

**IMPLEMENTATION OF A PROGRAMMATIC APPROACH TO SIMPLIFY
MASS SPECTROMETRY DATA ANALYSIS BY MACHINE LEARNING AND
APPLICATIONS IN BIOMARKER RESEARCH**

By

Leah Danielle Pfeifer

Submitted to the graduate degree program in Chemistry and the Graduate Faculty of the
University of Kansas in partial fulfillment of the requirements
for the degree of Master of Science.

Chair: Dr. Heather Desaire

Dr. Susan Lunte

Dr. Minae Mure

Date Defended: August 23, 2021

The thesis committee for Leah Danielle Pfeifer certifies that this is
the approved version of the following thesis:

**IMPLEMENTATION OF A PROGRAMMATIC APPROACH TO SIMPLIFY
MASS SPECTROMETRY DATA ANALYSIS BY MACHINE LEARNING AND
APPLICATIONS IN BIOMARKER RESEARCH**

Chair: Dr. Heather Desaire

Co-Chair: Dr. Susan Lunte

Co-Chair: Dr. Minae Mure

Date Approved: August 23, 2021

Abstract

The advancement of biomarker discovery and implementation will enable earlier disease detection, support clinical decision making and improve patient outcomes. A major challenge faced is the development of methods capable of detecting small changes in a biological sample; there is a need for analytical methods that measure changes selectively and sensitively, and data analysis methods that effectively identify those changes.

The use of glycans as a biomarker class has unique advantages. Due to their non-template production, their composition is dynamic with changes in the cellular environment. The composition of glycans is extremely heterogenous, so the use of glycans as biomarkers is analytically challenging but could reform the way disease is diagnosed and treated.

One way to approach this conundrum is to combine mass spectrometry and machine learning. Mass spectrometry experiments generate great amounts of data, and often times, it is not feasible to peruse in its native size or format. Implementing machine learning with mass spectrometry data allows the inclusion of more data and in turn, will enable advancements in biomarker discovery.

A software tool, LevR, has been developed to support mass spectrometrists implementing machine learning into their workflows; this tool provides automated formatting of mass spectrometry data into a machine learning ready format.

Acknowledgements

First, I wish to thank my parents for their endless love and support, which has been integral to my success as a student and person. I will never be able to express the amount of gratitude I have for both of you, partially because I would be a puddle of tears, but I hope I can begin by making you proud.

- *Mom*, thank you for supporting me as a young scientist by helping me create the coolest science projects. I am particularly fond of the project where we designed a tornado simulation using hot & cold water and food coloring, and the eukaryotic cell we made out of glass beads, metal wire, and resin, which is the best paper weight. Thank you for bringing me along to Relay for Life every year and showing me that hard work and empathy for others makes a difference.
- *Dad*, thank you for supporting me as a thinker by sharing with me your systematic approach to solving problems. Whenever I am having trouble, I feel a sense of relief when you arrive because I know you will help me reach a solution. And, thank you for making sure that I always have a safe mode of transportation that will get me to where I *want* to be.

Next, I wish to express my deepest gratitude to the people at KU who have supported me:

- My research advisor, Dr. Heather Desaire, thank you for being my mentor. I have learned quite a bit about mass spectrometry and 'omics, but what I value most from my experience at KU is being challenged by you.
- Dr. Eden Go, thank you for training me; I value the time you spent sharing your expertise in mass spectrometry.

- All of the Desaire group members, especially Hanna Nguyen, Dr. Milani Wijeweera Patabandige, and Dr. Josh Shipman, thank you for helping me find my footing in the lab. Aleesa Chua and Madeline Haga: although our time together was short, I have learned so much from mentoring such brilliant young scientists, so thank you.
- The University of Kansas Department of Chemistry faculty and staff, whose support has been unwavering.

To the friends I have made at KU and in Lawrence, you all are what has made Lawrence feel like home; the completion of this work would not have been possible without your companionship. I am looking forward to a time when we can gather for trivia night and enjoy a cocktail at JBUG.

Callaway Holt, my partner, thank you for your words of affirmation and your insistence that I learn to code... it turned out to be quite useful. My favorite respite from my studies, especially during the pandemic, has been going on adventures with you to look for bald eagles. For this and many other reasons, I know that I am the luckiest.

I want to express my appreciation for Matilda, my cat, for making sure I am never alone. She has taught me the importance of uninterrupted naps, getting fresh air, and always having treats available.

Aunt Kerry, thank you for being my confidant and my sounding board. You have impeccable timing when it comes to checking in on me.

To the many other people in my life who have certainly supported me reaching this milestone: Thank you for being the best company. You bring me great joy.

Table of Contents

Chapter 1 Introduction	1
1.1 N-linked glycosylation.....	1
1.2 Glycans identified as potential biomarkers.....	3
1.3 Analytical methods for studying glycosylation	3
1.4 Studying glycosylation and enhancing detection of glycans and glycopeptides	4
1.4.1 Analysis of released glycans.....	5
1.4.1.1 Fluorescence-based glycomics analysis.....	7
1.4.1.2 Mass spectrometry methods for analysis of released glycans	8
1.4.1.2.1 MALDI-TOF MS of released glycans	8
1.4.1.2.2 LC-ESI-MS of released glycans	9
1.4.2 Analysis of glycopeptides	10
1.4.2.1 LC-MS of glycopeptides.....	10
1.5 Glycans as biomarkers and the computational challenges in biomarker development	11
1.6 Summary of following chapters.....	14
1.7 Acknowledgements.....	16
1.8 References.....	16
Chapter 2 Leveraging R for fast analysis of mass spectrometry data with machine learning	22
2.1 Abstract.....	22
2.2 Introduction.....	23
2.3 Experimental Methods	26
2.3.1 Fingerprint Samples:	26
2.3.1.1 Fingerprint Collection and Preparation.....	26
2.3.1.2 ESI-MS conditions.....	27
2.3.2 Glycopeptide Samples:	27
2.3.2.1 Materials and Reagents:	27
2.3.2.2 Preparation of Native and Partially Defucosylated IgG Tryptic Digests.....	27
2.3.2.3 Preparation of Native and Mixed Samples for Analysis.....	29
2.3.2.4 Liquid Chromatography-Mass Spectrometry Analysis of IgG Glycopeptide Samples	29
2.3.2.5 Mass Spectrometry (MS) Conditions.....	29
2.3.3 .RAW file handling.....	30
2.3.4 Pipeline construction.....	31
2.3.5 Description of binning method	31

2.3.6	Specific settings used for fingerprint samples	32
2.3.7	Aristotle Classifier settings and submission to the Aristotle Classifier	32
2.3.7.1	Extracting features by high scores	32
2.3.7.2	Workflow accommodation for LC data	33
2.3.8	Specific settings for glycopeptide samples	33
2.3.9	Using the Aristotle Classifier to classify samples.....	33
2.3.10	Identification of features associated with glycopeptides	33
2.3.11	Classification of samples using subset of data.....	34
2.3.12	Using PCA as a comparison	34
2.4	Results and Discussion	34
2.4.1	Overview and Interface.....	34
2.4.2	Test set one: Fingerprints	37
2.4.3	Test set two: Glycopeptides	40
2.5	Conclusion	46
2.6	Acknowledgements.....	46
2.7	References.....	47
2.8	Appendix A: LevR for ESI-MS data	51
2.9	Appendix B: LevR for LC-MS data.....	55
Chapter 3	Future directions.....	60
3.1	Summary	60
3.2	Studying metabolic health by fingerprint samples.....	60
3.2.1	Fingerprint Applications for monitoring nourishment.....	60
3.2.2	Modeling changes in nourishment via fingerprints	62
3.3	Experimental Methods	62
3.3.1	Diet Regimen	62
3.3.2	Fingerprint Collection and Preparation.....	63
3.3.3	ESI-MS conditions.....	64
3.3.4	Specific settings used for fingerprint samples	64
3.3.5	Aristotle Classifier settings and submission to the Aristotle Classifier.....	64
3.4	Preliminary results	64
3.5	Discussion.....	66
3.6	Normalization of MS data to improve outcomes from merged data sets	67
3.7	Conclusion	69
3.8	Acknowledgements.....	69

3.9 References.....	70
---------------------	----

Chapter 1 Introduction

1.1 N-linked glycosylation

Glycans are complex carbohydrates that are involved in a common post-translational modification called glycosylation, and they play a major role in cellular activity, including protein function and stability, cell signaling, and disease development. They are composed of many monosaccharide units and their composition can be likened to a direct read-out of the cellular environment from which the glycoprotein originated; as the cell's environment changes throughout a day, or week, or year, the glycosylation profile also undergoes changes. Many factors can influence cellular environment conditions; for example, the extent of reaction for glycosidase or transferase enzymes is dependent on the biochemistry in their surroundings. This level of sensitivity is what gives rise to the extreme heterogeneity in glycoforms, where in addition to enzyme kinetics, the monosaccharide units, linkages and branching also contribute to their complexity.¹

While there exist multiple types and classes of glycoconjugates,²⁻⁵ this thesis is focused on protein glycosylation- specifically N-linked protein glycosylation and methods of analysis by mass spectrometry. N-linked glycosylation, whose naming convention is derived from the amino acid residue to which the glycan is attached- asparagine, single letter code “N”, occurs in a specific amino acid motif: Asparagine-X-Serine/Threonine/Cysteine, where X can be any amino acid except proline.^{6,7}

Glycans are synthesized in the endoplasmic reticulum and Golgi apparatus, where the precursor composition is $\text{Glc}_3\text{Man}_9\text{GlcNAc}_2$. As the glycan makes its way through the cellular environment, it is exposed to a series of enzymes which trim and/or replace monosaccharides.

Figure 1-1 depicts this precursor, its building blocks, and the types of glycans that originate from this precursor.

