ABSTRACT

PORTER, CALEB JOSEPH. The Development of Advanced Measurement Technologies and Analysis Strategies via LC-MS/MS to Measure Global Site Specific Modifications—Application in the Exposome. (Under the direction of Drs. Michael Bereman & David Muddiman).

The sequencing of the Human Genome has had a tremendous impact on basic biological research and ignited several new fields of study. However, the promises of the identification of the causes, or discovery of cures to complex diseases, have not come to fruition. Coined by Christopher Wild in 2005, the exposome was postulated to complement the genome and raise awareness of the importance of the “environment” on human health. A key component of the exposome is reactive electrophiles as these molecules directly damage biopolymers including DNA and proteins. To date, the majority of studies of protein adducts have focused on a single electrophile bound to bulk tissue or detection of a protein adduct derived from one molecule. Many challenges exist in the detection of protein adducts, including: their occurrence at sub-stoichiometric levels requiring enrichment strategies, lack of multiplex methods, and no ‘gold-standard’ for quantitation. While many techniques have been utilized in the study of protein adducts, mass spectrometry is playing an essential role in the study, characterization, and discovery of protein adducts. Mass spectrometry is a versatile tool used routinely in analysis of proteins due to the high sensitivity and molecular specificity. Data acquisition modes have been developed to investigate specific biological questions. Data independent acquisition mass spectrometry bypasses problems encountered in the study of protein adducts which used data dependent acquisition or targeted methodologies. Herein, is presented the development of a novel method termed Multiplex Adduct Peptide Profiling.
(MAPP) which based on data independent acquisition mass spectrometry and targeted data analysis for global characterization of protein adducts derived from biological fluids.
The Development of Advanced Measurement Technologies and Analysis Strategies via LC-MS/MS to Measure Global Site Specific Modifications—Application in the Exposome

by

Caleb Joseph Porter

A thesis submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the degree of Master of Science

Chemistry

Raleigh, North Carolina

2015

APPROVED BY:

______________________________  ______________________________
Michael Bereman Co-Chair       David Muddiman Co-Chair

______________________________  ______________________________
Gufeng Wang                   Edmond Bowden
DEDICATION

To my wife Michelle for her loving support. To my parents, brother, and sisters who helped and encouraged me along the way. Finally to my daughter Natalee, this is what daddy was working on for most of your life.
BIOGRAPHY

This work described herein was written and performed by Caleb Joseph Porter while a graduate student at North Carolina State University from August 2012-April 2015.
ACKNOWLEDGMENTS

I would like to express appreciation for the help of those who helped bring this work to fruition. My family members provided moral support which was invaluable. I would like to thank my advisor Dr. Michael Bereman for many helpful conversations and feedback. I would like to thank committee members Drs. David Muddiman, Ed Bowden, and Gufeng Wang for helpful feedback. Additionally I would like to thank Gina Hilton, Ian Murray, and Josh Beri for their support. Finally, I would like to thank the Center for Human Health and the Environment, Pilot Project funding, and North Carolina State University for funding this research.
TABLE OF CONTENTS

LIST OF TABLES ................................................................................................................ v

LIST OF FIGURES ............................................................................................................. viii

CHAPTER 1. GENERAL INTRODUCTION

1.1 Complex Disease and the Exposome ................................................................. 1
1.2 The Study of Protein Adducts ............................................................................. 9
1.3 Data Collection Techniques in Mass Spectrometry ........................................... 11
1.4 Conclusion ......................................................................................................... 16

CHAPTER 2. MULTIPLEX ADDUCT PEPTIDE PROFILING (MAPP)

2.1 Introduction ......................................................................................................... 17
2.2 Materials and Methods ...................................................................................... 22
2.3 Results and Discussion ....................................................................................... 24
2.4 Conclusion ......................................................................................................... 30

CHAPTER 3. CHARACTERIZATION AND OPTIMIZATION OF MULTIPLEX
ADDUCT PEPTIDE PROFILING: SENSITIVITY, NORMALIZED COLLISION
ENERGY, AND PROTEIN ISOLATION

3.1 Introduction ......................................................................................................... 32
3.2 Materials and Methods ...................................................................................... 36
3.3 Results and Discussion
3.3.1 Linearity and Sensitivity of MAPP ................................................................. 45
3.3.2 Albumin Isolation .......................................................................................... 47
3.3.3 HSA Adduct Enrichment DOE .................................................................... 51

v
3.3.4 NCE Optimization .................................................................53

3.4 Conclusion .................................................................................55

CHAPTER 4. DEVELOPMENT OF DATA PIPELINE FOR MULTIPLEX ADDUCT PEPTIDE PROFILING (MAPP)

4.1 Introduction .................................................................................57

4.2 Materials and Methods ............................................................64

4.3 Results and Discussion .............................................................68

4.4 Ongoing/Future Work ...............................................................74

4.5 Conclusion .................................................................................75

CHAPTER 5. APPLICATION TO COMPLEX BIOLOGICAL FLUIDS

5.1 Introduction .................................................................................77

5.2 Materials and Methods .............................................................79

5.3 Results and Discussion .............................................................82

5.4 Conclusion .................................................................................87

REFERENCES ......................................................................................88

APPENDIX ..........................................................................................100
LIST OF TABLES

Table 2.1 Putative adducts identified from smoker’s plasma by Multiplex Adduct Peptide Profiling (MAPP) ......................................................................................................................29

Table 3.1 Parameters and motivations for Design of Experiments (DOE) optimization of parameters affecting adducted albumin digestion and enrichment ........................................51

Table 3.2 Results from DOE optimization..............................................................................52

Table A.1 Parameters and motivations for DOE optimization of parameters affecting digestion using sodium deoxycholate (SDC)..........................................................................................................111

Supplemental Table.1 Results from the SDC DOE optimization.................................................122

Supplemental Table.2 Results from the detergent comparison study...........................................123
LIST OF FIGURES

Figure 1.1 Increase in incidence rates of testicular cancer, breast cancer, and diabetes in the United States ..................................................................................................................2

Figure 1.2 Factors of the exposome .................................................................................6

Figure 1.3 Structure of Human Serum Albumin (HSA) ..................................................8

Figure 1.4 Mechanism of adduction of electrophiles to cysteine ..................................9

Figure 1.5 Compounds from protein adduct studies .....................................................10

Figure 1.6 Limitations of Data Dependent Acquisition (DDA) mass spectrometry in the detection of protein adducts ......................................................................................14

Figure 2.1 Steps in Multiplex Adduct Peptide Profiling (MAPP) .....................................26

Figure 2.2 Unique ion signatures for several potential protein adducts .........................27

Figure 2.3 Multiplex Adduct Peptide Profiling applied to cysteinylated peptide ...........30

Figure 3.1 Experimental workflow used in sensitivity and linearity study .......................45

Figure 3.2 Results from sensitivity and linearity study ..................................................47

Figure 3.3 Experimental workflow for gel electrophoresis and protein concentration monitoring study ................................................................................................................48

Figure 3.4 Results from Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE) experiment ........................................................................................................49

Figure 3.5 Results from protein concentration study .....................................................50

Figure 3.6 Methods used in Normalized Collision Energy (NCE) study .........................53

Figure 3.7 Results from NCE study ...............................................................................55

Figure 4.1 Limitations of DDA methods in detection of protein adducts .......................59
Figure 4.2 Schematic depicting Fixed Step Selected Reaction Monitoring (FS-SRM) on a triple quadrupole ..............................................................................................................61
Figure 4.3 Data workflow used in FS-SRM for accepting peptide adducts .....................62
Figure 4.4 Transitions monitored in FS-SRM .....................................................................63
Figure 4.5 Uniqueness of peptide transitions monitored as related to the number of transitions monitored .............................................................................................................68
Figure 4.6 Data analysis pipeline for Multiplex Adduct Peptide Profiling (MAPP) ............70
Figure 4.7 Multiplex Adduct Peptide Profiling (MAPP) data analysis pipeline applied to cysteinylated peptide .............................................................................................................72
Figure 4.8 Multiplex Adduct Peptide Profiling (MAPP) data analysis pipeline applied to hemoglobin peptide adduct .............................................................................................................74
Figure 4.9 MAPP pipeline for mappmymod.com ..................................................................75
Figure 5.1 Experimental workflow used in comparison of smoker’s and non-smoker’s plasma .....................................................................................................................................82
Figure 5.2 Distribution of adducts to HSA detected in smoker’s and non-smoker’s Plasma .....................................................................................................................................83
Figure 5.3 Correlation plot comparing peak area of fragments of adducts to HSA detected in plasma from a smoker and a non-smoker .............................................................................84
Figure 5.4 Comparison between smoker and non-smoker peak area for peptide adducts to HSA .....................................................................................................................................86
Figure A.1 Experimental workflows used in SDC study ..........................................................110
Figure A.2 Results from DOE optimization of parameters affecting digestion and
recovery using SDC ................................................................. 114

Figure A.3 Results from the detergent comparison experiment ......................... 116

Figure A.4 Results from comparisons of peptide recovery performed with low
amounts of starting material (< 25 µg protein) ......................................................... 118

Supplemental Figure.1 Box plot of gravy score distribution of recovered
peptides .................................................................................................................. 124

Supplemental Figure.2 Plot of gravy score distribution of recovered peptides .......... 125
CHAPTER 1

GENERAL INTRODUCTION

1.1 Complex Disease and the Exposome

Complex human diseases ultimately manifest themselves through the interaction between genetics and environmental factors. Complex diseases include heart disease, diabetes, obesity, and cancer. The mortality rate for a complex disease can range from high for patients diagnosed with stage IV ovarian cancer (83%), to low for patients diagnosed with Crohn’s disease (0.53%). The number of lives lost to complex diseases underscores the importance of understanding their causes. Cancer is exceeded only by heart disease as the leading cause of death in the United States killing approximately 569,490 in 2010. Obesity, not far behind cancer, accounted for nearly 280,000 deaths in 2010. Another concern with complex diseases is the financial burden from decreased population productivity and increased medical expenses. The total estimated cost for diagnosing and treating diabetic patients in 2012 was in excess of 245 billion dollars. The prevalence of complex diseases is of great concern and stemming their prevalence should be a high priority.

Unfortunately over the years the prevalence and incidence rates of complex diseases have escalated in the United States of America. Obesity in children and adolescents has grown dramatically over the past 30 years from 7% (ages 6-11) and 5% (ages 12-19) in 1980 to 18% and 21% in 2010. Incidence rates of testicular and breast cancer in the population have risen by 40% and 25% respectively since 1975 as shown in Figure 1A-B. The number of diagnosed diabetics has climbed by 73% from 1980-2010 as shown in Figure 1C. Some of the increased cancer trends can be explained through improved methodologies for detection
and diagnosis of certain diseases, such as the use of mammograms in breast cancer detection.\cite{17} Therefore, the rise in the number of those afflicted with complex diseases throughout the population and associated cost of treatment has added motivation to discover cures or prevention measures.

**Figure 1A)** The increase in new cases of testicular cancer from 1970-2000 as plotted by age adjusted incidence per 100,000 in the population. From 1970 to 2000 the incidence rate climbed from 3.7/100,000 to 6.1/100,000. The incidence rate of testicular cancer had increased by ~40% in thirty years.\cite{13} **Figure 1B)** The increase in new cases of breast cancer from 1970-2000 as plotted according to the age adjusted incidence per 100,000. From 1970 to 2000 the incidence rate had risen from 105/100,000 to 142/100,000. The incidence rate of breast cancer had increased by over 25%.\cite{14} **Figure 1C)** The increase in new cases of diabetes diagnosed in millions from 1980-2010. The number of diabetics in the United States increased from 5.6 to 20.9 million. The number of diabetics had increased by over 73%.\cite{15}
The sequencing of the human genome was a major scientific milestone reached near the end of the last century.\[18\] Upon its completion President Bill Clinton remarked, “With this profound new knowledge, humankind is on the verge of gaining immense, new power to heal. It will revolutionize the diagnosis, prevention and treatment of most, if not all, human diseases.\[19\] The sequencing of the Human Genome has had a tremendous impact on basic biological research and ignited several new fields of study.\[20\] It has also led to an enormous effort in genetic research to develop early diagnostic procedures and cures to complex human diseases. However, continued studies strongly suggest that complex human disease results from the interactions of nature (i.e., DNA) and nurture (i.e., environment) on an individual.\[21, 22\]

Genome Wide Association Studies (GWAS) analyze genetic variations within a population to determine relationships between genes and complex diseases.\[23\] A landmark study by Lichtenstien et al. performed GWAS on 40,000 pairs of twins investigating the genetic contribution to cancer and observed a contribution of 10-30\%.\[24\] These data support others who assert that 50-70\% of complex disease risk is attributable to environment factors.\[21, 25\]

One environmental exposure that has been linked to complex disease is smoking and cancer. It is estimated 90\% of lung cancer cases in the United States are linked to cigarette smoke.\[26\] The relationship between environment and disease has led to the concept of the exposome. The exposome was introduced by Christopher Wild in 2005 to complement the genome and push for enhanced understanding of environmental effects on human health.\[27\] As explained by Wild, the exposome entails identifying and monitoring the chemical environment encountered from conception to death.\[27\] Over the past 10 years the definition has been modified and refined by Rappaport, Louis, and Miller to enhance the quantitative
potential of the exposome and account for internal responses to external stimuli.\cite{28-30} Today, the exposome represents the totality of both endogenous and exogenous exposures from conception onward. Regardless of the significant effect of the exposome on human health, the field has been largely neglected. Some of the neglect stems from challenges in studying a concept that is vast, complex, and amorphous, such as all chemicals contacted over a lifespan.

However, the path leading to a more complete understanding of complex disease etiologies must incorporate the effects of complex environmental exposures. The limitations of studying the exposome must be overcome. An opportunity for confronting the challenges lies in analytical technology development for the detection and identification of environmental exposures. To illustrate, toxicologists focused on epidemiology, utilize Genome Wide Association Studies (GWAS) to study genetic components of disease but investigate environmental exposures via surveys.\cite{31-33} Undoubtedly, there is room for improvement by supplying the field with methods which are reproducible, precise, quantitative, and have a broad range of applications. As science and technology continue to progress, new products and chemicals are developed which can have hazardous effects on human health, including various types of nanoparticles.\cite{34} Methods are needed which can broaden the understanding of the relationship between environmental exposures and complex diseases. By elucidating the exposome in pieces, the field can tackle this herculean task and improve the quality of life via the identification of environmental exposures that contribute to complex disease risk.

An individual’s exposome is complex and it is composed of chemicals found in food, exogenous pollutants, metals, endogenous species (e.g., reactive nitrogen species and gut flora metabolites), and pharmaceuticals (Figure 1.2). Certain molecules that enter the body become
reactive as part of the body’s innate metabolism. Methanol, for example, oxidizes to formic acid causing ocular toxicity.\cite{35} Other reactive chemical species are synthesized exogenously and enter in the food we eat, like acrylamide in fried foods.\cite{36} Reactive electrophiles represent a key component of the exposome and are believed to be a main contributor to complex diseases as they can directly modify biopolymers.\cite{37, 38} Modification of biopolymers, such as protein adduction, may result in loss of function or carcinogenesis for modified DNA.\cite{39} For instance parathion, developed as a pesticide in the 1940’s, binds irreversibly to serine 188 (the active site) in acetylcholinesterase leading to paralysis.\cite{40} As a result, detection and quantification of reactive electrophiles is extremely important in assessing the impact of the environment on disease risk. These molecules are difficult to detect by themselves because of their inherent reactivity but can be detected via the covalent modification of nucleophilic amino acids (e.g., cysteine or histidine).\cite{41} Blood protein adducts have extended residence times in biological fluids and provide a more precise estimate of internal dose of reactive electrophiles (e.g., glycated hemoglobin). As result, blood protein adducts are an attractive area for identifying and quantifying these species.\cite{42, 43}
Figure 1.2 Factors of the exposome representing the totality of physical forces and chemical species encountered from conception to death.

Reactive electrophiles are an important piece of the exposome and their toxicity has been shown to impair function from the molecular to organism level.\cite{44} These compounds can be generated in the body or ingested through oral, respiratory, and dermal routes.\cite{41, 45-48} Reactive electrophiles include carcinogens such as benzene, naphthalene, quinones, aldehydes, and acrolein which adduct to biopolymers including DNA and proteins.\cite{39} A problem encountered in the measurement of reactive electrophiles (and their metabolites) is their short residence times (hours or days) in the body.\cite{49} However, adducts of reactive electrophiles to proteins can be detected weeks after initial exposures due to the extended lifetime of proteins in the blood. HSA and hemoglobin have half-lives of 28 and 60 days.\cite{50} Limitations in the
study of protein adducts include their presence is typically at sub-stoichiometric levels (1:1000) compared to non-adducted protein levels, and the presence of reactive pollutants in the blood occur at concentrations ranging from $10^{-7}$-10^{-1} \( \mu \text{M} \).\[51\] Another limitation is the lack of methods to discover multiple peptide adducts of unknown mass. In fact, most mass spectrometry methods have been targeted to the measurement of protein adducts derived from one well-characterized compound at a time.\[39\]

Human Serum Albumin (HSA) is the most abundant protein in blood plasma and has a variety of functions.\[52\] One function of HSA is molecular transportation by binding drugs, hormones, fatty acids, metabolic products, and distributing these compounds throughout the body.\[53-57\] The concentration of HSA in plasma is 3.5-5 g HSA/dL plasma and HSA constitutes over 50% of total protein content in plasma.\[41, 58\] HSA was first identified as a blood protein in 1859 by Denis and it has been widely studied since.\[59\] The complete primary structure of HSA was elucidated in 1975 and the crystal structure was first reported in 1990.\[60, 61\] From these studies, HSA was observed to be 66.5 kDa and has 585 amino acids including 35 cysteines. Thirty-four of the cysteines are involved in di-sulfide linkages; however, HSA-Cys^{34} is an unbound thiol available to interact with a variety of substrates (Figure 1.3).\[50\]
The unbound cysteine at HSA-Cys$^{34}$ accounts for up to 80% of free thiols in blood plasma.$^{[62]}$ The structure of cysteine and the mechanism for nucleophilic attack on electrophilic substances (iodoacetamide) is shown in Figure 1.4. The free electrons on the sulfur attack the carbon (which has a partial positive charge) in iodoacetamide in an $S_N2$ fashion with the iodine functioning as a leaving group. Many compounds can serve as electrophiles to react in this fashion with free thiols to generate a protein adduct to human serum albumin.
1.2 The Study of Protein Adducts

The first report of a small molecule binding covalently to bulk tissue dates back to landmark work by the Millers’ at the University of Wisconsin in 1947.\cite{45} The effect of p-dimethylaminoazobenzene (Figure 1.5A), a carcinogenic yellow dye, on rat livers was studied to illustrate covalent binding of small molecules to liver proteins. The livers were extracted and homogenized, and the protein samples were isolated and washed to remove the unadducted dye. The adducted dye was liberated and the concentration of dye was measured by absorbance at 425 nm. The success of the Millers’ inspired others to measure and verify adducts to biopolymers. In the early 1960’s, Magee et al. investigated adducts to nucleic acids.\cite{63} They dosed rats with $^{14}$C labeled dimethylnitrosamine (Figure 1.5B), and measured enzymatically oxidized forms of the compound which modified nucleic acids. Organs were extracted, nucleic acids were precipitated, and samples were isolated for the measurement of radioactivity with a Geiger counter. Adducts to RNA derived from the liver were three fold higher than RNA derived from spleen, kidney, or pancreas.

