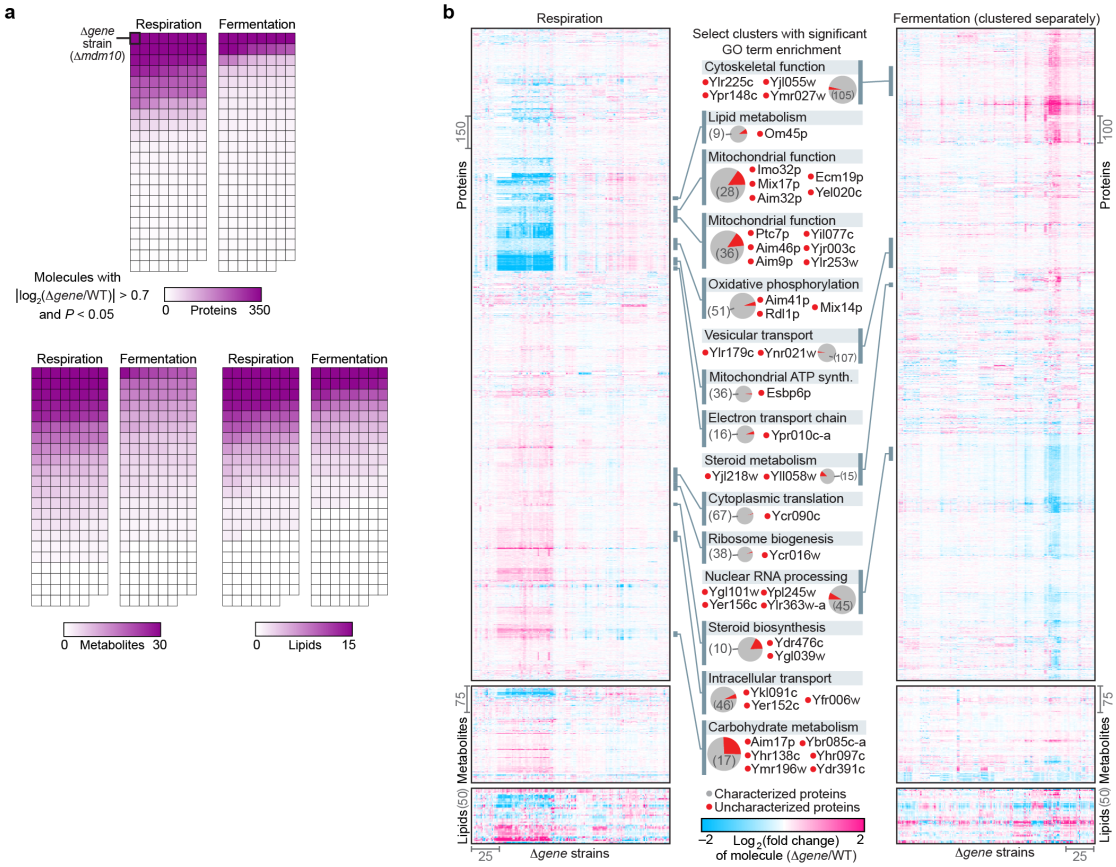
Supplementary Figure S6.1: (b) Density of yeast cultures in the respiratory growth condition (mean, n = 3) plotted in strain rank order (left) or against fermentation culture density (mean, n = 3) (right). **(c)** Optical density at 600 nm (OD_{600}) of yeast cultures (media with 3% [w/v] glycerol and 0.1% [w/v] glucose) indicating time points at which yeast were harvested during fermentation (F1–F3) or respiration (R4–R8). Time point R6 (25 h) was selected for the respiration culture condition of the larger study. **(d)** Whole-proteome plot of protein abundances at time points R5 and R8. **(e)** Pairwise whole proteome plot comparisons (as in d) across all eight time points (lower left) and linear regression analysis of each comparison (r^2, Pearson correlation coefficients) (upper right).

used to generate biological hypotheses **(Supplementary Figs. S6.5 and S6.6)**. We decided to investigate one of these observations at biochemical depth: a Δ*hfd1*-specific decrease in 4-hydroxybenzoate (4-HB), the CoQ headgroup precursor **(Fig. 6.2c)**.

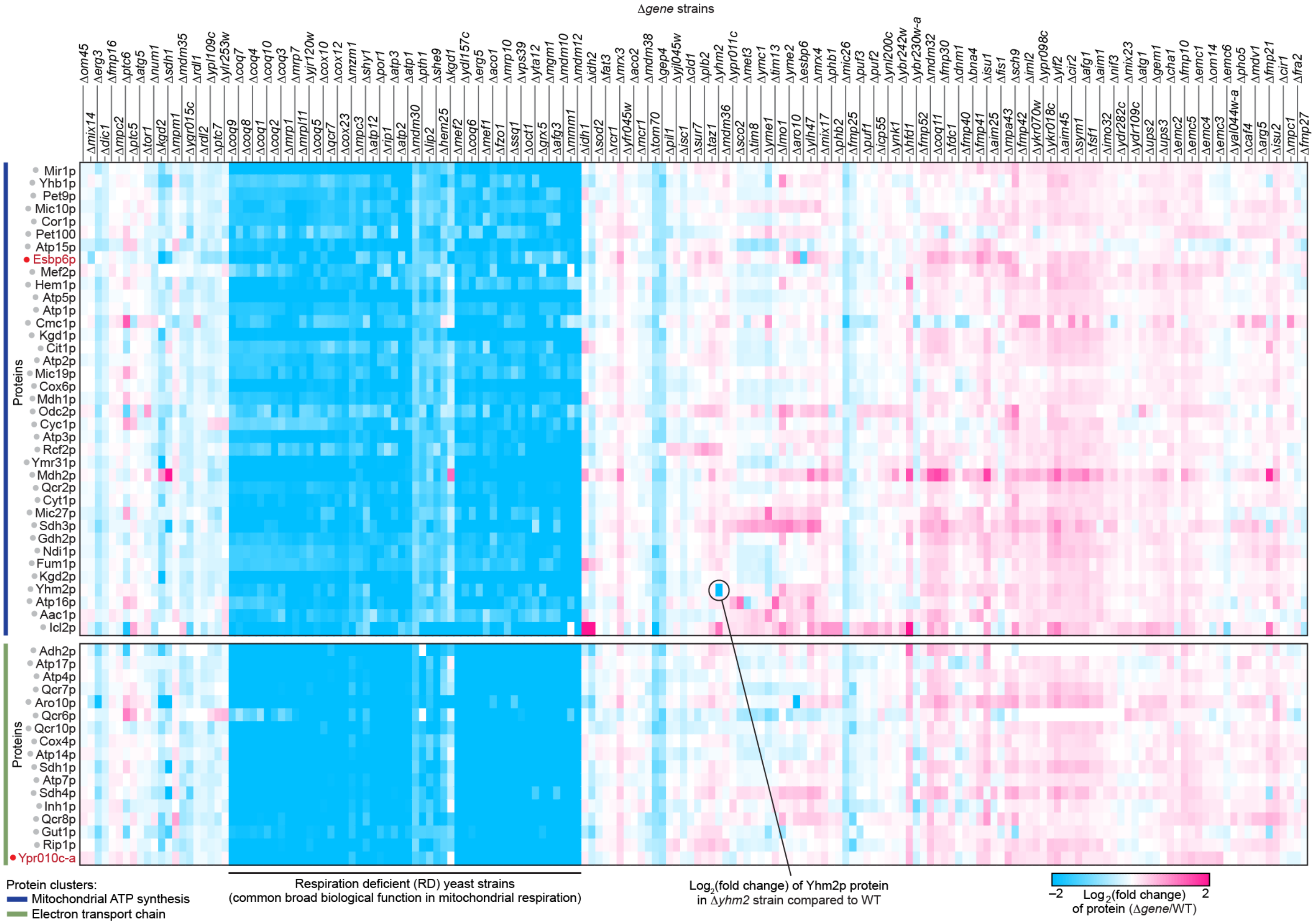
Though it has been known for decades that mammals can convert tyrosine (Tyr) into 4-HB for CoQ biosynthesis[19,20], the biochemical pathway has remained undefined in mammals and yeast **(Fig. 6.2c)**. The Y3K dataset reveals Δ*hfd1* yeast to be significantly deficient in both the metabolite 4-HB ($P < 0.001$) and the lipid CoQ intermediate 3-polyprenyl-4-hydroxybenzoate (PPHB) ($P < 10^{-5}$) **(Fig. 6.2c and Supplementary Fig. S6.7a)**. Despite the PPHB deficiency, Δ*hfd1* yeast have normal CoQ abundance **(Fig. 6.2c)**, likely because of increased flux through an alternative para-amino-benzoate (pABA)- dependent CoQ pathway[21,22], as suggested by elevation of the aminated analog of PPHB (PPAB) in Δ*hfd1* yeast **(Fig. 6.2c)**. This is in contrast to terminal CoQ biosynthesis genes (*coq3*–*coq9*), and some genes not previously linked to CoQ function (e.g. *oct1* and *fzo1*), whose deletion causes significant ($P < 0.05$) CoQ deficiency and accumulation of PPHB **(Fig. 6.2c)**. Because Hfd1p is predicted to be an aldehyde dehydrogenase[23], we hypothesized that it catalyzes



Supplementary Figure S6.2: Mass spectrometry analysis metrics and quality assessment. **(a)** Proteins, lipids, and metabolites quantified per Δ*gene* strain (mean ± s.d., n = 3). **(b)** MS experiments conducted per day (top) and phenotypes (molecules) quantified per day (bottom) for proteomics, lipidomics, and metabolomics. **(c)** Overview of the yeast protein extraction method optimized for this study compared to previous work. **(d)** Violin plots depicting the range of fold changes in molecule abundance ($\log_2[\Delta gene/\mathrm{WT}]$) across all molecule classes and metabolic states. **(e)** Density plots of the distribution of coefficients of variation (CVs) (%) for each molecule measured in biological triplicate across all mutants and growth conditions. **(f)** Venn diagrams depicting the average overlap of molecules quantified within individual Δ*gene* strains across fermentation and respiration growth conditions. **(g)** Average profile overlap between different Δ*gene* strains.



Supplementary Figure S6.3: Features of protein-lipid-metabolite perturbation profiles. (a) Heat maps depicting the number of molecules significantly perturbed within each Δ*gene* strain ($P < 0.05$; two-sided Student's t-test). **(b)** Hierarchical clusters of Δ*gene* strains and significantly perturbed molecules (relative abundances compared to WT quantified by MS; $P < 0.05$; two-sided Student's t-test). The center column annotates select clusters with significant functional (GO term) enrichments ($P < 0.05$; Fisher's exact test followed by Benjamini-Hochberg FDR correction for multiple hypothesis testing). Pie charts indicate proteins in clusters encoded by characterized (gray) or uncharacterized (red) genes.



Supplementary Figure S6.4: Expanded view of two protein clusters from the respiration Y3K dataset heat map (respiration profiles). Heat map indicates relative abundance of proteins in Δ*gene* strains compared to WT as quantified by MS. See **Supplementary Figure S6.3** for the full heat map.

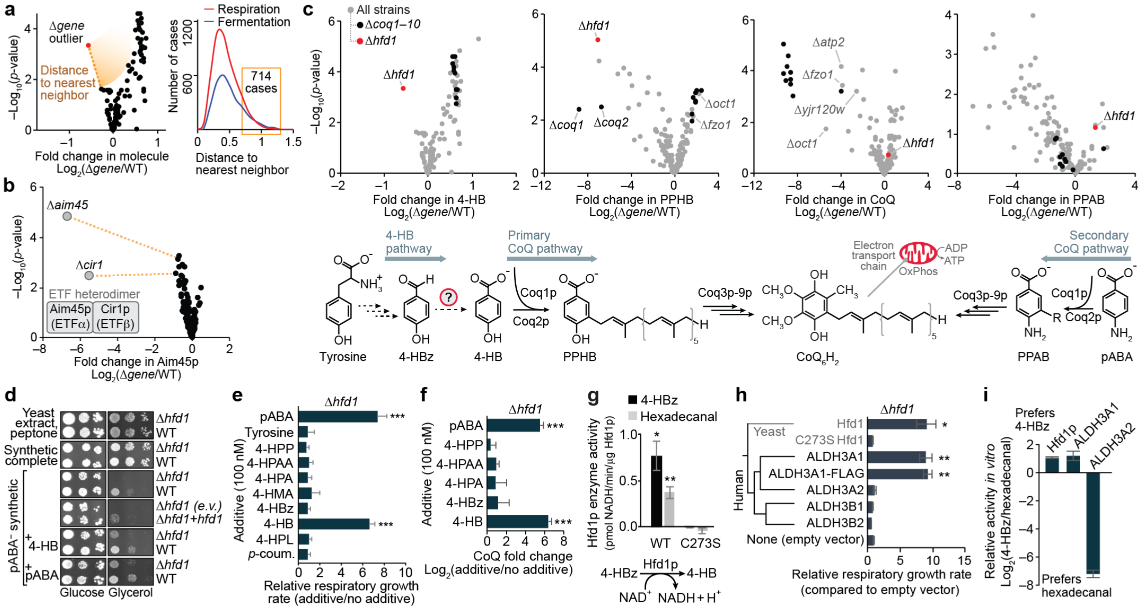
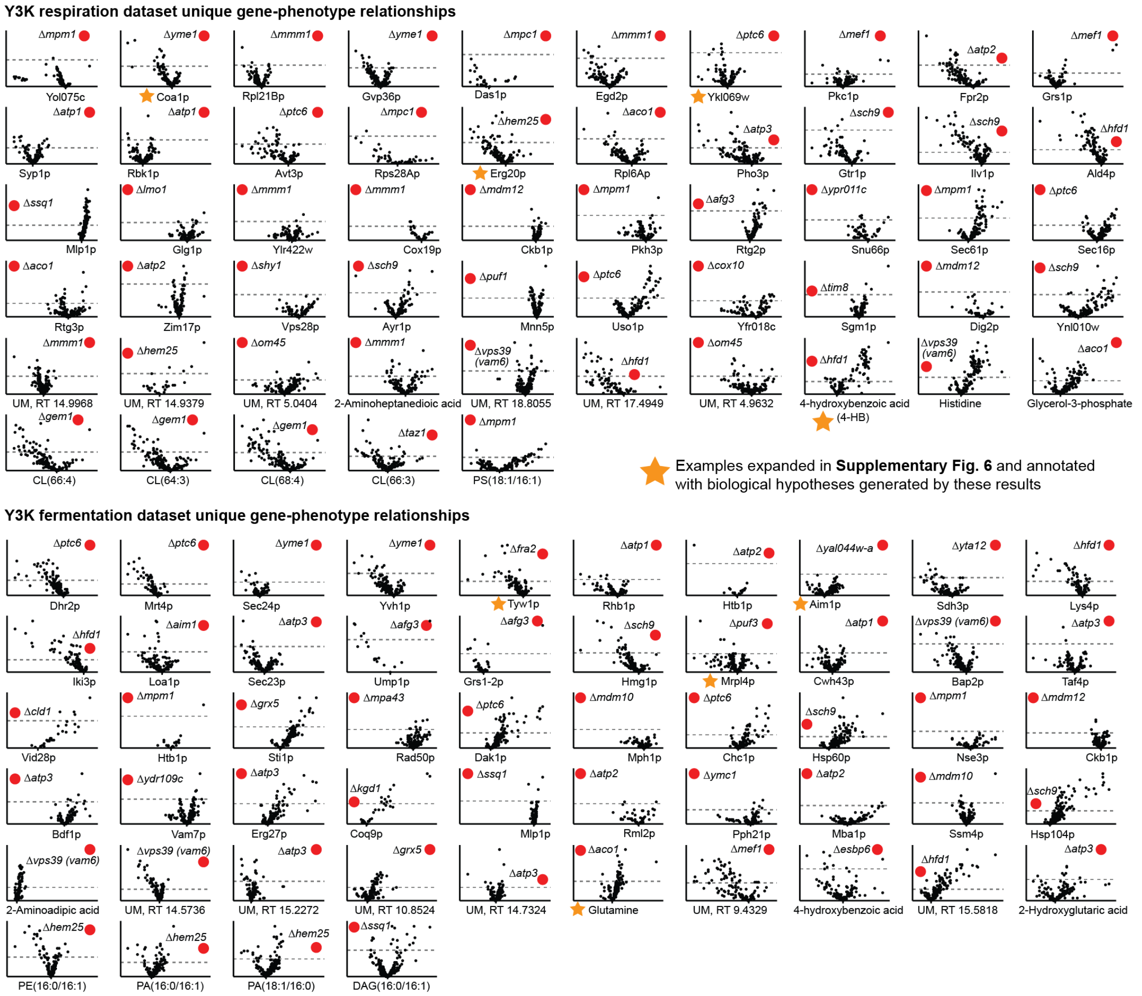
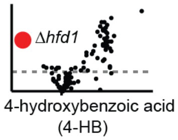
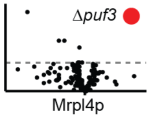
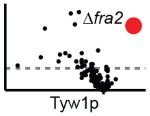
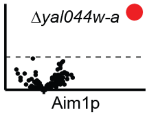
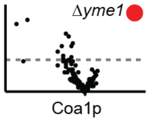
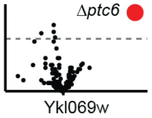
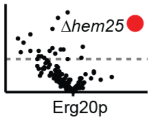
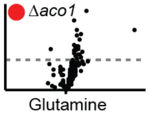