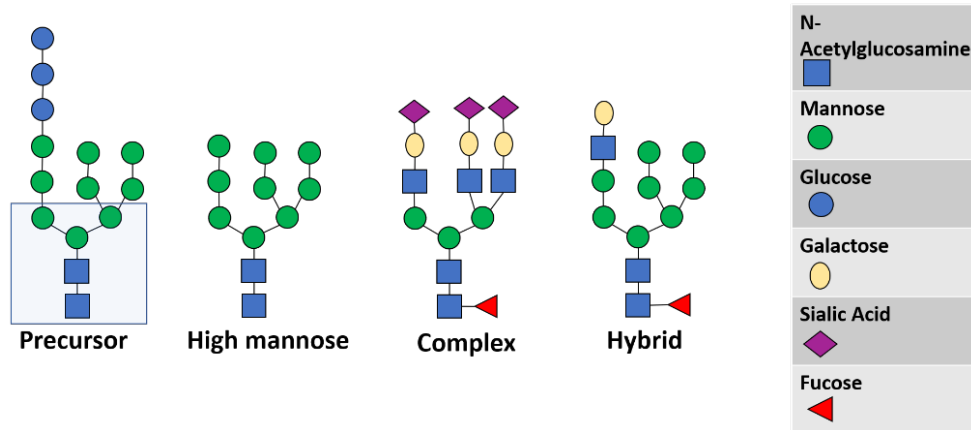


Figure 1-1. Glycans depicted with symbols, where each symbol represents a different sugar building block. The boxed portion on the precursor glycan highlights the conserved core. The three major types of N-linked glycans are high mannose, complex, and hybrid (which has characteristics of both high mannose and complex glycans).

Proteins that are differentially glycosylated possess different physical characteristics, including folding, interactions, structure and thus, function, in comparison to their native glycosylation profile. For example, IgG has one N-linked glycosylation site within the Fc region; when glycosylation changes occur, so too does the structure, conformation, and binding affinity⁸. This has implications for immune and inflammatory responses if IgG is no longer performing its essential duties. In fact, IgG has been of interest during the COVID-19 pandemic; studies suggest that patients who have core fucosylated anti-SARS-CoV-2 antibodies are less likely to experience severe disease when compared to patients whose antibodies were afucosylated,⁹⁻¹¹ or differentially glycosylated.¹² This change in glycosylation is thought to be related to the immune response, specifically an amplified cytokine presence, which is a proinflammatory response. Although much remains unknown with how many roles IgG plays in the progression of COVID-19 disease, the importance of studying and understanding glycosylation is ever apparent.

Beyond native biological changes that occur in relation to disease development, glycans are also critical in pharmaceutical development, as they determine efficacy, stability, and metabolism of biotherapeutics.¹³ For example, IgG glycosylation has been exploited to improve the efficacy of intravenous Ig (IVIg). By altering the glycan profile to be fully sialylated, the anti-inflammatory activity was increased 10X, when compared to traditional IVIg treatment.⁸

1.2 Glycans identified as potential biomarkers

A number of glycoproteins, and the changes that occur within their glycosylation profiles, have been identified as potential biomarkers for disease, as glycosylation has been found to undergo changes during disease pathogenesis,^{4, 14} chronic inflammation,¹⁵⁻¹⁷ and conditions associated with aging.^{18, 19} More specifically, abnormal glycosylation has been implicated in many disease states including Alzheimer's disease,¹⁹ galactosaemia,²⁰ acute lymphoblastic leukemia,²¹ and ovarian cancer,^{22, 23} among many other cancers,²⁴⁻²⁸ autoimmune disorders,^{29, 30} and infectious disease.¹⁵

In a disease state, glycosylation can be upregulated or downregulated, depending on the protein and glycan. For example, many rheumatoid arthritis patients experience remission during pregnancy and relapse post-partum; the glycosylation changes on IgG in serum included an increase in sialylation and galactosylation during remission, followed by a sharp decrease.³¹ In another glycomics-related study with cerebrospinal fluid from Alzheimer's disease patients, female patients were found to have higher instances of fucosylation and bisecting GlcNac structures, and in both males and females, high mannose structures were less abundant.³²

1.3 Analytical methods for studying glycosylation

Given the high diversity in structure due to their non-template production, glycans are challenging to study but contain a vast amount of information that could be leveraged for

diagnosing and prognosing disease, monitoring changes in health status and more. Beyond upregulation and downregulation of glycosylation, changes in branching and modifications to the core structure can be indicative of a change in the cellular environment.²⁴

Depending on the time scale and origin of the protein, these changes can be very subtle and hard to detect. Although these changes may be slight, the ability to monitor changes in glycosylation could be the key to advancing diagnostics, enabling earlier diagnosis of disease or other abnormal states. To harness the information contained in these sugars attached to proteins and detect subtle changes in a glycosylation profile, robust analytical methods are needed to generate interpretable results so that glycans can be used as biomarkers in a clinical setting.

Current analytical methods for monitoring changes in glycosylation have been reviewed extensively.^{5, 30, 33-41} The most popular methods employ fluorescence or mass spectrometry for detection of enzymatically released glycans and/or glycopeptides, specifically: analysis of released glycans by LC-Fluorescence, MALDI-TOF MS and LC-MS, and LC-MS of glycopeptides. Each method possesses a unique set of advantages and disadvantages. As such, there exists no absolute “best method”, but more so a “best selection” from the methods available; this includes both the selection of sample preparation methods and instrumental analysis methods. A brief discussion of these methods is included herein.

1.4 Studying glycosylation and enhancing detection of glycans and glycopeptides

Changes in glycosylation can be monitored and quantified in two ways: 1) the glycans are released from the protein prior to analysis, or 2) the glycoprotein is digested to generate glycopeptides prior to analysis. Selection of the best method will depend on the specifics of the study at hand. Figure 1-2 is a visual depiction of some common glycomics workflows.

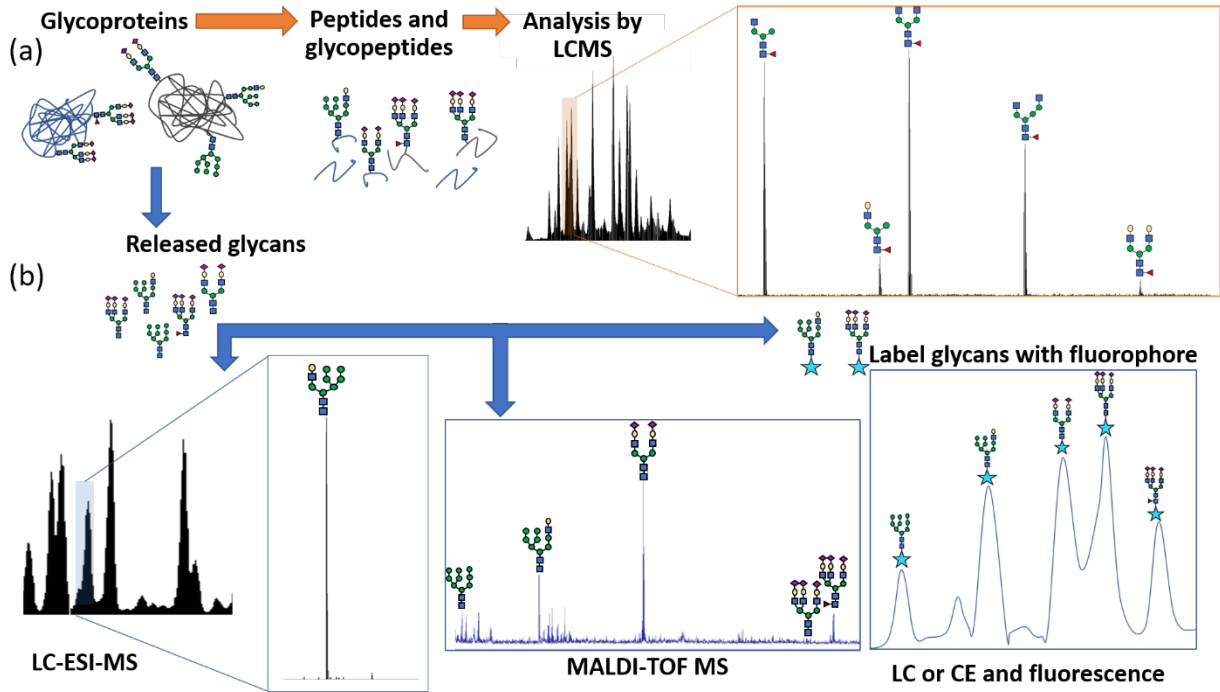


Figure 1-2. Common glycomics workflows. (a) Glycopeptides are generated enzymatically, either in crude mixture or after glycoproteins have been isolated from the sample, then analyzed by LCMS and tandem MS experiments. (b) Released glycans from either glycoproteins or glycopeptides can be analyzed by LC-ESI-MS, MALDI-TOF MS, and LC or CE with a fluorescence tag.

1.4.1 Analysis of released glycans

To obtain released N-linked glycans, PNGase F is commonly used; it cleaves the bond between the asparagine residue and the reducing end of the glycan and its digestion efficiency is highest when the glycoprotein has been denatured.⁴² Other enzymes may perform better, depending on the glycoform composition. For example, PNGase A is used for glycans with core alpha 1,3-fucosylation, as PNGase F is unable to cleave glycans with this character. However, PNGase A is not able to cleave glycans from glycoproteins, so glycopeptides must first be generated, using a proteolytic enzyme like trypsin.

Native released glycans cannot be detected by optical detection methods because they do not have chromophore or fluorophore moieties. Glycans can, however, be derivatized with a fluorophore or chromophore tag by reductive amination at the reducing end, which can improve

separation and detection by liquid chromatography or capillary electrophoresis coupled with an optical detection method.^{43, 44}

Additional chemistry can be performed at the reducing end of the glycan to derivatize with other reagents via the formation of hydrazones and oximes.⁴⁵⁻⁴⁸ With this chemistry, a variety of tags can be appended to the glycan to improve ionization, detection, and assignment of the glycans in mass spectrometry experiments. In addition to reducing end modifications, glycans can be chemically modified to either alter the chemistry or aid in conserving the composition. A common modification necessary is the neutralization and stabilization of sialic acid moieties, as they are labile and easily lost during mass spectrometry experiments. This is most commonly achieved using permethylation, where all free hydroxyl groups are converted into methyl esters, amide hydrogens into methyl esters, and carboxyl groups are esterified.⁴² These can bring many benefits to the analysis, including improved chromatographic separation, enhanced ionization efficiency, and stabilization of the labile sialic acid components of the glycans.