Brodie and coworkers were among the first who explored drugs as adducting species. They dosed rats with phenobarbital (Figure 1.5C) which has been used as a
Necrosis of rat liver tissue was compared in rats dosed with $^{14}$C labeled phenobarbital and bromobenzene and the radioactivity of isolated liver proteins was measured with a Geiger counter. In 1975, Ostermann-Golkar *et al.* helped the field branch out of the liver and into the blood using hemoglobin to monitor exposures to a reactive electrophilic compound by measuring adducted hemoglobin concentration levels. To test the hypothesis, they dosed rats with $^{14}$C labeled and unlabeled ethylene oxide (*Figure 1.5D*), a mutagenic compound.[47] They isolated the hemoglobin and compared the labeled to unlabeled samples by thin layer chromatography and measured radioactivity with a Geiger counter.

![Figure 1.5A-F](image)

*Figure 1.5A-F* Structures of reactive electrophiles from key studies which adduct to biopolymers include **A** p-dimethylaminoazobenzene, **B** dimethylnitrosamine, **C** phenobarbital, **D** ethylene oxide, **E** halothane, and **F** acetaminophen.

In 1985 Satoh and coworkers were among the first to use antibody enrichment to detect a modified membrane protein in rabbit livers.[64] Rabbits were dosed with $^{14}$C labeled
halothane (Figure 1.5E), an anesthetic, to monitor adducted membrane proteins in the liver. They developed an Enzyme-Linked Immunosorbent Assay (ELISA) to detect adducts to liver membrane proteins and confirmed radioactivity of liver proteins with a Geiger counter. In 1998, Qiu et al. reported one of the first studies using mass spectrometry based proteomic procedures where they used in-gel tryptic digestion coupled to MS (MALDI-TOF) and database searching.[65] Mice were dosed with $^{14}$C labeled acetaminophen (Figure 1.5F) and liver tissues were extracted, isolated, lysed, and separated by 2D Gel Electrophoresis. The radioactive spots were excised and digested using trypsin. They identified mitochondrial and cytosolic protein adducts involved in detoxification, including glutathione peroxidase. These seminal studies of protein adducts using mass spectrometry have been important to our understanding molecular sites of adduction and have underscored the need to develop advanced multiplex analytical technologies.[66]

1.3 Contemporary Proteomic techniques in Mass Spectrometry

Before discussion of data collection techniques used in mass spectrometry, a brief overview of principles and instrumentation will be covered. In mass spectrometry, analytes are ionized and measured in the gas phase using a combination of electric and or magnetic fields. The principle components of a mass spectrometer include: an ionization source, a mass analyzer (to measure $m/z$ of ions), and a detector which meters intensity according to the number of ions detected at a given $m/z$.[19] Electrospray ionization (ESI) and Matrix Assisted Laser Desorption/Ionization (MALDI) have revolutionized the field of proteomics and are used most frequently for ionization of large molecules.[67, 68] ESI and MALDI are the most
commonly used ionization sources in proteomic methods because they are ‘soft’ ionization techniques, thus allowing biomolecules to ionize and remain intact. In MALDI, gaseous peptide ions are created via sublimation of a solid matrix by laser pulses, whereas in ESI, a voltage is applied to solutions containing the analytes to generate ionized species. Fragmentation of peptides is typically accomplished by Collision Induced Dissociation (CID) as it is a low energy technique where inert gas (argon, nitrogen) molecules collide with ions increasing the internal energy of the ion. Once the energy reaches a threshold, then fragmentation occurs. Mass analyzers vary in sensitivity, resolution, and mass accuracy, and range from the ion trap (which traps, fragments, and measures ions based on electric fields) to Time-of-Flight (TOF which measures time for ions to travel a specific distance), orbitrap (which measures the frequency of axial oscillations of ions), and quadrupole (which acts as a mass filter by varying the magnitudes of the DC and RF).

In proteomics, two main mass spectrometric techniques exist for data acquisition of complex mixtures. Data Dependent Acquisition (DDA) has become the gold standard for protein identification due to improvements in instrumentation, liquid chromatography (LC) separation, software tools, and sample preparation. DDA methods are performed on hybrid instruments. The methods consist of a precursor scan which monitors all ions present within the mass range (400-1200 m/z), isolation, and fragmentation of ions (one at a time) which are most abundant. The isolation and fragmentation steps are repeated until cycling through the loop count (number of ions to interrogate usually 10-20) whereupon another precursor scan is taken. Dynamic exclusion is a parameter which allows the user to specify a time frame where the instrument will neglect species which have been isolated and
fragmented to prevent the acquisition of repetitive data.\[79] Normalized Collision Energy (NCE) is an instrument parameter which controls the fragmentation of isolated ions. Based upon the m/z of the ion, the instrument adjusts the energy imparted to dissociate the ion.\[80] In general, 3+ ions are less stable than 2+ ions, they require less energy for fragmentation and the NCE allows the instrument to adjust the energy imparted to ions based upon their respective charges. After the data is collected, it is searched against a database of known proteins (e.g. a fasta file from Uniprot). The search uses an algorithm which performs in-silico digestion of the database proteins with an enzyme (e.g. trypsin) and searches for matches in the predicted fragmentation of peptides and the fragmentation in the raw data.\[81] Proteins are then inferred based on peptide identifications. Four main limitations exist for the application of DDA methods in the discovery and reproducible profiling of protein adducts, two are shown in Figure 1.6A-B. These limitations include: 1) DDA is biased towards the identification of high abundant ions, 2) poor reproducibility of peptide identifications, 3) increased time required for searches involving multiple modifications, and 4) the population of a list of modifications to search (i.e. only known masses can be searched which stifles discovery).\[82-84] DDA methods utilize a precursor scan taken of all ions within the mass range (e.g., 400-1200 m/z), followed by isolation and fragmentation of the most abundant peptide ion. Since protein adducts occur at low stoichiometric levels, this could be problematic for analyses of these species by DDA. The aforementioned challenge leads to the second limitation: reproducibility. DDA runs on analytical replicates from large sample cohorts have 70-80% reproducibility of peptide identifications.\[84] Third, multiple modification searches are needed to identify multiple peptide adducts. This requires searching each peptide with every
modification and can exponentially increase search space.\cite{85} Search space and time increase significantly with the addition of multiple modifications to search. In some cases, using ProteinDiscoverer 1.4 (Thermo Fisher Bremen, Germany), the searches are limited to 5 dynamic modifications, which would necessitate performing multiple searches. The greatest limitation of DDA applied to discovery of protein adducts is the population of a list of modification masses to query the data. The masses of various reactive electrophile modifications may not be known, and thus limit DDA methods in the discovery and identification of protein adducts.

**Figure 1.6A-B** Limitations for use of DDA for profiling protein adducts to discover protein adducts are: **A)** Reproducibility of peptide identifications is between 70-80% for three analytical replicates. **B)** The exponential increase in search times is shown when searching for multiple modifications. Long search times in Comet for multiple modifications as plotted by search time in minutes relative to the number of modifications queried.

Data Independent Acquisition (DIA) mass spectrometry was developed by Venable in 2004 and represents a promising alternative to DDA.\cite{86} In contrast to targeted and DDA methods where ions are isolated and fragmented one at a time, all precursor ions within a pre-
specified isolation window are isolated and fragmented together using DIA. The isolation window width can vary from the full mass range (400-2000 m/z) to as narrow as 2-4 m/z.\[^{83}\] An important consideration regarding isolation windows is that wider widths increase the complexity of the data.\[^{87}\] Not only does this increase in complexity warrant the development of novel bioinformatics tools but it may decrease the sensitivity of any individual peptide ion in trapping instruments with automatic gain control. A variety of methods have been developed which use different isolation window widths.

\( \text{MS}^E \) and All Ion Fragmentation (AIF) are analogous methods that use the widest isolation windows, as they fragment all precursors in the full mass range (e.g., 400-1400 m/z). \( \text{MS}^E \) was developed by Waters (Waltham, MA) for use on the Q-TOF. The energy level in the collision cell alternates between low and high energy which enables passage of precursor ions at low energy and fragmentation of ions at high energy.\[^{88}\] AIF was developed by Thermo Fisher (Bremen, Germany) for the Orbitrap and is similar to \( \text{MS}^E \), though fragmentation occurs in the High Energy Collisional Dissociation (HCD) cell. Data collected from these methods can be searched using ProteinLynx Global Server (\( \text{MS}^E \)) and Max Quant (AIF) which use algorithms for noise reduction, apex retention time matching, and database searching.\[^{89, 90}\] A comparison between the two methods was performed on the Universal Proteomics Standard 1 (UPS 1) and reported protein identifications of 43 (\( \text{MS}^E \)) and 45 (AIF) out of 48 total proteins.\[^{88}\] Limitations of AIF and \( \text{MS}^E \) include the lack of false discovery rates for identified peptides, need for validation of results by DDA, and difficulty with isobaric labels or tags.\[^{83}\]

Other DIA methods include Sequential Window Acquisition of all Theoretical Mass Spectra (SWATH MS) and Multiplexed MS/MS (MSX).\[^{83, 91}\] SWATH uses 25 m/z isolation
windows to cover the mass range (400-1400 m/z) on Q-TOF mass spectrometers.[83, 91] SWATH differs from traditional DIA approaches by using foreknowledge of analyte peptide fragmentation and chromatographic elution to assist in data processing.[92] The duty cycle in SWATH is greater than in AIF or MS^E, as there are more scans taken to cover the same m/z space, but the spectra are less complex (due to smaller isolation window widths) aiding in quantification of peptides.[83] Targeted data extraction can be utilized in a similar fashion to SRM data extraction; however, one has the freedom to refine, expand, and re-mine transitions after data collection.[91] Limitations of SWATH include lower resolution and mass accuracy from the TOF, interferences from the 25 m/z isolation window which affect quantitation, and the data is not compatible with traditional database searches.[83] Other variations of DIA have been developed recently.[93]

1.4 Conclusion

Increases in complex diseases and their incidence rates represent a growing problem for societies to confront.[10, 15, 21] Current studies indicate the role of an individual’s local environment plays a significant part in human health.[25, 51] The exposome was introduced as a conceptual framework to elucidate relationships between exposures and health.[94] However, the exposome is incredibly complex and current methodologies are lacking to measure the various components of the exposome.[41] Reactive electrophiles are a toxic piece of the exposome which are known to have adverse effects on the biomolecules of individuals.[44] These compounds have residence times of days or hours in the blood, but can adduct to proteins (which have half-lives of 4-8 weeks) containing a free nucleophilic thiol such as HSA.[50]
Protein adducts have been widely studied over the years, however, these studies have been limited to a single well characterized compounds and targeted tissues.\cite{39, 44} Unfortunately, studies conducted under the aforementioned conditions do not mimic real life exposures or lend themselves to discovery of new protein adducting species.\cite{41} Current methods using DDA mass spectrometry have limitations in the discovery of protein adducts including poor reproducibility of peptide identifications, detection bias toward high abundant peptide ions, and the requirement of a list of known modification masses with which to search the data.\cite{83} DIA methods coupled to targeted data extraction in the form of Multiplex Adduct Peptide Profiling (MAPP) provide an opportunity to develop methods which are not biased toward high abundant species and enable discovery of protein adducts of unknown mass.

In Chapter 2 Multiplex Adduct Peptide Profiling (MAPP) is developed and discussed. Chapter 3 describes characterization of MAPP by assessment of sensitivity, comparison of Normalized Collision Energy (NCE), and the optimization of enrichment and digestion parameters affecting adducted HSA. The data processing workflow is discussed in Chapter 4 and examples are illustrate how the pipeline is used to identify adducts. In Chapter 5 MAPP is used to compare HSA peptide adducts detected in plasma samples derived from a smoker and a non-smoker. The Appendix describes a comparison of sodium deoxycholate to commercially available LC-MS/MS detergents.
CHAPTER 2
MULTIPLEX ADDUCT PEPTIDE PROFILING (MAPP)

2.1 Introduction

The sequencing of the Human Genome has had a tremendous impact on basic biological research and ignited several new fields of study.\textsuperscript{[20]} It has also led to an enormous effort in genetic research to develop early diagnostic procedures and cures to complex human diseases. However, continued studies strongly suggest that complex human disease ultimately manifests itself through an interplay between genetic (G) and environmental (E) factors (G x E).\textsuperscript{[21, 22]} In a study of 40,000 pairs of twins, primary genetic factors contributed to approximately 10-30\% of non-communicable complex diseases (e.g., cancer).\textsuperscript{[24]} These data support claims that between 50-70\% of complex disease risk is linked to an individual’s local environment (e.g., diet, smoking history, occupation, etc).\textsuperscript{[21, 25]} Coined by Wild in 2005\textsuperscript{[27]}, the exposome was created to raise awareness of the importance of the local environment on human health. The exposome represents the totality of both endogenous and exogenous exposures from conception onward. Despite its clear role in human health, the exposome has received little attention for discovering disease etiologies.

The study of the exposome is essential to investigate the overall burden of environmental factors on complex disease risk. However, research in this field is significantly hindered due to the paucity of methods for detecting environmental exposures on a global scale. For example, epidemiologists use advanced sequencing tools for genome wide associated studies; however, resort to questionnaires/surveys to “quantify” environmental exposures. Clearly there is an essential need for more precise, specific, yet global methods for
quantifying exposure such that key gene × environment interactions can be elucidated. As humans are exposed to new chemicals on a routine basis, a more fundamental understanding of the effects of the local environment on disease risk is imperative – with an ultimate goal of decreasing disease risk by limiting exposure.

An individual’s exposome is complex and composed of a variety of compound types from chemicals found in food, exogenous pollutants, metals, endogenous species (e.g., reactive oxygen/nitrogen species), and pharmaceuticals. Certain molecules that enter the body become reactive as part of the body’s innate metabolism or are created endogenously. Reactive electrophiles represent a key component of the exposome and are believed to be a main contributor to complex diseases as they directly modify biopolymers.\[37, 38\] As a result, detection and quantification of reactive electrophiles is extremely important in assessing the impact of the environment on disease risk. These molecules are difficult to detect by themselves because of their inherent reactivity but can be detected via the covalent modification of nucleophilic amino acids. Due to their high abundance and extended residence times in blood, protein adducts provide a more precise estimate of internal dose of reactive electrophiles (e.g., glycated hemoglobin)\[42, 43\] and are an attractive area for identifying and quantifying these species. Although the first report of a small molecule binding covalently to bulk tissue dates back to landmark work by the Millers’ in the 1940’s\[45\], the majority of subsequent studies using mass spectrometry have focused on the measurement of protein modifications resulting from a single molecule.\[66\] However, in order to elucidate the contribution of the exposome to disease risk, discover gene x environment interactions, and
investigate synergetic effects of chemical mixtures – multiplexed discovery methods are needed.

Recent research has focused on characterizing the sub-adductome of a tryptic peptide from human serum albumin using LC MS/MS.\textsuperscript{[95, 96]} Serum albumin is the most abundant protein in plasma and has one free cysteine that can be detected as a tryptic peptide. This nucleophilic cysteine acts as a “sink” to highly reactive electrophilic species. This problem presents a fundamental limitation of conventional mass spectrometry strategies which necessitates prior knowledge of the mass of the modification for identification of modified peptides via traditional database searches (i.e., multiple modification searches via data dependent acquisition strategies) or for the development of targeted assays based on selected reaction monitoring or targeted MS/MS.\textsuperscript{[97]} However, the precise chemicals an individual is exposed to or the metabolism of these species that lead to their inherent reactivity may not be known. Recently, Rappaport and coworkers\textsuperscript{[96]} developed a novel method using a triple quadrupole termed fixed step selected reaction monitoring (FS-SRM) in which Q1 was stepped at small $m/z$ increments ($m/z = 1.5$) around the targeted peptide’s $m/z$ and Q3 scanned a theoretical list of 4 fragments per precursor bin. Signals that were detected were classified as putative peptide adducts. Certain limitations exist for this FS-SRM method. The low resolution and low mass measurement accuracy inherent of the instrument could lead to false positives. The method is not easily applicable to online chromatography due to the need to scan hundreds of transitions in order to cover a broad mass range of possible adducts. In addition, FS-SRM still requires a theoretical list of parent ions which can limit the discovery of novel peptide-adducts.
Data independent acquisition (DIA)[98] coupled to targeted data extraction[99, 100] has offered an alternative pipeline for data acquisition and analysis of proteomic samples using LC MS/MS. DIA methods are based off fragment scans of wide isolation windows but can vary between 2 and greater than 100 m/z.[98,100-103] Conventional methods for protein identification and quantitation rely on data dependent acquisition methods coupled to protein sequence databases for protein identification. Although these traditional strategies are extremely powerful for surveying proteomes with state of the art instrumentation reaching sequencing capacities of 4000 protein identifications per hour[104], methods based on independent acquisition offer certain benefits. The principal advantages are improved reproducibility of peptide identifications due to the sequential scanning of isolation windows, potential improvements in quantitative accuracy due to improvements in signal to noise of MS2 versus MS 1 based label-free quantitation, and when coupled to targeted data extraction the capability to re-query a dataset for any peptide of interest. Essentially, one can propose questions and test those hypotheses in the data.

Herein, we present a 3 step LC MS/MS method named Multiplex Adduct Peptide Profiling (MAPP) to identify putative signals of interest, calculate the mass of the modification, and verify the adducted mass. The method relies on protein enrichment and targeted data extraction[99,100] coupled to data independent acquisition[98]. It can be used to investigate site specific heterogeneity of any targeted peptide of interest. We demonstrate its potential to profile global adducts of a single high abundant peptide due to covalent modifications by reactive electrophiles in biological fluids.
2.2 Materials and Methods

Materials

Formic acid (FA), ammonium bicarbonate (AB), hydrochloric acid (HCl), iodoacetamide (IAM), dithiothreitol (DTT), sodium deoxycholate (SDC) were obtained from Sigma Aldrich (St. Louis, MO). Trypsin was obtained from Promega (Madison, WI). HPLC grade acetonitrile, methanol and water were from Burdick & Jackson (Muskegon, MI). The hemoglobin peptide (GTFATLSELHCDK) was synthesized and purchased from Thermo Fisher Scientific GmbH (Ulm, Germany). Plasma obtained from smokers was purchased from BioreclamationIVT (New York, New York).