Figure 6.2: Δ*Gene*-specific phenotype detection links Hfd1p to production of 4-hydroxybenzoate for coenzyme Q biosynthesis. **(a)** Overview of the Δ*gene*-specific phenotype detection approach and number of Δ*gene*-specific phenotypes identified in the respiration and fermentation datasets (distance to nearest neighbor on a normalized scale, see Supplementary Note 2). **(b)** Relative abundance of Aim45p (mean, n = 3) versus statistical significance across strains. **(c)** Relative abundances of 4-HB, PPHB, CoQ, and PPAB (mean, n = 3) versus statistical significance across Δ*gene* strains. **(d)** Serial dilutions of yeast grown on variable solid medias. E.v., empty vector; +hfd1, hfd1 plasmid transformed. **(e)** Relative respiratory growth rates of Δ*hfd1* yeast in pABA$^-$ synthetic media with the additives shown (mean ± s.d, n = 3). 4-HPP, 4-hydroxyphenylpyruvate; 4-HPAA, 4-hydroxyphenylacetaldehyde; 4-HPA, 4-hydroxyphenylacetate; 4-HMA, 4-hydroxymandelate; 4-HPL, 4-hydroxyphenyllactate; p-coum., para-coumarate. **(f)** Relative CoQ abundance in Δ*hfd1* yeast cultured in pABA$^-$ media with the additives shown (mean ± s.d., n = 3). **(g)** Enzyme activity of recombinant MBP-Hfd1$^{C\Delta 25}$ in vitro against 4-HBz (200μM) or hexadecanal (200μM) (mean ± s.e.m., n = 3). **(h)** Phylogenetic relationship between yeast Hfd1p and the human ALDH3 family, and relative respiratory growth rates of Δ*hfd1* yeast transformed with plasmids encoding the proteins shown and cultured in pABA$^-$ synthetic media (mean ± s.d, n = 4). **(i)** Relative activity of the dehydrogenases shown against 4-HBz compared to hexadecanal (mean ± s.e.m., n = 3). *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$ (two-sided Student's t-test for all panels).



Supplementary Figure S6.5: Subsets of the Δ*gene*-specific phenotypes identified in this study. Relative abundances of individual molecules (mean $\log_2[\Delta \mathit{gene}/\mathrm{WT}]$, $n = 3$) (x-axes) versus statistical significance ($-\log_{10}[p\text{-value}]$; two-sided Student's t-test) (y-axes) as quantified by MS. The plots shown represent a subset of molecules identified as 'Δ*gene*-specific phenotypes' through an unbiased survey of the Y3K dataset (see Fig. 2a). The array here is limited to the most robust outliers (based on both statistical significance and fold-change, see Supplementary Note 2 and Methods)—the top 20 upregulated proteins, the top 20 downregulated proteins, the top 10 metabolites, and the top 4 or 5 lipids—excluding 'knocked out proteins' (e.g. Fmp52p in the Δ*fmp52* strain) and excluding a given Δ*gene* strain after it appeared twice on the rank list. Biological hypotheses surrounding gene-phenotype relationship were generated for the starred plots (see Supplementary Fig. 6).

Volcano plot	Biological hypothesis generated by Y3K gene-phenotype relationship
	Hypothesis: Hfd1p, a predicted dehydrogenase, catalyzes production of 4-HB. (This hypothesis is tested in detail in this manuscript, see Fig. 2)
	Hypothesis: The mRNA encoding Mrpl4p, a mitochondrial ribosome subunit, is directly bound by Puf3p, an mRNA-binding protein that enhances degradation of select mRNAs.
	Hypothesis: Tyw1p, an iron-sulfur protein, is upregulated in response to iron overload due to loss of Fra2p, a negative regulator of the iron regulon.
	Hypothesis: Yal044w-a, a BolA-like protein, is upregulated to compensate for the loss of Aim1p, a homologous BolA-like protein.

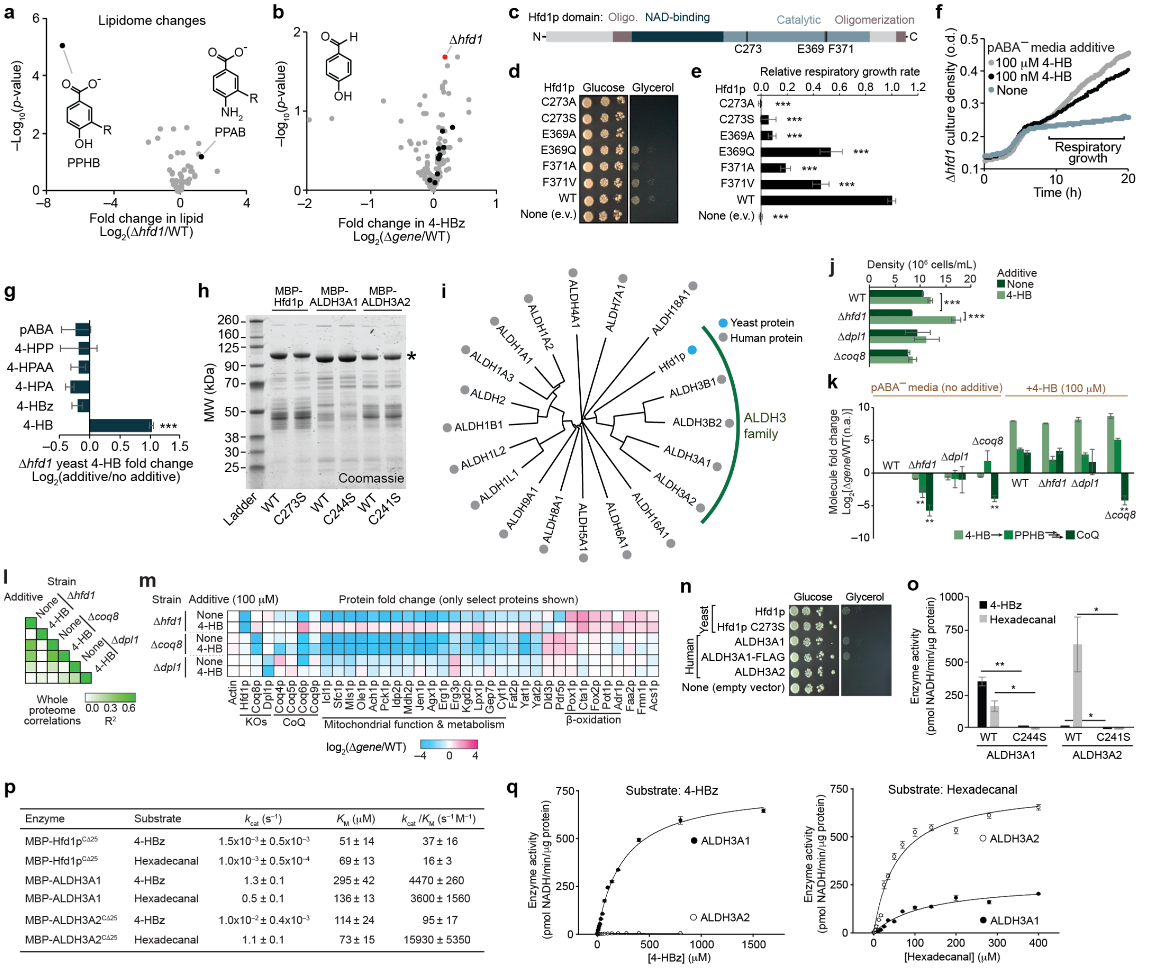
Volcano plot	Biological hypothesis generated by Y3K gene-phenotype relationship
	Hypothesis: Coa1p, a mitochondrial inner membrane OxPhos Complex IV assembly factor, is a substrate for the mitochondrial inner membrane protease Yme1p (i-AAA protease).
	Hypothesis: Ykl069w, a putative methionine-R-sulfoxide reductase, is upregulated in response to oxidative stress due to loss of Ptc6p, a mitochondrial type 2C protein phosphatase.
	Hypothesis: Erg20p, farnesyl pyrophosphate synthetase, is upregulated to compensate for a defect in Heme A (a farnesylated form of heme) biosynthesis due to loss of Hem25p.
	Hypothesis: Intracellular glutamine is decreased due to loss of Aco1p, aconitase, which enables glutamine biosynthesis and inhibition of which enhances glutaminolysis. Consistently, glutamine auxotrophy of the Δ*aco1* strain has already been reported.

Supplementary Figure S6.6: Examples of hypotheses that can be generated from a subset of the Δ*gene*-specific phenotypes identified in this study. Subset of Δ*gene*-specific phenotypes identified in the Y3K dataset. Volcano plots indicate relative molecule abundances (mean $\log_2[\Delta gene/\mathrm{WT}]$, n = 3) (x-axes) versus statistical significance ($-\log_{10}[p\text{-value}]$; two-sided Student's t-test) (y-axes) as quantified by MS. Hypotheses were developed to describe each Δ*gene*-phenotype relationship reported here.

dehydrogenation of 4-hydroxybenzaldehyde (4-HBz) to form 4-HB. Consistently, 4-HBz is elevated in Δ*hfd1* yeast **(Supplementary Fig. S6.7b)**.

We used chemical-genetics to test the proposed Hfd1p activity. Most culture media contain either 4-HB (in yeast extract) or pABA (in standard yeast nitrogen base), enabling yeast to bypass the Tyr-to-4-HB pathway, so we used a defined medium lacking pABA and 4-HB ("pABA$^-$"). Δ*hfd1* yeast exhibited striking respiration deficiency on pABA$^-$ media, a phenotype rescued by pABA, 4-HB, or WT Hfd1p, but not by Hfd1p with mutations to putative catalytic residues[24] **(Fig. 6.2d and Supplementary Fig. S6.7c–e)**. Testing a panel of potential intermediates in the pathway revealed that 4-HB, but not 4-HBz, can rescue the respiratory growth and CoQ production ofΔ*hfd1* yeast **(Fig. 6.2e,f and Supplementary Fig. S6.7f,g)**, supporting a role for Hfd1p in dehydrogenation of 4-HBz. To directly test this activity, we purified recombinant Hfd1p for enzyme assays **(Supplementary Fig. S6.7h)**. WT Hfd1p catalyzes NAD$^+$-dependent dehydrogenation of 4-HBz, but a C273S (catalytic residue) point mutant does not **(Fig. 6.2g)**. Together, these results demonstrate that Hfd1p dehydrogenates 4-HBz to produce 4-HB for CoQ biosynthesis.

Hfd1p is a member of the ancient aldehyde dehydrogenase (ALDH) superfamily, which is found across all three superkingdoms of life and includes 19 human homologs with diverse functions[25]. Based on phylogenetic analyses, Hfd1p is most similar to the human ALDH3 family **(Supplementary Fig. S6.7i)**. ALDH3A2 (FALDH) mutations cause Sjögren–Larsson Syndrome[26] due to defective fatty aldehyde metabolism. However, the endogenous functions of ALDH3A1, B1, and B2 remain obscure, and which of these human ALDH3



Enzyme	Substrate	k_{cat} (s^{-1})	K_M (μM)	k_{cat}/K_M ($s^{-1} M^{-1}$)
MBP-Hfd1p$^{C\Delta25}$	4-HBz	$1.5 \times 10^{-3} \pm 0.5 \times 10^{-3}$	51 ± 14	37 ± 16
MBP-Hfd1p$^{C\Delta25}$	Hexadecanal	$1.0 \times 10^{-3} \pm 0.5 \times 10^{-4}$	69 ± 13	16 ± 3
MBP-ALDH3A1	4-HBz	1.3 ± 0.1	295 ± 42	4470 ± 260
MBP-ALDH3A1	Hexadecanal	0.5 ± 0.1	136 ± 13	3600 ± 1560
MBP-ALDH3A2$^{C\Delta25}$	4-HBz	$1.0 \times 10^{-2} \pm 0.4 \times 10^{-3}$	114 ± 24	95 ± 17
MBP-ALDH3A2$^{C\Delta25}$	Hexadecanal	1.1 ± 0.1	73 ± 15	15930 ± 5350

Supplementary Figure S6.7: Hfd1p supports production of 4-HB for CoQ biosynthesis. **(a)** Relative lipid abundances (mean, n = 3) versus statistical significance (–$\log_{10}$[*p*-value]; two-sided Student's t-test) as quantified by MS. **(b)** Relative abundances of 4-HBz (mean, n = 3) versus statistical significance (–$\log_{10}$[*p*-value]; two-sided Student's t-test) across all Δgene strains in the study. **(c)** Protein domain structures of Hfd1p, highlighting residues involved in catalysis. **(d)** Serial dilutions ofΔ*hfd1* yeast transformed with plasmids encoding the indicated Hfd1p variants grown on pABA⁻ synthetic solid medias with glucose or glycerol. **(e)** Relative respiratory growth rates ofΔ*hfd1* yeast transformed with plasmids encoding the indicated Hfd1p variants and grown in pABA⁻ synthetic liquid media. **(f)** Growth curves showing the respiratory growth ofΔ*hfd1* yeast in pABA⁻ synthetic media with the additives shown. **(g)** Relative 4-HB abundance inΔ*hfd1* yeast cultured in pABA⁻ media with the additives shown (mean $\log_2$[additive/unsupplemented] ± s.d., n = 3). **(h)** SDS-PAGE analysis (Coomassie stained gel) of protein fractions from an isolation of MBP-Hfd1p(CΔ25), MBP-ALDH3A1, and MBP-ALDH3A2(CΔ25) (WT and catalytically dead mutant for each). **(i)** Phylogenetic tree of human ALDH superfamily members and yeast Hfd1p.

Supplementary Figure S6.7: **(j)** Density of yeast (upon harvest) cultured in pABA⁻ media ± 4-HB (mean ± s.d., n = 3). **(k)** Relative abundances of 4-HB, PPHB, and CoQ compared to WT yeast cultured in pABA⁻ media (mean $\log_2[\Delta\text{gene/WT with no additive}]$ ± s.d., n = 3) as quantified by MS. **(l)** Whole proteome correlation map for yeast grown in pABA⁻ media ± 4-HB (mean, n = 3). **(m)** Relative abundances of select proteins as quantified by MS (mean $\log_2[\Delta\text{gene/WT}]$, n = 3) analysis of yeast cultured in pABA⁻ media ± 4-HB. **(n)** Serial dilutions of Δ*hfd1* yeast transformed with plasmids encoding the proteins shown and cultured on solid pABA⁻ synthetic media plates. **(o)** Enzyme activity of MBP-ALDH3A1 or MBP-ALDH3A2(CΔ25) against 4-HBz (200μM) or hexadecanal (200μM) (mean ± s.e.m., n = 3). **(p)** Table of enzyme kinetic parameters for MBP-Hfd1p(CΔ25), MBP-ALDH3A1, and MBP-ALDH3A2(CΔ25) (mean ± s.e.m., n = 3). **(q)** Representative enzyme kinetic curves for MBP-ALDH3A1 and MBP-ALDH3A2(CΔ25). $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$ (two-sided Student's t-test).

functions are conserved in Hfd1p has not been completely defined. Previous work showed that sphingolipid metabolism is perturbed in Δ*hfd1* yeast due to a defect in dehydrogenation of hexadecanal, and this defect can be rescued by ALDH3A2, but not by ALDH3A123[27]. However, a separate sphingolipid pathway defect (Δ*dpl1*) does not disrupt the 4-HB-CoQ pathway **(Supplementary Fig. S6.7j–m and Supplementary Note 3)**, suggesting that the two pathways are otherwise independent. Consistent with the idea that Hfd1p is a dual-function protein that supports both sphingolipid metabolism and CoQ biosynthesis, we observed Hfd1p activity in vitro with hexadecanal, similar to that observed with 4-HBz **(Fig. 6.2g)**. However, in contrast to rescue of the sphingolipid metabolism defect, we found that ALDH3A1, but not ALDH3A2, rescues the pABA⁻ respiratory growth phenotype of Δ*hfd1* yeast **(Fig. 6.2h and Supplementary Fig. S6.7n)**. Moreover, while ALDH3A2 shows a strong substrate preference for hexadecanal over 4-HBz, Hfd1p and ALDH3A1 show a preference for 4-HBz **(Fig. 6.2i and Supplementary Fig. S6.7o–q)**. These results suggest that the dual functions of yeast Hfd1p have diverged in human ALDH3A1 and ALDH3A2. Collectively, these results demonstrate a major cellular function for the aldehyde dehydrogenase Hfd1p in the Tyr-to-4-HB pathway and strongly suggest that ALDH3A1 plays a similar role in human CoQ biosynthesis.