Stable isotopes can be incorporated into glycans by various strategies to quantify the relative abundances of glycan species from distinct sample groups (e.g., disease vs healthy) in a mass spectrometry experiment. Dual labeling enables simultaneous quantification of neutral and acidic glycans.⁴⁹⁻⁵¹ This approach involves incorporation of isotopic species at both ends of the glycan. Notably, this method provides added value through stabilization of sialic acid moieties, simplified sialic acid counting, and enhanced ionization.^{51, 52} Various tags have been designed, implementing this strategy to improve glycans' detection and identification by mass spectrometry. For example, the mdSUGAR tag uses reductive amination and periodate oxidation to derivatize the reducing end and sialic acids, respectively, which introduces a larger mass

difference and renders this method amenable to lower resolution instruments.⁵³ Isotopic PFBHA and isotopic methylamine have been used in a similar fashion, and they bring the added benefit of facilitating glycan enrichment via Fluorous SPE.^{52, 54} Other methods include enzymatic ¹⁸O labeling,⁵⁵ isotopic permethylation,⁵⁶ and metabolic labeling.⁵⁷

In the case of all tags, efficient derivatization is critical to achieving reliable quantification. If 100% of glycans in a given sample are not derivatized, the sample cannot be reliably quantified because not all glycans have been tagged. As discussed, glycans are typically derivatized at the reducing end. As derivatization is a chemical reaction, many factors contribute to a reaction's ability to move to completion including temperature and chemical environment. Derivatization inefficiency continues to hinder reliable quantification, particularly when the reaction requires a certain pH, temperature, or concentration to proceed. The ideal derivatization workflow is one that can move to completion at any volume in a simple solvent like water.

Tagging glycans can provide many benefits to the workflow when used appropriately. Tag-specific enrichment and simultaneous elimination of untagged components of a sample (e.g., undesirable salts and proteins) can improve signal during analysis. However, poor selectivity during enrichment can result in sample loss, so it is critical to use a selective enrichment method that effectively captures the analyte of interest.

1.4.1.1 Fluorescence-based glycomics analysis

In fluorescence-based glycomics analyses, the glycans are released enzymatically and derivatized with a reagent to enhance detection. The tag often serves dual purposes- for improved detection by fluorescence and to enhance the separation of different glycans. Since fluorescence on its own does not provide information on the composition of the glycans detected, the chromatogram generated is often utilized to match each peak to a glycan composition. To assign peaks in a chromatogram, a calibration method can be used. For example, a standard

dextran ladder can be used to normalize the data acquired by the separation method. A glucose unit is assigned to each peak, which is representative of the number of monosaccharides and linkages present in a glycan, and enables the determination of the theoretical position in the chromatogram (LC) or electropherogram (CE).⁵⁸ This can then be compared to the experimental chromatogram to assign glycan compositions to various peaks. Although this methodology can be useful in glycomics research, supplemental data is needed to completely characterize the glycan structures; this can be done using highly specific enzymes (exoglycosidase) and additional analytical experiments or tandem mass spectrometry methods. Nevertheless, fluorescence-based methods are attractive for clinical assay design, particularly for high-throughput applications with a minimized cost. A method for translation to clinical applications was recently described,⁵⁹ in anticipation of advancements with glycan biomarkers.

1.4.1.2 Mass spectrometry methods for analysis of released glycans

Mass spectrometry is the leading analytical technique to study glycans, as it provides the ability to identify, profile, and quantify glycans when coupled with appropriate sample introduction and separation methods. For the analysis of released glycans, there are two dominant mass spectrometry methods: MALDI-TOF MS or HPLC-ESI-MS. The first step of either method is generating released glycans; isolating the glycoprotein of interest or performing the enzymatic release in a crude mixture, using PNGase F to cleave the glycan from the protein. Then, depending on the detection method, glycans can be modified and enriched in various ways prior to the mass spectrometry experiment.^{28, 60, 61}

1.4.1.2.1 MALDI-TOF MS of released glycans

During a MALDI-TOF MS experiment, the sample is ionized by a laser, which necessitates the sample be dissolved in a compatible matrix. Although dihydroxy-benzoic acid

(DHB) is the most common matrix used in glycan analysis, there are other matrices available that may be more amenable to the sample of interest.^{62, 63} Once in the matrix, the sample is spotted onto a MALDI plate and is irradiated with a laser, ions are generated and detected by the Time-of-Flight mass spectrometer. Since this method does not require a separation method prior to MS analysis, it is capable of being very high throughput which is attractive for translation to clinical glycomics. However, this method has potential shortcomings: quantification of different glycans within a sample is difficult due to varying ionization and technical expertise is required to operate the mass spectrometer and interpret the results.

1.4.1.2.2 LC-ESI-MS of released glycans

For the analysis of released glycans by LC-ESI-MS, the separation method development requires consideration of the glycan compositions and any tags used. Additional considerations include the ionization and detection method. Tags are frequently used to improve separation and detection, similar to fluorescence-based glycomics methods. An additional reason for using a tag in mass spectrometry experiments is to improve ionization efficiency during electrospray ionization or to aid in compositional assignment of the glycans. To enhance ionization efficiency of glycans, as native glycans ionize poorly due to their hydrophilicity, hydrophobic moieties are commonly used.^{22, 64}

First, the glycans are released in a crude mixture or after the glycoprotein of interest has been isolated, followed by enrichment, potentially derivatizing with a tag to aid in LC separation, ionization, and subsequent detection by mass spectrometry.

The analysis of released glycans has limitations since the glycan is no longer attached to its glycosylation site; this is particularly problematic when there are multiple glycoproteins in the sample or there are multiple glycosylation sites on a single glycoprotein. In a scenario where the

protein of interest is not enriched prior to enzymatic release of the glycans, it is difficult to confidently monitor changes in the glycosylation profile, since the changes could be an artifact of a change in the glycoprotein expression rather than a change in the glycosylation profile.

1.4.2 Analysis of glycopeptides

1.4.2.1 LC-MS of glycopeptides

The third and most comprehensive analysis in glycomics research is LC-MS of glycopeptides. The glycoprotein must first be isolated, denatured, reduced, alkylated, and purified, all before performing a digestion to generate glycopeptides and peptides. Various enzymes can be used in either a targeted or untargeted manner to cleave at different amino acid sequences to generate different peptides. The advantage to using enzymes with high specificity is that the peptide fragments it generates can be predicted, provided that the amino acid sequence of the protein is known. For example, trypsin can be used to cleave the protein between lysine and arginine on the C-terminal side, but missed cleavages are possible if the surrounding amino acids inhibit the enzyme by steric hindrance.⁶⁵ Alternatively, non-specific digestion can be implemented with proteases like proteinase K, which results in many more, shorter peptides that can be enriched or purified by other methods to capture only the glycopeptides. The challenge with this approach is that since the enzyme cleaves rather indiscriminately, the peptides and glycopeptides generated are not predictable.

The LC method prior to MS detection aids in separating the peptides, generated during the digestion, from the glycopeptides. This is an important step, as co-eluting peptides reduce the ionization efficiency of the glycopeptides.

To begin to assign glycan compositions, tandem mass spectrometry experiments are necessary. These experiments fragment the carbohydrate moieties into smaller pieces. During collision-induced dissociation (CID), diagnostic oxonium ions are generated, which indicate that

there are glycopeptides present. Additionally, the product ions can be used to infer the glycan components of the glycopeptides. To assign the composition of the peptide, electron transfer dissociation (ETD) is an alternative fragmentation approach that primarily provides fragmentation information about the peptide component. Both experiments provide complementary information as to the composition and structure of the glycan. These data, in addition to the high-resolution mass of the precursor ion, are used to assign the glycans present.

While the preparation and analysis of glycopeptides is more involved, the data acquired can be more informative, since site-specific and protein-specific information is retained. Arriving at this level of information can be challenging, however, because both the peptide and glycan need to be characterized and assigned compositions. Once the useful glycopeptides are identified, multiple reaction monitoring (MRM) is typically used for targeted quantification of the glycopeptides across multiple sample sets.^{34, 66, 67} This method is used widely in biomarker discovery phases, however, it requires an experienced mass spectrometrists to conduct the experiments and interpret the results. In order to use this approach, the glycopeptides to be quantified must be known, their approximate retention times must be known, and their key product ions must be known. Thus, it is not an approach that can be applied without considerable up-front effort.

1.5 Glycans as biomarkers and the computational challenges in biomarker development

The use of glycans as biomarkers requires the ability to reliably monitor glycans in a simple, effective, high throughput manner. This is particularly important in clinical settings where a fast and reliable readout is critical to patient outcomes. Considerations for which is the best method have been reviewed, where each method has figures of merit.⁶⁸ While significant

progress has been made, some remaining obstacles need to be addressed, including the ability to process large mass spectrometry data sets.

Mass spectrometric analysis of changes in glycosylation can be complicated and often times, the data has to be reduced to a more manageable size and format. This can be particularly problematic in untargeted analyses, where the pertinent data is not known *a priori*. Traditional unsupervised data analysis methods, like PCA, do not yield useful results when applied to dense data sets, including glycomics. This is due to a mismatch in the data set and analysis process. If the process does not match the characteristics of the data set, it fails to capture the complexity within the samples. For example, applying PCA to glycomics data is potentially problematic because PCA only identifies the most impactful components contributing to the variability within the sample set. The method does not explicitly look for variability due to the sample type (healthy vs. disease state, for example.). In glycomics samples, the changes and differences are subtle and numerous, so dimension reduction strategies are inappropriate to use as they can (unintentionally) omit important aspects of the data that help distinguish between the two samples. To advance the possibility of using changes in glycosylation as a clinical readout, strategies are needed that are built to handle mass spectrometry data sets with many features.

An alternative data analysis approach to PCA, which has been developed and used increasingly over the last few years,⁶⁹⁻⁷⁴ is doing supervised machine learning on mass spectrometry data sets. While many different supervised classification algorithms exist, our research group has developed one specifically for use on glycomics data. This machine learning algorithm, the Aristotle Classifier,^{72, 74, 75} operates on the assumption that retaining a more complete set of features, regardless of their known or perceived importance, improves classification accuracy. In the context of diagnostics and disease/healthy samples, many features

within a sample contribute to its overall character, rather than an “all or nothing” approach. In Figure 1-3, this strategy is pictorially depicted. In Figure 1-3a, 6 squares from the original portrait are presented in no particular order. In Figure 1-3b, the 6 squares have been arranged in their original position, such that they are in context of one another. In Figure 1-3c, the full portrait is provided. The representation in Figure 1-3b illustrates how having the squares in context of one another helps the onlooker discern what the full portrait might be.