NanoLC MS/MS

Two µL of sample were injected onto a self-made 4 cm trap using an Easy nanoLC 1000 coupled to a Q Exactive Plus mass spectrometer (Bremen Germany). PicoFrit columns from New Objective (Woburn, MA) were packed to 25 cm in house with 3 μm C18 silica particles (Dr. Maisch, Entringen, Germany). Mobile phase B was 99.9 % acetonitrile with 0.1% formic acid and mobile phase A was 98 % water, 2 % acetonitrile, and 0.1 % formic acid. A 90 minute LC MS/MS method was used and consisted of a linear gradient from 0-40 % B over 70 minutes followed by a ramp to 80 % B in one minute. The column was washed at 80% B for 9 minutes and regenerated at 0 % B for 10 minutes.

Albumin Isolation/Enrichment

Albumin isolation and adduct enrichment were performed based upon the method described elsewhere.[50] In brief, ammonium sulfate was added to 4 mL of plasma to create a 60% saturated solution. Albumin was acid precipitated using HCl from the supernatant.
Isolated albumin (500 µg) was added to 75 mg of dry Activated Thiol Sepharose in a spin column (Thermo product #69725) and the flow through fraction (nonmercaptoalbumin) was collected after 18 hours of enrichment. The enriched albumin fraction was then prepared using the filter-aided sample preparation procedure for shotgun proteomics\[105\] using a 4 hour tryptic digestion at an enzyme to substrate ratio of 1:50.

DIA Methods were created in Skyline\[106\]. The 21 amino acid tryptic peptide from human serum albumin that contained the free Cys\(^{34}\) residue was entered into skyline (K.ALVLIAFAQYLQQ\(^\text{C}\)PFEDHVKL). An isolation list was created in Skyline that used 12 10 \(m/z\) isolation windows which spanned \(m/z\) 810 to 930 but still contained the triply charged unmodified peptide precursor (\(m/z\) 811.7594). This range of precursor detection included the mass of the albumin free cysteine unadducted peptide to a mass addition (i.e., adduct) of 360 Daltons which encompasses the masses of all reported protein adducts aside from Satratoxin G\[66\].

The isolation list was imported into the method editor of Q Exactive Plus and an MS1 scan was added from \(m/z\) 550 to 1300 for a total 13 scans per cycle with an average duty cycle of 1.3 seconds. For the MS1 scan the resolving power was set to 70,000, an AGC of 1e6, and a max injection time of 50 ms was used. For the DIA scans, parameters were set to the given levels: AGC of 1e6, 30 NCE (normalized collision energy), max injection time of 50 ms, and a resolving power of 17,500.

**Data Analysis**

Peptides with fake modifications were created in Skyline by adding an artificial modification to the cysteine residue of the Cys\(^{34}\) tryptic peptide such that a single precursor
Putative signals of interest (SOI) were identified by extracting common fragment ions from the Cys$^{34}$ peptide and finding the instances in retention time space where all co-eluted with high mass measurement accuracy (<3 ppm). In addition, the proportion of relative abundances of each fragment to total abundance was used as a method to eliminate false positives. Once an SOI was identified, precursors were filtered based on the isolation window from which the SOI originated and matched in retention time space with the SOI using Xcalibur 2.2. This matching was based on the principle that since the MS1 precursor and associated fragment ions all belong to the same chemical species then the chromatographic profile of both should be theoretically identical.

2.3 Results and Discussion

Figure 2.1 displays the overall workflow for global peptide adduct analysis by data independent acquisition mass spectrometry using a targeted analysis. The method consists of a 1-dimensional reversed phase liquid chromatographic separation, followed by data independent acquisition with a dedicated MS1 full scan (Figure 2.1A). Relatively large isolation windows at fixed incremental $m/z$ increases are used to detect signals of interest (SOI). Data analysis and interpretation then follows a 3 step process. In the first step, a fragment ion tag is extracted from the data that consists of all b and y fragment ions that do not contain the amino acid that is believed to be modified. The point in retention time at which these fragment ions co-elute in the chromatogram represents a signal of interest. In step 2, MS1 signals are overlaid with the signal of interest to match the precursor with corresponding fragment ions. The ambiguity in the correct MS1 feature is greatly reduced by only considering
the precursors resulting from the isolation window at which the signals of interest (i.e., fragment ions) were generated. The degree of this complexity reduction is inversely proportional to the width of the isolation window. In addition, the assumption that an adducted peptide would have a higher \( m/z \) than the unmodified peptide excludes many features. Next, based on retention time apex and peak shape, a putative precursor \( m/z \) representing the modified peptide can be identified. The signal can be further confirmed if the M, M+1 and M+2 precursor ions are extracted. Then the mass of the modification can be calculated using Equation 1.

**Equation 1:** \( \text{Mod} = [\text{Experimental} \ (m/z) - \text{Theoretical unmodified} \ (m/z)] \times z \)

Finally, in **step 3** the mass of the modification can be verified by extracting the fragment ions that contain the modification calculated from **step 2**. If the mass is calculated correctly, then co-elution of additional adducted fragment ions should be observed. It is important to note that this step assumes that the modification is stable upon collision induced dissociation which may not be the case in all instances. **Figure 2.1B** summarizes the overall workflow using the tryptic peptide which contains a free cysteine from hemoglobin that has been carbamidomethylated.
Figure 2.1: An overview of the method used for Multiplex Adduct Peptide Profiling (MAPP). The scan cycle consisted of a full MS1 scan followed by 12 m/z isolation windows that sequentially increased from the peptide’s unadducted mass. Data analysis consisted of three steps for the identification of global peptide modifications. Step 1 identifies potential signals of interest (SOI) by extracting the masses of unmodified fragment ions (i.e., tag). Step 2 calculates the mass of the modification by matching co-eluting MS1 features to the fragment ion tag extracted in step 1. Step 3 verifies the calculation in step 2 by extracting the modified fragment ions. An alkylated tryptic peptide from the beta chain of hemoglobin was used to illustrate the workflow.

This procedure represents a powerful strategy to identify both the number of putative modifications of a targeted peptide in complex mixtures or enriched samples and identification of those modifications through accurate mass analysis. In theory each modification changes the hydrophobicity of the peptide which then can be separated in time by reversed phase chromatography. By counting the number of different peaks resulting from co-elution of the unmodified fragment ions one can determine the putative number of targeted peptide forms.

Figure 2.2 displays data from the LC MS/MS analysis of human plasma after enrichment of modified mercaptoalbumin (Cys\textsuperscript{34}) using the method described. Unmodified fragment ions were extracted and three of the isolation windows are shown. Several signals
of interest were observed as identified by co-elution and accurate mass (< 3 ppm) of the unmodified fragment ions. For the majority, the adducts had little effect on retention time using our standard 90 minute run and eluted between 65 and 76 minutes. The ratio of the unmodified fragment ions can be used to help ensure positive identification of a putative adduct and limit false positives. All 10 putative peptide adducts observed from these isolation windows, had similar relative fragment ion abundances as illustrated in Figure 2.2B. The most abundant fragment ion was the b3 ion followed by the y7 ion. Interestingly the Cys34 tryptic peptide contains a proline amino acid (y7) which is known to preferentially fragment (N-terminal) upon collision induced dissociation.\cite{107,108}

**Figure 2.2:** A-C) A display of chromatograms from three isolation windows where unmodified y and b fragment ions were extracted from the LC MS/MS analysis of the Cys34 albumin peptide. Within these three windows, 10 signals of interest (SOI) were identified. D) Comparison of the relative abundance of each fragment ion across the 10 signals of interest shows a similar fragmentation signature.
It is important to note at this point in the data analysis procedure that we have identified the number of potentially modified forms of a single peptide – a feat that is difficult to perform by current technologies. However, we add to these new capabilities, by elucidating the mass of the modification by overlaying MS1 features and matching the monoisotopic mass of the precursor ion to the SOI in retention time space. An exact mass of the adduct can then be calculated using Equation 1. In theory, as the field of adductomics advances and databases are readily developed this mass could then be searched and identified through a database.

Figure 2.3 illustrates the utility of this step in which the signal of interest (# 5 from Figure 2.2) was matched to the precursor mass of 851.4274 in retention time space. Upon substituting the correct masses into Equation 1, a mass of 119.004 was calculated. Searching previous literature, it was discovered that this mass is a common albumin adduct in biological fluids in which a thiol bond is created between two cysteine residues – known as cysteinylation. The verification step adds this mass to the Cys34 residue and then overlays the modified fragment ions. Co-elution of multiple fragment ions extracted with the elucidated mass provides confirmation that the peptide is believed to be adducted with the mass calculated. Interestingly, increased human serum albumin Cys34 cysteinylation has been shown to be a marker for oxidative stress related diseases such as diabetes mellitus, chronic liver diseases, and kidney diseases.\textsuperscript{[109]}
Table 2.1 Known adducts observed from smokers’ plasma using MAPP

<table>
<thead>
<tr>
<th>Observed m/z</th>
<th>Theoretical m/z</th>
<th>Calculated Adduct Mass (Da)</th>
<th>MMA (PPM)</th>
<th>Ret Time (min)</th>
<th>Putative Adduct</th>
</tr>
</thead>
<tbody>
<tr>
<td>811.7589</td>
<td>811.7594</td>
<td></td>
<td>0.6 (± 0.2)</td>
<td>73.2 (± 0.2)</td>
<td>Unadducted</td>
</tr>
<tr>
<td>817.0928</td>
<td>817.0910</td>
<td>16.0002</td>
<td>1.3 (± 1.3)</td>
<td>70.6 (± 0.2)</td>
<td>Oxidation</td>
</tr>
<tr>
<td>822.4227</td>
<td>822.4229</td>
<td>31.9899</td>
<td>0.8 (± 0.1)</td>
<td>72.2 (± 0.2)</td>
<td>Dioxidation</td>
</tr>
<tr>
<td>827.7537</td>
<td>827.7546</td>
<td>47.9829</td>
<td>0.8 (± 0.1)</td>
<td>72.6 (± 0.2)</td>
<td>Trioxidation</td>
</tr>
<tr>
<td>851.4261</td>
<td>851.4272</td>
<td>119.0001</td>
<td>0.9 (± 0.2)</td>
<td>68.5 (± 0.2)</td>
<td>Cysteinylation</td>
</tr>
<tr>
<td>856.0988</td>
<td>856.0993</td>
<td>133.0197</td>
<td>1.0 (± 0.4)</td>
<td>68.6 (± 0.2)</td>
<td>Homocysteinylation</td>
</tr>
<tr>
<td>913.4483</td>
<td>913.4487</td>
<td>305.0682</td>
<td>0.6 (± 0.4)</td>
<td>68.7 (± 0.2)</td>
<td>Glutathione</td>
</tr>
</tbody>
</table>

Thirty-eight signals of interest were observed which represent putative adducts of Cys^{34} albumin tryptic peptide using co-elution of fragment ions and accurate mass analysis across the 12 isolation windows. A list of peptide adducts that were identified are given in Table 2.1. Several of these identified adducts have been previously reported in the literature.\textsuperscript{110} The list consists of varying degrees of cysteine oxidation including sulfenic (SOH), sulfinic (SO\textsubscript{2}H), and sulfonic acids (SO\textsubscript{3}H) which are known to be potential modifications due to oxidative stress from reactive oxygen species.\textsuperscript{111, 112} Several SOIs could not be identified due to the limited resources and databases available for known adducts. Also it is worth noting that some of these adducts may be a result of storage or in-vitro processing which would increase the complexity in accurate identification. Regardless, MAPP presents a powerful strategy to profile targeted peptide site specific modifications in complex samples and will be used for
future applications to investigate the adductome in relation to disease due to believed environmental causes.

**Figure 2.3:** A) SOI was found to match the precursor 851.4261 in retention time space which was due to cysteinylation. B) Adducted fragment ions were then extracted from the data to verify this assignment.

### 2.4 Conclusion

A novel method termed Multiplex Adduct Peptide Profiling (MAPP) which uses data independent acquisition coupled to targeted data extraction has been developed to investigate global site specific modifications of targeted peptides of interest. The authors note while the initial application has focused on adductomics, the method could be applied to other in vivo modifications as well, given the protein of interest could be efficiently isolated and enriched. Enrichment is key in these experiments since both in vivo and in vitro modifications typically
occur at sub-stoichiometric levels. However, we will also investigate its use to identify in vivo modifications of peptides in more complex mixtures. Future experiments will explore optimization of the method including statistics to match MS2 and MS1 signals, as well as evaluation of the linearity of the Cys$^{34}$ peptide when spiked with multiple reactive species.
CHAPTER 3
CHARACTERIZATION AND OPTIMIZATION OF MULTIPLEX ADDUCT PEPTIDE PROFILING: SENSITIVITY, NORMALIZED COLLISION ENERGY, AND PROTEIN ISOLATION

3.1 Introduction

During the development of new technologies, it is important to explore the experimental space to identify and optimize parameters that affect response. A response that is routinely investigated in the development of new analytical technologies includes the sensitivity of the method. Identification and optimization of the parameters that affect these fundamental analytical figures of merit are essential, prior to confidently applying the technique to carefully collected plasma sample cohort. Data Independent Acquisition (DIA) has strengths which surpass Data Dependent Acquisition (DDA) and targeted MS for the discovery of protein adducts which have been previously noted in Chapter 1 Section 1.3.

While DDA is a powerful technique for surveying proteomes, it has limitations in the global discovery of protein modifications.[113] Limitations of DDA for discovery of protein adducts include: 1) DDA is biased towards the identification of high abundant ions, 2) poor reproducibility of peptide identifications, 3) increased time required for searches involving multiple modifications, and 4) the population of a list of modifications to search (i.e. only known masses can be searched which inhibits discovery).[82-84] The latter may be most noteworthy as it biases identification toward only known species which stifles discovery. DIA used in concert with targeted data extraction is a promising alternative to DDA analysis of
samples containing protein modifications. DIA methods take a full mass range precursor scan (400-1300 m/z) followed by fragment scans at pre-specified isolation window widths varying from 2-1000 m/z. Principal advantages of using DIA and targeted data extraction include: 1) enhanced reproducibility of peptide identifications from sequential scanning of isolation windows, 2) potential improvements in quantitative accuracy due to improvements in signal to noise of fragment to precursor based label-free quantitation, and 3) the capability to query the dataset for any peptide. These advantages enable hybrid investigations which can be exploratory (e.g., modification) in a targeted (e.g., peptide) fashion.

Reactive electrophiles have been studied for many years and have been shown to have toxic effects upon cells, tissues, organs, and individuals. Sources of reactive electrophiles can include the environment or they can be generated in the body through a variety of processes such as metabolism of exogenous compounds such as drugs or oxidation of lipids etc. Electrophilic compounds include carcinogens such as benzene, naphthalene, and acrolein. These species and their metabolites are difficult to detect due to their inherent reactivity and short residence times in the body (hours or days). Also, reactive electrophiles do adduct to biopolymers including DNA and proteins causing damage and loss of function. However, adducts of reactive electrophiles to proteins can be detected weeks (28 days for HSA and 60 days for hemoglobin) after initial exposures due to the lifetime of these proteins in the blood. Difficulties in the study of protein adducts include their occurrence at substoichiometric levels (1:1000) which necessitate enrichment strategies. Also, another difficulty is the requirement of a proteotypic peptide that contains the modified amino acid. Another limitation is the lack of methods to discover unknown mass adducts. Most mass spectrometry methods have been
targeted to the measurement of protein adducts derived from one well characterized compound at a time. Yet, DIA coupled to targeted data extraction can be used to discover protein adducts and overcomes several of the limitations. In DIA methods, ions are fragmented and detected within the isolation window which allows sampling of low and high abundant species. An m/z space is chosen which can detect subsequent mass additions from the unmodified precursor. The data can be mined to identify and verify the presence of adducted peptides. These features of DIA make it a powerful technique to enhance investigations of protein adducts.

In this study, we investigated three important areas in the detection of albumin adducts using Multiplex Adduct Peptide Profiling (MAPP), including: 1) Normalized Collision Energy (NCE), 2) sensitivity, and 3) adducted albumin isolation. In DDA methods, NCE adjusts the collision energy according to the charge of the ion. Rice-Ramsperger-Kassel-Marcus (RRKM) theory forms the basis upon which collision induced dissociation (CID) occurs and RRKM theory treats the molecule as a set of harmonic oscillators which freely interact and exchange energy. With the RRKM equation written as a function of internal energy, the equation yields the rate constant for a molecule with a given energy (E) and activation energy (e)

\[
k(E) = \frac{\sigma N^+[E - e]}{h \rho(E)}
\]

\(\sigma\) is the reaction degeneracy, \(N^+[E - e]\) is the transition state sum of states from 0 to \(E-e\), \(h\) is Planck’s constant, and \(\rho(E)\) is the parent ion density of states at an energy \(E\). The rotational energy states are excluded, so only the vibrational energy states are considered. As the internal energy of the ion increases, the precursor dissociates to fragment ions. A peptide ion has
a multitude of vibrational modes and if the precursor is modified, the number of available modes can vary based upon the structure of the adduct. For example, a peptide containing a monoxidized cysteine would contain less vibrational modes than the same peptide modified with an addition of glutathione. This difference in vibrational modes could lead to varied fragmentation depending on the collisional energy imparted upon impact of the gas molecule. Furthermore, the role of NCE has not been fully explored in relation to DIA methods.

The sensitivity and dynamic range of Multiplex Adduct Peptide Profiling (MAPP) were investigated. In analytical chemistry sensitivity is defined as ‘the change in the response of a system for a small change of the stimulus causing the response’.[115] The slope of a line generated from the measured outputs of an analytical technique against a set of standards of known concentration is a function of the sensitivity of the analytical technique. In DIA mass spectrometry, the isolation window width impacts the sensitivity of the method.[83] The wider an isolation window width will result in coverage of a larger m/z space, but it will also introduce more complexity to the spectra by way of interfering species. Multiplex Adduct Peptide Profiling sensitivity was tested with small (2 m/z), intermediate (5 m/z), and wide isolation widths (25 m/z). The sensitivity of an analytical method is defined in Equation 2 as:

\[ S(x, y) = \frac{dy}{dx} \]

Where S is sensitivity, y is the measured output, and x is the concentration of standard.[116]

Design of Experiments (DOE) was used to map the experimental space and to identify parameters which significantly affected adducted albumin enrichment and digestion. DOE was pioneered by R. Fisher and has been utilized in applications to mass spectrometry.[80, 117, 118] DOE has been employed by researchers to save time and money in industrial and academic
settings. The use of statistically designed experiments to analyze a system helps one approach the optimum for an experimental space because DOE measures the effect of interactions on response. Other principles used in DOE are replication, sample blocking, sequence randomization, and fractional factorials. In bottom-up proteomic sample preparation protocols and in protein isolation procedures, there are opportunities to optimize a host of parameters. S. Rappaport et al. at UC Berkeley has studied albumin adducts to styrene derivatives, benzoquinone derivatives, and other compounds for the past 20 years.\[^{41, 119, 120}\] The procedure utilized for albumin isolation from plasma relies upon ammonium sulfate precipitation of immunoglobulins, re-suspension in tris/NaCl, and enrichment of adducted albumin by use of thiol resin. The enrichment of modified albumin was the most time consuming step of the sample preparation used in Fixed Step-SRM and MAPP studies with albumin.\[^{41}\] S. Rappaport et al. used a protocol that called for the isolated albumin to interact with the thiol resin for eighteen hours at room temperature with agitation. The albumin to resin ratio used for enrichment was 1 mg isolated albumin: 100 mg thiol resin. However, the literature is not suggestive of this procedure every being optimized. We investigated five parameters using DOE which included: albumin to thiol resin ratio (1 mg albumin: 5 mg resin, 1 mg albumin: 100 mg resin), enrichment on protein or peptide level, agitation during enrichment, enrichment time (2, 18 hr), and digestion time (0.5, 4 hr).