Regression analysis of global perturbation profiles. While molecular changes unique to a given Δ*gene* strain can be functionally informative, similarities between Δ*gene* strains can also assist characterization. In our second analysis approach, we examined Δ*gene*–Δ*gene* correlations through pairwise comparisons of global Δ*gene* perturbation profiles. Deletion of functionally related genes, such as the cytochrome c oxidase genes *cox12* and *cox23*, caused highly similar whole proteome perturbations **(Fig. 6.3a)**. Notably, highly correlated phenotype changes were also observed in Δ*cox12* and Δ*cox23* metabolomes and lipidomes **(Fig. 6.3a)**. However, deletion of unrelated genes, such as *cox12* and *mic26*, generated uncorrelated phenotype changes **(Fig. 6.3a)**. Examination of Δ*gene*–Δ*gene* correlations across the entire study indicated numerous functional relationships, with stronger correlations observed in respiration **(Fig. 6.3b)**.

A group of respiration-deficient (RD) strains showed robust correlations across all three omes **(Fig. 6.3b)**, reflecting their similar broad biological functions in mitochondrial OxPhos and suggesting that they share a universal "respiration deficiency response" (RDR). Multi-omic principle component and GO term analyses revealed a coordinated RDR that provides biological insight into respiration defects—a common feature of many diseases

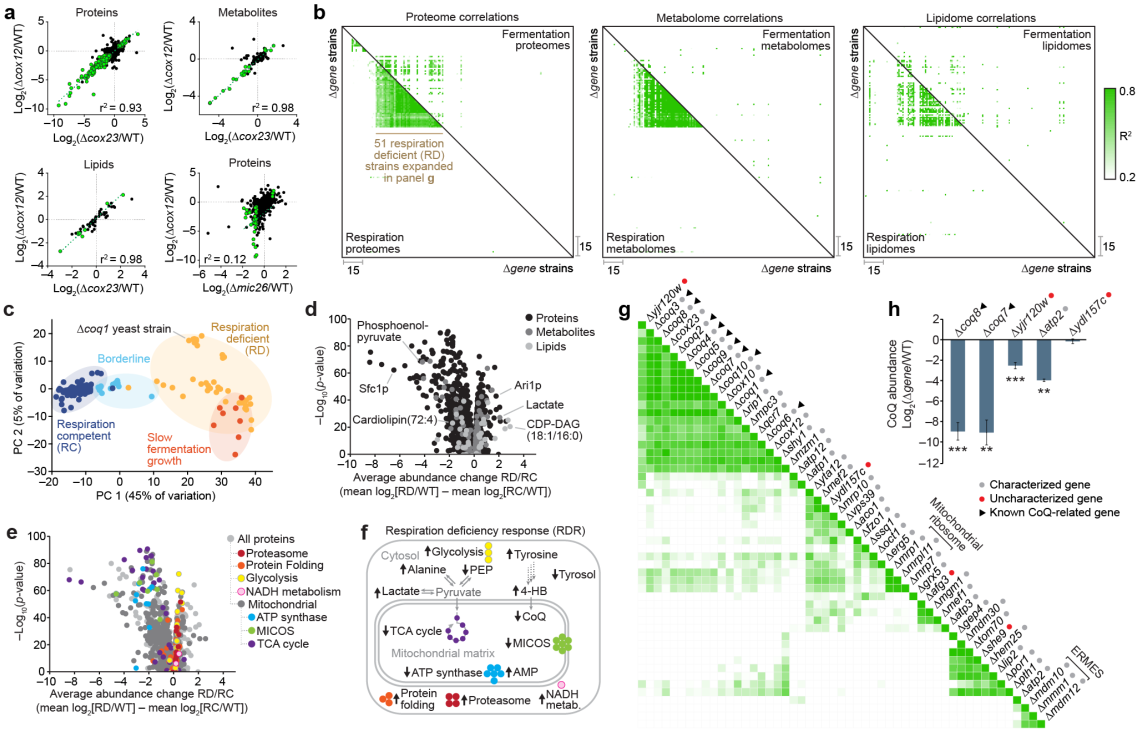
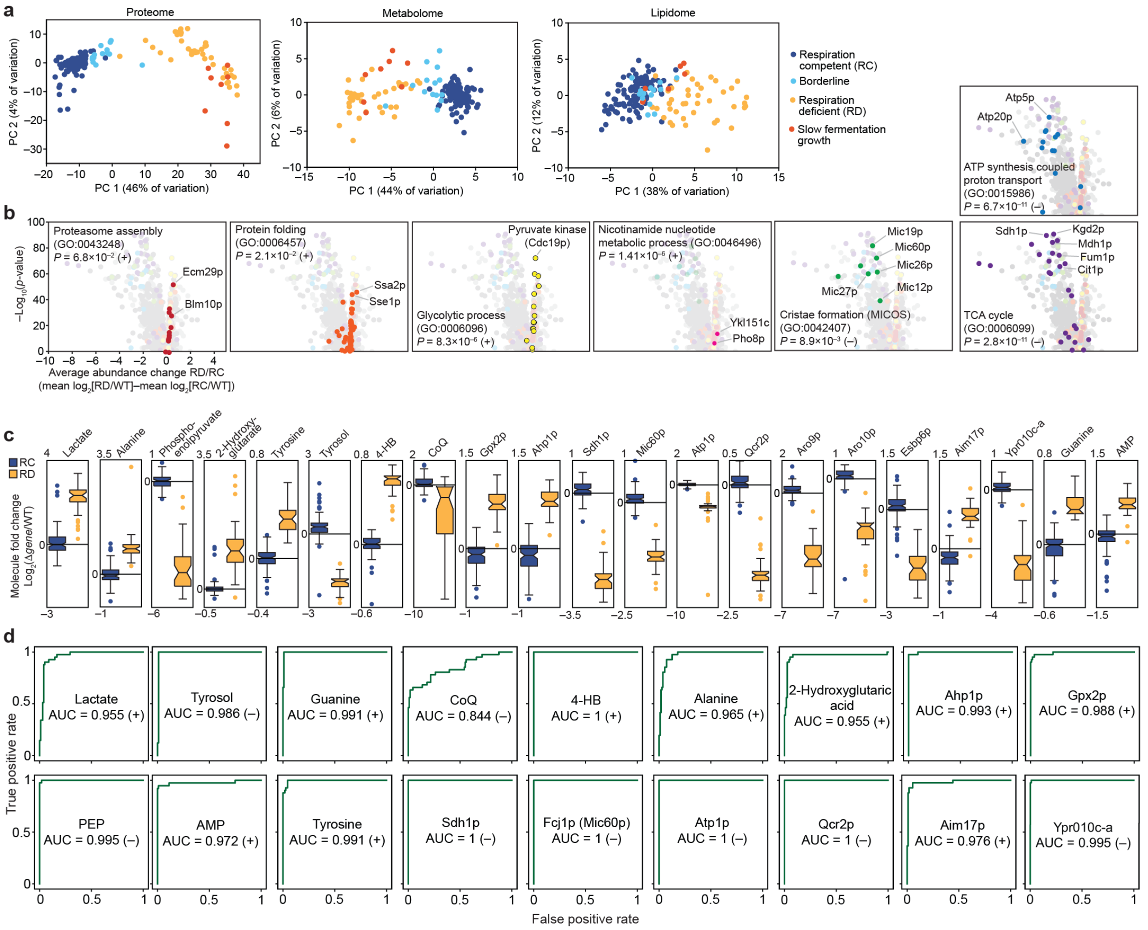


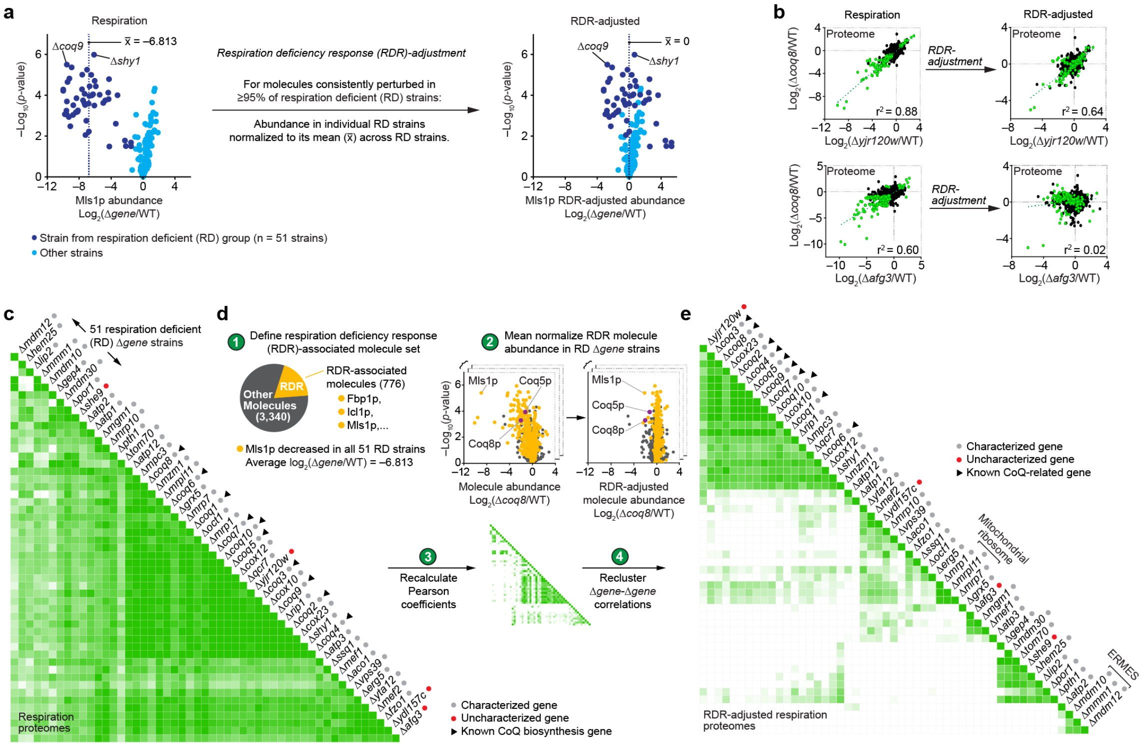
Figure 6.3: Functional correlations through perturbation profile regression analysis. (a) Plots comparing relative molecule abundances between pairs of Δ*gene* strains. Strain-strain similarity assessed by linear regression analysis of Δ*gene* perturbation profiles. Green points indicate molecules significantly perturbed in both mutants ($|\log_2[FC]| > 0.7$, $P < 0.05$; two-sided Student's t-test). **(b)** Maps of Pearson correlation coefficients (r^2) for pairs of Δ*gene* perturbation profiles across omes and metabolic conditions. Strains are clustered based on respiration proteome correlations, and this strain order is held consistent across all 6 maps. **(c)** Projection of respiration competent (RC) and deficient (RD) strains onto the plane defined by principal component (PC) axes 1 and 2 (full multi-omic respiration dataset). **(d)** Average fold change in molecule abundances (mean $\log_2$[RD strains/RC strains]) versus statistical significance (–$\log_{10}$[p-value, Bonferroni corrected two-sided t-test]). **(e)** RD versus RC proteome perturbation volcano plot (as in d) showing select functional groups (GO terms) significantly enriched in either upregulated or downregulated proteins. **(f)** Scheme of RDR pathways. **(g)** Re-clustered respiration proteome strain-strain correlation map following RDR- adjustment. (h) CoQ abundance changes in select Δ*gene* strains (mean ± s.d., n = 3); $^{**}P < 0.01$; $^{***}P < 0.001$ (two-sided Student's t-test).

including cancer—and suggests that a multi-omic biomarker fingerprint could afford a specific diagnostic for mitochondrial disease **(Fig. 6.3c–f, Supplementary Fig. S6.8, and Supplementary Note 4)**. However, stress responses such as the RDR also pose a barrier to biochemical investigations because they can obscure functionally-informative phenotypes. To assess more specific biochemical roles for individual proteins, we normalized for the RDR across RD strains **(Supplementary Fig. S6.9 and Supplementary Note 5)**. Across all of our RD strains, 776 molecules were identified as being consistently perturbed. The individual measurements of these RDR-associated molecules were mean normalized ("RDR-adjusted") to reveal characteristic deviations from the general RDR and to enable visualization of Δ*gene*-specific changes.

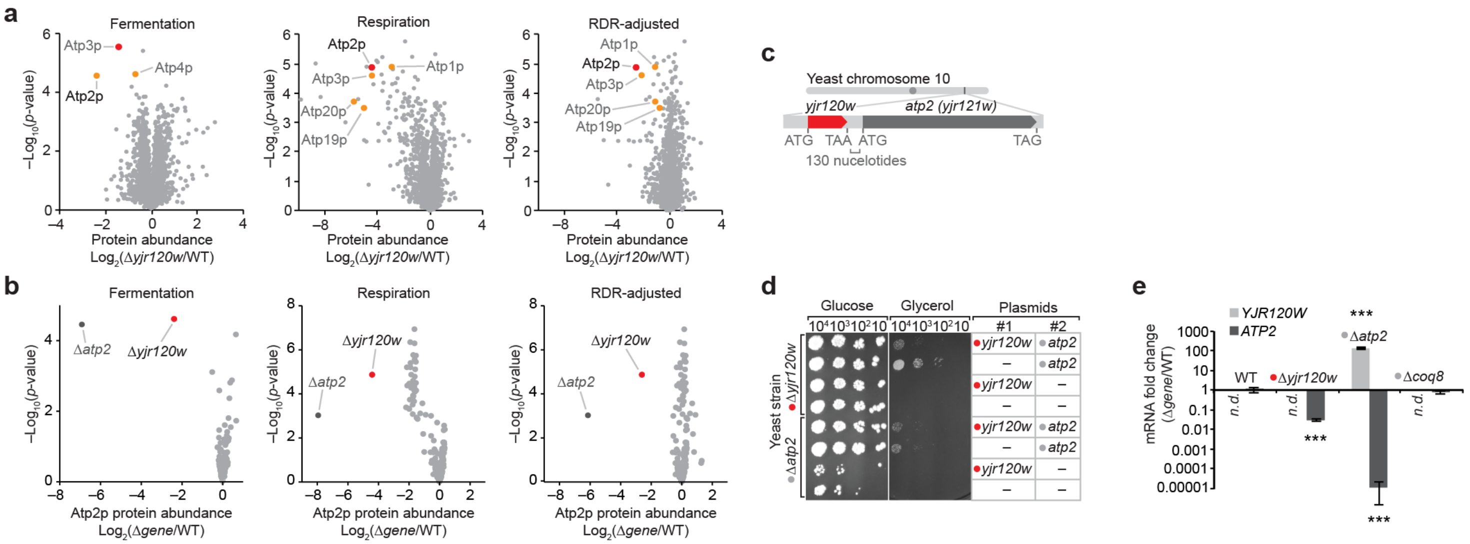
Recalculating Δ*gene*–Δ*gene* correlation coefficients with RDR-adjusted plots strikingly reduces correlations between more functionally disparate genes **(Supplementary Fig. S6.9c–e)**. Reclustering Δ*gene*–Δ*gene* correlations reveals new clusters of genes with similar biochemical functions **(Fig. 6.3g)**. For example, known CoQ biosynthesis genes were brought into a tighter cluster that also includes the uncharacterized gene *yjr120w* **(Fig. 6.3g)**, suggesting that *yjr120w* might support CoQ biosynthesis. Consistently, we observed CoQ deficiency in Δ*yjr120w* yeast **(Fig. 6.3h)**, the molecular basis of which we determined to include loss of Atp2p, an ATP synthase subunit **(Supplementary Fig. S6.10 and Supplementary Note 6)**. These results show that specific ATP synthase subunits support CoQ biosynthesis and, more broadly, demonstrate how global mass spectrometry profiling can reveal functional links between genes.