Figure 1-3. A visual representation of the classifier’s approach to classifying samples. (a) 6 squares from the portrait, in no particular order. It is challenging to discern what the full portrait is. (b) the 6 squares from panel a, in their original arrangement. (c) The full portrait of “*Ballet Scene*” by Edgar Degas, c. 1907, obtained via Creative Commons- National Gallery of Art (NGA46491).

The use of machine learning facilitates interpretable biological information to be obtained from large amounts of complex non-linear data; this is in contrast to traditional methods that tend to perform well with linearly related data. With supervised machine learning strategies, the machine learning algorithm is first trained with samples that are labeled as belonging to Group 1 or Group 2 (i.e., healthy or disease). Then, new data is submitted to the algorithm to be classified.

To develop robust machine learning tools, large amounts of data are needed to sufficiently train the model. When small data sets are used, there is a significant risk of overfitting the data; this is a common bioinformatics problem, frequently referred to as the big p (many measurements per subject), little n (not many subjects) problem. As the ratio of the number of features to the number of samples increases, the higher the likelihood that the algorithm will train on noise, which will be detrimental in validation studies.⁷⁶ An obvious way to resolve this is to increase the number of samples in the data set, as any classifier will perform better with more training data, but this may not be possible if the workflow is not scalable or if the sample type is rare/difficult to obtain. The Aristotle Classifier evades this problem better than other classifiers because it is less likely to over-emphasize a single feature or two in its scoring algorithm, and rather, uses many features to assign the sample to a given category.

In glycomics research, obtaining large data sets is challenging due to the sample preparation, time requirement, and level of necessary expertise. Furthermore, often times in glycomics, there are a limited number of samples due to the nature of the disease of interest, including its rare incidence. Therefore, an opportunity exists to benefit both the fields of machine learning and glycomics/biomarker research by developing tools at the interface of these two fields. This thesis occupies that intellectual niche.

1.6 Summary of following chapters

The software tool created to facilitate machine learning on glycomics samples and other mass spectrometry data sets is called *LevR*, named for its utility in quickly formatting mass spectrometry data into a machine learning ready format **L**everaging **R** programming. It is intended to make machine learning tools accessible to all mass spectrometrists, regardless of their level of experience with programming/coding; it enables 1) easy submission of MS data for

analysis by machine learning methods 2) fast application of machine learning methods on large MS data sets, and 3) straightforward interpretability to enable useful conclusions and actions. This workflow is scalable and does not sacrifice data quality or quantity in the process. The key contribution is the increased accessibility to machine learning tools for all mass spectrometrists: a simple, yet effective, method for configuring data into a machine learning-ready format.

The second contribution is the development of a model data set using fingerprints that can be used to mimic the subtle changes that occur in glycosylation, but which does not require the tedious tasks that glycomics samples demand. This model data set is tunable, in that variability can be introduced and controlled in various ways. This characteristic is particularly useful when developing machine learning tools, as it enables the researcher to easily generate more challenging data sets without significant time or material investment.

In the following chapter, these come together to provide a straightforward solution for mass spectrometrists who wish to implement machine learning strategies into their workflow, but who have minimal experience in formatting their data for ML-compatibility. The acquisition of a model data set via latent fingerprints is also described, which is particularly amenable to machine learning tool development. In the final chapter, future directions are discussed, including the continuation of fingerprints as a potential sample type for health status readout.

1.7 Acknowledgements

This work was supported by NIH Grant R35GM130354 to HD.

1.8 References

1. Dwek, R. A., Glycobiology: Toward Understanding the Function of Sugars. *Chemical reviews* **1996**, *96* (2), 683-720.
2. Flynn, R. A.; Pedram, K.; Malaker, S. A.; Batista, P. J.; Smith, B. A. H.; Johnson, A. G.; George, B. M.; Majzoub, K.; Villalta, P. W.; Carette, J. E.; Bertozzi, C. R., Small RNAs are modified with N-glycans and displayed on the surface of living cells. *Cell* **2021**.
3. Flynn, R. A.; Smith, B. A. H.; Johnson, A. G.; Pedram, K.; George, B. M.; Malaker, S. A.; Majzoub, K.; Carette, J. E.; Bertozzi, C. R., Mammalian Y RNAs are modified at discrete guanosine residues with N-glycans. *bioRxiv* **2019**, 787614.
4. Kailemia, M. J.; Xu, G.; Wong, M.; Li, Q.; Goonatilleke, E.; Leon, F.; Lebrilla, C. B., Recent Advances in the Mass Spectrometry Methods for Glycomics and Cancer. *Analytical Chemistry* **2018**, *90* (1), 208-224.
5. Dong, X.; Huang, Y.; Cho, B. G.; Zhong, J.; Gautam, S.; Peng, W.; Williamson, S. D.; Banazadeh, A.; Torres-Ulloa, K. Y.; Mechref, Y., Advances in mass spectrometry-based glycomics. *Electrophoresis* **2018**, *39* (24), 3063-3081.
6. Zhu, Z.; Desaire, H., Carbohydrates on proteins: site-specific glycosylation analysis by mass spectrometry. *Annual Review of Analytical Chemistry* **2015**, *8*, 463-483.
7. Zhu, Z.; Go, E. P.; Desaire, H., Absolute Quantitation of Glycosylation Site Occupancy Using Isotopically Labeled Standards and LC-MS. American Chemical Society: 2014.
8. Cobb, B. A., The history of IgG glycosylation and where we are now. *Glycobiology* **2020**, *30* (4), 202-213.
9. Larsen, M. D.; de Graaf, E. L.; Sonneveld, M. E.; Plomp, H. R.; Nouta, J.; Hoepel, W.; Chen, H.-J.; Linty, F.; Visser, R.; Brinkhaus, M.; Šuštić, T.; de Taeye, S. W.; Bentlage, A. E. H.; Toivonen, S.; Koeleman, C. A. M.; Sainio, S.; Kootstra, N. A.; Brouwer, P. J. M.; Geyer, C. E.; Derksen, N. I. L.; Wolbink, G.; de Winther, M.; Sanders, R. W.; van Gils, M. J.; de Bruin, S.; Vlaar, A. P. J.; Rispens, T.; den Dunnen, J.; Zaaier, H. L.; Wuhrer, M.; Ellen van der Schoot, C.; Vidarsson, G., Afucosylated IgG characterizes enveloped viral responses and correlates with COVID-19 severity. *Science* **2021**, *371* (6532), eabc8378.
10. Larsen, M. D.; de Graaf, E. L.; Sonneveld, M. E.; Plomp, H. R.; Linty, F.; Visser, R.; Brinkhaus, M.; Šuštić, T.; de Taeye, S. W.; Bentlage, A. E. H.; Nouta, J.; Natunen, S.; Koeleman, C. A. M.; Sainio, S.; Kootstra, N. A.; Brouwer, P. J. M.; Sanders, R. W.; van Gils, M. J.; de Bruin, S.; Vlaar, A. P. J.; Zaaier, H. L.; Wuhrer, M.; van der Schoot, C. E.; Vidarsson, G., Afucosylated immunoglobulin G responses are a hallmark of enveloped virus infections and show an exacerbated phenotype in COVID-19. *bioRxiv* **2020**, 2020.05.18.099507.
11. Hoepel, W.; Chen, H.-J.; Geyer, C. E.; Allahverdiyeva, S.; Manz, X. D.; de Taeye, S. W.; Aman, J.; Mes, L.; Steenhuis, M.; Griffith, G. R.; Bonta, P. I.; Brouwer, P. J. M.; Caniels, T. G.; van der Straten, K.; Golebski, K.; Jonkers, R. E.; Larsen, M. D.; Linty, F.; Nouta, J.; van Roomen, C. P. A. A.; van Baarle, F. E. H. P.; van Drunen, C. M.; Wolbink, G.; Vlaar, A. P. J.; de Bree, G. J.; Sanders, R. W.; Willemsen, L.; Neele, A. E.; van de Beek, D.; Rispens, T.; Wuhrer, M.; Bogaard, H. J.; van Gils, M. J.; Vidarsson, G.; de Winther, M.; den