3.2 Materials and Methods

Formic acid (FA), ammonium bicarbonate (AB), hydrochloric acid (HCl), iodoacetamide (IAM), dithiothreitol (DTT), β-mercapto ethanol, benzoquinone (BQ),
naphthoquinone (NQ), sodium chloride (NaCl) tris, and maleimide were obtained from Sigma Aldrich (St. Louis, MO). Trypsin was obtained from Promega (Madison, WI). HPLC grade acetonitrile, methanol, and water were from Burdick & Jackson (Muskegon, MI). The hemoglobin peptide (GTFATLSELHCDK) was synthesized and purchased from Thermo Fisher Scientific GmbH (Ulm, Germany). Activated Thiol Sepharose 4B resin was purchased from GE Healthcare Life Sciences (Piscataway, NJ). Pooled plasma obtained from smokers was purchased from BioreclamationIVT (New York, New York). SDS-PAGE materials including coomassie stain (part #161-0786), running buffer 10x (part #161-0732), laemmli buffer (part #161-0737), and molecular weight ladder (part #161-0317) were obtained from BioRad (Hercules, CA).

**Linearity and Sensitivity**

A 7 mM hemoglobin peptide (GTFATLSELHCDK) solution was prepared with 500 μL distilled water. Dithiothreitol was added to a final concentration of 5 mM and the sample was incubated at 60 °C for 30 minutes to reduce disulfide linkages. Next, four 60 μL aliquots were reacted with the following electrophilic compounds: iodoacetamide, maleimide, naphthoquinone, and benzoquinone. Each of these compounds were added to achieve a final concentration of 30 mM and incubated in the dark for an hour at room temperature. Agitation was provided by placing samples on a belly dancer during the adduction step. Solid phase extraction (SPE) columns were used to remove excess reactive electrophiles from the samples. MCX SPE columns (30mg LP extraction cartridges part# 186000782) from Oasis (Milford, MA) were mounted onto a MCX manifold (Waters). Columns were conditioned with one mL of methanol, one mL of 10% NH₄OH in water, two mL of methanol, and three mL of 0.1% FA
in water. Samples were added to SPE columns and washed with 0.1% FA in water to remove salts. Neutral species were removed with one mL 0.1% FA in methanol. Peptides were eluted and collected in one mL 10% NH₄OH in methanol. Samples were lyophilized with a Savant SPD313DDA speed vacuum (Thermo Fisher). Peptides were reconstituted with mobile phase A (98% water, 2% acetonitrile, and 0.1% FA) and diluted into digested kidney cell lysate (prepared as described in the methods section of the Appendix under detergent comparison) such that in theory the concentration of each adducted peptide was 200, 100, 50, 25, 12, and 6 femtomoles adducted peptide/µL of solution.

DIA Methods were created in Skyline (daily 2.6.1.6899).[106] The 13 amino acid tryptic peptide from human hemoglobin subunit β that contained the free Cys⁹³ residue was entered into skyline (K.GTFATLSELHCDK). An isolation list was created in Skyline that used 12 2, 5, and 25 m/z isolation windows which span contained the masses of all four adducts to the peptide and still contained the doubly charged unmodified peptide precursor (m/z 711.3401). The isolation list was imported into the method editor of Q Exactive Plus, and an MS1 scan was added from m/z 550 to 1300 for a total 13 scans per cycle with an average duty cycle of 1.3 seconds. For the MS1 scan the resolving power was set to 70,000, an AGC of 1e6, and a max injection time of 50 ms was used. For the DIA scans, parameters were set to the given levels: AGC of 1e6, NCE of 30, max injection time of 50 ms, and a resolving power of 17,500.

Adducted Albumin Recovery and Isolation Assessment by SDS-PAGE & A280

Albumin isolation and adduct enrichment were performed based upon the method previously reported by S. Rappaport.[50] The albumin was isolated from 1 mL of raw plasma instead of 4 mL. During various steps in sample manipulation (raw plasma, supernatant
albumin, de-salted albumin, and enriched albumin) 40 µL aliquots were taken for gel electrophoresis and for measurement of protein concentration by absorbance on a Nanodrop 2000 spectrophotometer (Thermo Fisher). Two technical replicates were prepared and taken through the procedure. The first aliquots were taken and ammonium sulfate was added to 1 mL of plasma to create a 60% saturated solutions. The samples were centrifuged at 5000 rpm for 10 minutes to pellet immunoglobulins, and aliquots were taken for analysis of albumin recovery. The supernatant containing albumin was taken and albumin was acid precipitated from the supernatant by adding HCl to a final concentration of 250 mM. The albumin pellet was re-suspended in buffer containing 500 mM NaCl and 100 mM Tris and the third set of aliquots were taken for analysis of albumin recovery. Isolated albumin was added to 75 mg of dry Activated Thiol Sepharose in a spin column (Thermo product #69725) and the flow through fraction (nonmercaptoalbumin) was collected after 18 hours of enrichment. The final aliquots were taken for analysis of adducted albumin recovery.

**SDS-PAGE on Albumin Isolation and Enrichment**

SDS-PAGE was performed using a PowerPac Basic Criterion from BioRad (Hercules, CA). Running buffer was prepared by dilution of 100 mL 10 x stock from BioRad with 900 mL of distilled water. Gels (part # 345-0042) were purchased from BioRad. The comb and tape were removed from the gel and the electrophoresis cell was assembled. The inner and outer buffer chambers were filled with running buffer. Ten µL of molecular weight ladder (BioRad) spanning a range from 20-250kDa was injected on the first and last lane of the gel. Three µg of albumin was loaded as a control for qualitative comparison of albumin recovery during the sample processing steps. Blanks were run in between samples taken from different
steps in the procedure. Samples were prepared in the following manner: 1 μL sample/(50 mM ammonium bicarbonate for blanks), 9 μL 50 mM ammonium bicarbonate buffer, 9.5 μL Laemmli buffer, and 0.5 μL (355 mM final concentration) β-mercapto ethanol. Samples were heated for 30 minutes at 60 °C to denature proteins. The electrophoresis cell was connected to the power supply and the run conditions were 200 V for 60 minutes. After the run finished, the gel was removed and stained by performing 3 five minute rinses in distilled water. Next, approximately 50 mL of Coomassie stain (BioRad) was added, the gel was covered, and shaken for 1 hr on a belly dancer from IBI Scientific (Peosta, IA). Then, the gel was rinsed with distilled water for 30 minutes. The gel was imaged using the trans-UV setting on a Gel Doc XR+ (BioRad).

**DOE on Enrichment Parameters**

Design Expert software version 9.0.3.1 (Stat-Ease inc. Minneapolis, MN) was used to design the adducted albumin enrichment DOE experiment. The experimental design had a resolution of 4 allowing for main effects and 2nd order interactions to be determined which resulted in 16 conditions to analyze 5 factors. The factors and levels tested were: enrichment on the protein level vs enrichment on the peptide level (e.g. digestion before or after enrichment), agitation vs no agitation during enrichment, enrichment time (2, 18 hr), digestion time (0.5, 4 hr), and albumin to thiol resin ratio (1 mg albumin: 5 mg resin, 1 mg albumin: 100 mg resin). A 3.0 mL sample of plasma was used to isolate albumin. Saturated ammonium sulfate was added to the plasma to a final concentration of 60 % ammonium sulfate (w/v). The plasma was centrifuged at 5000 rpm for 10 minutes to pellet immunoglobulins and the supernatant containing albumin was de-salted by acid precipitation (HCl was added to a final
concentration of 250 mM) of albumin from the supernatant. The albumin pellet was re-
suspended in buffer containing 500 mM NaCl and 100 mM Tris. Solutions of human serum 
albumin (HSA) adducted maleimide and HSA adducted N-ethyl maleimide were prepared such 
that each adduct was present at 9.167 µg adducted HSA/µL tris NaCl buffer solution. At this 
point, 16 solutions of 200 µL each were prepared containing 400 µg isolated HSA, 50 µg of 
maleimide adducted HSA, and 50 µg of N-ethyl maleimide adducted HSA. The samples were 
added at a 1 mg albumin: 5 mg resin or 1 mg albumin: 100 mg resin (Activated Thiol 
Sepharose) in a spin column (Thermo product #69725). Half of the samples were agitated on 
a belly dancer while the other samples received no agitation. The flow through fraction 
(nonmercaptoalbumin) was collected after a 2 or 18 hour enrichment. The enrichment was 
performed after digestion for half the samples and before digestion for the other half. All 
samples were digested using filter-aided sample preparation for shotgun proteomics\textsuperscript{105} using 
a 0.5 or 4 hr tryptic digestion at an enzyme to substrate ratio of 1:50.

DIA Methods were created in Skyline.\textsuperscript{106} The 21 amino acid tryptic peptide from 
human serum albumin that contained the free Cys\textsuperscript{34} residue was entered into Skyline 
(K.AVLIAFAQYLQQ\textsuperscript{C}PFEDHVKL). An isolation list was created in Skyline that used 12 10 
m/z isolation windows which spanned m/z 810 to 930 but still contained the triply charged 
unmodified peptide precursor (m/z 811.7594). This range of precursor detection included the 
mass of the albumin free cysteine unadducted peptide to a mass addition (i.e., adduct) of 360 
Daltons which encompasses the masses of all reported protein adducts aside from Satratoxin 
G\textsuperscript{66}. 
The isolation list was imported into the method editor of Q Exactive Plus and an MS1 scan was added from \( m/z \) 550 to 1300 for a total 13 scans per cycle with an average duty cycle of 1.3 seconds. For the MS1 scan, the resolving power was set to 70,000, an AGC of 1e6, and a max injection time of 50 ms. For the DIA scans, parameters were set to the following levels: AGC of 1e6, 30 NCE (normalized collision energy), max injection time of 50 ms, and a resolving power of 17,500.

*NCE Optimization*

A volume of 50 µL enriched adducted albumin sample (digested) was prepared as described in Chapter 2 Materials and Methods Albumin Isolation and Enrichment. In brief, ammonium sulfate was added to 4 mL of plasma to create a 60% saturated solution. Albumin was acid precipitated using HCl from the supernatant. Isolated albumin (500 µg) was added to 75 mg of dry Activated Thiol Sepharose in a spin column (Thermo product #69725) and the flow through fraction (nonmercaptoalbumin) was collected after 18 hours of enrichment. The enriched albumin fraction was then prepared using the filter-aided sample preparation procedure for shotgun proteomics\[105\] using a 4 hour tryptic digestion at an enzyme to substrate ratio of 1:50.

The DIA methods used in this experiment were prepared in the same manner described in the DOE study, but modifications to the HSA-Cys\(^{34}\) peptide were maleimide, sulfonic acid (dioxidation), cysteine, and iodoacetamide. The instrument parameters used were the same as those described in the DOE analysis with the exception of the NCE used. NCE (normalized collision energy) values were 20, 22, 24, 26, 28, and 30. Analytical replicates were used by running each method three times on the same sample.
NanoLC MS/MS

A sample volume of 2 µL was injected onto a self-made 4 cm trap using an Easy nanoLC 1000 coupled to a Q Exactive Plus mass spectrometer (Bremen Germany). PicoFrit columns from New Objective (Woburn, MA) were packed to 25 cm in house with 3 µm C18 silica particles (Dr. Maisch, Entringen, Germany). Mobile phase B was 99.9 % acetonitrile with 0.1% formic acid and mobile phase A consisted of 98 % water, 2 % acetonitrile, and 0.1 % formic acid. A 90 minute LC MS/MS method was used for the linearity and sensitivity study and consisted of a linear gradient from 0-40 % B over 70 minutes followed by a ramp to 80 % B in one minute. The column was washed at 80% B for 9 minutes and regenerated at 0 % B for 10 minutes. For the NCE and DOE studies the LC MS/MS method was 40 minutes and utilized a 20 minute gradient from 0-40% B.

Data Analysis

In the linearity study, peptides with maleimide, benzoquinone, naphthoquinone, and iodoacetamide modifications were created in Skyline by adding the modifications to the cysteine residue of the Cys\textsuperscript{93} tryptic peptide. Adducted hemoglobin peptides were identified by extracting MS1 features, adducted, and common fragment ions from the Cys\textsuperscript{93} peptide, and finding the instances in retention time space where all co-eluted with high mass measurement accuracy (<3 ppm). The precursor areas were exported from Skyline and analyzed in Microsoft Excel.

The protein absorbance concentrations at 280 nm were multiplied by volumes to give total protein for each of the following: raw plasma, supernatant depleted of immunoglobulins, de-salted and re-suspended HSA, and enriched adducted HSA.
The DOE the HSA peptides with maleimide and N-ethyl maleimide modifications were created in Skyline by adding the modification to the cysteine residue of the Cys$^{34}$ tryptic peptide. Maleimide adducted HSA and N-ethyl maleimide adducted HSA were identified by extracting MS1 features, adducted, and common fragment ions from the Cys$^{34}$ peptide, and by finding the instances in retention time space where all co-eluted with high mass measurement accuracy (<3 ppm). The precursor areas were exported from Skyline and imported into Design Expert (version 9.0.3.1). Design Expert analyzed factors and interactions for significance.

In the NCE optimization, HSA peptides with carbamidomethylation, trioxidation, cysteinylation, and maleimide modifications were created in Skyline by adding the modification to the cysteine residue of the Cys$^{34}$ tryptic peptide. Adducted HSA peptides were identified by extracting MS1 features, adducted, and common fragment ions from the Cys$^{34}$ peptide and finding the instances in retention time space where all monitored fragments and MS1 features co-eluted with high mass measurement accuracy (<3 ppm). The fragment ion peaks from b4-b8 and y4-y8 transitions, and precursor peaks were integrated and exported to excel (Microsoft Excel 2013 version 15.0.4693.1000). In excel, the peak areas of the transitions were summed and divided by the precursor peak area for each modified HSA peptide. As three analytical replicates were run at each NCE value tested, an average value was calculated for each modified HSA type at each NCE level. The data was analyzed by LOWESS (Locally Weighted Scatter-plot Smoothing) in RStudio version 0.98.1102.$^{[121]}$
3.3 Results and Discussion

3.3.1 Linearity and Sensitivity

Figure 3.1 describes the experimental workflow utilized in the linearity and sensitivity study. The hemoglobin peptide from the β subunit which contains the free thiol under physiological conditions (Hemβ Cys)$^{93}$ was adducted with two environmentally relevant carcinogenic compounds (benzoquinone and napthoquinone) and two alkylating reagents commonly used in proteomic experiments (maleimide and iodoacetamide). The adducted peptides were spiked into a 293 kidney cell lysate at concentrations of 200, 100, 50, 25, 12, and 6 fmole of adducted peptide/µL solution for each type of modified peptide. The injection volume used in this study was 2 µL, thus 400, 200, 100, 50, 25, and 12 fmoles of each adducted peptide were injected on-column. The samples were analyzed by MAPP with isolation window widths of 2, 5, and 25 $m/z$.

Figure 3.1 The experimental workflow utilized in the linearity and sensitivity study. The adduct standard consisted of the hemoglobin peptide (GTFATLSELHCDK) adducted to benzoquinone, naphthoquinone, iodoacetamide, and maleimide.
**Figure 3.2** describes the sensitivity of MAPP with 2, 5, and 25 \( m/z \) isolation windows for hemoglobin adducted with benzoquinone and naphthoquinone. The sensitivity of an analytical technique corresponds to the slope of the line generated from analysis with the method. Precursor peak area for each adducted peptide was plotted against the fmoles of the adducted peptide injected on-column. The slope of lines generated with 2 and 5 \( m/z \) isolation windows were significantly [p-values < 0.0001 (BQ), 0.003 (NQ), 0.11 (IAM), 0.20 (Mal)] steeper than that of the 25 \( m/z \) isolation window line as calculated by analysis of covariance at alpha = 0.05. This observation is noteworthy, because it shows that a 25 \( m/z \) isolation window introduces significant noise and complexity to the signal that decreases the sensitivity of the method. Another interesting observation from this work was that a 5 \( m/z \) window does not significantly differ in sensitivity from a targeted \( m/z \) isolation window. The use of a 5 \( m/z \) isolation window (rather than 2 \( m/z \)) enables 50 % greater \( m/z \) space coverage compared to a targeted method.
Figure 3.2 Plots of peak area of Hem-Cys\textsuperscript{93} adducted with benzoquinone and naphthoquinone spiked into kidney cell lysate against fmoles adducted peptide injected on-column. The differences in slope shows varied sensitivity depending upon isolation window width chosen. The slopes of lines generated with a 25 m/z isolation window were significantly different from those of both the 2 and 5 m/z isolation windows [p-values < 0.0001 (BQ), 0.003 (NQ)]. However, the slopes of lines derived from targeted and 5 m/z isolation windows were not significantly different. Analysis was performed using analysis of covariance at $\alpha = 0.05$.

3.3.2 Verify Albumin Isolation

Figure 3.3 describes the workflow used in the SDS-PAGE and protein concentration experiments which monitored isolation and enrichment of adducted albumin. Raw plasma from a male smoker was depleted of immunoglobulins by addition of a saturated solution of ammonium sulfate, the supernatant containing albumin was removed, and the solution was desalted by acid precipitation of albumin and re-suspension in buffer. The isolated albumin was enriched for adducted albumin with a thiol affinity resin. Two technical replicates were prepared and at each step (circled in red in Figure 3), and two aliquots were taken from each replicate. One aliquot was taken for protein concentration determination by absorbance at 280 nm, and the other aliquot was saved for SDS-PAGE.
Figure 3.3 The experimental workflow used for isolation and enrichment of human serum albumin. Steps circled in red indicate points where aliquots were taken for SDS-PAGE and protein absorbance experiments.