Supplementary Figure S6.8: Identification of respiration deficiency response pathways and potential biomarkers. **(a)** Projection of RC and RD strains onto the planes defined by principal component (PC) axes 1 and 2 for separate proteome, metabolome, and lipidome PC analyses. **(b)** RD versus RC proteome perturbation volcano plots (as in Fig. 3e) showing select functional groups (GO terms) significantly enriched (Bonferroni corrected p-values shown in figure) in either upregulated or downregulated proteins. **(c)** Box plots depicting median molecule fold changes for RC and RD strains ($\log_2$[RD or RC average/WT]) (n = 111 for RC, 41 for RD). Notch indicates 95% c.i. **(d)** Receiver operating characteristic (ROC) curves for select molecules depicting the false positive rates and true positive rates for prediction of respiration deficiency associated with particular molecule fold changes. AUC, area under the curve.



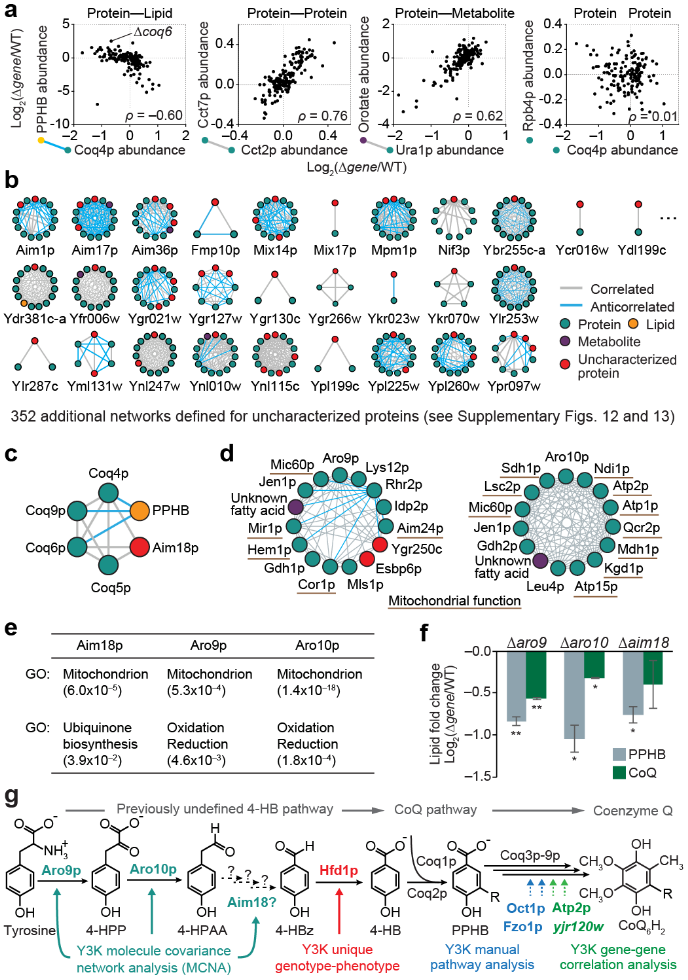
Supplementary Figure S6.9: Subtraction of shared responses to reveal deeper biochemical insight. (a) RDR-abundance adjustment of a representative molecule (Mls1p) by subtraction of the average fold change in abundance (mean $\log_2[\Delta gene/\mathrm{WT}]$, n = 3) across respiration deficient (RD) strains. This adjustment was only performed within RD strains. **(b)** Plots comparing relative protein abundances between pairs of Δ*gene* strains. Linear regression analysis of pairs of perturbation profiles before (left) and after (right) RD-abundance adjustment. Green points indicate molecules significantly perturbed in both mutants ($|\log_2(\mathrm{FC})| > 0.7$; $P < 0.05$; two-sided Student's t-test) prior to RDR-adjustment. **(c)** Expanded view of highly correlated strains in the respiration proteomes correlation map (see Fig. 3b). **(d)** Procedure for normalization of the RDR. **(e)** Re-clustered respiration proteome strain-strain correlation map following RDR-adjustment (also shown in Fig. 3g).



Supplementary Figure S6.10: Molecular perturbations of yeast lacking *yjr120w*. **(a)** Relative protein abundances (mean $\log_2[\Delta yjr120w/\mathrm{WT}]$, n = 3) versus statistical significance ($-\log_{10}[p\text{-value}]$; two-sided Student's t-test) as quantified by MS. (b) Relative Atp2p protein abundance (mean $\log_2[\Delta gene/\mathrm{WT}]$, n = 3) versus statistical significance ($-\log_{10}[p\text{-value}]$; two-sided Student's t-test) across all mutants in the study. (c) Genomic organization of *yjr120w* and *atp2*. (d) Serial dilutions of yeast transformed with the indicated plasmids grown on agar plates with glucose (to enable fermentation) or glycerol (to force respiration). (e) Fold changes in mRNA abundances (mean $\Delta gene/\mathrm{WT}$, n = 3) as quantified by real time polymerase chain reaction (RT-PCR) analysis. Yjr120w mRNA was not detected (n.d.) in WT yeast, so imputation of this missing value was used to calculate the fold increase in yjr120w mRNA shown for the Δ*atp2* strain. $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$ (two-sided Student's t-test).

Molecule covariance network analysis. Similarly, in our third analysis approach, we leveraged the multi-omic nature of our mass spectrometry profiles to determine pairwise covariance between proteins, metabolites, and lipids. This approach is similar to mRNA coexpression profiling, which can be used to predict gene function[28–30], but it integrates three complementary classes of molecules. Perturbations for functionally related molecules, such as the protein Coq4p and the lipid CoQ intermediate PPHB, show strong positive or negative correlations, while those of unrelated molecules, such as Coq4p and Rpb4p, lack correlations **(Fig. 6.4a)**. Correlated molecules include proteins in complexes, such as the cytosolic TRiC/CCT chaperonin complex (Cct2p and Cct7p), and enzyme-product pairs (e.g. Ura1p and orotic acid) **(Fig. 6.4a)**.

Examining correlations across all 4,505 molecules in the Y3K dataset through this multi-omic molecule covariance network analysis (MCNA) reveals numerous functional relationships, which can be visualized as networks of molecules (nodes) and correlations (edges) **(Fig. 6.4b and Supplementary Fig. S6.11a)**. After applying strict correlation thresholds (Bonferroni-adjusted p-value < 0.001, $|\rho| \geqslant 0.58$), 237,342 edges remain among 2,382 nodes in the respiration dataset **(Supplementary Fig. S6.11a–f)**. Many edges were observed between RDR-associated molecules **(Supplementary Fig. S6.11g)**, reflecting their common relationship to mitochondrial metabolism. As described above for Δ*gene* correlations, we deepened the molecular insight of the MCNA by RDR-adjustment, which reduced overall connectivity and increased the selectivity of functionally related molecule sub-networks **(Supplementary Fig. S6.11g)**. For example, the selectivity of the mitochondrial ribosome



	Aim18p	Aro9p	Aro10p
GO:	Mitochondrion (6.0×10^{-5})	Mitochondrion (5.3×10^{-4})	Mitochondrion (1.4×10^{-18})
GO:	Ubiquinone biosynthesis (3.9×10^{-2})	Oxidation Reduction (4.6×10^{-3})	Oxidation Reduction (1.8×10^{-4})

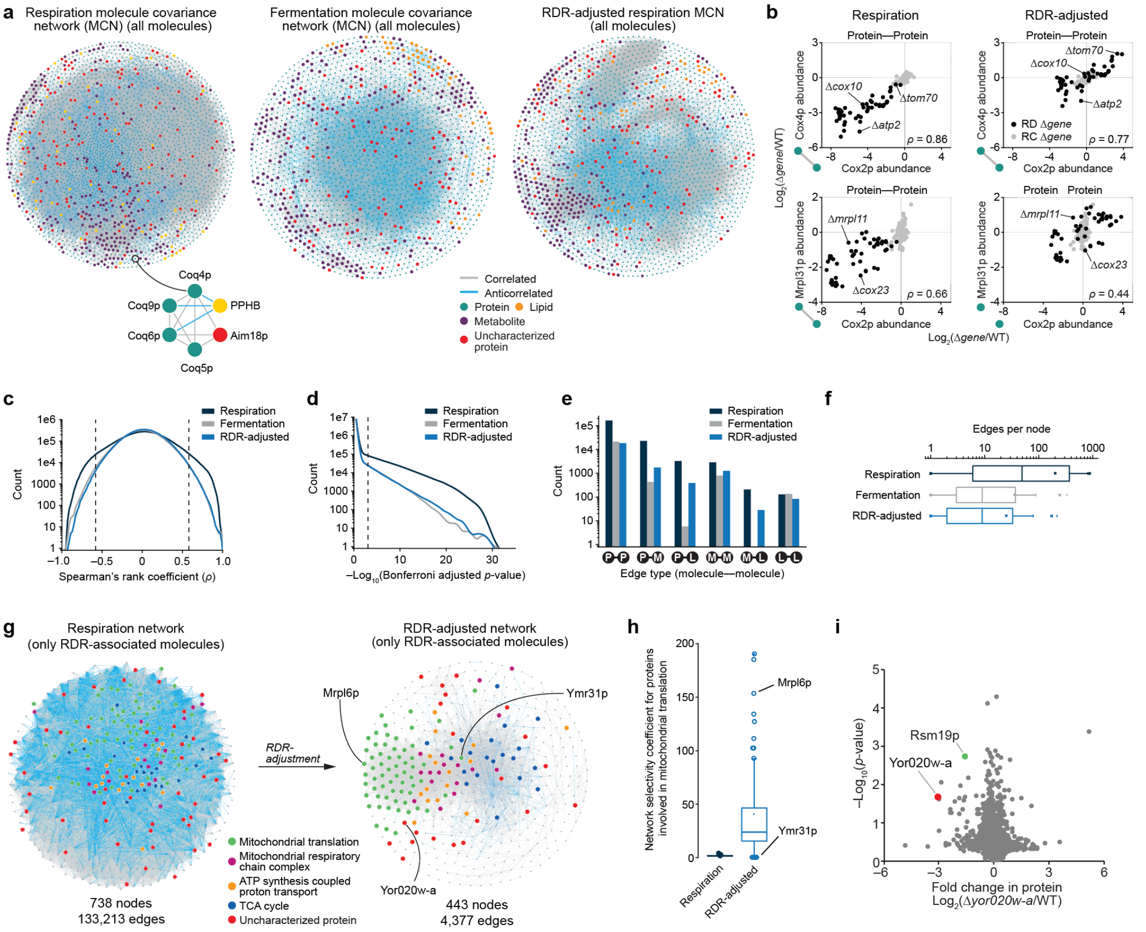
Figure 6.4: Multi-omic molecule covariance network analysis assists functional characterization. **(a)** Relative abundances of molecule pairs across Δ*gene* strains. Covariance assessed by Spearman's rank coefficient (ρ). **(b)** Nearest neighbor molecule covariance networks for a representative subset of uncharacterized proteins. **(c)** Network for Coq4p in the RDR-adjusted respiration dataset. **(d)** Networks showing the 14 molecules most strongly correlated to Aro9p or Aro10p in the RDR-adjusted respiration dataset. **(e)** GO term analyses of the Aim18p, Aro9p, and Aro10p networks (*p*-values). **(f)** Relative abundances of CoQ and PPHB (mean $\log_2[\Delta gene/\mathrm{WT}]$, n = 2) in Δ*aro9*, Δ*aro10*, and Δ*aim18* strains compared to WT yeast cultured in pABA$^-$ media; $^*P < 0.05$; $^{**}P < 0.01$ (two-sided Student's t-test). **(g)** Y3K-enabled characterization of proteins that support the CoQ pathway.

sub-network increased 16-fold **(Supplementary Fig. S6.11h)**. These RDR-adjusted networks associated the MXP Yor020w-a with the mitochondrial ribosome **(Supplementary Fig. S6.11g)**. To test this association, we examined the proteome of Δ*yor020w-a* yeast, which showed a significant decrease in the mitochondrial ribosome protein Rsm19p **(Supplementary Fig. S6.11i)**, suggesting that Yor020w-a is linked to mitochondrial translation.

Hundreds of additional uncharacterized proteins were linked to characterized molecules by our MCNA, providing a foundation for generating hypotheses about their functions **(Fig. 6.4b, Supplementary Figs. S6.12 and S6.13)**. For example, the MXP Aim18p was linked to a network of CoQ biosynthesis proteins, and Aro9p and Aro10p were linked to numerous mitochondrial proteins that support OxPhos **(Fig. 6.4c–e)**. Based on domain homology and predicted enzymatic functions, we hypothesized that Aim18p, Aro9p, and Aro10p could function in the Tyr- to-4-HB pathway **(Supplementary Fig. S6.14 and Supplementary Note 7)**. Consistently, when cultured in a pABA$^-$ media, Δ*aim18*, Δ*aro9*, and Δ*aro10* yeast are deficient in both CoQ and PPHB **(Fig. 6.4f)**. This work shows how global mass spectrometry profiling can be used to generate biological hypotheses and characterize protein functions through distinct multi-omic data analysis approaches **(Fig. 6.4g)**.

Discussion

A constant challenge in biology is to comprehensively monitor and understand the molecular effects of a defined alteration (e.g., a disease mutation, a drug treatment, or a gene deletion). Mass spectrometry (MS) has become central to answering this challenge.



Supplementary Figure S6.11: Features of multi-omic molecule covariance networks. Network of all covariant molecules observed in each dataset ($|\rho| \geqslant 0.58$, Bonferroni-adjusted $P < 0.001$; two-sided Student's t-test). **(b)** Regression analysis of pairs of RDR-associated molecules before and after RDR adjustment using Spearman's rank coefficient (ρ). Points corresponding to RD and RC Δ*gene* strains are indicated. **(c)** Distribution of calculated Spearman coefficients for all pairwise molecule covariance comparisons (ρ cutoff at ±0.58 used throughout the study is indicated). **(d)** Distribution of Bonferroni-adjusted *p*-values from all pairwise molecule comparisons (*p*-value cutoff at 0.001 used throughout the study is indicated). **(e)** Bar chart indicating number of protein–protein (P–P), protein–metabolite (P–M), protein–lipid (P–L), metabolite–metabolite (M–M), metabolite–lipid (M–L), and lipid–lipid (L–L) edges in each dataset. **(f)** Box plots indicating the number of edges per node in the respiration, fermentation, and RDR-adjusted networks.

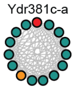
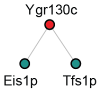
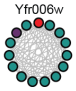
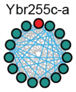
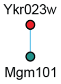
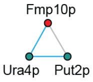
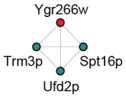
Supplementary Figure S6.11: (g) Network of all covariant RDR-associated molecules ($|\rho| \geqslant 0.58$, Bonferroni-adjusted $P < 0.001$; two-sided Student's t-test) generated using the respiration (left) and RDR-adjusted (right) datasets. Nodes are highlighted according to GO category. **(h)** Box plots indicating the molecule covariance network (MCN) specificity coefficient for all nodes involved in mitochondrial translation in both the respiration and RDR-adjusted respiration RDR-associated molecule networks (shown in panel G). **(i)** Relative protein abundances (mean $\log_2[\Delta \mathit{yor020w\text{-}a}/\mathrm{WT}]$, n = 2) versus statistical significance ($-\log_{10}[p\text{-value}]$; two-sided Student's t-test) as quantified by MS.

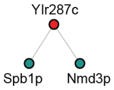
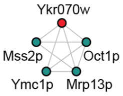
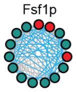
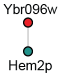
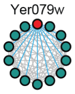
Here, we leveraged a subset of our multi-omic dataset to investigate gaps in knowledge of CoQ biosynthesis. Despite CoQ's essential function in the mitochondrial electron transport chain, role as a key cellular antioxidant, and link to numerous human diseases (e.g., ataxias, myopathies, and nephrotic syndromes), multiple steps in CoQ biosynthesis remain uncharacterized[17,31,32]. In particular, enzymes involved in the initial stage of CoQ biosynthesis— wherein the headgroup precursor 4-HB is produced—were previously undefined in mammals and yeast.

Our Δ*gene*-specific phenotype detection approach suggested a role for the ancient aldehyde dehydrogenase superfamily member Hfd1p in 4-HB biosynthesis. Biochemical and genetic studies confirmed this role for Hfd1p in yeast and further demonstrated that the human homolog ALDH3A1 can also catalyze production of 4-HB *in vivo* and *in vitro* **(Fig. 6.2)**, thereby highlighting ALDH3A1 as a candidate disease gene for primary CoQ deficiency.