- Dunnen, J., High titers and low fucosylation of early human anti-SARS-CoV-2 IgG promote inflammation by alveolar macrophages. *Science Translational Medicine* **2021**, *13* (596), eabf8654.
12. Petrović, T.; Alves, I.; Bugada, D.; Pascual, J.; Vučković, F.; Skelin, A.; Gaifem, J.; Villar-Garcia, J.; Vicente, M. M.; Fernandes, Â.; Dias, A. M.; Kurolt, I.-C.; Markotić, A.; Primorac, D.; Soares, A.; Malheiro, L.; Trbojević-Akmačić, I.; Abreu, M.; Sarmiento e Castro, R.; Bettinelli, S.; Callegaro, A.; Arosio, M.; Sangiorgio, L.; Lorini, L. F.; Castells, X.; Horcajada, J. P.; Pinho, S. S.; Allegri, M.; Barrios, C.; Lauc, G., Composition of the immunoglobulin G glycome associates with the severity of COVID-19. *Glycobiology* **2021**, *31* (4), 372-377.
 13. Jefferis, R., Glycosylation as a strategy to improve antibody-based therapeutics. *Nature reviews. Drug discovery* **2009**, *8* (3), 226-34.
 14. Keser, T.; Tijardović, M.; Gornik, I.; Lukić, E.; Lauc, G.; Gornik, O.; Novokmet, M., High-throughput and site-specific N-glycosylation analysis of human alpha-1-acid glycoprotein offers a great potential for new biomarker discovery. *Molecular & cellular proteomics : MCP* **2021**, *20*, 100044.
 15. Plomp, R.; Bondt, A.; de Haan, N.; Rombouts, Y.; Wuhrer, M., Recent Advances in Clinical Glycoproteomics of Immunoglobulins (Igs). *Molecular & Cellular Proteomics* **2016**, *15* (7), 2217.
 16. Vanderschaeghe, D.; Meuris, L.; Raes, T.; Grootaert, H.; Van Hecke, A.; Verhelst, X.; Van de Velde, F.; Lapauw, B.; Van Vlierberghe, H.; Callewaert, N., Endoglycosidase S Enables a Highly Simplified Clinical Chemistry Procedure for Direct Assessment of Serum IgG Undergalactosylation in Chronic Inflammatory Disease. *Mol Cell Proteomics* **2018**, *17* (12), 2508-2517.
 17. Oswald, D. M.; Jones, M. B.; Cobb, B. A., Modulation of hepatocyte sialylation drives spontaneous fatty liver disease and inflammation. *Glycobiology* **2020**, *30* (5), 346.
 18. Ruhaak, L. R.; Uh, H. W.; Beekman, M.; Hokke, C. H.; Westendorp, R. G.; Houwing-Duistermaat, J.; Wuhrer, M.; Deelder, A. M.; Slagboom, P. E., Plasma protein N-glycan profiles are associated with calendar age, familial longevity and health. *J Proteome Res* **2011**, *10* (4), 1667-74.
 19. Lundstrom, S.; Yang, H. Q.; Lyutvinskiy, Y.; Rutishauser, D.; Herukka, S.; Soininen, H.; Zubarev, R., Blood Plasma IgG Fc Glycans are Significantly Altered in Alzheimer's Disease and Progressive Mild Cognitive Impairment. *J. Alzheimers Dis.* **2014**, *38* (3), 567-579.
 20. Coss, K. P.; Hawkes, C. P.; Adamczyk, B.; Stöckmann, H.; Crushell, E.; Saldova, R.; Knerr, I.; Rubio-Gozalbo, M. E.; Monavari, A. A.; Rudd, P. M.; Treacy, E. P., N-Glycan Abnormalities in Children with Galactosemia. *Journal of Proteome Research* **2014**, *13* (2), 385-394.
 21. Feng, Y.; Chen, B.; Yu, Q.; Zhong, X.; Frost, D. C.; Ikonomidou, C.; Li, L., Isobaric Multiplex Labeling Reagents for Carbonyl-Containing Compound (SUGAR) Tags: A Probe for Quantitative Glycomic Analysis. *Analytical Chemistry* **2019**, *91* (4), 3141-3146.
 22. Hecht, E. S.; Scholl, E. H.; Walker, S. H.; Taylor, A. D.; Cliby, W. A.; Motsinger-Reif, A. A.; Muddiman, D. C., Relative Quantification and Higher-Order Modeling of the Plasma Glycan Cancer Burden Ratio in Ovarian Cancer Case-Control Samples. *Journal of Proteome Research* **2015**, *14* (10), 4394-4401.
 23. Miyamoto, S.; Stroble, C. D.; Taylor, S.; Hong, Q.; Lebrilla, C. B.; Leiserowitz, G. S.; Kim, K.; Ruhaak, L. R., Multiple Reaction Monitoring for the Quantitation of Serum Protein

- Glycosylation Profiles: Application to Ovarian Cancer. *Journal of Proteome Research* **2018**, *17* (1), 222-233.
24. Thomas, D.; Rathinavel, A. K.; Radhakrishnan, P., Altered glycosylation in cancer: A promising target for biomarkers and therapeutics. *Biochimica et biophysica acta. Reviews on cancer* **2021**, *1875* (1).
 25. Sethi, M. K.; Hancock, W. S.; Fanayan, S., Identifying N-Glycan Biomarkers in Colorectal Cancer by Mass Spectrometry. *Accounts of chemical research* **2016**, *49* (10), 2099-2106.
 26. Ruhaak, L. R.; Barkauskas, D. A.; Torres, J.; Cooke, C. L.; Wu, L. D.; Stroble, C.; Ozcan, S.; Williams, C. C.; Camorlinga, M.; Rocke, D. M.; Lebrilla, C. B.; Solnick, J. V., The serum immunoglobulin G glycosylation signature of gastric cancer. *EuPA open proteomics* **2015**, *6*, 1-9.
 27. Ruhaak, L. R.; Taylor, S. L.; Stroble, C.; Nguyen, U. T.; Parker, E. A.; Song, T.; Lebrilla, C. B.; Rom, W. N.; Pass, H.; Kim, K.; Kelly, K.; Miyamoto, S., Differential N-Glycosylation Patterns in Lung Adenocarcinoma Tissue. *Journal of Proteome Research* **2015**, *14* (11), 4538-4549.
 28. Liu, S.; Cheng, L.; Fu, Y.; Liu, B. F.; Liu, X., Characterization of IgG N-glycome profile in colorectal cancer progression by MALDI-TOF-MS. *J Proteomics* **2018**, *181*, 225-237.
 29. Ming-Chu, C.; I-Lin, T.; San-Yuan, W.; Chung-Hsuan, C.; Yu-Ting, C., High accuracy differentiating autoimmune pancreatitis from pancreatic ductal adenocarcinoma by immunoglobulin G glycosylation. *Clinical Proteomics (Online)* **2019**, *16* (1).
 30. Reiding, K. R.; Bondt, A.; Hennig, R.; Gardner, R. A.; O'Flaherty, R.; Trbojevic-Akmacic, I.; Shubhakar, A.; Hazes, J. M. W.; Reichl, U.; Fernandes, D. L.; Pucic-Bakovic, M.; Rapp, E.; Spencer, D. I. R.; Dolhain, R.; Rudd, P. M.; Lauc, G.; Wuhrer, M., High-throughput Serum N-Glycomics: Method Comparison and Application to Study Rheumatoid Arthritis and Pregnancy-associated Changes. *Mol Cell Proteomics* **2019**, *18* (1), 3-15.
 31. van de Geijn, F. E.; Wuhrer, M.; Selman, M. H.; Willemsen, S. P.; de Man, Y. A.; Deelder, A. M.; Hazes, J. M.; Dolhain, R. J., Immunoglobulin G galactosylation and sialylation are associated with pregnancy-induced improvement of rheumatoid arthritis and the postpartum flare: results from a large prospective cohort study. *Arthritis Res Ther* **2009**, *11* (6), R193.
 32. Cho, B. G.; Veillon, L.; Mechref, Y., N-glycan profile of cerebrospinal fluids from Alzheimer's Disease patients using LC-MS. *J Proteome Res* **2019**.
 33. Dalpathado, D. S.; Desaire, H., Glycopeptide analysis by mass spectrometry. *The Analyst* **2008**, *133* (6), 731-8.
 34. Goldman, R.; Sanda, M., Targeted methods for quantitative analysis of protein glycosylation. 2015; Vol. 9, pp 17-32.
 35. Huffman, J. E.; Pucic-Bakovic, M.; Klaric, L.; Hennig, R.; Selman, M. H.; Vuckovic, F.; Novokmet, M.; Kristic, J.; Borowiak, M.; Muth, T.; Polasek, O.; Razdorov, G.; Gornik, O.; Plomp, R.; Theodoratou, E.; Wright, A. F.; Rudan, I.; Hayward, C.; Campbell, H.; Deelder, A. M.; Reichl, U.; Aulchenko, Y. S.; Rapp, E.; Wuhrer, M.; Lauc, G., Comparative performance of four methods for high-throughput glycosylation analysis of immunoglobulin G in genetic and epidemiological research. *Mol Cell Proteomics* **2014**, *13* (6), 1598-610.
 36. Liu, H.; Zhang, N.; Wan, D.; Cui, M.; Liu, Z.; Liu, S., Mass spectrometry-based analysis of glycoproteins and its clinical applications in cancer biomarker discovery. *Clin Proteomics* **2014**, *11* (1), 14.

37. Mechref, Y.; Dong, X.; Huang, Y.; Cho, B. G.; Zhong, J.; Gautam, S.; Peng, W.; Williamson, S. D.; Banazadeh, A.; Torres-Ulloa, K. Y.; Mechref, Y., Advances in mass spectrometry-based glycomics. *Electrophoresis* **2018**, *39* (24), 3063.
38. Peng, W.; Zhao, J.; Dong, X.; Banazadeh, A.; Huang, Y.; Hussien, A.; Mechref, Y., Clinical application of quantitative glycomics. *Expert review of proteomics* **2018**, *15* (12), 1007-1031.
39. Yang, Y.; Franc, V.; Heck, A. J. R., Glycoproteomics: A Balance between High-Throughput and In-Depth Analysis. *Trends in Biotechnology* **2017**, *35* (7), 598-609.
40. Zhang, Y.; Yin, H.; Lu, H., Recent progress in quantitative glycoproteomics. *Glycoconjugate journal* **2012**, *29* (5-6), 249-58.
41. Zhou, S.; Veillon, L.; Dong, X.; Huang, Y.; Mechref, Y., Direct comparison of derivatization strategies for LC-MS/MS analysis of N-glycans. *The Analyst* **2017**, *142* (23), 4446-4455.
42. Zhang, Y.; Peng, Y.; Yang, L.; Lu, H., Advances in sample preparation strategies for MS-based qualitative and quantitative N-glycomics. *TrAC Trends in Analytical Chemistry* **2018**, *99*, 34-46.
43. Ruhaak, L. R.; Zauner, G.; Huhn, C.; Bruggink, C.; Deelder, A. M.; Wührer, M., Glycan labeling strategies and their use in identification and quantification. *Analytical and Bioanalytical Chemistry* **2010**, *397* (8), 3457-3481.
44. Szabo, Z.; Guttman, A.; Rejtar, T.; Karger, B. L., Improved sample preparation method for glycan analysis of glycoproteins by CE-LIF and CE-MS. *Electrophoresis* **2010**, *31* (8), 1389-95.
45. Kölmel, D. K.; Kool, E. T., Oximes and Hydrazones in Bioconjugation: Mechanism and Catalysis. *Chemical reviews* **2017**, *117* (15), 10358-10376.
46. Kool, E. T.; Crisalli, P.; Chan, K. M., Fast Alpha Nucleophiles: Structures that Undergo Rapid Hydrazone/Oxime Formation at Neutral pH. *Organic Letters* **2014**, *16* (5), 1454-1457.
47. Kameyama, A.; Dissanayake, S. K.; Thet Tin, W. W., Rapid chemical de-N-glycosylation and derivatization for liquid chromatography of immunoglobulin N-linked glycans. *PloS one* **2018**, *13* (5), e0196800.
48. Hahne, H.; Neubert, P.; Kuhn, K.; Etienne, C.; Bomgarden, R.; Rogers, J. C.; Kuster, B., Carbonyl-Reactive Tandem Mass Tags for the Proteome-Wide Quantification of N-Linked Glycans. *Analytical Chemistry* **2012**, *84* (8), 3716-3724.
49. Zhou, H.; Warren, P. G.; Froehlich, J. W.; Lee, R. S., Dual Modifications Strategy to Quantify Neutral and Sialylated N-Glycans Simultaneously by MALDI-MS. *Analytical Chemistry* **2014**, *86* (13), 6277-6284.
50. Wang, C.; Wu, Y.; Zhang, L.; Liu, B. F.; Lin, Y.; Liu, X., Relative quantitation of neutral and sialylated N-glycans using stable isotopic labeled d0/d5-benzoyl chloride by MALDI-MS. *Analytica chimica acta* **2018**, *1002*, 50-61.
51. Wei, L.; Cai, Y.; Yang, L.; Zhang, Y.; Lu, H., Duplex Stable Isotope Labeling (DuSIL) for Simultaneous Quantitation and Distinction of Sialylated and Neutral N-Glycans by MALDI-MS. *Analytical Chemistry* **2018**, *90* (17), 10442-10449.
52. Wang, L.; Yang, L.; Zhang, Y.; Lu, H., Dual isotopic labeling combined with fluorosolid-phase extraction for simultaneous discovery of neutral/sialylated N-glycans as biomarkers for gastric cancer. *Analytica chimica acta* **2020**, *1104*, 87-94.