Figure 3.4 illustrates the results of the SDS-PAGE experiment. The results confirmed the presence of albumin in each step of the isolation and provided a qualitative assessment of protein isolation purification. Some lower and higher molecular weight protein bands were detected in the lanes from the isolation steps; however, the enrichment step removed most of these species.
Figure 3.4 The imaged gel used in the SDS-PAGE experiment. The molecular weight ladder was run in lanes 1 and 12. The middle lanes contained from left to right: lane 2 a control comprised of albumin from Sigma Aldrich, lane 3 a blank, lane 4 enriched adducted albumin, lanes 5-6 duplicates of acid precipitated (de-salted) isolated albumin, lanes 7-8 duplicates of IgG depleted plasma, and lanes 10-11 duplicates of raw plasma.

Figure 3.5 illustrates isolated and enriched adducted albumin from the plasma of a male smoker during the protein concentration experiment. Protein concentration values were measured by absorbance of light at 280 nm and were subsequently multiplied by the sample volume to calculate the total protein.
Figure 3.5 The results from the protein concentration experiment conducted during isolation and enrichment of HSA. The total protein was calculated by multiplying absorbance with sample volume (1 mL of plasma) and plotted against isolation steps. Two technical replicates were utilized and a decreasing trend was observed as albumin was isolated and enriched. For this individual, the percent recovery of adducted albumin from total protein was 2.3%.

Initially the plasma samples contained approximately 60,000 µg of protein, and a loss of protein occurred in each step. After the enrichment, approximately 1400 µg of albumin was recovered (approximately 2.3% of total protein in raw plasma). Albumin constitutes between 50-60% of protein in plasma. The following calculation shows how one can estimate the amount of adducted albumin in this individual:

1. $HSA \, \mu g = (Total \, protein \, \mu g) \times 0.55 = 33000 \, \mu g \, HSA$

2. $\frac{1400 \, \mu g \, adducted \, HSA}{33000 \, \mu g \, HSA} \times 100 = 4.3\% \, adducted \, HSA$

After the observation of decrease in protein recovery through the procedure, we confirmed in a qualitative manner the presence of albumin in each step of the isolation and enrichment. After verification of the sample preparation procedure, we noted a host of parameters which could be optimized using design of experiments (DOE).
3.3.3 DOE on Parameters Affecting Enrichment

Table 3.1 summarizes the parameters selected for optimization and motivations for investigating each factor. Some parameters were chosen to save time and money, such as the enrichment which was the rate limiting step in the procedure at 18 hours, or the amount of thiol resin to use as the cost is over $600/15 grams. Other parameters chosen were exploratory, such as the need of agitation during enrichment, or whether enrichment is favored on the peptide or protein level.

Table 3.1 Parameters and motivations selected for optimization.

<table>
<thead>
<tr>
<th>Parameter (levels)</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enrichment time (2, 18 hr)</td>
<td>Does longer enrichment result in decreased mercaptoalbumin?</td>
</tr>
<tr>
<td>Digestion time (0.5, 4 hr)</td>
<td>Does digestion time affect digestion efficiency of adducted albumin?</td>
</tr>
<tr>
<td>Enrichment Agitation (shake, no shake)</td>
<td>Does shaking during digestion lead to greater depletion of mercaptoalbumin?</td>
</tr>
<tr>
<td>Thiol resin: albumin ratio (1:100, 1:5)</td>
<td>What amount of thiol resin is sufficient to maximize mercaptoalbumin capture?</td>
</tr>
<tr>
<td>Enriched species (peptide, protein)</td>
<td>Is it preferable to enrich at the protein or peptide level?</td>
</tr>
</tbody>
</table>

Table 3.2 summarizes results obtained from analysis of the enrichment parameters through DOE. The output measured in the optimization experiment was MS1 peak area for the tryptic peptide containing Cys$^{34}$ of HSA adducted with N-ethyl maleimide (NEM) and maleimide. The peak areas for maleimide and N-ethyl maleimide adducted HSA were obtained from Skyline and were input into Design Expert. Parameter p-values were calculated
by analysis of variance (ANOVA) performed at alpha = 0.05. Both NEM and maleimide adducted HSA enrichment was most efficient at the protein level \( [p\text{-value} = 0.016 \text{ (Mal)}, 0.011 \text{ (NEM)}] \). The higher amount of thiol resin (1:100) required was found to be significant for maleimide adducted HSA \( (p\text{-value} = 0.044) \) but marginally significant for NEM adducted HSA \( (p\text{-value} = 0.21) \). The shorter digestion time was marginally significant for maleimide adducted HSA \( (p\text{-value} = 0.120) \) but was not found to be significant for NEM adducted HSA \( (p\text{-value} = 0.308) \). Other parameters were not found to be significant including the longer enrichment time (18 hr), so the protocol was altered in future experiments to a 2 hour enrichment time. The modified protocol obtained after exploring the experimental space with DOE enabled subsequent sample preparation which previously required 2-3 days to be accomplished in 1-2 days.

\textbf{Table 3.2} Summarizes results from the optimization with MS1 peak area as the measured output. P-values correspond to alpha = 0.05. Favorable conditions were underlined and include enrichment on the protein level, higher thiol resin to albumin ratio (1:100), and 4 hour digestion.

<table>
<thead>
<tr>
<th>Parameter (levels)</th>
<th>P-value (Mal)</th>
<th>P-value (NEM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enrichment time (2, 18 hr)</td>
<td>0.840</td>
<td>0.543</td>
</tr>
<tr>
<td>Digestion time (0.5, 4 hr)</td>
<td>0.120</td>
<td>0.308</td>
</tr>
<tr>
<td>Enrichment Agitation (shake, no shake)</td>
<td>0.929</td>
<td>0.686</td>
</tr>
<tr>
<td>Thiol resin: albumin ratio (1:100, 1:5)</td>
<td><strong>0.044</strong></td>
<td><strong>0.0210</strong></td>
</tr>
<tr>
<td>Enriched species (peptide, protein)</td>
<td><strong>0.016</strong></td>
<td><strong>0.011</strong></td>
</tr>
</tbody>
</table>
3.3.4 NCE Optimization

Figure 3.6 describes the methods used in the NCE optimization procedure. The range of NCEs investigated spanned from 20-30 at increments of two. Three analytical replicates were performed with each method, and the sequence was block randomized. The measured outputs were MS1 (precursor) and MS2 (y4-y8 and b4-b8 ions) features from four adducts to the tryptic peptide containing the free thiol on HSA (Cys34). The modifications measured were cysteine, maleimide, iodoacetamide, and sulfonic acid (dioxidation).

![Repeats every 1.5 s]

**Figure 3.6** The methods created for NCE analysis. Six methods were created spanning a range of 20-30 NCE were used. Analytical replication was used (3x/NCE tested). The sequence was block randomized. The peak areas of HSA adducted to cysteine, maleimide, iodoacetamide, and sulfonic acid (dioxidation) were monitored at each NCE.

The results from the NCE optimization are summarized in Figure 3.7 as a LOWESS (Locally Weighted Scatterplot Smoothing) regression plot (analyzed in RStudio). Each NCE was tested three times and monitored four modifications (cysteine, iodoacetamide,
maleimide, and sulfonic acid). The precursor and fragment ions were identified, extracted from Skyline, and sorted according to the NCE used for fragmentation. The peak area for the monitored fragments for each adduct were summed (e.g. b4-b8 and y4-y8 for maleimide HSA) and divided by the precursor peak area of the corresponding adducted peptide. As there were three analytical replicates, an average was calculated for total MS2 peak area/MS1 peak area features for each adducted peptide at each NCE level. The averaged ratios of MS2 peak area/MS1 peak area for each adduct were used in the LOWESS regression. This ratio accounted for any errors prior to the mass spectrometry analysis (i.e., autosampler injection). An interesting observation was the presence of a maximum at an NCE of 24. The methods were adjusted in future analysis to an NCE of 24 for more efficient fragmentation.
Figure 3.7 The LOWESS regression plot summarizes results from the NCE optimization study. The NCEs tested were: 20, 22, 24, 26, 28, and 30 and are plotted against ratios of total peak area of 10 fragments (b4-b8 and y4-y8) divided by the precursor peak area. The albumin peptide above was monitored with the four modifications on the right. Three analytical replicates were used at each NCE tested which enabled averaging. A maximum was observed at an NCE of 24 and the dark grey region corresponds to the 95% confidence limit.

3.4 Conclusion

Herein, the characterization of MAPP has been described through assessment of the sensitivity and linearity of the technique using an adduct standard spiked into a complex cell lysate. The sensitivity was compared across three isolation windows: 2, 5, and 25 m/z. From these results a 5 m/z isolation window was found to be comparable to a 2 m/z isolation window. The 25 m/z window was significantly less sensitive due to the presence of interfering species.
Determination of the sensitivity of different $m/z$ isolation windows will enable the development of optimal MAPP methods for the future detection of protein adducts.

Before MAPP could be used to explore the $m/z$ space of adducts in plasma, sample preparation procedures had to be investigated. DOE parameters from the enrichment and digestion steps were optimized to enhance protein adduct recovery and the protocol was modified according to the results. After we optimized the sample preparation, the fragmentation of modified peptides was examined.

The effect of NCE on peptide fragmentation in DIA is poorly understood. In DDA, the role of the NCE is clearer because one ion is isolated at a time and the NCE adjusts to impart sufficient energy for optimal fragmentation. In DIA, the relationship between the NCE and analytes with varied vibrational modes is not well understood. We tested a variety of NCE values (3x per NCE tested) on one sample and analyzed which energy gave the best fragmentation for four protein adducts. An NCE of 24 was determined by LOWESS regression to be optimal in the range of NCEs tested. Enhanced fragmentation is critical as the identification of precursors using MAPP depends upon the elution profiles of MS2 features.
CHAPTER 4

DEVELOPMENT OF A DATA PIPELINE FOR MULTIPLEX ADDUCT PEPTIDE PROFILING (MAPP)

4.1 Introduction

Bioinformatics is a burgeoning field of research due to the emergence of big data and systems biology.\textsuperscript{123, 124} It encompasses the science of collecting and analyzing complex biological data, and is an especially active area of research in the field of proteomics.\textsuperscript{125, 126} Data files collected over a three hour gradient performing LC-MS/MS can be as large as 2-3 gigabytes. The ultimate goal is the deconvolution of data into biological meaningful results with appropriate statistical confidence.

Traditional discovery proteomic workflows using data dependent acquisition (DDA) LC-MS/MS to identify protein modifications require prior knowledge of the mass of the modification and variable modification searches. The reason foreknowledge is required can be appreciated by an understanding of a DDA data analysis pipeline (in this example a pipeline using ProteomeDiscoverer analysis software is given). A variety of freeware and proprietary software tools to aid in analysis of DDA data are available including: ProteinDiscoverer (Thermo), Sequest, OPENMS, Percolator, and Preview.\textsuperscript{127-130} The DDA workflow will include the raw data file, the corresponding proteome database for the data to be searched against (e.g. FASTA file from Uniprot), an algorithm used to query the data (SEQUST, MSAMANDA). Within the search node, one inputs the mass of modification and the amino acid believe to be modified to query as a dynamic modification. The software then considers each amino acid to be both modified and unmodified (i.e., dynamic). Due to the exponential increase
in search space most algorithms limit the number of modifications to less than five (e.g., protein discoverer). This approach has 3 principal limitations, two of which are illustrated in Figure 4.1, in elucidating unknown modifications to peptides: 1) Multiple modification searches exponentially increase data analysis times limiting most searches to a few known modifications (Figure 4.1A); 2) Identification is ultimately dependent on the modified peptide being sampled by data dependent acquisition – a process known to be semi-stochastic and biased towards the identification of abundant species (Figure 4.1B) and; 3) These workflows only identify known protein modifications that are included as variable modifications in the database search.[131] The latter may be the most noteworthy limitation as it biases identification and stifles discovery of novel or unanticipated modifications. Other methods include the burgeoning field of targeted proteomics [132, 133] but this technique also requires prior knowledge of the elemental composition of the modification such that appropriate precursor-fragment transitions can be chosen.
Figure 4.1 Two limitations of DDA LC-MS/MS in the detection of protein adducts include A) Reproducibility, identification of peptides is 70-80% for three analytical replicates. B) The increase in search time with the number of modifications searched.

Data Independent Acquisition methods \(^{98}\) coupled to targeted data extraction \(^{99,100}\) has offered an alternative workflow for analysis of proteomic samples using LC MS/MS. Briefly DIA differs from DDA in the selection of ions for fragmentation. DDA selects the most abundant peptide species (the number selected depends on the loop count 10-20), then isolates and fragments each ion one at a time. However, in DIA after a precursor scan is taken, all ions within a pre-specified \(m/z\) isolation window are isolated and fragmented together. The isolation window width can range from 2 \(m/z\) to the entire precursor scan range. Although conventional data dependent acquisition strategies are extremely powerful for surveying proteomes\(^{104}\), methods based on independent acquisition offer certain benefits. The principal advantages are improved reproducibility of peptide identifications due to the sequential scanning of isolation windows, potential improvements in quantitative accuracy due to improvements in signal to noise of MS2 versus MS1 based label-free quantitation, and when
coupled to targeted data extraction the capability to re-query a dataset for any peptide of interest. Essentially, one can propose questions and test hypotheses in the data.

Some variations of DIA methods include MS\textsuperscript{E} (Waters Q-TOF) and All-ion Fragmentation (AIF Thermo Orbitrap) both of which alternate between low energy and high energy state scans enabling the transmission of all precursors followed by their fragments. However, a disadvantage of MS\textsuperscript{E} and AIF is the complexity of the MS2 spectra collected which complicates matching fragment ions to precursors.\textsuperscript{[83]}

In Sequential Window Acquisition of all Theoretical Mass Spectra (SWATH-MS ABSciex Q-TOF), the first quadrupole sequentially moves in 25 m/z window across the mass range of interest, ions are fragmented, and analyzed by a TOF mass analyzer. For SWATH-MS the duty cycle is longer than that of MS\textsuperscript{E} and AIF and the method requires knowledge of the m/z space to target, but the spectra collected contain less interfering species.\textsuperscript{[134]}

In spite of the aforementioned data acquisition techniques and bioinformatics, there remains a paucity of bioinformatic tools for the discovery of protein modifications and data collected by DIA.

S. Rappaport and coworkers were among the first to attempt discovery of protein adducts using a novel method called Fixed Step Single Reaction Monitoring (FS SRM) shown in Figure 4.2.\textsuperscript{[41]}
Figure 4.2 The schematic used for FS-SRM. In Q1 ions were selected within an isolation window of 1.5 m/z. Seventy-six steps covering 815-929 m/z were used to find potential adducts. In q2 ions were fragmented by collision induced dissociation, and in Q3 the b4, y15, y16, and y17 ions were selected. The mass of the b4 ion is not specific to HSA Cys34 peptide (with the human proteome as a background) according to SRM Collider indicating more unadducted transitions would need to be monitored to suggest a modification.\textsuperscript{[135]}

They studied reactive electrophilic adducts to the nucleophilic cysteine 34 on human serum albumin. The free thiol on cysteine 34 accounts for approximately 80\% of free thiols in plasma and provides a “sink” for reactive electrophilic species.\textsuperscript{[50]} The importance of studying protein adducts as a metric in exposure is due to the stability of protein adducts relative to the reactive electrophiles. Human serum albumin has a half-life of 28 days.\textsuperscript{[41]} To process data collected utilizing FS SRM a data pipeline with 3-4 potential steps was developed in-house to assign adduct masses as shown in Figure 4.3. Control data and sample data were collected and converted to MzXML.
Figure 4.3 The involved data workflow developed by Rappaport for accepting potential cysteine 34 albumin adducts using FS SRM.

Additionally, geometric means and a 95% confidence limit were calculated for the monitored adducted (y15-y17) and unadducted (b4) transitions in Figure 4. The first rule for determining an adduct was the experimental lower confidence limit of intensity must be less than the control upper confidence limit, if this was not the case the potential adduct was discarded. Other rules involved comparison of intensity ratios of y15 and y16 relative to numeric threshold values.
Figure 4.4 The transitions monitored in FS-SRM are shown in red. The $b_4^+$ ion was common to all potentially modified species. The $y_{15}^- - y_{17}^+$ were modified.

The data pipeline utilized for FS SRM was rather complicated, the method was based on low mass measurement accuracy (600 ppm) and the threshold values used in the criteria for determining an adduct lack sufficient justification. Another significant limitation of FS-SRM is the monitoring of one unadducted transition to suggest the presence of an adduct. SRM Collider (software tool developed by the Aebersold lab) can assess the specificity of transition masses within a digested proteome. SRM Collider analysis of the HSA Cys34 peptide has shown one transition ($b_4$) is not a unique mass in the tryptically digested human proteome.[135]

We have developed a simplified DIA MS data processing pipeline for assignment of potential adducts using high mass accuracy (< 5 ppm). The contribution of high mass measurement accuracy to MAPP provides significantly greater confidence in protein adduct identification. The MAPP data pipeline has four steps from the collection of raw data file to adduct identification. The 4 steps include: 1) identification of potential peptide adduct utilizing Skyline to monitor unadducted peptide transitions, 2) extraction of the precursor mass from the raw file, 3) filtration of precursor masses according to the isolation window where the adduct of interest originated, and 4) precursors and product ions are graphed and matched.
according to elution profile. While the method is not yet automated, it provides a great starting point for future bioinformatics software development and applications.

### 4.2 Materials and Methods

**Materials**

Formic acid (FA), ammonium bicarbonate (AB), hydrochloric acid (HCl), iodoacetamide (IAM), dithiothreitol (DTT), sodium deoxycholate (SDC) were obtained from Sigma Aldrich (St. Louis, MO). Trypsin was obtained from Promega (Madison, WI). HPLC grade acetonitrile, methanol and water were from Burdick & Jackson (Muskegon, MI). The hemoglobin peptide (GTFATLSELHCDK) was synthesized and purchased from Thermo Fisher Scientific GmbH (Ulm, Germany). Pooled plasma obtained from smokers was purchased from BioreclamationIVT (New York, New York).