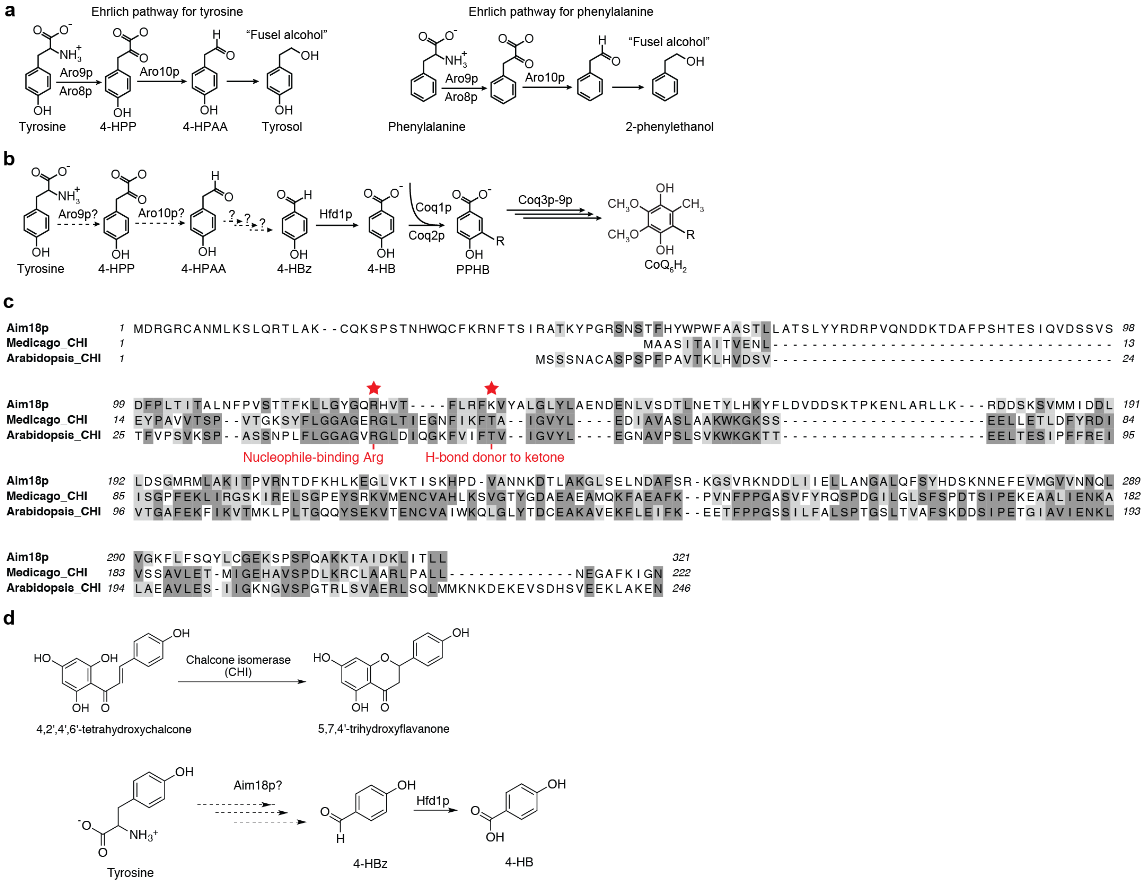
Distinct Y3K dataset analyses placed additional proteins into the CoQ biosynthesis pathway. MCNA showed unexpected connections between Aro9p, Aro10p, and mitochon-

Supplementary Figure S6.12: Molecule covariance networks for uncharacterized proteins. 'Nearest neighbor' molecule covariance networks for all uncharacterized proteins observed across the respiration, fermentation, and RDR-adjusted respiration datasets ($|\rho| \geqslant 0.58$, Bonferroni-adjusted $P < 0.001$; two-sided Student's t-test). If more than 14 correlated molecules were present in a given covariance network, only the top 14 correlated molecules (nearest neighbors) are displayed.

Molecule covariance network	GO terms (for larger networks) or brief protein descriptions (for smaller networks)	Biological hypothesis generated by Y3K MCNA
Respiration networks		
	Mitochondrion (2.6x10^{-13}) Mitochondrial respiratory chain complex III (4.7x10^{-7})	Hypothesis: Ydr381c-a, an uncharacterized protein, supports mitochondrial OxPhos Complex III.
	Eis1p: Eisosome component Tfs1p: Phospholipid binding protein that regulates protein kinase signaling pathways	Hypothesis: Ygr130c regulates eisosome activity by linking it to a protein kinase signaling network
	Golgi apparatus (4.5x10^{-3}) Intracellular protein transport (7.8x10^{-5})	Hypothesis: Yfr006w, a putative peptidase, supports intracellular peptide transport.
	Mitochondrion (2.0x10^{-14}) Oxidation-reduction process (1.7x10^{-4})	Hypothesis: Ybr255c-a, an uncharacterized protein, supports mitochondrial OxPhos.
	Mas2p: Subunit of the mitochondrial processing protease	Hypothesis: Mix17p, an uncharacterized mitochondrial protein, is involved in mitochondrial protein import and processing.
	Mgm101: Mitochondrial DNA repair protein.	Hypothesis: Ykr023w, an uncharacterized mitochondrial protein, supports mitochondrial genome maintenance.
	Put2p: delta-1-pyrroline-5-carboxylate dehydrogenase	Hypothesis: Fmp10p, an uncharacterized mitochondrial protein, is involved in proline-to-glutamate metabolism.
	Hxt5p: Hexose transporter with high affinity for glucose.	Hypothesis: Ydl199c, a predicted membrane transporter, is involved in hexose transport.
	Spt16p: subunit of FACT complex that regulates chromatin structure.	Hypothesis: Ygr266w, an uncharacterized mitochondrial outer membrane protein that is upregulated due to DNA replication stress, affects Spt16p activity.

Molecule covariance network	GO terms (for larger networks) or brief protein descriptions (for smaller networks)	Biological hypothesis generated by Y3K MCNA
Respiration networks (continued)		
	Mitochondrion (1.4x10^{-25}) Oxidation-reduction process (2.5x10^{-7})	Hypothesis: Ylr253w, an uncharacterized mitochondrial kinase, supports mitochondrial OxPhos.
	Spb1p: Ribosome processing Nmd3p: Ribosome processing	Hypothesis: Ylr287c supports cytosolic ribosome maturation.
	Nucleolus (1.3x10^{-8}) Ribosome biogenesis (4.4x10^{-8})	Hypothesis: Ynl247w supports cytosolic ribosome maturation.
	Mss2p: CIV assemblyfactor Oct1p: Mitochondrial intermediate peptidase Ymc1p: Putative mitochondrial glycine transporter that supports heme biosynthesis.	Hypothesis: Ykr070w, an uncharacterized mitochondrial protein, supports mitochondrial Complex IV function.
Fermentation networks		
	Mitochondrion (9.8x10^{-8}) Metabolic process (3.7x10^{-3})	Hypothesis: Fsf1p supports mitochondrial metabolism
	Nucleolus (1.6x10^{-2}) 90S preribosome (2.0x10^{-2})	Hypothesis: Ybl028c supports cytosolic ribosome maturation.
	Hem2p: Aminolevulinate dehydratase	Hypothesis: Ybr096w supports heme biosynthesis
	Nucleolus (3.9x10^{-10}) Ribosome biogenesis (2.8x10^{-8})	Hypothesis: Yer079w is a negative regulator of ribosome biogenesis

Supplementary Figure S6.13: Examples of hypotheses that can be generated from a subset of the molecule covariance network analyses in this study. Nearest neighbor molecule covariance networks from uncharacterized proteins containing more than four connected nodes were tested for GO term enrichment using a Fisher's exact test with Benjamini–Hochberg FDR adjustment to account for multiple hypothesis testing. Networks containing four or fewer connected nodes were analyzed manually for functionally related molecules. Based on these MCNA results, biological hypotheses about the functions of the uncharacterized proteins shown were developed.



Supplementary Figure S6.14: Hypothesized pathways for Aro9p, Aro10p, and Aim18p. **(a)** Putative biochemical functions of Aro9p and Aro10p in catabolism of tyrosine and phenylalanine. **(b)** Predicted functions for Aro9p and Aro10p in the Tyr-to-4-HB-to-CoQ pathway. **(c)** Protein sequence alignments of Aim18p (*S. cerevisiae*) and chalcone isomerases (CHI) from *Medicago* and *Arabidopsis* highlighting conservation of putative catalytic residues (starred residues). **(d)** Example of a CHI catalyzed reaction (upper scheme) and the hypothesized pathway of Aim18p action (lower scheme).

drial OxPhos proteins, which helped place Aro9p and Aro10p into the Tyr-to-4-HB pathway **(Fig. 6.4)**. Similarly, links between Aim18p and known CoQ biosynthesis enzymes also connected Aim18p to CoQ biosynthesis. Furthermore, Y3K gene-gene correlation analyses and manual pathway analyses linked CoQ biosynthesis to other proteins whose molecular functions in this pathway are not yet fully defined (e.g. Atp2p, Fzo1p, and Oct1p). Disruption of the mammalian Fzo1p homolog, MFN2—a protein essential for mitochondrial fusion that harbors causative mutations in Charcot-Marie-Tooth disease[33]—was recently shown to cause CoQ deficiency through an unclear molecular mechanism[34]. Our results suggest that this unexpected relationship between MFN2 and CoQ biosynthesis is evolutionarily conserved, and establish yeast as a model system for further probing its mechanism.

Our Y3K dataset provides many additional leads for further biochemical studies of numerous metabolic pathways that impact human health and disease, and we expect that the open access web utility (http://y3kproject.org/) will enable others to generate their own hypotheses. With demand for multi-omic dataset analysis approaches increasing, we also hope that our multifaceted, data visualization website will serve as a useful model for future studies.

We anticipate that the multi-omic Y3K dataset will provide a resource for broader systems biology inquiries. For example, our definition of the yeast respiration deficiency response (RDR) **(Fig. 6.3)** may assist studies of how cells broadly respond to defects in OxPhos, which are observed in diverse diseases including many cancers. Our RDR work also suggests that a multi-omic fingerprint of numerous molecules could provide a highly specific biomarker panel.

Methods

Yeast strains and cultures. The parental (WT) *Saccharomyces cerevisiae* strain for this study was the haploid MATalpha BY4742. Single gene deletion (Δ*gene*) derivatives of BY4742 were either obtained through the gene deletion consortium[30] or made in-house using a *KanMX* deletion cassette to match those in the consortium collection. All gene deletions were confirmed by either proteomics (significant decrease in the encoded protein) or a PCR assay. Δ*gene* strains made in-house were also confirmed by gene sequencing.

Single lots of yeast extract ('Y') (Research Products International, RPI), peptone ('P') (RPI), agar (Fisher), dextrose ('D') (RPI), glycerol ('G') (RPI), and G418 (RPI) were used for all medias. YP and YPG solutions were sterilized by automated autoclave. G418 and dextrose were sterilized by filtration (0.22 μm pore size, VWR) and added separately to sterile YP or YPG. YPD+G418 plates contained yeast extract (10 g/L), peptone (20 g/L), agar (15 g/L), dextrose (20 g/L), and G418 (200 mg/L). YPD media (fermentation cultures) contained yeast extract (10 g/L), peptone (20 g/L), and dextrose (20 g/L). YPGD media (respiration cultures) contained yeast extract (10 g/L), peptone (20 g/L), glycerol (30 g/L) and dextrose (1 g/L).

Yeast from a –80 °C glycerol stock were streaked onto YPD+G418 plates and incubated (30 °C, ~60 h). Starter cultures (3 mL YPD) were inoculated with an individual colony of yeast and incubated (30 °C, 230 rpm, 10–15 h). A WT culture was included with each set of Δ*gene* strain cultures (usually 19 Δ*gene* cultures and 1 WT culture). Cell density was determined by optical density at 600 nm (OD_{600}) as described[35]. YPD or YPGD media (100 mL media at ambient temperature in a sterile 250 mL Erlenmeyer flask) was inoculated with 2.5×10^6 yeast cells and incubated (30 °C, 230 rpm). Samples of the YPD cultures were harvested 12 h after inoculation, a time point that corresponds to early fermentation (logarithmic) growth. Samples of YPGD cultures were harvested 25 h after inoculation, a time point that corresponds to early respiration growth.

Liquid chromatography tandem mass spectrometry (LC-MS/MS) proteomics. 1×10^8 yeast cells were harvested by centrifugation (3,000 *g*, 3 min, 4 °C), the supernatant was removed, and the cell pellet was flash frozen in $N_{2(l)}$ and stored at –80 °C. Yeast pellets were resuspended in 8 M urea, 100 mM tris (pH = 8.0). Yeast cells were lysed by the addition of methanol to 90%, followed by vortexing (~30 s). Proteins were precipitated by centrifugation (12,000 *g*, 5 min). The supernatant was discarded, and the resultant protein pellet was resuspended in 8 M urea, 10 mM tris(2-carboxyethyl)phosphine (TCEP), 40 mM chloroacetamide (CAA) and 100 mM tris (pH = 8.0). Sample was diluted to 1.5 M urea with 50 mM tris and digested with trypsin (Promega) (overnight, ~22 °C) (1:50, enzyme:protein). Samples were desalted using Strata X columns (Phenomenex Strata-X Polymeric Reversed Phase, 10 mg/mL). Strata X columns were equilibrated with one column volume of 100% acetonitrile (ACN), followed by 0.2% formic acid. Acidified samples were loaded on column, followed by washing with three column volumes of 0.2% formic acid or 0.1% TFA. Peptides were eluted off the column by the addition of 500 µL 40% ACN with either 0.2% formic acid or 0.1% TFA and 500 µL 80% ACN with either 0.2% formic acid or 0.1% TFA. Peptide concentration was measured using a quantitative colorimetric peptide assay (Thermo). LC-MS/MS analyses were performed using previously described methodologies1, 2.

LC/MS data analysis. Raw data files were acquired in batches of 60 (3 biological replicates of 19 Δ*gene* strains and 1 WT strain) with time between LC-MS analyses minimized to reduce run-to-run variation. Batches of raw data files were subsequently processed using MaxQuant[36] (Version 1.5.0.25). Searches were performed against a target-decoy[37] database of reviewed yeast proteins plus isoforms (UniProt, downloaded January 20, 2013) using the Andromeda[38] search algorithm. Searches were performed using a precursor search tolerance of 4.5 ppm and a product mass tolerance of 0.35 Da. Specified search parameters included fixed modification for carbamidomethylation of cysteine residues and a variable modification for the oxidation of methionine and protein N-terminal acetylation, and a maximum of 2 missed tryptic cleavages. A 1% peptide spectrum match (PSM) false discovery rate (FDR) and a 1% protein FDR was applied according to the target-decoy method. Proteins were identified using at least one peptide (razor + unique). Proteins were quantified using MaxLFQ with an LFQ minimum ratio count of 2. LFQ intensities were calculated using the match between runs feature, and MS/MS spectra were not required for LFQ comparisons. Missing values were imputed where appropriate for proteins quantified in $\geqslant$ 50% of MS data files in a batch. Proteins not meeting this requirement were omitted from subsequent analyses. Imputation was performed on a replicate-by-replicate basis. For each replicate MS analysis a normal distribution with mean and standard deviation equivalent to that of the lowest 1% of measured LFQ intensities was generated. Missing values were filled in with values drawn from this distribution at random. Approximately 4.05% and 4.53% of quantitative measurements were imputed in the respiration and fermentation proteomic datasets, respectively. Replicate protein LFQ values from corresponding Δ*gene* or WT strains were pooled, $\log_2$ transformed, and averaged (mean $\log_2$[strain], n = 3). Average Δ*gene* LFQ intensities were normalized against their appropriate WT control (mean $\log_2[\Delta gene/\mathrm{WT}]$, n = 3) and a 2-tailed t-test (homostatic) was performed to obtain P values.

To control for batch-specific effects, proteins having unexpected and characteristic misregulation across a majority of Δ*gene* strains processed together were identified and omitted from the dataset. For each protein quantified within a batch of Δ*gene* strains a distribution of protein fold-changes (intra-batch) was generated. The analogous distribution of protein fold-changes from all other Δ*gene* strains processed separately (inter-batch) was created. These two distributions were compared against each other using a Kolmogorov-Smirnov test (2-tailed) to obtain P values. If a significant difference existed at $P < 0.05$ (Bonferroni-adjusted) protein abundance measurements were omitted from the batch in question. This process of comparing intra-batch and inter-batch protein fold change distributions was carried iteratively and to exhaustion and resulted in the omission of an average 165 proteins/Δ*gene* strain (~4.8% of quantified proteins) for respiration, and 188 proteins/Δ*gene* strain (~5.9%) for fermentation.