53. Feng, Y.; Li, M.; Lin, Y.; Chen, B.; Li, L., Multiplex Quantitative Glycomics Enabled by Periodate Oxidation and Triplex Mass Defect Isobaric Multiplex Reagents for Carbonyl-Containing Compound Tags. *Analytical Chemistry* **2019**, *91* (18), 11932-11937.
54. Yang, L.; Du, X.; Peng, Y.; Cai, Y.; Wei, L.; Zhang, Y.; Lu, H., Integrated Pipeline of Isotopic Labeling and Selective Enriching for Quantitative Analysis of N-Glycome by Mass Spectrometry. *Analytical Chemistry* **2019**, *91* (2), 1486-1493.
55. Zhang, W.; Wang, H.; Tang, H.; Yang, P., Endoglycosidase-Mediated Incorporation of ¹⁸O into Glycans for Relative Glycan Quantitation. *Analytical Chemistry* **2011**, *83* (12), 4975-4981.
56. Alvarez-Manilla, G.; Warren, N. L.; Abney, T.; Atwood, J., III; Azadi, P.; York, W. S.; Pierce, M.; Orlando, R., Tools for glycomics: relative quantitation of glycans by isotopic permethylation using ¹³CH₃I. *Glycobiology* **2007**, *17* (7), 677-687.
57. Orlando, R.; Lim, J.-M.; Atwood, J. A.; Angel, P. M.; Fang, M.; Aoki, K.; Alvarez-Manilla, G.; Moremen, K. W.; York, W. S.; Tiemeyer, M.; Pierce, M.; Dalton, S.; Wells, L., IDAWG: Metabolic incorporation of stable isotope labels for quantitative glycomics of cultured cells. *J Proteome Res* **2009**, *8* (8), 3816-3823.
58. Vreeker, G. C. M.; Wuhrer, M., Reversed-phase separation methods for glycan analysis. *Analytical and Bioanalytical Chemistry* **2017**, *409* (2), 359-378.
59. Shipman, J. T.; Nguyen, H. T.; Desaire, H., So You Discovered a Potential Glycan-Based Biomarker; Now What? We Developed a High-Throughput Method for Quantitative Clinical Glycan Biomarker Validation. *ACS Omega* **2020**, *5* (12), 6270-6276.
60. Shubhakar, A.; Kozak, R. P.; Reiding, K. R.; Royle, L.; Spencer, D. I.; Fernandes, D. L.; Wuhrer, M., Automated High-Throughput Permethylation for Glycosylation Analysis of Biologics Using MALDI-TOF-MS. *Anal Chem* **2016**, *88* (17), 8562-9.
61. Wei, L.; Cai, Y.; Yang, L.; Zhang, Y.; Lu, H., Duplex Stable Isotope Labeling (DuSIL) for Simultaneous Quantitation and Distinction of Sialylated and Neutral N-Glycans by MALDI-MS. *Anal Chem* **2018**, *90* (17), 10442-10449.
62. Snovida, S. I.; Chen, V. C.; Perreault, H., Use of a 2,5-Dihydroxybenzoic Acid/Aniline MALDI Matrix for Improved Detection and On-Target Derivatization of Glycans: A Preliminary Report. *Analytical Chemistry* **2006**, *78* (24), 8561-8568.
63. Rohmer, M.; Meyer, B.; Mank, M.; Stahl, B.; Bahr, U.; Karas, M., 3-Aminoquinoline Acting as Matrix and Derivatizing Agent for MALDI MS Analysis of Oligosaccharides. *Analytical Chemistry* **2010**, *82* (9), 3719-3726.
64. Walker, S. H.; Budhathoki-Uprety, J.; Novak, B. M.; Muddiman, D. C., Stable-Isotope Labeled Hydrophobic Hydrazide Reagents for the Relative Quantification of N-Linked Glycans by Electrospray Ionization Mass Spectrometry. *Analytical Chemistry* **2011**, *83* (17), 6738-6745.
65. Ruhaak, L. R.; Xu, G.; Li, Q.; Goonatilleke, E.; Lebrilla, C. B., Mass Spectrometry Approaches to Glycomic and Glycoproteomic Analyses. *Chemical reviews* **2018**, *118* (17), 7886-7930.
66. Song, E.; Pyreddy, S.; Mechref, Y., Quantification of glycopeptides by multiple reaction monitoring liquid chromatography/tandem mass spectrometry. *Rapid communications in mass spectrometry : RCM* **2012**, *26* (17), 1941-1954.
67. Yuan, W.; Wei, R.; Goldman, R.; Sanda, M., Optimized Fragmentation for Quantitative Analysis of Fucosylated N-Glycoproteins by LC-MS-MRM. *Analytical Chemistry* **2019**, *91* (14), 9206-9212.

68. Patabandige, M. W.; Pfeifer, L. D.; Nguyen, H. T.; Desaire, H., Quantitative clinical glycomics strategies: A guide for selecting the best analysis approach. *Mass Spectrom Rev* **2021**, n/a (n/a).
69. Xie, Y. R.; Castro, D. C.; Bell, S. E.; Rubakhin, S. S.; Sweedler, J. V., Single-Cell Classification Using Mass Spectrometry through Interpretable Machine Learning. *Analytical chemistry (Washington)* **2020**, *92* (13), 9338-9347.
70. Ulf, W. L.; An, N. T. P.; Malvika, S.; Karthik, R.; Lars, M. B., Machine Learning Applications for Mass Spectrometry-Based Metabolomics. *Metabolites* **2020**, *10* (6), 243.
71. Cao, W.-Q.; Jiang, B.-Y.; Huang, J.-M.; Zhang, L.; Liu, M.-Q.; Yao, J.; Wu, M.-X.; Zhang, L.-J.; Kong, S.-Y.; Wang, Y.; Yang, P.-Y., Straightforward and Highly Efficient Strategy for Hepatocellular Carcinoma Glycoprotein Biomarker Discovery Using a Nonglycopeptide-Based Mass Spectrometry Pipeline. *Analytical Chemistry* **2019**.
72. Hua, D.; Desaire, H., Improved Discrimination of Disease States Using Proteomics Data with the Updated Aristotle Classifier. *Journal of Proteome Research* **2021**.
73. Hua, D.; Liu, X.; Go, E. P.; Wang, Y.; Hummon, A. B.; Desaire, H., How to Apply Supervised Machine Learning Tools to MS Imaging Files: Case Study with Cancer Spheroids Undergoing Treatment with the Monoclonal Antibody Cetuximab. *Journal of the American Society for Mass Spectrometry* **2020**, *31* (7), 1350-1357.
74. Hua, D.; Patabandige, M. W.; Go, E. P.; Desaire, H., The Aristotle Classifier: Using the Whole Glycomic Profile To Indicate a Disease State. *Analytical Chemistry* **2019**, *91* (17), 11070-11077.
75. Desaire, H.; Hua, D., Adaption of the Aristotle Classifier for Accurately Identifying Highly Similar Bacteria Analyzed by MALDI-TOF MS. *Analytical Chemistry* **2020**, *92* (1), 1050-1057.
76. Vabalas, A.; Gowen, E.; Poliakoff, E.; Casson, A. J., Machine learning algorithm validation with a limited sample size. *PloS one* **2019**, *14* (11), e0224365-e0224365.

Chapter 2 Leveraging R for fast analysis of mass spectrometry data with machine learning

2.1 Abstract

Applying machine learning strategies to interpret mass spectrometry data has the potential to revolutionize the way in which disease is diagnosed, prognosed, and treated. A persistent and tedious obstacle, however, is relaying mass spectrometry data to the machine learning algorithm. Given the native format and large size of mass spectrometry data files, preprocessing is a critical step. To ameliorate the challenging steps that hinder mass spectrometrists incorporating machine learning strategies into their workflows, we sought to create an easy-to-use, continuous pipeline that runs from data acquisition to the machine learning algorithm.

Here, we present a start-to-finish pipeline designed to facilitate analysis of mass spectrometry data by machine learning. The input can be any ESI data set collected by LC-MS or flow injection, and the output is a machine learning ready matrix, in which each row is a feature (an abundance of a particular m/z), and each column is a sample. This workflow provides automated handling of large mass spectrometry data sets for researchers seeking to implement machine learning strategies but who lack expertise in programming/coding to rapidly format the data. We demonstrate how the pipeline can be used in conjunction with machine learning using two different mass spectrometry data sets: 1) ESI-MS of fingerprint lipid compositions acquired by direct infusion and, 2) LC-MS of IgG glycopeptides. This workflow is uncomplicated and provides value via its simplicity and effectiveness.