**NanoLC MS/MS**

A sample volume of 2 µL was injected onto a self-made 4 cm trap using an Easy nanoLC 1000 coupled to a Q Exactive Plus mass spectrometer from Thermo Scientific (Bremen Germany). PicoFrit columns from New Objective (Woburn, MA) were packed to 25 cm in-house with 3 µm C18 silica particles (Dr. Maisch, Entringen, Germany). Mobile phase B composition was 99.9 % acetonitrile with 0.1% formic acid, and mobile phase A composition was 98 % water, 2 % acetonitrile, and 0.1 % formic acid. A 90 minute LC MS/MS method was used and consisted of a linear gradient from 0-40 % B over 70 minutes, followed by a ramp to 80 % B in one minute. The column was washed at 80% B for 9 minutes and regenerated at 0 % B for 10 minutes.
Albumin isolation and adduct enrichment were performed based upon the method described in chapter 2. In brief, ammonium sulfate was added to 4 mL of plasma to create a 60% saturated solution. Albumin was acid precipitated using HCl from the supernatant. Isolated albumin (500 µg) was added to 75 mg of dry Activated Thiol Sepharose in a spin column and the flow through fraction (non-mercaptoalbumin) was collected after 18 hours of enrichment. The enriched albumin fraction was then prepared using the filter-aided sample preparation procedure for shotgun proteomics[105] using a 4 hour tryptic digestion at an enzyme to substrate ratio of 1:50.

DIA Methods were created in Skyline.[106] The 21 amino acid tryptic peptide from human serum albumin that contained the free Cys34 residue was entered into skyline (K.ALVLIAFAQYLQQ\textsubscript{C}PFEDHVK.L). An isolation list was created in Skyline that used 12 10 m/z isolation windows which spanned 810 to 930 m/z but still contained the triply charged unmodified peptide precursor (m/z 811.7594). This range of precursor detection included the mass of the albumin free cysteine unadducted peptide to a mass addition (i.e., adduct) of 360 Daltons which encompasses the masses of all reported protein adducts aside from Satratoxin G[66].

The isolation list was imported into the method editor of Q Exactive Plus and an MS1 scan was added from 550 to 1300 m/z for a total 13 scans per cycle with an average duty cycle of 1.3 seconds. For the MS1 scan, the resolving power was set to 70,000, an AGC of 1e6, and a max injection time of 50 ms. For the DIA scans, parameters were set to the given levels:
AGC of 1e6, 30 NCE (normalized collision energy), max injection time of 50 ms, and a resolving power of 17,500.

**Hemoglobin Isolation/Enrichment**

Pooled smoker’s blood was used for the hemoglobin analysis. The procedure was performed based on a previously reported method.\[^{136}\] Briefly, an aliquot of 150 μL of whole blood was taken and red blood cells were isolated from whole blood by centrifugation at 1400*g for 20 min. Isolated red blood cells were washed 3x with 300 μL 0.9% (w/v) saline solution. The red blood cells were lysed by the addition of 450 μL distilled water. The samples were centrifuged at 5600*g for 60 min and liberated hemoglobin was collected in the supernatant. The hemoglobin was enriched and digested using thiol chemistry and the FASP method analogous to the albumin enrichment (see methods). A similar MAPP method was used for LC MS/MS analysis of hemoglobin adduct which consisted of an 11 10 m/z isolation window and a full precursor scan.

**Data Analysis**

SRM Collider was used to assess uniqueness of transitions monitored from the triply charged HSA Cys\[^{34}\] peptide (ALVLIAFAQYLQQCPFEDHVK). The parameters in collider were: SSRCalculation window (10), Q1 mass window (10 m/z), Q3 mass window (0.050 m/z), mass range for transitions (300-1500 m/z), isotopes to consider (up to 3 amu), missed cleavages (0), find unique ion signature (up to order 2), and ions to monitor (y and b). The transitions monitored include y8-y3 and b3-b6 and uniqueness was assessed through SRM Collider by monitoring combinations of 1, 2, 3, and 4 transitions.
Peptides with artificial modifications were created in Skyline by adding a modification to the cysteine residue of the Cys\textsuperscript{34} tryptic peptide such that a single precursor m/z occupied each isolation window from 810-930 m/z. Putative signals of interest (SOI) were identified by extracting common fragment ions from the Cys\textsuperscript{34} peptide and finding the instances in retention time space where all co-eluted with high mass measurement accuracy (<3 ppm). In addition, the proportion of relative abundances of each fragment to total abundance was used as a method to eliminate false positives. Once an SOI was identified, precursors were filtered based on the isolation window from which the SOI originated and matched in retention time space with the SOI using Xcalibur 2.2 (Thermo). This matching was based on the principle that the MS1 precursor and associated fragment ions all belong to the same chemical species, thus the chromatographic profile of both should be theoretically identical.

*Data Dependent Acquisition for Confirmation of Hemoglobin Adducts*

Hemoglobin adducts were confirmed using a traditional DDA workflow using Protein Discoverer 1.4 and the Sequest HT algorithm using a variable modification search. Files were percolated\textsuperscript{[137]} to increase sensitivity and q value <0.01 was enforced. Precursor mass tolerance was set to 10 ppm while MS 2 mass tolerance was set 0.6 Da. Variable Mods on the cysteine residue included dioxidation, trioxidation, and the putative adduct mass (170.1028). Spectra which identified the different adducts were at similar retention times as identified in the MAPP method and had a posterior error probability (PEP)\textsuperscript{[138]} of less than 0.01.
4.3 Results and Discussion

SRM Collider was used to verify the uniqueness of the monitored transitions (y3-y8 and b3 to b6) to the HSA Cys\textsuperscript{34} peptide\textsuperscript{[135]} The number of transitions and combined uniqueness outputs varied with the combinations as shown in Figure 4.5. When only one transition was monitored, all nine ions (e.g. b9, y4, or b5) were not found to be unique. Monitoring at least two transitions allowed for 36 combinations of which eight were not unique representing 22.2\% non-unique combinations out of the total available combinations. However, when three transitions from the HSA Cys\textsuperscript{34} peptide were utilized to determine unique combinations only the b3, b4, and b5 combination was not unique. The monitoring of four transitions provided only unique ion combinations mapping back to the HSA Cys\textsuperscript{34} peptide.

![Figure 4.5](image)

**Figure 4.5** The number of non-unique combinations as related to the number of transitions monitored using SRM Collider\textsuperscript{[135]} Transitions included y3-y8 and b3-b6. It is important to note that FS-SRM only monitored 1 unmodified fragment ion which is not specific in the human proteome.
Many modifications were identified as adducts to cysteine 34 from human serum albumin extracted from a male smoker using Multiplex Adduct Peptide Profiling (MAPP) and the data pipeline shown in Figure 4.6. The unmodified transitions for these modifications (which included oxidation, dioxidation, trioxidation, and cysteinylation) were found in the 810-820, 820-830, and 850-860 m/z isolation windows in Skyline. In each isolation window, the elution profile contained multiple signals of interest. These potential adducts had peaks composed of five or more unadducted transitions with a mass accuracy of less than 3 ppm. The nominal mass of precursors in the MS1 scan that were ± 1 min were exported from the raw file into an excel file in order find the mass of the precursor. The precursors and their intensities were organized according to the retention time. Then the data was filtered to include only masses within the isolation window where the SOI was observed, e.g. 820-830 m/z for dioxidation.
Figure 4.6: Data pipeline for accepting cysteine 34 adduct on albumin using MAPP. The steps in the pipeline are: 1) Identification of a potential adduct by extraction of unadducted transitions, 2) Extraction of precursors from raw file spanning retention time of signal of interest, 3) Filtration of extracted precursors according to isolation window m/z values, and 4) Plot of transitions with precursors to match elution profile.

Next, the transition intensities detected were plotted against the intensities of precursors to match the elution profile and expected MS1 intensity (a potential adduct must have greater precursor intensity than product ion intensity). After identification of a potential precursor which matched the elution profile, the raw file was viewed to check the charge and isotopic distribution. If the peptide was triply charged and had an isotopic distribution, the monoisotopic mass was used in Equation 1 to calculate the mass of the adduct.
Equation 1: Mod = \[\text{Experimental (m/z)} - \text{Theoretical unmodified (m/z)}\] \times z

Finally, the masses of the y ions containing the adducted mass were then extracted. The mass accuracy threshold for accepting an adduct was within 5 ppm.

**Figure 4.7** displays this process for a SOI detected from the plasma derived from an individual who smoked. The retention time of interest is obtained from the tag (SOI = 68.5 min). It is known from SRM collider that detection of this tag is unique (as over 4 unadducted transitions were extracted) to the albumin peptide. The full MS1 mass spectrum is inspected at that particular retention time (**Figure 4.7A**). One can gain further specificity by only considering the MS 1 features in the m/z range from the isolation window that created the SOI (**Figure 4.7B**). In this particular example, the m/z range was not complicated and the precursor m/z that resulted in the SOI was readily determined to be 851.4266 and verified by investigating the elution profiles of the tag (SOI) and precursor (**Figure 4.7C**). It’s important to note the efficacy of this step is heavily dependent on sample complexity, isolation window width, and peak capacity. More robust tools and statistics are needed to provide a confidence in this match. Once the precursor m/z is determined, an exact mass of the adduct can then be calculated using **Equation 1**.

The mass of the modification was calculated to be 119.004 Da. Searching previous literature, it was discovered that this mass is a common albumin adduct in biological fluids in which a thiol bond is created between two cysteine residues – known as cysteinylation. The verification step adds this mass to the Cys\(^{34}\) residue and then overlays the modified fragment ions. Co-elution of multiple fragment ions extracted with the elucidated mass provides confirmation that the peptide is believed to be adducted with the mass calculated. It is
important to note that this final step will not be possible for some modifications and is
dependent on the stability of the modification upon collision induced dissociation.
Interestingly, increased human serum albumin Cys\textsuperscript{34} cysteinylation has been shown to be a
marker for oxidative stress related diseases such as diabetes mellitus and chronic liver and
kidney diseases.\textsuperscript{[109]} Through the developed data pipeline we were able to analyze and identify
the albumin peptide from smokers’ plasma modified with homocysteinylation, monoxidation,
dioxidation, and trioxidation.

**Figure 4.7:** A) An expanded base peak chromatogram (66 to 73 min) of the MS 1 filtered
scans. The signals of interest indicate the retention time to investigate the precursor scan in
efforts to match the SOI with its precursor \( m/z \). B) The full mass spectrum at 68.5 min is shown
which corresponded to the SOI. The full scan can be further filtered to only show the
precursors from the isolation window (inset) from which the SOI was generated (\( m/z 850-860 \)).
C) A comparison between the elution profile of the putative precursor and SOI can be used to
provide confidence in the match. The y axis was normalized to the maximum intensity of the
summed fragment ions for the SOI and the extracted monoisotopic mass in order access
similarity in elution profiles. The SOI was found to match the precursor 851.4261 in retention
time space which was due to cysteinylation.
We also used MAPP to investigate the peptide that contains a free cysteine (GTFATLSELHCDK) from the beta chain of hemoglobin (Hemβ Cys93) after isolation and enrichment from pooled smoker’s blood. A high abundant putative adduct with a modification of 170.1208 Da was detected, but could not readily be identified based on previous literature reports and databases (Figure 4.8). This example exemplifies the power of MAPP to detect peptides with unknown modifications. Future applications will utilize MAPP to investigate the adductome in relation to disease due to environmental causes. In addition, we will investigate the ability to provide an adduct fingerprint in efforts to classify exposures to complex substances (cigarette smoke, air pollution etc.).
Figure 4.8 A tag (b4-b10) was extracted from the isolation of the doubly charged hemoglobin peptide from the beta chain (GTFATLSELHCDK). The MS 1 precursor was matched to the tag in retention time space. The mass of the adduct was calculated and placed back on the cysteine (170.1028). The adducted ions were then extracted indicating that the mass was identified correctly. Upon searching the literature and database we were unable to identify this cysteine modification.

4.4 Ongoing/Future work

While the current data analysis pipeline for MAPP has many advantages over FS-SRM, there remains room for improvement. Currently the pipeline involves the usage of three different software programs: Xcalibur, Skyline, and Microsoft Excel. The data analysis is manual, time consuming, and provides no descriptive statistics for the confidence of an adduct and the corresponding transitions. All of these concerns are being addressed in current/future work. Figure 4.9 displays the website (mappmymod.com) we would like to develop to aid
researchers in data analysis using MAPP. The pipeline would be on a single platform and allow researchers to select parameters for peptides of interest.

Figure 4.9 The website mappymod.com which we are working to develop. Within the site one would upload the data, input the peptide sequence of interest, modified amino acid, and unadducted transitions to extract. The next steps would overlay the MS1 features, calculate and identify modification masses in Unimod, and export results to Panorama.

4.5 Conclusion

A novel method termed Multiplex Adduct Peptide Profiling (MAPP) provides a framework to investigate the global modifications of a single amino acid along a peptide backbone has been developed. The method offers several improvements to FS-SRM including
the use of high mass measurement accuracy and a simplified data pipeline for accepting a peptide adduct. While the initial application has focused on adductomics, it could be applied to other in modifications (e.g., histone modifications, protein glycosylation heterogeneity, etc) given that isolation and enrichment strategies are employed for the protein of interest. Enrichment is crucial as modifications occur at sub-stoichiometric levels. As with any method, certain limitations do exist. The pipeline is not automated which makes the analysis time consuming. The specificity of the tag may restrict the method’s use for small peptides in complex mixtures; although, the HSA Cys34 peptide tag specificity was evaluated using SRM collider to determine uniqueness of monitored transition mass combinations.\textsuperscript{[139]} In addition, complexity of the precursor isolation window which is dependent on the separation efficiency (i.e., peak capacity), sample complexity, and the width of the isolation window is important to consider as it will affect the efficacy of matching the correct MS 1 feature to the fragment ion tag of interest. More advanced DIA techniques that improve precursor specificity by isolating nonadjacent windows could be utilized.\textsuperscript{[100]} Future experiments will explore optimization of the method including statistics to match MS2 and MS1 signals as well as evaluation of the linearity of the Cys34 peptide when spiked with multiple reactive species.
CHAPTER 5
APPLICATION TO COMPLEX BIOLOGICAL FLUIDS

5.1 Introduction

A main objective in the environmental health community is the identification/discovery of biomarkers specific to the presence of toxic exposures or their effects. Biomarkers encompass almost any measurement reflecting an interaction between a biological system and a potential hazard, which may be chemical, physical, or biological. The measured response may be functional and physiological, biochemical at the cellular level, or a molecular interaction. Many chemicals become reactive after metabolism by the phase I enzymes and due to their inherent reactivity have short residence times in the body before exerting deleterious effects on biopolymers which can range from DNA adduction to loss of protein function (commonly referred to as protein damage). These reactive electrophiles can have residence times in blood or urine which vary from hours to days. Due to the short residence time of toxic electrophiles, the measurement of protein adducts derived from modifications of these reactive compounds is an attractive alternative.

Both hemoglobin and albumin have an extended residence times in blood of 60 and 28 days respectively. As a result, the covalent modification of these proteins by reactive species presents a promising avenue to obtain a mean dose level of these chemicals over time.

Both hemoglobin and albumin are abundant proteins (albumin constitutes over 50% protein content in plasma at 3.5-5 g/dL and hemoglobin is present in blood at concentrations 15g/dL) found in blood which contain a free thiol in their native conformations (Hb-Cysβ, HSA-Cys34) and these thiols readily react with toxic electrophilic species forming adducted
species.\textsuperscript{[41, 58, 146-148]} Of the nucleophilic free thiols in plasma, up to 80% can be attributed to HSA-Cys\textsuperscript{34}.\textsuperscript{[149, 150]} HSA-Cys\textsuperscript{34} has been shown to adduct aldehydes, metal ions (Hg\textsuperscript{+}, Ag\textsuperscript{+}), quinones, and reactive oxygen species.\textsuperscript{[151, 152]} While adducts to HSA-Cys\textsuperscript{34} have been detected and reported, human plasma is a complex biological matrix and adducted protein levels are sub-stoichiometric. The low levels of protein adducts present in biological fluids and matrices necessitates enrichment strategies. One recently proposed enrichment strategy for depleting unmodified HSA (enriching for adducted albumin) uses a thiol-affinity resin.\textsuperscript{[50]} The unmodified albumin thiols bind to the resin while adducted species are not retained by the resin. The utilization of enrichment strategies to detect protein adducts is imperative to enable sensitive detection.

We chose to apply our Multiplex Adduct Peptide Profiling method to a complex environmental exposure (i.e., cigarette smoke) in order to assess its potential to detect multiple adducts in a single analysis. Cigarettes represent a harmful, complex matrix containing over 4000 identified chemicals including carcinogens and reactive electrophiles such as: benzene derivatives, poly-aromatics, and aldehydes.\textsuperscript{[153]} Tobacco usage is the leading cause of preventable deaths in America leading to an estimated 480,000 premature deaths each year.\textsuperscript{[154]} According to the Center for Disease Control, 18% of the adult population in the United States smokes.\textsuperscript{[154]} Over 90% of lung cancers in the United States are linked to smoking.\textsuperscript{[154]} To study the effect of each compound individually on HSA would grossly oversimplify the human exposure experienced with every puff of smoke. Certain compounds could preferentially bind the HSA-Cys\textsuperscript{34} site based upon their affinity for the nucleophilic thiol. In consideration of these observations regarding cigarette smoke, it would be of great interest to apply a method
developed for the discovery and detection of protein adducts which could compare plasma samples from individuals which are smokers and non-smokers.

Multiplex Adduct Peptide Profiling (MAPP) has been demonstrated as a method which can aid in the discovery of protein adducts. The method has been described previously in Chapters 1-2. Herein, we have applied MAPP to investigate differences in HSA-Cys$^{34}$ adduct distribution and abundance between pooled smoker and nonsmoker plasma.

5.2 Materials and Methods

Materials

Formic acid (FA), ammonium bicarbonate (AB), hydrochloric acid (HCl), iodoacetamide (IAM), dithiothreitol (DTT), sodium deoxycholate (SDC) were obtained from Sigma Aldrich (St. Louis, MO). Trypsin was obtained from Promega (Madison, WI). HPLC grade acetonitrile, methanol and water were from Burdick & Jackson (Muskegon, MI). The hemoglobin peptide (GTFATLSELHCDK) was synthesized and purchased from Thermo Fisher Scientific GmbH (Ulm, Germany). Pooled plasma obtained from smokers was purchased from BioreclamationIVT (New York, New York).

NanoLC MS/MS

Two µL of sample were injected onto a self-made 4 cm trap using an Easy nanoLC 1000 coupled to a Q Exactive Plus mass spectrometer (Bremen Germany). PicoFrit columns from New Objective (Woburn, MA) were packed to 25 cm in house with 3 µm C18 silica particles (Dr. Maisch, Entringen, Germany). Mobile phase B was 99.9 % acetonitrile with 0.1% formic acid and mobile phase A was 98 % water, 2 % acetonitrile, and 0.1 % formic acid.
A 90 minute LC MS/MS method was used and consisted of a linear gradient from 0-40 % B over 70 minutes followed by a ramp to 80 % B in one minute. The column was washed at 80% B for 9 minutes and regenerated at 0 % B for 10 minutes.