Gas chromatography-mass spectrometry (GC-MS) metabolomics. 1×10^8 yeast cells yeast cells were isolated by rapid vacuum filtration onto a nylon filter membrane (0.45 μm pore size, Millipore) using a Glass Microanalysis Filter Holder (Millipore), briefly washed with phosphate buffered saline (1 mL), and immediately submerged into ACN/MeOH/H2O (2:2:1, v/v/v, 1.5 mL, pre-cooled to –20 °C) in a plastic tube. The time from sampling yeast from the culture to submersion in cold extraction solvent was less than 30 s. Tubes with the extraction solvent, nylon filter, and yeast were stored at –80 °C prior to analysis.

Tubes with yeast extract (also still containing insoluble yeast material and the nylon filter) were thawed at room temperature for 45 min., vortexed (~15 s), and centrifuged at room temperature (6400 rpm, 30 s) to pellet insoluble yeast material. Yeast extract (25 μL aliquot) and internal standards (25 μL aqueous mixture of isotopically labelled alanine-2,3,3,3-d_4, adipic acid-d_{10}, and xylose-$^{13}C_5$ acid, 5 ppm in each) were aliquoted into a 2 mL plastic tube and dried by vacuum centrifuge (~1 hr). The dried metabolites were resuspended in pyridine (25 μL) and vortexed. 25 μL of N-methyl-N-trimethylsilyl]trifluoroacetamide (MSTFA) with 1% trimethylchlorosilane (TMCS) was added, and the sample was vortexed and incubated (60 °C, 30 min). Samples were then transferred to a glass autosampler vials and analyzed using a GC/MS instrument comprising a Trace 1310 GC coupled to a Q Exactive Orbitrap mass spectrometer. For the yeast metabolite extracts a linear temperature gradient ranging from 50 °C to 320 °C was employed spanning a total runtime of 30 minutes. Analytes were injected onto a 30 m TraceGOLD TG-5SILMS column (Thermo) using a 1:10 split at a temperature of 275 °C and ionized using electron ionization (EI). The mass spectrometer was operated in full scan mode using a resolution of 30,000 ($m/\Delta m$) relative to 200 *m/z*.

GC/MS data analysis. The resulting GC-MS data were processed using an in-house developed software suite (https://github.com/coongroup/Y3K-Software). Briefly, all *m/z* peaks are aggregated into distinct chromatographic profiles (i.e., feature) using a 10 ppm mass tolerance. These chromatographic profiles are then grouped according to common elution apex (i.e., feature group). The collection of features (i.e., *m/z* peaks) sharing a common elution apex, therefore, represent an individual EI-MS spectrum of a single eluting compound. The EI-MS spectra were then compared against a matrix run and a background subtraction was performed. Remaining EI-MS spectra are then searched against the NIST 12 MS/EI library and subsequently subjected to a high resolution filtering (HRF) technique as described elsewhere. EI-MS spectra that were not identified were assigned a numeric identifier. Feature intensity, which was normalized using total metabolite signal, was used to estimate metabolite abundance. Following initial processing, raw data files were re-analyzed to extract metabolite signals which were not successfully deconvolved and registered as missing values in the dataset. This process provided measurements for ~1.87%, and 2.25% of metabolites quantified in the respiration and fermentation datasets, respectively. Remaining missing values were imputed using the same imputation strategy as described in the proteomic data processing section. Quantitative values imputed using this process account for ~0.17% and 0.13% of metabolites in the respiration and fermentation datasets, respectively.

Replicate metabolite intensities from corresponding Δ*gene* or WT strains were pooled, $\log_2$ transformed, and averaged (mean $\log_2$[strain], n = 3). Average Δ*gene* metabolite intensities were normalized against their appropriate WT control (mean $\log_2$[Δ*gene*/WT], n = 3) and a 2-tailed t-test was performed to obtain *P* values. To account for batch-specific effects the same Kolmogorov–Smirnov testing approach as described in the proteomic data processing section was used. Distributions of inter-batch and intra-batch metabolite fold changes were compared iteratively and those that were significantly different at $P < 0.05$ (Bonferroni-adjusted) resulted in metabolite abundance measurements being omitted from the batch in question (~15 metabolites/Δ*gene* strain (~5.0%) from respiration and ~21 metabolites/Δ*gene* strain (~5.9%) from fermentation).

Δ*Gene*-specific phenotype detection. For each profiled molecule (in both respiration and fermentation growth conditions) we separated potential Δ*gene*-specific measurements into two groups: positive $\log_2$ fold change ($\log_2$[Δ*gene*/WT]) and negative $\log_2$ fold change. These two sets were then plotted individually with $\log_2$ fold change and $-\log_{10}$(p-value [two-sided Student's t-test]) along the x- and y- axes, respectively. Data were normalized such that the largest $\log_2$ fold change and largest $-\log_{10}$(*p*-value) were set equal to 1. Considering the three largest fold changes where $P < 0.05$, we calculated the Euclidean distance to all neighboring data points and stored the smallest result. A requirement was imposed that all considered 'neighbors' have a smaller fold change than the data point being considered. It is anticipated that data points corresponding to Δ*gene*-specific phenotypes will be outliers in the described plots and have large associated nearest-neighbor Euclidean distances. The described routine yielded three separate distances, the largest of which was stored for further analysis. We set a cutoff for classification as a 'Δ*gene*-specific phenotype' at a Euclidean distance of 0.70.

Regression analysis of Δ*gene*–Δ*gene* perturbation profiles. For all pairwise combinations of Δ*gene* strains from the same growth condition linear regression analysis was conducted on protein, lipid, and metabolite perturbation profiles, respectively. Fold change measurements (mean $\log_2[\Delta gene/\mathrm{WT}]$, n = 3) from molecules where FC > 0.7 and $P < 0.05$ were used and a minimum of 20 proteins, 10 metabolites, and 5 lipids, respectively, were required. These measurements were fit to a line and the associated Pearson correlation coefficient was reported. Coefficients carrying negative signs were set to 0. For pairs of Δ*gene* strains lacking a sufficient number of molecules that met the aforementioned criteria, the Pearson coefficient was reported as 0. Hierarchical clustering of Δ*gene*–Δ*gene* correlations was performed as described below.

Respiration deficiency response (RDR) abundance adjustment. All Δ*gene* strains grown under respiration conditions were classified as respiration deficient (RD) (51) or respiration competent (RC) (123) based on observation of a common perturbation profile signature. For all molecules profiled within RD Δ*gene* strains an RDR score was calculated. This metric represents the proportion of RD Δ*gene* strains over which the molecule was consistently perturbed, relative to all RD Δ*gene* strains where the molecule was quantified. Considering all RD Δ*gene* strains, 776 molecules produced an RDR score > 0.95 (consistently perturbed across more than 95% of RD Δ*gene* strains where quantified) and were subsequently classified as RDR-associated. For each RDR-associated molecule, individual RD Δ*gene* strain measurements were mean normalized and stored. These RDR-adjusted measurements were then used in described respiration–RDR analyses.

Regression analysis of RDR-adjusted Δ*gene*–Δ*gene* perturbation profiles. For all RD Δ*gene* strains linear regression analysis was performed pairwise on RDR-adjusted protein perturbation profiles. Fold change measurements from molecules where FC > 0.7 and $P < 0.05$ (p-value prior to RDR adjustment) were used and a minimum of 20 proteins was required. Correlations and clustering were otherwise conducted as described above.

Hierarchical clustering. All hierarchical clustering performed in this study was done in Perseus. For all clustering operations Spearman correlation was used with average linkage, preprocessing with k-means, and the number of desired clusters set to 300 for both rows and columns.

For clustering of Δ*gene* perturbation profiles, clustering was performed separately for fermentation and respiration datasets, and column-wise cluster order for fermentation and respiration datasets was generated using only protein fold change profiles. Column ordering was then applied to metabolite and lipid fold change datasets from the corresponding growth condition and row-wise clustering was conducted. GO term enrichment was performed in Perseus. P values were obtained from a Fisher's exact test, adjusted for multiple hypothesis testing[39] and reported where $P < 0.05$.

For the analysis of Δ*gene*–Δ*gene* correlations, clustering was performed on respiration protein perturbation profile correlation data and the resultant ordering was applied to Δ*gene*–Δ*gene* correlation datasets from all other omes and growth conditions for parallel visual display. The same clustering process was carried out for the analysis of Δ*gene*–Δ*gene* correlations of RD Δ*gene* strains following RDR-adjustment.

Generation of Δ*gene* strains and cloning of genes and mutants for follow-up studies. *S. cerevisiae* (BY4742) gene deletion strains for *hfd1*, *atp2*, *ypr010c-a*, and *yjr120w* were generated using a PCR deletion strategy in which the open reading frames were replaced by a KanMX cassette from the pFA6a-kanMX6 plasmid. Briefly, KanMX was amplified with primers containing sequence homologous to sequence just upstream of the ATG and just downstream from the terminal codon for each ORF. Amplicons were transformed into BY4742, and yeast were plated onto YEPD plates containing 100 μg/mL G418. Knockouts were confirmed by PCR and sequencing.

To generate plasmid yeast gene constructs, *S. cerevisiae hfd1*, *atp2*, and *yjr120w* were amplified by Accuprime Pfu polymerase (Invitrogen, USA) with primers generating a SpeI site (forward) and SalI (reverse) (BamHI forward and EcoRI reverse for *yjr120w*). The *hfd1*, *atp2*, and *yjr120w* amplicons and the yeast expression vectors p426GPD and p423GPD were digested with SpeI and SalI or BamHI and EcoRI. *Hfd1* and *yjr120w* were ligated to p426GPD, *atp2* was ligated to p423GPD, and each ligation was transformed into DH5α *E. coli*. Plasmid minipreps were performed and recombinants were confirmed by sequencing. *Hfd1* mutants were generated via standard site-directed mutagenesis, and mutations were confirmed by sequencing.

To generate plasmid human gene constructs, *Homo sapiens ALDH3A1* and *ALDH3A2* were amplified by Accuprime Pfu polymerase with primers generating a SpeI site (forward) and SalI (reverse). The *ALDH3A1* and *ALDH3A2* amplicons and the yeast expression vector p426GPD were digested with SpeI and SalI. *ALDH3A1* and *ALDH3A2* were ligated to p426GPD and each ligation was transformed into DH5α *E. coli*. Plasmid minipreps were performed and recombinants were confirmed by sequencing.

***Yjr120w* molecular biology studies—yeast growth assays.** Δ*atp2* and Δ*yjr120w* yeast were transformed with p426GPD plasmids (either encoding for Yjr120w or empty vector) and p423GPD (either encoding for Atp2p or empty vector) and grown on Ura$^-$, His$^-$ plates containing 2% glucose. Starter cultures were inoculated with individual colonies of yeast and incubated (30 °C, ~16 h, 230 rpm). To assay Δ*atp2* and Δ*yjr120w* yeast growth on agar plates, serial dilutions of yeast from a starter culture were prepared in Ura$^-$, His$^-$

media lacking glucose. 10-fold serial dilutions of yeast cells were dropped onto Ura$^-$, His$^-$ agar media plates containing either glucose (2%, w/v) or glycerol (3%, w/v) and incubated (30 °C, 4 d).

***Yjr120w* molecular biology studies—mRNA quantitation.** BY4742 WT, Δ*coq8*, Δ*atp2*, and Δ*yjr120w* yeast were grown overnight in 3 mL YEPD. From the overnight culture, 2.5×10^6 cells were used to inoculate 100 mL YPGD media. 1 mL of culture was collected after 25 hours and total RNA was isolated using Masterpure Yeast RNA Purification Kit (Epicentre). 1 μg of RNA was reverse transcribed using Superscript III first strand synthesis kit (Thermo). Using the resultant cDNA as template, set up QPCR reactions: 2 μL cDNA, 12.5 μL Power Sybr Green Master Mix (Thermo), and 300 nmol/L forward and reverse primers. Primers amplifying the following targets were used: *atp2*, *yjr120w*, and *ubc6* (reference gene). QPCR cycled as follows: After an initial 2 minute incubation at 50 °C, template was denatured at 95 °C for 10 minutes, cycled 40 times: 95 °C for 15 s, 60 °C for 1 minute. RNA abundance was calculated using the $\Delta\Delta Ct$ method.

Hfd1p and ALDH3A1 biochemical studies—media lacking pABA. A specially formulated synthetic media lacking pABA ('pABA$^-$') was used for numerous follow-up studies in this project. This media consisted of CSM Mixture; Complete, 790 mg/L (# DCS0019, Formedium LTD, Hunstanton, U.K.) and yeast nitrogen base without amino acids and para-amino benzoic acid, 6.9 g/L (# CYN4102, Formedium LTD, Hunstanton, U.K.).

Hfd1p and ALDH3A1 biochemical studies—yeast growth assays. Δ*hfd1* yeast transformed with p426GPD plasmids encoding for Hfd1p variants were grown on uracil drop-out (Ura^-) synthetic media plates containing glucose (2%, w/v). Individual colonies of yeast were used to inoculate starter cultures of synthetic media lacking pABA ($pABA^-$) but containing 20 g/L glucose. To assay WT and Δ*hfd1* yeast growth on agar plates, serial dilutions of yeast from a starter culture were prepared in $pABA^-$ media lacking glucose. 10^4, 10^3, or 10^2 yeast cells were dropped onto agar media plates containing either glucose (2%, w/v) or glycerol (3%, w/v) and incubated (30 °C, 4 d). The base medias for the agar plates consisted of either YEP (rich media), synthetic complete, $pABA^-$, $pABA^-$ supplemented with 100 μM 4-hydroxybenzoic acid, or pABA– supplemented with 100 μM pABA.

To assay yeast growth in liquid media, yeast from a $pABA^-$ starter culture were swapped into $pABA^-$ media with glucose (0.1%, w/v) and glycerol (3%, w/v) (base medium) at an initial density of 5×10^6 cells/mL. To interrogate the rescue efficacy of various compounds, 100 nM (final concentrations) of pABA, tyrosine, 4-HPP, 4-HPAA, 4-HPA, 4-HMA, 4-HBz, 4-HB, 4HPL, or *p*-coumarate were added to the base medium. The cultures were incubated in a sterile 96 well plate with an optical, breathable coverseal (shaking at 1140 rpm). Optical density readings (OD_{600}) were obtained every 10 min. Respiratory growth rates were determined by fitting a linear equation to the respiratory growth phase and determining the slope of the line. Relative respiratory growth rates were determined by comparing cultures with additives to those without additive.

Hfd1p and ALDH3A1 biochemical studies—Quantitation of CoQ and 4-HB in pABA⁻ Δ*hfd1* yeast cultures. 2.5×10^6 Δ*hfd1* yeast cells from a pABA⁻ (2% w/v glucose) starter culture were used to inoculate 100 mL of pABA⁻ media with glucose (0.1%, w/v), glycerol (3%, w/v), and potential rescue compound (100 nM pABA, 4-HPP, 4-HPAA, 4-HPA, 4-HBz, 4-HB, or none). These 100 mL cultures were incubated (30 °C, 230 rpm). After 25 h (analogous to the primary respiration culture system used for this study), 1×10^8 yeast cells were harvested for lipidomic or metabolomic analyses, and CoQ and 4-HB were quantified by mass spectrometry as described above. These cultures and analyses were conducted in biological triplicate.

Hfd1p and ALDH3A1 biochemical studies—Hfd1p phylogenetics. The amino acid sequences of the 19 known *Homo sapiens* ALDH proteins25 and *S. cerevisiae* Hfd1p (NP_013828.1) were aligned by MUSCLE[40], analyzed by ClustalW2 Phylogeny[41], and visualized in iTOL[42].

Hfd1p and ALDH3A1 biochemical studies—Mass spectrometry profiling of pABA⁻ yeast cultures (WT, Δ*hfd1*, Δ*dpl1*, and Δ*coq8*). 2.5×10^6 yeast cells from a pABA⁻ (2% w/v glucose) starter culture were used to inoculate 100 mL of pABA⁻ media with glucose (0.1%, w/v), glycerol (3%, w/v), and rescue compound (100 μM 4-HB or none). These 100 mL cultures were incubated (30 °C, 230 rpm). After 25 h, 1×10^8 yeast cells were harvested for lipidomic, metabolomics, and proteomic analyses by mass spectrometry as described in the main Methods section. These cultures and analyses were conducted in biological triplicate.