2.2 Introduction

The value of machine learning is best realized with large amounts of data; thus, a prime data type for machine learning is generated by mass spectrometry experiments. Applying machine learning strategies to mass spectrometry has yielded many advancements in the realm of human health; early detection of cancer,¹⁻³ clinical decision support,^{2, 4-6} monitoring treatment response,^{7, 8} facilitating the discovery of novel drugs,^{9, 10} identifying microbial strains and screening for antibiotic resistance,¹¹ and classifying single-cell types.¹² A significant challenge, however, is the reproducible and comprehensive transfer of the data from the mass spectral files to the machine learning algorithm. In most of the above-mentioned cases, researchers first select a class of compounds of interest within the sample, identify and quantify them, then build data sets that are amenable to machine learning. But this process requires that researchers know in advance which peaks to select for analysis. Alternatively, all the MS data can be extracted for study, without identifying compounds of interest *a priori*. Due to the sheer amount of data, the required memory, and the need for interpretable results, mass spectrometrists have struggled to implement machine learning strategies into their workflow.¹³ Preprocessing methods tend to omit large parts of the valuable data, often using peak picking to reduce the number of features and increase interpretability.^{12, 14} By this omission, cryptic patterns and slight, nevertheless important, changes between sample types can be lost and the purpose of machine learning is defeated. If the goal is to detect subtle differences between highly similar samples (i.e., healthy vs. early-stage disease) in a high-throughput manner, a pipeline for mass spectrometry data from spectral files to a ML-ready format could be preferable in contrast to doing learning on a vastly slimmed-down data set. To support mass spectrometrists in implementing machine learning into their workflows, we developed a start-to-finish pipeline to relay hundreds of mass spectral files

from their native format to a ML-ready format in a matter of minutes using a binning approach where every peak in the mass spectrum is included in the data matrix.

The functionality of the tool we developed herein is most similar to XCMS,¹⁵ but it differs in many notable ways. We expect that in many cases, our approach will provide a significant benefit to a fraction of the MS community that wants a rapid solution to their data formatting problems. XCMS has functionality to align LC-MS data by retention time, identify molecular features within each LC-MS chromatogram, and export the identified features into a data matrix, which can then be used for machine learning. However, each of these steps requires its own code input, making the XCMS package a set of functions accessible to experienced programmers, rather than a tool designed for mass spectrometry experts (who have beginner skills in programming) to readily use. Furthermore, the approach that XCMS uses to build its data matrices is fundamentally different than what is described here; in the former tool, the product attempts to define “features”, which are compounds with a unique mass and retention time. This approach requires the chromatograms and spectra be aligned in both the time and m/z dimensions. As a radically simpler tool, LevR simply defines bins in the m/z domain, and each of these bins becomes a feature in the data matrix; no spectral alignment is done in advance, as the tool is predominantly envisioned to be used on either direct infusion experiments or LC-MS experiments where a short time segment is chosen for study.

Initially, this pipeline was developed for ESI-MS data of extracted lipids from latent fingerprint samples. Analysis of latent fingerprints by mass spectrometry is an emerging research area showing potential, particularly in the field of forensics¹⁶⁻²³. By taking advantage of the natural chemical changes that occur over time, the age of a fingerprint can be determined with analysis by mass spectrometry²⁴. For example, unsaturated lipid molecules present in sebum are

susceptible to ozonolysis and over time, their amount decreases²⁴⁻²⁷. Additionally, fingerprints may be able to assist law enforcement in developing a profile, as their composition can possibly indicate identifying characteristics like age, sex, and lifestyle^{23, 28-32}. Fingerprints have also been considered for clinical applications³³, such as assays for diagnosing and monitoring metabolic disorders like diabetes^{34, 35}. To harness the full power and potential of fingerprint analysis, machine learning tools need to be incorporated into the analysis workflow.

From a machine learning method development perspective, fingerprints are also an appealing sample type because they are dynamic and heterogenous. They can be used to generate many samples, and their biochemical composition can be modulated by, for example, varying the amount of time exposed to ambient air conditions prior to their extraction into organic solvents. The use of fingerprints enables non-invasive collection of a dynamic biological sample, easy preparation, and a relatively high-throughput MS method using direct infusion. These are ideal characteristics enabling the acquisition of samples that are highly similar with subtle differences,^{36, 37} thereby mimicking the key challenges faced in classification problems today. Following the development of the pipeline for direct infusion mass spectral data, we sought to enhance the approach to also accommodate LC-MS files, which are larger files and have the added complexity of peaks eluting at various retention times; these aspects necessitate significantly higher memory on a computer. After adapting the pipeline, we tested it using a data set of IgG2 glycopeptides that were present in two different forms, a native form and one that was slightly altered via the use of a glycosidase enzyme, to mimic the changes that occur in a glycosylation profile in the beginning stages of disease,³⁷

Here, we present the pipeline and show its utility using two different data sets. The output is compatible with machine learning strategies, like the Aristotle Classifier,^{7, 36-38} which makes

use of the many features within a spectrum that can all contribute to identifying a disease state. This tool will serve mass spectrometrists who have previously lacked accessibility to apply machine learning strategies to their data sets. LevR will enable enhanced data analysis and advance mass spectrometry research as a means for improving human health.

2.3 Experimental Methods

2.3.1 Fingerprint Samples:

2.3.1.1 Fingerprint Collection and Preparation

The collection and preparation of fingerprint samples was performed by adapting previously described methods.²⁴⁻²⁹ A single donor was used, and prior to fingerprint deposition, the donor swiped her fingertips over regions of the face that typically have high sebum secretion prior to depositing the fingerprints onto aluminum foil. These groomed fingerprints were collected over a series of days, limited to 6 fingerprint deposits (3 from each hand) per collection period, where two collection periods occurred ~ 1 hour apart. For each collection period, half of the samples were prepared immediately, while the other half were placed on a large watch glass for 24 hours on the lab bench, exposed to ambient air.

Immediately after fingerprint deposition or after the 24-hour aging period, the aluminum foil squares containing the fingerprints were rolled loosely using clean tweezers and placed into individual 2 mL screw thread sample vials with PTFE closure. 200 μ L dichloromethane was added to each, and the vials were vortexed for 1 minute, followed by 1 minute of rest, and removal of the foil. Then, to each vial, 200 μ L deionized water was added, vortexed for 1 minute, followed by 1 minute of rest, prior to liquid-liquid extraction. The aqueous layer was removed, and the organic layer was kept in the vial with an additional 200 μ L dichloromethane. All samples were stored -20 °C until analysis, such that only one thaw cycle occurred. Gas-tight Hamilton syringes were used throughout the experiment. For analysis, an aliquot of 88 μ L of the

fingerprint sample solution described above was diluted with 500 μ L dichloromethane and 400 μ L NH_4OAc in MeOH to achieve 5 mM ammonium acetate in the final solution.

2.3.1.2 ESI-MS conditions

Direct infusion ESI-MS analysis of the extracted fingerprint lipid samples was performed using an Orbitrap Fusion Tribrid mass spectrometer (ThermoScientific, San Jose, CA). The mass spectrometer was operated in negative ion mode with a sample injection flow rate of 3 μ L/min. The heated-electrospray source was held at 2.3 kV while the ion transfer tube temperature, sweep, aux, and sheath gas flow rates were set at 300 $^\circ\text{C}$, 2, 5, and 10 Arb units, respectively. The full MS scans for the m/z range of (150-600) were acquired in the Orbitrap with a resolution of 60k. The AGC target value for the full MS scan was 5×10^4 , and the maximum injection time was 100 ms. For each sample, 30 scans were averaged for each file. Between analysis of every sample, a methanol/dichloromethane mixture was injected at 10 μ L/min for approximately 10 minutes or until the total ion count had returned to its baseline, established at the beginning of the experiment.

2.3.2 Glycopeptide Samples:

2.3.2.1 Materials and Reagents:

Human serum IgG, ammonium bicarbonate, guanidine hydrochloride (GdnHCl), dithiothreitol (DTT), iodoacetamide (IAM), formic acid and HPLC grade acetonitrile and methanol were purchased from Sigma Aldrich (St. Louis, MO). Sequencing grade trypsin was from Promega (Madison, WI), and α 1-2,3,4,6 fucosidase, 10X glycobuffer (pH 5.5), 100X BSA, was from New England BioLabs (Ipswich, MA). Ultrapure water was obtained from a Direct-Q water purification system (MilliporeSigma, Darmstadt, Germany).

2.3.2.2 Preparation of Native and Partially Defucosylated IgG Tryptic Digests

IgG glycoprotein (160 μ g) was dissolved in 50 mM NH_4HCO_3 buffer at pH 8.0, to give a 4 mg/mL concentrated glycoprotein solution; then, the glycoprotein solution was denatured by

adding GdnHCl (at 6 M final concentration). To reduce the disulfide bonds, DTT was added to the glycoprotein solution to a 10 mM final concentration, followed by sample incubation at room temperature for 1 h. Thereafter, disulfide bonds were alkylated by adding IAM to a final concentration of 25 mM, and this reaction was carried out in the dark, at room temperature for 1 h. After the alkylation step, the excess IAM was neutralized by adding DTT to the reaction mixture (at a 30 mM final concentration), and the reaction was continued for 30 mins at room temperature. The resultant glycoprotein solution was filtered through a 10 kD MWCO filter and buffer exchanged two times with the NH_4HCO_3 buffer at pH 8.0. Subsequently, the glycoprotein concentrate was collected through reverse spin ($1000 \text{ g} \times 2 \text{ min}$) and diluted with the buffer to give a $1 \text{ } \mu\text{g}/\mu\text{L}$ final concentration prior to the trypsin digestion. Then, trypsin was added to the glycoprotein solution at a protein-to-enzyme ratio of 30:1 and incubated for 20 h at $37 \text{ }^\circ\text{C}$. After the trypsin digestion, the pH of the IgG tryptic digest was adjusted to pH 5.5 by using 0.01% formic acid; then, the tryptic digest was filtered through 10 kD MWCO filters to remove trypsin, and the filtrate was collected. The filtrate that contains a mixture of IgG glycopeptides and peptides was aliquoted into two fractions; both aliquots ($67 \text{ } \mu\text{L}$ each) were treated with equal volumes ($7.6 \text{ } \mu\text{L}$ of each) of 10X glycobuffer and 10X BSA, which was diluted from 100X BSA stock solution. To obtain partially defucosylated IgG, α 1-2,3,4,6 fucosidase enzyme ($10 \text{ } \mu\text{L}$) was added to one treated aliquot, while the other fraction was treated with an equal volume ($10 \text{ } \mu\text{L}$) of 10X glycobuffer to obtain a native (control) sample. Both aliquots were incubated at $37 \text{ }^\circ\text{C}$ for 1 week. The aliquots were filtered through 10 kD MWCO filters separately, to remove BSA and fucosidase enzyme. Then, the filtrates were collected and acidified with 0.1% FA. Both aliquots-native and partially defucosylated- were diluted to result in IgG glycopeptide stock solutions of concentration $0.9 \text{ } \mu\text{g}/\mu\text{L}$ and were then stored at $-20 \text{ }^\circ\text{C}$ prior to analysis.