Albumin Isolation/Enrichment

Albumin isolation and adduct enrichment were performed based upon the revised protocol optimized in Chapter 3. In brief two technical replicates for plasma from a male smoker and a male non-smoker, ammonium sulfate was added to 1 mL of plasma to create a 60% saturated solution. Albumin was acid precipitated using HCl from the supernatant. Isolated albumin (500 µg) was added to 75 mg of dry Activated Thiol Sepharose in a spin column (Thermo product #69725) and the flow through fraction (nonmercaptoalbumin) was collected after 2 hours of enrichment. The enriched albumin fraction was then prepared using the filter-aided sample preparation procedure for shotgun proteomics[105] using a 4 hour tryptic digestion at an enzyme to substrate ratio of 1:50.

DIA Methods were created in Skyline.[106] The 21 amino acid tryptic peptide from human serum albumin that contained the free Cys34 residue was entered into skyline (K.ALVLIAFAQYLQQC PfEDHVKL). An isolation list was created in Skyline that used 12 10 m/z isolation windows which spanned m/z 810 to 930 but still contained the triply charged unmodified peptide precursor (m/z 811.7594). This range of precursor detection included the mass of the albumin free cysteine unadducted peptide to a mass addition (i.e., adduct) of 360 Daltons which encompasses the masses of all reported protein adducts aside from Satratoxin G[66]. The sequence of samples run was block randomized.
The isolation list was imported into the method editor of Q Exactive Plus and an MS1 scan was added from \( m/z \) 550 to 1300 for a total 13 scans per cycle with an average duty cycle of 1.3 seconds. For the MS1 scan the resolving power was set to 70,000, an AGC of 1e6, and a max injection time of 50 ms were used. For the DIA scans, parameters were set to the given levels: AGC of 1e6, 24 NCE (normalized collision energy), max injection time of 50 ms, and a resolving power of 17,500.

Data Analysis

Peptides with fake modifications were created in Skyline by adding a mass to the cysteine residue of the Cys\(^{34}\) tryptic peptide such that a single precursor \( m/z \) occupied each isolation window. Putative signals of interest (SOI) were identified by extracting at least four common fragment ions from the Cys\(^{34}\) peptide and finding the instances in retention time space where all co-eluted with high mass measurement accuracy (< 5 ppm). In addition, the proportion of relative abundances of each fragment to total abundance was used as a method to eliminate false positives. Once an SOI was identified, precursors were filtered based on the isolation window from which the SOI originated and matched in retention time space with the SOI using Xcalibur 2.2. This matching was based on the principle that since the MS1 precursor and associated fragment ions all belong to the same chemical species then the chromatographic profile of both should be theoretically identical. The distribution of adducts detected and the peak area for detected peptide modifications was compared for smoker plasma versus non-smoker plasma in Excel.
5.3 Results and Discussion

Figure 5.1 describes the experimental workflow used in this study. Two technical replicates were prepared for both the smoker’s and the non-smoker’s plasma. Adducted albumin was isolated and enriched based upon the optimizations described in Chapter 3. Filter Aided Sample Preparation (FASP) was used for sample cleanup and tryptic digestion. Samples were then analyzed by MAPP.\cite{155}

![Workflow Diagram](image)

Figure 5.1 shows the experimental workflow used in comparison of smoker’s and non-smoker’s plasma. Isolation, enrichment, digestion, and Multiplex Adduct Peptide Profiling were performed as described in the Methods.

Using Skyline, peptides with fake modifications were created in each isolation window. Unadducted fragment ions of the HSA-Cys\textsuperscript{34} peptide were extracted. Within the various isolation windows several signals of interest were observed. However, only signals of interest
containing at least 4 unadducted fragment ions from the HSA-Cys\textsuperscript{34} peptide were used for further processing. The distribution of adducts detected in smoker’s and non-smoker’s plasma is illustrated in Figure 5.2. Interestingly, the distribution of adducts was the same for both conditions (11 adducts detected in the same isolation windows). Pollutant concentrations in blood range from µM-fM, whereas food, drugs, and endogenous species are present in blood at concentrations of mM-pM.\cite{51} Thus, a possible explanation for the similar adducts detected could be due to adducts from food, endogenous species, and drugs. Other reasons for the resemblance between the two distributions could be linked to a low number of cigarettes smoked/day, drugs or medications, or occupational exposures which contain smoke/pollution (refinery, welding), etc..\cite{156} After analyzing the distributions of adducts detected, the correlation among fragment peak area was investigated.

**Figure 5.2** The distributions of HSA-Cys\textsuperscript{34} adducts detected in non-smoker’s (blue) and smoker’s (orange) plasma. Two technical replicates were used and each adduct contained a minimum of 4 unadducted fragment ions from the HSA-Cys\textsuperscript{34} peptide. Similar adducts were detected from both samples.

The peak area of unadducted peptide fragments were exported from Skyline to Excel. The peak areas for each signal of interest were averaged among the technical replicates for
both smoker’s and non-smoker’s plasma. **Figure 5.3** describes the correlation observed between the average total unadducted fragment peak areas (plotted by smoker’s vs non-smoker’s plasma) of the observed adducts. The calculated $R^2$ value (0.92) indicates high correlation among the total unadducted fragment peak areas for these two individuals. The high correlation observed could again be due to similar exposures of these two individuals. Further analysis was performed exploring which adduct levels detected differed most between the smoker and non-smoker.

![Correlation Plot](image)

**Figure 5.3** The correlation plot of the average total MS2 peak area from unadducted fragments detected in smoker’s and non-smoker’s plasma. The Pearson correlation coefficient $R$ value was higher than 0.95 indicating high correlation.

**Figure 4A-B** illustrates the individual comparison between the smoker and non-smoker of total fragment peak areas for adducted HSA-Cys$^{34}$ according to the isolation window which detected the modification. **Figure 5.4A** shows the modifications to HSA-Cys$^{34}$ which were detected at higher levels in the smoker. Six adducts to HSA-Cys$^{34}$ were detected at levels greater than 30% in the smoker. Among the modifications were dioxidation and trioxidation,
which were previously identified using MAPP. These oxygen adducts in the smoker were observed at levels which were 32% (dioxidation) and 53% (trioxidation) higher than samples derived from the non-smoker. Oxygen modifications to cysteine in proteins have been associated with oxidative stress.\textsuperscript{[157, 158]} \textbf{Figure 5.4B} shows the modifications to HSA-Cys\textsuperscript{34} which were detected at similar levels and higher levels in the non-smoker.
Figure 4 The comparison of average MS2 peak area of adducts to the HSA-Cys$^{34}$ peptide detected in smoker’s and non-smoker’s plasma. Dioxidation and trioxidation were previously identified using MAPP and were higher in the smoker. Other putative adducts are labeled on the x-axis according to the isolation window from which they originated (e.g. 810 = 810-820) and the number of the adduct detected in the particular isolation window (as two were detected in the window 870-880 these are labeled 870_1 and 870_2). Future work will identify the masses of the adducts through the data analysis pipeline discussed in Chapter 4. **A)** Displays HSA-Cys$^{34}$ adducts with higher peak areas detected in the smoker. **B)** Displays HSA-Cys$^{34}$ adducts where similar peak areas were observed in both as well as adducts which were detected with higher peak areas in the non-smoker.
5.4 Conclusion

The first application of Multiplex Adduct Peptide Profiling (MAPP) to discover and detect differences in protein adducts to HSA-Cys$^{34}$ between plasma samples from a smoker and non-smoker has been reported. The peak areas for six of the eleven modifications were elevated over 30% in the samples derived from the smoker relative to the non-smoker. Among the higher adduct levels in the smoker were dioxidation and trioxidation which have been linked to oxidative stress which is known to occur in smokers.$^{[157]}$ However, the distribution of adducts and the correlation between the detected fragment peak areas were similar (the same number of adducts detected in both samples). This observation could potentially be attributed to identification of high abundant food, drug, or endogenously produced adducts or a lack of biological and technical replication. Future work will be to identify the putative adducts detected, increase the number of individuals monitored and the technical replication in order to establish thresholds for adduct levels in smokers and non-smokers, and apply the method to the Hem-Cys$^{93β}$ peptide to elucidate differences in adduct levels of smokers relative to non-smokers.
REFERENCES


[133] Nat Meth 2013, 10, 1.


In this study, we compared the performance of sodium deoxycholate (SDC) with several commercially available LC MS/MS compatible detergents for digestion of complex proteomic mixtures. First, the parameters affecting in-solution digestion using SDC were investigated with a full factorial experimental design. Metrics explored included trypsin ratio, digestion time, and concentration of SDC. These parameters were not found to be statistically associated with total peptide identifications in the experimental space investigated. However, in terms of digestion efficiency, time was highly significant (p = 0.0095) as determined by the percent of peptides identified with missed cleavages. The optimized protocol for peptide identification and throughput was used to compare the performance of SDC with various commercially available LC MS/MS compatible surfactants namely Invitrosol, RapiGest, and PPS Silent Surfactant. The detergents were found to be similar through comparisons of the total identified peptides and the hydrophobicity of recovered peptides. We found suitable recovery across a large range of SDC concentrations determined from a bicinchoninic acid (BCA) assay. In a spike down experiment, no distinct differences in total number of peptide identifications were discovered when comparing PPS (Silent Surfactant) and SDC for preparation of peptide samples derived from low protein amounts (< 20 µg). Combined, these results indicate that SDC is a cost effective alternative to other commonly used LC MS/MS compatible surfactants.

**Keywords:** Detergent; Sodium Deoxycholate; Shotgun Proteomics; Mass Spectrometry; Sample Preparation
A.1 Introduction

A continued goal in both discovery and targeted bottom-up proteomics is developing sample preparation procedures that maximize cost effectiveness, reproducibility, and throughput without sacrificing performance \[^{159-162}\]. The success of any shotgun proteomics experiment is dependent on the robustness (i.e., reproducibility, recovery) of sample preparation from protein solubilization, stabilization, digestion, and sample clean-up. Different procedures exist for preparation of complex mixtures in shotgun/bottom-up proteomics. Two main methods involve in-solution digestion and filter aided digestion \[^{105, 163}\] both of which are being continually improved upon \[^{164}\]. Sample preparation in proteomics experiments via LCMS/MS is an area where exact procedures, including reagents, vary widely amongst laboratories and sample type. Protein identification is initially predicated on the degree of protein solubilization and thus detergent selection is an important component in sample preparation procedures. Sodium dodecyl sulfate (SDS) is the “gold standard” detergent used ubiquitously in biological research. However, SDS is incompatible with LC MS/MS and requires specific procedures for near complete removal \[^{105, 164}\].

Acid cleavable commercial LC MS/MS detergents have been developed as an alternative to SDS for sample preparation in proteomics \[^{165}\]. They fill a niche of increasing the solubility of hydrophobic/membrane proteins and peptides without the need for specific steps for removal. Acid cleavable detergents can be easily removed prior to LC MS/MS analysis by lowering the pH of solution—a necessary step regardless following trypsin digestion. Early work with acid cleavable detergents in proteomics was done comparing acid labile surfactant performance to SDS in 2002 by Meng et al.\[^{166}\]. Other studies tested the use
multiple detergents simultaneously and differential detergent fractionation for enhancement of protein digestion and peptide recovery [167-171]. Yates and coworkers compared the peptide recovery of tryptic digestions with Invitrosol, RapiGest, and PPS Silent Surfactant. Their study showed MS compatible detergents increase the number of unique peptides and proteins observed in mammalian tissues [172, 173]. Waas et al. expanded studies to include Progenta in their testing and found increases in peptide identifications with the patented detergents [174]. However, these commercial detergents are relatively expensive which can limit resources especially for large biomarker and system biology studies.

Deoxycholate is an anionic water soluble acid produced in the liver and poses a less expensive alternative to commercially available LC MS/MS detergents. Acidification of solutions containing sodium deoxycholate causes the formation of a white deoxycholic acid precipitate. This enables easy and rapid removal of the detergent prior to LC MS/MS. One of the early reported proteomic workflows using SDC was by Zhou et al. where they concluded SDC enhanced the solubility of peptides and proteins. They observed an increase in peptide and membrane protein identifications using SDC compared to SDS [175]. A study comparing chaotropic agents and surfactants by Proc et al. used urea, SDS, and SDC. Their study verified that SDC had higher digestion efficiency, peptide recovery, and did not foul the instrument [176]. Further work by Yong et al. assessed trypsin activity at various concentrations of SDC to ascertain optimal conditions for digestion compared to urea, methanol, and SDS. They found trypsin activity is compromised significantly at concentrations higher than 5% SDC using a standard five protein mix [177]. Recent work has been reported comparing SDC to other detergents in eFASP [178] in addition to optimizing peptide recovery from SDC precipitates.
using the phase transfer or acid wash mechanisms[179]. However, it is still unknown the degree at which common parameters affect in-solution digestion of complex mixtures using SDC and how the peptide recovery of SDC compares to other commercially available LC MS/MS compatible detergents.

Herein, we investigated various conditions (e.g., digestion length, trypsin ratio, SDC concentration) that affect in-solution digestion using SDC with a full factorial experimental design. We then compared the performance of SDC to commercially available mass spectrometry compatible detergents. Based upon our findings and comparisons using these metrics, SDC is a cost effective alternative to LC MS/MS commercial detergents even at low protein starting amounts < 20 µg.

A.2 Materials & Methods

Materials

Formic acid (FA), ammonium bicarbonate, ammonium hydroxide, dithiothreitol (DTT), SDC, hydrochloric acid (HCl), bovine serum albumin (BSA), and iodoacetamide, were obtained from Sigma Aldrich (St. Louis, MO). Proteomics grade trypsin was purchased from Promega (Madison, WI). Invitrosol was purchased from Invitrogen (Carlsbad, CA), RapiGest from Waters (Milford, MA), and PPS Silent Surfactant from Expedeon (San Diego, CA). HPLC grade acetonitrile, methanol, and water were purchased from Burdick & Jackson (Muskegon, MI). The Micro BCA Protein Assay Kit (Part # 23235) was purchased from Thermo Fisher (San Jose, CA).
Methods

Sample preparation– Design of Experiment with SDC

The 293 kidney cell line was grown and harvested following the procedure described in Supplemental under Cell Culture. The cells were split into 50 µL aliquots each containing ~10^6 cells. Samples were lysed with an ultra-sonicator model #CL-34 (Thermo Fisher) in three 20 second bursts at 20% power in 50 mM AB with 1% or 3% SDC. Protein concentrations were determined via absorbance at 280 nm on a Nanodrop 2000 spectrophotometer (Thermo Fisher). Aliquots of 100 µg of protein were taken and were diluted to 100 µL with 50 mM AB with 1% or 3% SDC. DTT was added to final concentration of 5 mM and the samples were incubated for 30 minutes at 60 °C. Samples were alkylated with iodoacetamide added to achieve a final concentration of 15 mM and samples were incubated at room temperature for 20 minutes in the dark. Trypsin was added to the samples in a ratio of 1 µg trypsin: 100 µg total protein or 1 µg trypsin: 50 µg total protein. Samples were incubated for 4 or 12 hours at 37 °C. To quench tryptic digestion and remove SDC, HCl was added to the samples to achieve a final concentration of 250 mM. After precipitation of deoxycholic acid samples were centrifuged at 14 x g for two minutes. The supernatant was collected for further processing. To improve peptide recovery deoxycholic acid precipitates were acid washed [175]. The acid wash was performed with 100 µL of 250 mM HCl. The samples were briefly vortexed and were centrifuged at 14 x g for two minutes. The acid wash supernatant was also collected for further processing. Solid phase extraction (SPE) columns were used to remove neutrals and salts from the samples. MCX SPE columns (30mg LP extraction cartridges part# 186000782) from Oasis (Milford, MA) were mounted onto a MCX
Columns were conditioned with one mL of methanol, one mL of 10% NH₄OH in water, two mL of methanol, and three mL of 0.1% FA in water. Samples were added to SPE columns and washed with 0.1% FA to remove salts. Neutral species were removed with one mL 0.1% FA in methanol. Peptides were eluted and collected in one mL 10% NH₄OH in methanol. Samples were lyophilized with a Savant SPD313DDA speed vacuum (Thermo Fisher). Peptides were reconstituted with mobile phase A (98 % water, 2 % acetonitrile, and 0.1 % formic acid) to achieve a final concentration of 0.25 µg/µL of digested peptide in solution.

Sample Preparation – Detergent Comparison

For the detergent analysis, samples were prepared in PPS, RapiGest, Invitrosol, SDC, and a control containing no detergent. Three technical replicates were taken through the procedure for each detergent tested (including the control). For samples containing RapiGest, Invitrosol, and PPS the manufacturers’ suggested concentrations were used (0.1% w/v, diluted 1:5, and 0.1% w/v), whereas for samples containing SDC 1% w/v was the concentration used. The samples were digested for 4 hours at a ratio of 1 µg trypsin:50 µg total protein. After tryptic digestion and acidification, the samples containing commercial detergents were allowed to incubate for 45 minutes at room temperature.

SDC Peptide Recovery

SDC peptide recovery was tested at various concentrations using a BCA assay. Six aliquots of ~ 25 µg of peptide material were taken from digested kidney cell lysate. Aliquots were diluted to 60 µL of total volume with varied concentration of SDC spiked into each aliquot. Mixtures contained (final concentration) 0, 0.5, 1, 2, 3, and 5 % SDC. These solutions
were acidified with HCl added to a final concentration of 250 mM. The solutions were centrifuged at 14 x g for two minutes and the supernatant was collected. The pellets were acid washed as previously described. However, the volume of 250 mM HCl used in the wash was 40 µL instead of 100 µL. Peptide recovery was evaluated using the manufacturers’ protocol for the BCA assay and absorbance was measured at 562 nm on a BioMate 3S (Thermo Fisher).

*Detergent Comparison with Low Starting Material*

For the detergent comparison using low starting amounts of material, protein concentrations were determined via absorbance at 280 nm. Aliquots were taken of 20, 10, 5, and 1 µg of total protein and aliquots were diluted to 50 µL with 50 mM AB with detergent (1% SDC or 0.1% w/v for PPS). The samples were taken through the same reduction, alkylation, digestion, and clean up steps as used in the detergent study. After lyophilization, samples were reconstituted with appropriate volumes such that each sample was in theory at 0.2µg/µL of digested peptide in mobile phase A. This allowed the same amount of protein to be injected on column (assuming 100% peptide recovery) for each sample.