Hfd1p and ALDH3A1 biochemical studies—Hfd1p, ALDH3A1, and ALDH3A2 expression and purification. PIPE cloning was used to generate pVP68K vectors encoding ALDH3A1, Hfd1p$^{C\Delta25}$, or ALDH3A2$^{C\Delta25}$ (Hfd1p or ALDH3A2 lacking their C-terminal 25 amino acids, which comprise putative transmembrane domains) fused to an 8His-cytoplasmically-targeted maltose-binding protein with a linker including a tobacco etch virus protease recognition site (8His-MBP-[TEV]-ALDH3A1, 8His-MBP-[TEV]-Hfd1p$^{C\Delta25}$, or 8His-MBP-[TEV]-ALDH3A2$^{C\Delta25}$). These constructs were expressed in *E. coli* (BL21[DE3]-RIPL strain) by autoinduction. Cells were isolated and resuspended in lysis buffer (50 mM HEPES, 300 mM NaCl, 10% glycerol, 5 mM BME, 0.25 mM PMSF, 1 mg/mL lysozyme (Sigma), pH 7.5). Cells were lysed by sonication (4 °C, 2 × 20 s), and the lysate was clarified by centrifugation (15,000 g, 30 min, 4 °C). The clarified lysate was mixed with cobalt IMAC resin (Talon resin) and incubated (4 °C, 1 h). The resin was pelleted by centrifugation (700 g, 2 min, 4 °C) and washed three times (10 resin bed volumes each) with wash buffer (50 mM HEPES, 300 mM NaCl, 10% glycerol, 5 mM BME, 0.25 mM PMSF, 10 mM imidazole, pH 7.5). His-tagged protein was eluted with elution buffer (50 mM HEPES, 300 mM NaCl, 10% glycerol, 5 mM BME, 0.25 mM PMSF, 100 mM imidazole, pH 7.5). The eluted protein was concentrated with a 50-kDa MW-cutoff spin filter (Merck Millipore Ltd.) and exchanged into storage buffer (50 mM HEPES, 300 mM NaCl, 10% glycerol, 5 mM BME, 0.25 mM PMSF, pH 7.5). Protein concentrations were determined by absorbance at 280 nm. The MBP-fusion proteins were aliquoted, frozen in $N_{2(l)}$, and stored at –80 °C.

Hfd1p and ALDH3A1 biochemical studies—Hfd1p, ALDH3A1, and ALDH3A2 enzymology. Enzyme activity assays were conducted in groups of three replicate 100 μL reactions, each containing MBP-fusion protein (0.2–25 μg), 1 mM NAD^+, and 200 μM substrate (4-HBz or hexadecanal (Avanti 857458M)) in an aqueous buffer (50 mM Tris pH 8.0, 150 mM NaCl, 0.1% Triton X-100). NADH production was observed by monitoring fluorescence (356 nm excitation, 460 nm emission) over a 30–60 minute period with a Cytation 3 Imaging Reader (BioTek). K_M and k_{cat} values were determined by measuring reaction rates in the linear range at varying substrate (4-HBz or hexadecanal) concentrations. Curve fitting to generate Michaelis-Menten parameters was performed using SigmaPlot (Systat Software, San Jose, CA). Reported activity represents the mean of three separate protein purifications.

Molecule Covariance Network Analysis For all pairwise combinations of molecules quantified within a particular growth condition, regression analysis was conducted using fold change measurements from all Δ*gene* strains having a measurement for both molecules in the pair. Spearman's regression analysis was performed to obtain correlation coefficients (ρ). From these test statistics *P* values were calculated using a two-sided Student's t-test. All *P* values were corrected for multiple hypothesis testing (Bonferroni) and correlations where $|\rho| \geqslant 0.58$ and $P < 0.001$ were reported. For RDR-adjusted regression analysis, the RDR adjustment procedure was carried out as described in the 'Respiration deficiency response (RDR) abundance adjustment' section (above). All pairs of covariant molecules are visualized as networks generated using the Gephi open graph visualization platform (version 0.9.0). Complete respiration, fermentation and RDR-adjusted respiration network layouts were generated using the Fruchterman–Reingold graph-drawing algorithm with area set to 10,000 and gravity set to 30. Gene Ontology terms were obtained from the Saccharomyces Genome Database (SGD). To calculate network selectivity the following equation was used:

$$S_{\text{MCN}} = [E_{\text{Obs,In}}/E_{\text{Tot,In}}]/[(E_{\text{Obs,Out}} + 1)/E_{\text{Tot,Out}}]$$

Where S_{MCN} represents the selectivity coefficient for the molecule covariance network (MCN) surrounding an individual node of interest, $E_{\text{Obs,In}}$ is the number edges observed within a pathway of interest, $E_{\text{Tot,In}}$ is the number of total possible edges within the pathway of interest, $E_{\text{Obs,Out}}$ is the number of edges observed to molecules outside the pathway of interest, and $E_{\text{Tot,Out}}$ is the number total possible edges to molecules outside the pathway of interest.

Gene ontology (GO) term enrichment analysis was performed using a Fisher's exact test with subsequent Benjamini-Hochberg FDR adjustment39 to account for multiple hypothesis testing.

Proteomic analysis of Δ*yor020w-a* yeast 2.5×10^6 yeast cells from a pABA⁻ (2% w/v glucose) starter culture (Δ*yor020w-a* or WT) were used to inoculate 100 mL of pABA⁻ media with glucose (0.1%, w/v) and glycerol (3%, w/v). These 100 mL cultures were incubated (30 °C, 230 rpm). After 25 h, 1×10^8 yeast cells were harvested for proteomic analyses by mass spectrometry as described in the main Methods section. These cultures and analyses were conducted in biological duplicate.

Quantitation of CoQ and PPHB in pABA⁻ Δ*aro9*, Δ*aro10*, Δ*aim18*, and WT yeast cultures 2.5×10^6 yeast cells from a pABA⁻ (2% w/v glucose) starter culture were used to inoculate 100 mL of pABA⁻ media with glucose (0.1%, w/v) and glycerol (3%, w/v). These 100 mL cultures were incubated (30 °C, 230 rpm). After 25 h, 1×10^8 yeast cells were harvested for lipid analysis, and CoQ and PPHB were quantified by mass spectrometry as described in the Main methods section. These cultures and analyses were conducted in biological duplicate.

Respiration deficiency response analysis The densities of Δ*gene* cultures were compared to those of WT cultures (2-tailed T-test). Strains with slow growth in fermentation cultures ($\Delta gene/\mathrm{WT} \leqslant 0.2$ and $P < 0.05$) were categorized as 'slow fermentation growth' strains (8 strains). Remaining strains were grouped into three categories based on their growth rates in respiration cultures. Strains with significantly decreased respiration growth ($\Delta gene/\mathrm{WT} < 0.6$ and $P < 0.05$) were considered respiration deficient (RD) (41 RD strains). Strains with borderline respiration growth ($0.6 \leqslant \Delta gene/\mathrm{WT} < 0.8$) were categorized as 'borderline respiration' (14 strains). Strains with respiration growth rates near WT or better than WT ($0.8 \leqslant \Delta gene/\mathrm{WT}$) were categorized as respiration competent (RC) (111 RC strains).

For PCA, average $\log_2(\Delta gene/\mathrm{WT})$ values for each protein, metabolite, and lipid measured in the respiration condition were analyzed using Perseus PCA software. PCA projections were exported from Perseus.

For volcano plot analyses, average $\log_2$(RD/RC) values were calculated as [mean $\log_2$(RD Δ*gene* strains/WT)] – [mean $\log_2$(RC Δ*gene* strains/WT)]. A t-test (2-tailed, homostatic) was performed to obtain P values. P values were corrected for multiple hypothesis testing by multiplying each P value obtained by the number of biomolecules included in this analysis (4,116) (Bonferroni correction).

For GO term analyses, proteins were separated as increasing in RD strains (positive $\log_2$[RD/RC]) or decreasing in RD strains (negative $\log_2$[RD/RC]). Proteins with Bonferroni-corrected $P < 1\times10^{-20}$ were collected from each group and subjected to GO term enrichment analysis (http://geneontology.org/page/go-enrichment-analysis). Select GO terms were highlighted because they were significantly enriched (Bonferroni corrected $P < 0.05$) in proteins that were reduced (–) or increased (+) in RD strains. Boxplots of select molecules were generated using matplotlib in python to compare particular molecules across all RD and RC strains.

For ROC analysis, RD strains were considered positive examples whereas RC cells were considered negative examples. Using the $\log_2$(Δ*gene*/WT) values for individual biomolecules as a discriminator, ROCs were generated by calculating false positive rate (FPR) and true positive rate (TPR) for values that fall above a particular cutoff for molecules that are increased in RD strains relative to WT and below that cutoff for molecules that are decreased in RD strains relative to WT. A + sign indicates that an increase in that molecule is predictive of RD whereas a – sign indicates that a reduction in that molecule is predictive of RD.

Supplementary Notes

Development of a stable and reproducible respiration culture condition. To profile diverse yeast strains during respiratory growth, when mitochondrial OxPhos is highly active, we first needed to develop a distinct respiration condition suitable for large-scale investigation. Early log phase fermentation cultures repress mitochondrial respiration, cultures containing solely non-fermentable sugars preclude growth of respiration deficient yeast, and high glucose cultures grown past the diauxic shift are too biologically dynamic to allow reproducible sampling across a large scale study[43,44]. To overcome these problems, we developed a culture system that includes low glucose (1 g/L) and high glycerol (30 g/L), enabling a short fermentation phase followed by a longer respiration phase. This respiration condition affords steady growth and a stable biological state—as reflected by a proteome that is constant over multiple hours **(Supplementary Fig. S6.1c–e)**—and, thus, an essential window for reproducible sample harvesting.

Δ*Gene*-specific phenotype detection. To identify Δ*gene*-specific phenotypes, we broadly surveyed our data for characteristic outlier abundance measurements. For each profiled molecule (in both respiration and fermentation growth conditions) we separated potential Δ*gene*-specific measurements into two groups: positive $\log_2$ fold change ($\log_2[\Delta gene/\mathrm{WT}]$)

and negative $\log_2$fold change. These two sets were then plotted individually with $\log_2$ fold change and $-\log_{10}$(p-value [two-sided Student's t-test]) along the x- and y- axes, respectively. Data were normalized such that the largest $\log_2$ fold change and largest $-\log_{10}(p\text{-value})$ were set equal to 1. Considering the three largest fold changes where $P < 0.05$, we calculated the Euclidean distance to all neighboring data points and stored the smallest result. A requirement was imposed that all considered 'neighbors' have a smaller fold change than the data point being considered. It is anticipated that data points corresponding to Δ*gene*-specific phenotypes will be outliers in the described plots and have large associated nearest-neighbor Euclidean distances. The described routine yielded three separate distances, the largest of which was stored for further analysis. The results of this analysis and representative examples are highlighted **(Fig. 6.2, Supplementary Figs. S6.5 and S6.6)**. We observed maximal Euclidean distances across a range of 0.006 to 1.25. We set a cutoff for classification as a 'Δ*gene*-specific phenotype' at 0.70 and report 714 molecules (4.6% of considered cases across both culture conditions) which exceed this threshold. This procedure provided a useful 'first pass' analysis and afforded a truncated set of leads, which were used to develop biological hypotheses.

Lack of effect of Dpl1p disruption on the Tyr-to-4-HB-to-CoQ pathway. To test the idea that the CoQ biosynthesis and sphingolipid catabolism pathways are independent, we examined Δ*dpl1* yeast, which lack a known dihydrosphingosine phosphate lyase. Δ*dpl1* yeast show neither a pABA^- respiratory growth phenotype nor CoQ deficiency **(Supplementary**

Fig. S6.7j,k). These results demonstrate that disruption of the Tyr-to-4-HB pathway in Δ*hfd1* yeast is not downstream of a defect in sphingolipid metabolism. Furthermore, proteome analyses showed that Δ*hfd1* cultured without 4-HB and pABA are similar to Δ*coq8* yeast—but not Δ*dpl1* yeast—and adding 4-HB to Δ*hfd1* cultures returns their proteomes to WT-like profiles **(Supplementary Fig. S6.7l,m)**.

Quantitative definition of the respiration deficiency response (RDR). To quantitatively define the RDR, we categorized strains as respiration deficient (RD) or competent (RC) and examined differences between these two groups. Principal component analysis of the Y3K respiration dataset revealed marked separation of RD and RC strains **(Fig. 6.3c and Supplementary Fig. S6.8a)**. The underlying phenotype changes that distinguish RD and RC strains include proteins, lipids, and metabolites **(Fig. 6.3d)**. RDR perturbations include significant decreases in ATP synthase, TCA cycle, and MICOS proteins **(Fig. 6.3e,f and Supplementary Fig. S6.8b)**, likely to decrease allocation of useless proteome mass to dysfunctional mitochondria[45]. Importantly, the RDR also includes a positive response, and numerous proteins—including protein folding, NADH metabolism, and proteasome assembly proteins—are significantly upregulated in RD strains **(Fig. 6.3e,f)**. Numerous individual molecules—including lactate, alanine, 2-hydroxyglutarate, tyrosol, 4-HB, Gpx2p, and Ahp1p, among many others—are significantly perturbed in RD strains and strongly predictive of respiration deficiency **(Supplementary Fig. S6.8c,d)**. Our quantitative assessment of the RDR highlights biochemical features of the cellular response to defects in mitochondrial respiration, and suggests that a multi-omic assessment of proteins, lipids, and metabolites could afford a highly specific biomarker panel for diseases affected by OxPhos deficiency.

RDR normalization procedure. Δ*gene* strains were classified as RD (51) or respiration competent (RC) (123) based on observation of a common perturbation profile signature in the respiration culture condition. For each molecule we calculated an RDR score. This metric represents the proportion of RD Δ*gene* strains over which the molecule was consistently perturbed, relative to all RD Δ*gene* strains where the molecule was quantified. Across all RD Δ*gene strains*, 776 molecules were identified as having an RDR score > 0.95 (consistently perturbed across more than 95% of RD Δ*gene* strains where quantified) and classified as RDR-associated. The individual measurements of these RDR-associated molecules were then mean normalized ('RDR-adjusted') using abundance values from RD Δ*gene* strains. This normalization procedure revealed characteristic deviations from the general RDR **(Supplementary Fig. S6.9)**. Importantly, this procedure enables visualization of Δ*gene*-specific changes. For example, prior to RDR normalization, the expected decrease in Coq8p in Δ*coq8* yeast is obscured by RDR-associated proteins with large abundance changes **(Supplementary Fig. S6.9d)**. RDR normalization not only uncovers the decrease in Coq8p, but a significant decrease in Coq5p, a functionally-related CoQ biosynthesis protein, also becomes readily apparent **(Supplementary Fig. S6.9d)**.

Molecular defects of Δ*yjr120w* yeast. To examine the molecular basis for the CoQ deficiency of Δ*yjr120w* yeast, we inspected our proteomics dataset, which revealed significant decreases in ATP synthase proteins, especially Atp2p **(Supplementary Fig. S6.10a)**. Compared to other strains, the large decrease in Atp2p is unique to Δ*yjr120w* and Δ*atp2* **(Supplementary Fig. S6.10b)**. A relationship between *yjr120w* and *atp2* is also suggested by their genetic proximity **(Supplementary Fig. S6.10c)**. Plasmid overexpression of *atp2* rescues the Δ*yjr120w* respiratory growth defect **(Supplementary Fig. S6.10d)**, indicating a functional relationship between *atp2* and *yjr120w in vivo*. A decrease in *atp2* mRNA in the Δ*yjr120w* strain is a component of the underlying mechanism **(Supplementary Fig. S6.10e)**. Interestingly, CoQ deficiency was also observed in Δ*atp2* yeast **(Fig. 6.3h)**.

Predicted enzymatic functions of Aim18p, Aro9p, and Aro10p. Since 1907, yeast have been known to catabolize amino acids into fusel (German for 'bad liquor') alcohols through the Ehrlich pathway[46,47], but the physiological roles for the enzymes involved—such as Aro9p and Aro10p—are not fully understood. Aro9p and Aro10p were previously thought to provide a simple catabolic route for extracting nitrogen from aromatic amino acids[48] **(Supplementary Fig. S6.14a)**, but our MCNA unexpectedly indicated strong correlations between Aro9p, Aro10p, and proteins involved in mitochondrial respiration **(Fig. 6.4d,e)**, suggesting a more complicated biological function that supports OxPhos. We hypothesized that this function might be in the Tyr-to-4-HB-to-CoQ pathway **(Supplementary Fig. S6.14b)**, given the putative enzymatic activities of Aro9p and Aro10p in tyrosine and phenylalanine metabolism. Consistently, when cultured in pABA$^-$ media, Δ*aro9* and Δ*aro10* yeast are deficient in CoQ and PPHB **(Fig. 6.4f)**.

Aim18p is a protein of undefined molecular function that has been detected in mitochondria[49] and potentially linked to mitochondrial inheritance (Altered Inheritance of Mitochondria, 'AIM') by large-scale studies in yeast[50]. Protein sequence alignments show that Aim18p contains a chalcone-flavone isomerase (CHI)-like domain **(Supplementary Fig. S6.14c)**, whose homologs in plants typically function on aromatic small molecules (chalcones) **(Supplementary Fig. S6.14d)**[51–53]. Given the potential for this protein domain to catalyze modifications of aromatic small molecules, we hypothesized that Aim18p might function in the Tyr-to-4-HB pathway to produce the CoQ headgroup (Supplementary Fig. 14d). Consistently, when cultured in pABA$^-$ media, we observed deficiency of PPHB in Δ*aim18* yeast **(Fig. 6.4f)**.

References

[1] 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," *Molecular & Cellular Proteomics*, vol. 13, pp. 339–347, 2014.

[2] 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," *Nature Protocols*, vol. 10, pp. 701–714, 2015.

[3] A. C. Peterson, J. P. Hauschild, S. T. Quarmby, D. Krumwiede, O. Lange, R. A. S. Lemke, F. Grosse-Coosmann, S. Horning, T. J. Donohue, M. S. Westphall, J. J. Coon, and J. Griep-Raming, "Development of a GC/quadrupole-orbitrap mass spectrometer, Part I: Design and characterization," *Analytical Chemistry*, vol. 86, pp. 10036–10043, 2014.

[4] N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P. Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda,

N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori, and M. Tomita, “Multiple high-throughput analyses monitor the response of E. coli to perturbations.,” *Science (New York, N.Y.)*, vol. 316, pp. 593–7, 2007.

[5] J. M. Buescher, W. Liebermeister, M. Jules, M. Uhr, J. Muntel, E. Botella, B. Hessling, R. J. Kleijn, L. Le Chat, F. Lecointe, U. Mäder, P. Nicolas, S. Piersma, F. Rügheimer, D. Becher, P. Bessieres, E. Bidnenko, E. L. Denham, E. Dervyn, K. M. Devine, G. Doherty, S. Drulhe, L. Felicori, M. J. Fogg, A. Goelzer, A. Hansen, C. R. Harwood, M. Hecker, S. Hubner, C. Hultschig, H. Jarmer, E. Klipp, A. Leduc, P. Lewis, F. Molina, P. Noirot, S. Peres, N. Pigeonneau, S. Pohl, S. Rasmussen, B. Rinn, M. Schaffer, J. Schnidder, B. Schwikowski, J. M. Van Dijl, P. Veiga, S. Walsh, A. J. Wilkinson, J. Stelling, S. Aymerich, and U. Sauer, “Global network reorganization during dynamic adaptations of Bacillus subtilis metabolism.,” *Science (New York, NY)*, vol. 335, pp. 1099–1103, 2012.

[6] E. G. Williams, Y. Wu, P. Jha, S. Dubuis, P. Blattmann, C. A. Argmann, S. M. Houten, T. Amariuta, W. Wolski, N. Zamboni, R. Aebersold, and J. Auwerx, “Systems proteomics of liver mitochondria function.,” *Science (New York, N.Y.)*, vol. 352, p. aad0189, 2016.

[7] J. M. Chick, S. C. Munger, P. Simecek, E. L. Huttlin, K. Choi, and M. Daniel, “Defining the consequences of genetic variation on a proteome-wide scale,” *Nature*, vol. 534, pp. 500–505, 2016.

[8] J. Nunnari and A. Suomalainen, “Mitochondria: In sickness and in health,” 2012.

[9] W. J. Koopman, P. H. Willems, J. A. M. Smeitink, and D. Ph, “Monogenic mitochondrial disorders,” *The New England journal of medicine*, vol. 366, pp. 1132–1141, 2012.

[10] S. B. Vafai and V. K. Mootha, “Mitochondrial disorders as windows into an ancient organelle.,” *Nature*, vol. 491, pp. 374–83, 2012.

[11] D. J. Pagliarini, S. E. Calvo, B. Chang, S. A. Sheth, S. B. Vafai, S. E. Ong, G. A. Walford, C. Sugiana, A. Boneh, W. K. Chen, D. E. Hill, M. Vidal, J. G. Evans, D. R. Thorburn, S. A. Carr, and V. K. Mootha, “A Mitochondrial Protein Compendium Elucidates Complex I Disease Biology,” *Cell*, vol. 134, pp. 112–123, 2008.

[12] S. E. Calvo, K. R. Clauser, and V. K. Mootha, “MitoCarta2.0: An updated inventory of mammalian mitochondrial proteins,” *Nucleic Acids Research*, vol. 44, pp. D1251–D1257, 2016.

[13] A. Sickmann, J. Reinders, Y. Wagner, C. Joppich, R. Zahedi, H. E. Meyer, B. Schönfisch, I. Perschil, A. Chacinska, B. Guiard, P. Rehling, N. Pfanner, and C. Meisinger, "The proteome of Saccharomyces cerevisiae mitochondria.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 13207–12, 2003.

[14] E. D. Green and M. S. Guyer, "Charting a course for genomic medicine from base pairs to bedside.," *Nature*, vol. 470, pp. 204–13, Feb. 2011.

[15] D. J. Pagliarini and J. Rutter, "Hallmarks of a new era in mitochondrial biochemistry," 2013.

[16] B. J. Floyd, E. M. Wilkerson, M. T. Veling, C. E. Minogue, C. Xia, E. T. Beebe, R. L. Wrobel, H. Cho, L. S. Kremer, C. L. Alston, K. A. Gromek, B. K. Dolan, A. Ulbrich, J. A. Stefely, S. L. Bohl, K. M. Werner, A. Jochem, M. S. Westphall, J. W. Rensvold, R. W. Taylor, H. Prokisch, J.-J. P. Kim, J. J. Coon, and D. J. Pagliarini, "Mitochondrial Protein Interaction Mapping Identifies Regulators of Respiratory Chain Function.," *Molecular cell*, vol. 63, pp. 621–32, 2016.

[17] C. M. Quinzii and M. Hirano, "Coenzyme Q and mitochondrial disease," 2010.

[18] A. Kalén, E. L. Appelkvist, and G. Dallner, "Age-related changes in the lipid compositions of rat and human tissues.," *Lipids*, vol. 24, pp. 579–584, 1989.

[19] R. Bentley and V. Ramsey, "The origin of the benzoquinone ring of coenzyme Q 9 in the rat," *Biochemical and Biophysical Research Communications*, vol. 5, no. 6, pp. 443–446, 1961.

[20] S. Merle, D. J. Robbins, and H. Emerson, "Phenolic Acid Metabolites of Tyrosine," vol. 236, no. 9, 1960.

[21] F. Pierrel, O. Hamelin, T. Douki, S. Kieffer-Jaquinod, U. Mühlenhoff, M. Ozeir, R. Lill, and M. Fontecave, "Involvement of mitochondrial ferredoxin and para-aminobenzoic acid in yeast coenzyme q biosynthesis," *Chemistry and Biology*, vol. 17, pp. 449–459, 2010.

[22] B. Marbois, L. X. Xie, S. Choi, K. Hirano, K. Hyman, and C. F. Clarke, "para-aminobenzoic acid is a precursor in coenzyme Q6 biosynthesis in Saccharomyces cerevisiae," *Journal of Biological Chemistry*, vol. 285, pp. 27827–27838, 2010.

[23] K. Nakahara, A. Ohkuni, T. Kitamura, K. Abe, T. Naganuma, Y. Ohno, R. A. Zoeller, and A. Kihara, "The Sjögren-Larsson Syndrome Gene Encodes a Hexadecenal Dehydrogenase of the Sphingosine 1-Phosphate Degradation Pathway," *Molecular Cell*, vol. 46, pp. 461–471, 2012.

[24] Z. J. Liu, Y. J. Sun, J. Rose, Y. J. Chung, C. D. Hsiao, W. R. Chang, I. Kuo, J. Perozich, R. Lindahl, J. Hempel, and B. C. Wang, "The first structure of an aldehyde dehydrogenase reveals novel interactions between NAD and the Rossmann fold.," *Nature structural biology*, vol. 4, pp. 317–326, 1997.

[25] B. Jackson, C. Brocker, D. C. Thompson, W. Black, K. Vasiliou, D. W. Nebert, and V. Vasiliou, "Update on the aldehyde dehydrogenase gene (ALDH) superfamily.," *Human genomics*, vol. 5, pp. 283–303, 2011.

[26] V. De Laurenzi, G. R. Rogers, D. J. Hamrock, L. N. Marekov, P. M. Steinert, J. G. Compton, N. Markova, and W. B. Rizzo, "Sjögren-Larsson syndrome is caused by mutations in the fatty aldehyde dehydrogenase gene.," *Nature genetics*, vol. 12, pp. 52–7, 1996.

[27] T. Kitamura, T. Naganuma, K. Abe, K. Nakahara, Y. Ohno, and A. Kihara, "Substrate specificity, plasma membrane localization, and lipid modification of the aldehyde dehydrogenase ALDH3B1.," *Biochimica et biophysica acta*, vol. 1831, pp. 1395–401, 2013.

[28] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. a. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend, "Functional Discovery via a Compendium of Expression Profiles," *Cell*, vol. 102, pp. 109–126, 2000.

[29] P. Kemmeren, K. Sameith, L. A. L. Van De Pasch, J. J. Benschop, T. L. Lenstra, T. Margaritis, E. O'Duibhir, E. Apweiler, S. Van Wageningen, C. W. Ko, S. Van Heesch, M. M. Kashani, G. Ampatziadis-Michailidis, M. O. Brok, N. A. C. H. Brabers, A. J. Miles, D. Bouwmeester, S. R. Van Hooff, H. Van Bakel, E. Sluiters, L. V. Bakker, B. Snel, P. Lijnzaad, D. Van Leenen, M. J. A. Groot Koerkamp, and F. C. P. Holstege, "Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors," *Cell*, vol. 157, pp. 740–752, 2014.

[30] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K.-D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Güldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kötter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai-Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C.-y. Wang, T. R. Ward, J. Wilhelmy, E. a. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke,

M. Snyder, P. Philippsen, R. W. Davis, and M. Johnston, “Functional profiling of the Saccharomyces cerevisiae genome.,” *Nature*, vol. 418, pp. 387–391, 2002.

[31] L. N. Laredj, F. Licitra, and H. M. Puccio, “The molecular genetics of coenzyme Q biosynthesis in health and disease,” 2014.

[32] U. C. Tran and C. F. Clarke, “Endogenous synthesis of coenzyme Q in eukaryotes,” *Mitochondrion*, vol. 7, 2007.

[33] S. Züchner, I. V. Mersiyanova, M. Muglia, N. Bissar-Tadmouri, J. Rochelle, E. L. Dadali, M. Zappia, E. Nelis, A. Patitucci, J. Senderek, Y. Parman, O. Evgrafov, P. D. Jonghe, Y. Takahashi, S. Tsuji, M. a. Pericak-Vance, A. Quattrone, E. Battaloglu, A. V. Polyakov, V. Timmerman, J. M. Schröder, and J. M. Vance, “Mutations in the mitochondrial GTPase mitofusin 2 cause Charcot-Marie-Tooth neuropathy type 2A.,” *Nature genetics*, vol. 36, pp. 449–451, 2004.

[34] A. Mourier, E. Motori, T. Brandt, M. Lagouge, I. Atanassov, A. Galinier, G. Rappl, S. Brodesser, K. Hultenby, C. Dieterich, and N. G. Larsson, “Mitofusin 2 is required to maintain mitochondrial coenzyme Q levels,” *Journal of Cell Biology*, vol. 208, pp. 429–442, 2015.

[35] A. S. Hebert, A. E. Merrill, J. a. Stefely, D. J. Bailey, C. D. Wenger, M. S. Westphall, D. J. Pagliarini, and J. J. Coon, “Amine-reactive neutron-encoded labels for highly plexed proteomic quantitation.,” *Molecular & cellular proteomics : MCP*, vol. 12, pp. 3360–9, 2013.

[36] J. Cox and M. Mann, “MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.,” *Nature biotechnology*, vol. 26, pp. 1367–72, 2008.

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

[38] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, and M. Mann, “Andromeda: A peptide search engine integrated into the MaxQuant environment,” *Journal of Proteome Research*, vol. 10, pp. 1794–1805, 2011.

[39] T. Author, Y. Benjamini, Y. Hochberg, and Y. Benjaminit, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society*, vol. 57, pp. 289–300, 1995.

[40] R. C. Edgar, “MUSCLE: Multiple sequence alignment with high accuracy and high throughput,” *Nucleic Acids Research*, vol. 32, pp. 1792–1797, 2004.

[41] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. Mcgettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins, "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23, pp. 2947–2948, 2007.

[42] I. Letunic and P. Bork, "Interactive Tree of Life v2: Online annotation and display of phylogenetic trees made easy," *Nucleic Acids Research*, vol. 39, 2011.

[43] P. Picotti, B. Bodenmiller, L. N. Mueller, B. Domon, and R. Aebersold, "Full Dynamic Range Proteome Analysis of S. cerevisiae by Targeted Proteomics," *Cell*, vol. 138, pp. 795–806, 2009.

[44] A. Casanovas, R. R. Sprenger, K. Tarasov, D. E. Ruckerbauer, H. K. Hannibal-Bach, J. Zanghellini, O. N. Jensen, and C. S. Ejsing, "Quantitative analysis of proteome and lipidome dynamics reveals functional regulation of global lipid metabolism," *Chemistry and Biology*, vol. 22, pp. 412–425, 2015.

[45] M. Basan, S. Hui, H. Okano, Z. Zhang, Y. Shen, J. R. Williamson, and T. Hwa, "Overflow metabolism in Escherichia coli results from efficient proteome allocation," *Nature*, vol. 528, pp. 99–104, 2015.

[46] F. Ehrlich, "Über die Bedingungen der Fuselölbildung und über ihren Zusammenhang mit dem Eiweissaufbau der Hefe," *Berichte der deutschen chemischen Gesellschaft*, vol. 40, pp. 1027–1047, 1907.

[47] L. H. Hazelwood, J.-M. G. Daran, A. van Maris, J. T. Pronk, and J. R. Dickinson, "The Ehrlich Pathway for Fusel Alcohol Production: a Century of Research on Saccharomyces cerevisiae Metabolism," *Applied and Environmental Microbiology*, vol. 74, pp. 2259–2266, 2008.

[48] M. M. Kneen, R. Stan, A. Yep, R. P. Tyler, C. Saehuan, and M. J. McLeish, "Characterization of a thiamin diphosphate-dependent phenylpyruvate decarboxylase from Saccharomyces cerevisiae," 2011.

[49] J. Reinders, R. P. Zahedi, N. Pfanner, C. Meisinger, and A. Sickmann, "Toward the complete yeast mitochondrial proteome: Multidimensional separation techniques for mitochondrial proteomics," *Journal of Proteome Research*, vol. 5, pp. 1543–1554, 2006.

[50] D. C. Hess, C. Myers, C. Huttenhower, M. A. Hibbs, A. P. Hayes, J. Paw, J. J. Clore, R. M. Mendoza, B. S. Luis, C. Nislow, G. Giaever, M. Costanzo, O. G. Troyanskaya, and A. A. Caudy, "Computationally driven, quantitative experiments discover genes required for mitochondrial biogenesis," *PLoS Genetics*, vol. 5, 2009.

[51] M. Gensheimer and A. Mushegian, "Chalcone isomerase family and fold: no longer unique to plants," *Protein Sci*, vol. 13, pp. 540–544, 2004.

[52] M. N. Ngaki, G. V. Louie, R. N. Philippe, G. Manning, F. Pojer, M. E. Bowman, L. Li, E. Larsen, E. S. Wurtele, and J. P. Noel, “Evolution of the chalcone-isomerase fold from fatty-acid binding to stereospecific catalysis.,” *Nature*, vol. 485, pp. 530–3, May 2012.

[53] J. M. Jez, M. E. Bowman, R. a. Dixon, and J. P. Noel, “Structure and mechanism of the evolutionarily unique plant enzyme chalcone isomerase.,” *Nature structural biology*, vol. 7, pp. 786–791, 2000.

COLOPHON

This document was typesetted with LaTeX 2_ε using the MiKTeX project. It is based on the University of Wisconsin dissertation template created by William C. Benton (available at `https://github.com/willb/wi-thesis-template`).