*Quality Control*

To ensure optimal data quality, every fifth injection a simple BSA digest (i.e., quality control standard) was analyzed every fifth injection. The QC runs were collected and imported into Skyline (UW Seattle, WA) to check retention times, peak area ratio for precursors and products, and mass accuracy using Statistical Process Control in Proteomics \[180\].
NanoLC-MS/MS

A sample volume of two µL was injected and analyzed with an Easy nanoLC 1000 coupled to a Q Exactive Plus mass spectrometer (Thermo Fisher Bremen, Germany). PicoFrit columns (New Objective Woburn, MA) were packed with 3 µm C18 silica particles (Dr. Maisch, Germany) to a length of 25 cm in house. A four cm trap was packed in house and was placed prior to the column. The composition of mobile phases were 99.9% acetonitrile with 0.1% FA for mobile phase B and 98% water, 2% acetonitrile, and 0.1% FA for mobile phase A. LC MS/MS methods consisted of either a 60 minute (SDC DOE) or a 180 minute (detergent comparisons) linear gradient from 0-40% B followed by a ramp to 80% B in one minute and a column wash at 80% B for ten minutes. The column was regenerated at 0% B for ten minutes. A data dependent acquisition method was used for both gradients in which a full mass spectrum was acquired from m/z 400 to 1400 at an automatic gain control (AGC) of 1*e6. Following each full scan, twelve data dependent scans were acquired at an AGC of 5*e4 in which the top twelve most abundant precursors were selected for isolation and fragmentation. Dynamic exclusion was set for thirty seconds to limit interrogation of abundant peptide species. Instrument peptide match was set to preferred.

Peptide Identification

The database searches were conducted using Proteome Discoverer (PD) version 1.4 (Thermo Fisher) and the Sequest hyper-threaded algorithm. Data were searched against the Homo sapiens Swiss Prot protein database (number of sequences: 26,148 date accessed: 10/17/2013). The peptide spectrum matches were post processed using percolator to enforce a peptide spectral match threshold with a q value < 0.01 [181]. The following modifications
were included: carboxamidomethylation of cysteine as a static modification, and dynamic modifications including methionine oxidation, and de-amidation of glutamine and asparagine. The SDC DOE samples were searched allowing identification of peptides containing up to two missed cleavages and a second time allowing identification of peptides containing zero missed cleavages. The number of peptides, number of missed cleavages, peptide sequences, and precursor charge of identified peptides were obtained from PD output files.

Statistical Analysis

For the SDC DOE, data containing the number of peptides were analyzed in JMP Pro 10.0.0 from SAS (Cary, NC). JMP was used to design a two-level three factor full factorial which enabled interactions to be determined. The trypsin ratio (1:100, 1:50), the digestion time (4, 12 hours), and the concentration of SDC (1, 3%) were tested at the given levels. The DOE study sequence was randomized and the detergent study was block randomized. In a spike down study samples were analyzed from the lowest amount of starting material to highest amount of starting material to minimize any biases associated with peptide carry-over. For the detergent comparison, one-way analysis of variance (ANOVA) and least significant difference (LSD) were used to assess significance of parameters at $\alpha = 0.05$.

A.3 Results and Discussion

Figure A.1 describes the experimental workflows used for this study. First, designed experiments explored the experimental space of the parameters that affect in-solution digestion with SDC (Figure A.1A). Next, the optimized total peptide identifications protocol for the digestion with SDC was compared to three commercially available LC MS/MS compatible
surfactants (Figure A.1B). A BCA assay was used to evaluate SDC peptide recovery for samples prepared from low starting protein amounts (Figure A.1C).

Figure A.1: The experimental workflow for the: A) SDC DOE, B) Detergent comparison, and C) SDC peptide recovery using a BCA assay.

Experiments were statistically designed to investigate and optimize the parameters that affect in-solution digestion with SDC. Table A.1 summarizes the experimental parameters and levels tested. The parameters tested in the full factorial DOE included trypsin ratio, detergent concentration, and digestion time. The combination of these factors could affect peptide recovery and digestion efficiency, thus affecting the total number of peptides identified by LC MS/MS. The parameter levels chosen sought to cover a range of common trypsin to
protein ratios and digestion times utilized in typical bottom-up proteomic experiments. The primary goal for the SDC concentration optimization was to find an optimal range where tryptic activity was not compromised and to enable maximum peptide recovery.

Table A.1: The parameters, levels, and the motivations for each factor utilized for the SDC DOE optimization.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Low</th>
<th>High</th>
<th>Primary Motivation</th>
<th>Motivation of Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDC Concentration</td>
<td>1 %</td>
<td>3 %</td>
<td>Peptide recovery at different SDC concentrations can vary</td>
<td>Does high SDC concentration yield higher or lower trypsin activity?</td>
</tr>
<tr>
<td>Trypsin Ratio</td>
<td>1 : 100</td>
<td>1 : 50</td>
<td>More or less trypsin can impact digestion efficiency</td>
<td>Does longer trypsin digestion allow for lower amount of enzyme? Possible reduction of cost/sample by lowering enzyme concentration</td>
</tr>
<tr>
<td>Digestion Time</td>
<td>4 hr</td>
<td>12 hr</td>
<td>Digestion times can affect digestion efficiency</td>
<td>Does optimal SDC concentration vary with time?</td>
</tr>
</tbody>
</table>

Figure A.2 illustrates half normal plots which describe the impact of a parameter on response (i.e., total peptide identifications). In half normal plots, the vertical axis (absolute contrast) represents the difference in the response of the means at and low levels for the various parameters. The horizontal axis shows the corresponding normal quantile for each factor and interaction coordinate. Parameters which are on or near the diagonal line represent effects that would be observed based on a normal distribution and are not significant. As the distance from the line increases, the significance of the parameter does as well. Figure A.2A-B summarizes the results obtained from the SDC DOE experiment. The raw data is found in Supplemental Table 1. According to the DOE analysis, none of the parameters were significant. There were trends of higher trypsin ratio (14% increased peptide identifications), shorter digestion time
(18% increased peptide identifications), and lower percentage of SDC (9.4% increased peptide identifications), towards augmented numbers of peptide identifications. The percent increase was calculated from the difference in average peptide identifications for each condition (e.g. 4 hour vs 12 hour digestions). It is possible the study was conducted near the optimum for this experimental space, thus the parameter levels did not yield a significant difference relative to each other. However, these results do indicate that SDC is a robust detergent and operates well over a broad range of preparation conditions. Additionally, the percent of peptides identified containing missed cleavages were analyzed as a metric for assessing digestion efficiency. The only factor found to be statistically significant was digestion time (p =0.0095) (Figure A.2C-D). Longer digestion times with SDC showed higher digestion efficiency (i.e. lower percent of peptides identified containing missed cleavages). Longer digestion time was also implicated with a lower number peptide identifications – although this did not reach significance (p =0.18). One potential explanation for this observation is that the identification of more peptides in samples digested during the shorter time period correlates to the identification of a higher number of peptides containing missed cleavages. To investigate whether the total number of peptides identified containing missed cleavages was a function of the difference in the total number of peptide identifications as related to digestion time, the files were searched with zero missed cleavages. The results showed the higher number of peptide identifications still correlated to shorter digestion times. However, this observation did not reach significance (p= 0.20). As a result, increased peptide identifications with shorter digestion times were not simply due to a larger number of peptide identifications with missed cleavages.
Interestingly, these results are suggestive of the need for different digestion procedures dependent on the experimental goals. If one is performing a global discovery experiment, one would want to maximize peptide identifications and throughput by choosing a shorter digestion time. If one is performing absolute quantification of targeted peptides using protein cleavage isotope dilution mass spectrometry \cite{182} one would want to minimize targeted peptides containing missed cleavages. The reduction of peptides containing missed cleavages could be accomplished by selecting a longer digestion time. Recent work by Chiva et al. corroborates these results, indicating that different techniques and digestion enzymes can be customized to specific studies and that a one size fits all digestion approach for both discovery and targeted proteomic studies is not ideal \cite{183}.
Figure A.2: A) Results from the DOE factorial study for maximization of the number of peptides identified. B) Half normal plot for peptides identified. Half normal plots describe the impact of each parameter and interaction tested on response. The vertical axis (absolute contrast) represents the difference in response of the mean of the high level against the mean of the low level for a parameter. The horizontal axis shows the corresponding normal quantile for each factor and interaction coordinate. C) Results for minimization of the percent of peptides identified containing missed cleavages. D) Half normal plot for minimization of the percent of peptides identified containing missed cleavages.

Using this optimized procedure for global peptide identifications and throughput, we compared SDC with various commercially available LC MS/MS compatible surfactants. Three technical replicates were prepared for a control (i.e. no detergent was used) and for each detergent tested. It was important to have the lysis step executed multiple times using the same detergent, as we wanted to compare the variability in the whole procedure. The total number
of peptide identifications obtained from the output files was compared amongst the detergents as illustrated in **Figure A.3A**. One-way analysis of variance (ANOVA) followed by least significant difference (LSD) testing were used to determine significant methods \(^{[184]}\). The ANOVA results showed significant difference among the means for the different experimental conditions. As one would expect, there was a significant increase in the number of peptides identified due to the presence of a detergent (\(p < 0.025\)). The number of peptides identified was not significantly different across the detergents. However, the percentage of identifications of peptides with missed cleavages was significantly higher (\(p < 0.05\)) in the control samples and samples digested with RapiGest. Similar trends to those observed for the number of total peptide identifications were reflected in the total number of protein groups identified and the total number of PSMs as shown in **Supplemental Table 2**.
Figure A.3: A) The average number of identified peptides based upon detergent type. The ANOVA and LSD calculations show that the control was significantly lower for number of peptides detected (p=0.024). B-C) Percentage of charge state ions for each detergent was compared and RapiGest was significantly lower (p= 0.005) in the percentage of 2+ charged peptides identified, but significantly higher in the 4+ charged peptides (p=7.7e-7). D) The percentage of missed cleavages is plotted against detergent condition. The control and RapiGest have significantly higher percentage of missed cleavages (p <0.05). Error bars represent standard error in each sample.

Additionally the charge states of peptides identified were analyzed to compare the digestion efficiency of each detergent. A sample with optimal digestion efficiency would show a higher percentage of doubly and triply charged peptides. However, a less efficient digestion would result in the identification of peptides in higher charged states and less doubly and triply charged peptides. We examined the charge state distribution of the total identified peptides from each detergent. In the samples digested with RapiGest, the proportion of doubly charged
peptides identified was significantly lower (p= 0.0055) than the others and significantly higher for the 4+ (p=7.73e-7) as shown in Figure A.3 B-C which may indicate less efficient digestion.

Additionally, we compared the distribution of gravy scores (Grand Average Hydropathicity Index) for the total number of peptides identified for each detergent tested\[185\]. Gravy scores were calculated using S. Fuch’s website\[186\]. The gravy scores of each replicate were combined yielding five distributions containing between 40,000-55,000 gravy scores. The raw gravy scores were formatted as boxplots showing the distributions and median gravy scores as displayed in Supplemental Figure 1. The distributions of gravy scores for all samples were similar; a bias was observed toward peptides that were slightly hydrophilic as previously reported \[164\] for digestion of complex mixtures. Larger numbers of hydrophobic peptides were identified in samples prepared with detergent compared to the control. No distinct differences were observed for hydrophobic peptides identified using these three detergents.

Increases in instrument sensitivity combined with targeted biological questions have led to a significant trend in proteomics towards the preparation of proteomic samples with low amounts of starting material (e.g., Co-IP, LCM) \[187-192\]. A concern with using SDC for in-solution digestion with low protein starting amounts is the possibility of peptides being lost in the precipitation which are not recovered in the acid wash or ethyl acetate transfer \[175\]. In the previous experiments, high amounts (100 µg) of starting material were used. The goal of this experiment was to assess the peptide recovery of SDC from low amounts of starting material in order to identify whether peptides were lost due to precipitation of deoxycholic acid. We tested the performance of SDC in recovery of peptides spanning an order of magnitude (0.5-5
Suitable recovery of peptides was found across a wide range of SDC concentrations. Interestingly, it was observed that peptide recovery was lowest in the absence of SDC as shown in Figure A.4A. This observation may be due to peptides which had greater affinity for the plastics used during sample manipulation (e.g., pipet tips, Eppendorf tubes) in the absence of a detergent (i.e., carrier).

**Figure A.4: A** Results from the bicinchonic acid (BCA) assay performed on a complex lysate after precipitation of SDC. Results show peptide recovery is greater with SDC present and increases slightly with lower concentrations of SDC. However, without SDC present peptide recovery is lowest. **B** Performance of PPS and SDC at recovering peptides with low levels of starting protein material ranging from 1 µg-20 µg. No peptides for either detergent were identified from starting materials of less than 1 µg of protein.
To verify the performance of SDC peptide recovery with low amounts of starting material further experimentation was performed which compared PPS and SDC. PPS is an acid cleavable detergent that does not precipitate from solution which makes it suitable for comparing peptide recovery that utilizes samples with low amounts of starting material (i.e. 1-20 µg protein). In Figure A.4B total number of identified peptides was plotted against starting amount of protein in samples. The data illustrate the observed number of peptide identifications using PPS were similar compared to SDC at every protein amount level tested. This observation indicates that there was efficient peptide recovery from the SDC pellet. It is interesting to note that 1 µg of protein was deemed insufficient to recover peptides with this particular procedure. The comparable performance of SDC with PPS demonstrates that SDC is a robust detergent for preparation of peptide samples derived from high and low protein starting amounts.

Conclusions

We have reported a study comparing peptide identifications, peptide characteristics, and digestion efficiency amongst SDC and other commonly used commercial LC MS/MS compatible detergents. DOE was used to map the experimental space of parameters affecting in-solution digestion with SDC through a full factorial approach. The peptide recovery was comparable in the head-to-head comparison of detergents to SDC. Further, two experiments which utilized independent detection techniques, LC MS/MS and absorbance with a BCA assay, were used to assess SDC peptide recovery. In conclusion, SDC is a cost efficient LC
MS/MS compatible reagent and performs similarly to other more expensive detergents with low and high amounts of protein starting material.

Acknowledgements

The authors acknowledge support from the Pilot Project Award from the Center for Human Health and the Environment (850 Main Campus Dr. Raleigh, NC 27606). We also appreciate the help of Drs. Jonathan Hall and Rob Smart with preparation of the HEK 293 cell line.

Abbreviations

SDC, sodium deoxycholate; BCA, bicinchoninic acid; SDS, sodium dodecyl sulfate; FA, formic acid; DTT, dithiothreitol; HCl, hydrochloric acid; SPE, solid phase extraction; DOE, design of experiment; AGC, automatic gain control; PSM, peptide spectral match; ANOVA, analysis of variance; LSD, least significant difference

Supplemental

Cell Culture

Cells provided from North Carolina State University were frozen in a pellet. Cell culture was performed utilizing a HEPA lab Conco purifier class 2 biosafety cabinet delta series. Cells were thawed and plated in a T-75 flask from Fisher Scientific with DMEM from Invitrogen Life Technology (Carlsbad, CA). 10 % fetal bovine serum from Life Technology was added to flasks and cells were incubated in Napco Model 6200 incubator (Thermo) at 37 °C and 3 % carbon dioxide. After three days cells were split, passaged, and monitored daily
to check confluence microscopically with an Olympus CK2 microscope (Lake Success, NY). Once cells were ~ 90% confluent they were passaged by aspirating old media and washing with 10 mL PBS solution (Life Technology) aspirating PBS solution, and trypsinizing cells with 1 mL of trypsin (Life Technology) then adding 9 mL of DMEM. Trypsinized cells and new media were aspirated and combined with 10 mL of DMEM to yield a cell solution of 30 mL. In 15 new T-75 flasks 16 mL of DMEM, 2 mL of fetal bovine serum, and 2 mL of cell solution were added. Incubation of cells was accomplished using the aforementioned conditions for three days. Cells were harvested by aspirating media and rinsing cells with 10 mL of DMEM. Cells were centrifuged into a pellet, and supernatant was decanted. Cells were counted using a hemocytometer Cole-Parmer (Vernon Hills, IL). The pellet was estimated to contain ~2*e8 cells. Pellet was flash frozen in liquid nitrogen then re-suspended in 10 mL 50 mM ammonium bicarbonate. Ten 1 mL aliquots containing 20*e6 were taken and split into 50 µL aliquots containing 1*e6 cells. These aliquots of cells were used in the analysis.
Supplemental Table.1: The results of the SDC DOE experiment searched allowing identified peptides with up to two missed cleavages.

<table>
<thead>
<tr>
<th>Run #</th>
<th>Peptide Identifications</th>
<th>% SDC (1,3)</th>
<th>Digestion time (4,12)</th>
<th>Trypsin ratio (1:100, 1:50)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9559</td>
<td>3</td>
<td>12</td>
<td>1:50</td>
</tr>
<tr>
<td>2</td>
<td>9222</td>
<td>1</td>
<td>12</td>
<td>1:100</td>
</tr>
<tr>
<td>3</td>
<td>10364</td>
<td>1</td>
<td>4</td>
<td>1:100</td>
</tr>
<tr>
<td>4</td>
<td>8188</td>
<td>3</td>
<td>12</td>
<td>1:100</td>
</tr>
<tr>
<td>5</td>
<td>11413</td>
<td>1</td>
<td>12</td>
<td>1:50</td>
</tr>
<tr>
<td>6</td>
<td>12722</td>
<td>1</td>
<td>4</td>
<td>1:50</td>
</tr>
<tr>
<td>7</td>
<td>11306</td>
<td>3</td>
<td>4</td>
<td>1:100</td>
</tr>
<tr>
<td>8</td>
<td>10912</td>
<td>3</td>
<td>4</td>
<td>1:50</td>
</tr>
</tbody>
</table>
Supplemental Table 2: The average performance of detergents based upon PSMs, total number of identified peptides, and total number of protein groups. The control is significantly lower for each metric.

<table>
<thead>
<tr>
<th>Detergent</th>
<th>Peptides</th>
<th>PSMs</th>
<th>Protein groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Invitrosol</td>
<td>17583 ± 1424</td>
<td>26374 ± 2772</td>
<td>3475 ± 164</td>
</tr>
<tr>
<td>PPS</td>
<td>16521 ± 996</td>
<td>24302 ± 1145</td>
<td>3412 ± 100</td>
</tr>
<tr>
<td>RapiGest</td>
<td>16677 ± 847</td>
<td>24932 ± 1394</td>
<td>3374 ± 104</td>
</tr>
<tr>
<td>SDC</td>
<td>18408 ± 1063</td>
<td>27380 ± 2110</td>
<td>3550 ± 129</td>
</tr>
<tr>
<td>Control</td>
<td>13404 ± 1154</td>
<td>20064 ± 1800</td>
<td>2882 ± 181</td>
</tr>
</tbody>
</table>
**Supplemental Figure 1:** Box plots of the gravy score distribution for control and detergents. The median gravy score is shown in white. The total number of peptides recovered, from each set of replicates, is listed in the legend.
Supplemental Figure.2: Comparison of distributions of gravy scores for each of detergent. The distribution of the detergents was compared after normalizing to the control. This was accomplished by dividing the range of gravy scores into 0.1 sized bins from -3.5 to +2.5, and taking the quotient of total peptides for each detergent divided by those of the control in the corresponding bin.