2.3.2.3 Preparation of Native and Mixed Samples for Analysis

Native IgG glycopeptide samples at 0.1 $\mu\text{g}/\mu\text{L}$ were prepared by simply diluting the 0.9 $\mu\text{g}/\mu\text{L}$ IgG native glycopeptide stock solution, prepared in the previous section, with deionized water. The IgG partially defucosylated glycopeptide stock solution, also prepared in the previous section, was diluted three-fold with deionized water to obtain a stock solution at 0.3 $\mu\text{g}/\mu\text{L}$. Then, appropriate volumes of this solution (0.3 $\mu\text{g}/\mu\text{L}$) and the original IgG native glycopeptide stock solution (at 0.9 $\mu\text{g}/\mu\text{L}$) were mixed to generate IgG 20% defucosylated sample, with a final glycopeptide concentration of 0.1 $\mu\text{g}/\mu\text{L}$.

2.3.2.4 Liquid Chromatography-Mass Spectrometry Analysis of IgG Glycopeptide Samples

IgG glycopeptide samples were separated in a reverse phase C18 capillary column (3.5 μm , 300 μm i.d. \times 10 cm, Agilent Technologies, Santa Clara, CA) connected online to a Waters Acquity high performance liquid chromatography system (Milford, MA) followed by mass spectrometric (MS) data acquisition using an Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific, San Jose, CA). For each run, 3 μL of sample volume was injected into the C18 column with a mobile phase flow rate of 10 $\mu\text{L}/\text{min}$. A gradient elution was performed to separate IgG glycopeptides with mobile phase A and mobile phase B; mobile phase A consists of 99.9% of water with 0.1% formic acid while the mobile phase B consists of 99.9% acetonitrile with 0.1% of formic acid. The gradient included column equilibration by running 5% of mobile phase B for 3 mins, followed by linear increase of B from 5% to 20% in 22 min to separate the glycopeptides. Then B was ramped to 90% in 20 min for glycopeptide elution, followed by decrease of B to 5% in 5 min, and re-equilibrating the column at 5% B for another 10 mins.

2.3.2.5 Mass Spectrometry (MS) Conditions

Electrospray ionization (ESI)-MS in the positive ion mode with a heated ion source, which was held at 2.3 kV was used. The temperature of the ion transfer tube and the vaporizer

was set as 300 °C and 20 °C, respectively. Full MS scans were acquired with the Orbitrap resolution at 60 k (at m/z 200) and the scan range was set at m/z range of 400 – 2000. The AGC target and the maximum ion injection time were set at 4×10^5 and 50 ms, respectively. Data dependent MS/MS data were acquired to confirm the glycopeptide compositions; collision-induced dissociation (CID) data were collected by selecting the first five most abundant peaks from the full MS run. CID spectra were collected in the ion trap with a rapid scan rate, exclusion duration was set at 30s with a repeated count of one. For CID, AGC target of 2×10^3 and maximum injection time of 300 ms was used. Furthermore, during the MS/MS data acquisition, 2 Da isolation width was used for parent ion selection, and the selected precursor ions were fragmented by applying 35% of collision energy for 10 ms.

The data were acquired on two different days over a period of three weeks. For group 1 (IgG native glycopeptides) and group 2 (IgG 20% defucosylated glycopeptides) samples, a small data set with five sample runs for each group were acquired on the first day. Blank runs were included in between each sample run. A larger data set was acquired 3 weeks later, where 14 sample runs were included for each group, and blank runs were performed after each pair of sample runs.

2.3.3 .RAW file handling

The data, in .RAW format, was converted to .MS1 files using RawConverter (Scripps, Version 1.2.0.1).³⁹ The settings used were the default selections after launching the software. The number of decimal places was set to match the output from the mass spectrometer. Once the files were in .MS1 format, they were relocated to a single folder in the working directory. This conversion process was the same for both data sets.

2.3.4 Pipeline construction

The pipeline was built to run in R, and all code is confirmed to function in RStudio (version 1.4.1106-5) and R (version 4.0.3).⁴⁰ The pipeline relies on the following packages to function: here,⁴¹ tidyverse,⁴² readr,⁴³ dplyr,⁴⁴ data.table,⁴⁵ and ggplot.⁴⁶ These dependencies are included in the script to be installed and loaded.

2.3.5 Description of binning method

Software Overview: The code used for all analyses in this manuscript is in Appendix A (ESI-MS data) and Appendix B (LC-MS data) at the end of the chapter. The entire text should be copied and pasted into the RStudio IDE as an RMarkdown (.Rmd) file. Included are basic operating instructions and guidelines. The script has six key sections: 1) reading in the data files, 2) cleaning up the files, 3) compiling all data from all files in a single list, 4) creating bins whose size is specified by the user, 5) binning all data, and 6) outputting the binned data in a matrix format. From this output, the data can be submitted to the Aristotle Classifier or other analysis methods, like PCA. A descriptive overview of each component follows, as well as suggestions for appropriate parameters to input.

Housing the files: A file folder within the working directory in the R environment should contain all .MS1 files the user intends to use during the experiment. Each file must contain at least m/z values and their corresponding peak intensities and/or relative abundances; however, additional information, such as scan headers, can also be present in the text files, and they will not interfere. This script is written specifically to process the standard output from RawConverter, which leaves header information in the file. The lines at the header, and between each scan are removed during file processing. For optimal machine learning results, the data housed in any single folder should have identical acquisition parameters, including the m/z range,

resolution, and other parameters described below. This ensures that the data's variability is not an artifact of a difference in experimental conditions.

Adjusting parameters: After the user has moved the data files into a folder within the working directory, the parameters specific to the experiment are entered. When the user opens the RStudio window and opens the .Rmd file, a Knit button with an arrow will appear on the top bar above the script window; in the dropdown menu, the user selects "Knit with Parameters". A graphical user interface (GUI) then appears that is self-explanatory, requiring no programming or coding experience to operate. Parameters that are data-specific can be input, like the m/z range, the number of empty observations allowed for any given feature, and the bin width. The input parameters used in the experiments herein are reported in the specific settings section, below. After all the parameters are set as desired and the MS files are present in the working directory, the user selects the "Knit" function and the software script will proceed to produce the requested data matrix.

2.3.6 Specific settings used for fingerprint samples

The settings used for the analyses in this manuscript were as follows: 25% empty cells allowed, 20 lines in header, Lower m/z : 150, Upper m/z : 600, Bin width: 0.0125 Da.

2.3.7 Aristotle Classifier settings and submission to the Aristotle Classifier

The output matrix generated by LevR in the previous section was modified by the addition of a row of 1's to last row of the matrix, as required by the Aristotle Classifier.³⁷ K (repeats) value was set to 1000, and X value was set to 6.

2.3.7.1 Extracting features by high scores

After analyzing the fingerprint and glycopeptide data with the Aristotle Classifier, the highest-contributing features were identified. To do this, the absolute value of each feature score for each sample was extracted. Then, the total score for each feature was calculated by summing

by row. This gives the total magnitude each feature contributed to distinguishing the samples. Next, the features were sorted in descending order.

This process can be particularly useful for cases in which no a priori knowledge of the samples exists. Using the process outlined here, the feature scores can be extracted from the Aristotle Classifier; they can then be used retroactively to determine which features best distinguish between the samples.

2.3.7.2 Workflow accommodation for LC data

The original pipeline was modified to accommodate LC data- by the simple addition of two lines of code- to handle the significantly larger data files and dictate a narrow retention time range.

2.3.8 Specific settings for glycopeptide samples

The settings used for the analyses in this manuscript were as follows: 50% empty cells allowed, 20 lines in header, Lower m/z : 800, Upper m/z : 2000, Bin width: 0.10 Da., Retention time start: 21.3, Retention time end: 22.6.

2.3.9 Using the Aristotle Classifier to classify samples

The binned data, in the matrix format output from the LevR pipeline, was submitted to the Aristotle Classifier,⁴⁷ after the addition of a row of 1's to the last row of the matrix. The parameters were K value (repeats)=1000, and X value=4.

2.3.10 Identification of features associated with glycopeptides

A table of possible IgG glycopeptides- both native and partially defucosylated- was built. Included were the glycan composition and the theoretical m/z values for the first 8 isotopic peaks expected to appear in the spectrum. Bins were created to capture each m/z value present in the table, then, the data from the glycopeptide experiment were binned according to m/z value. Only the data that fit within the bins (associated with glycopeptides) were retained. This subset of data only contains data from the original matrix whose m/z values fit into glycopeptide bins.

2.3.11 Classification of samples using subset of data

Only the data associated with the glycopeptides was retained, which was then submitted to the Aristotle Classifier. The parameter inputs were not changed from the previous classification of the same data set.

2.3.12 Using PCA as a comparison

The factextra⁴⁸ package was used to generate all PCA plots in this work.

2.4 Results and Discussion

2.4.1 Overview and Interface

The overall goal of this research is to develop a pipeline for performing supervised classification and other machine learning techniques on ESI-MS data. While we^{37, 47} and others^{2, 23, 49} have already demonstrated that machine learning on ESI-MS data is possible, and indeed, quite useful, one of the major bottlenecks is processing the mass spectral data files into a data matrix, which is a prerequisite for applying these advanced mathematical techniques to the data. Normally, the data matrix is developed by users who first identify interesting features in their MS data and then quantify the relevant peaks in each of the samples. For example, we identified all the glycopeptides for IgG from two different glycosylation states then quantified each relevant glycoform across a set of samples prior to machine learning. While this approach was effective for generating a data set that could be classified by machine learning tools, the data set generation process is laborious and has inherent limitations. Alternatively, particularly in the field of metabolomics, many researchers turn to existing open-source software like XCMS, which can build a machine-learning ready data matrix from the mass spectrometry files. Yet, learning to correctly use and apply this complex academic software, which does not come with user manuals, requires a considerable up-front time investment. Furthermore, we aimed to retain all of the data, avoiding feature identification as is used in tools like XCMS. We envisioned an

alternative route forward, where the data matrix preparation could be done in a single step, after users selected a few parameters from a graphical user interface; this process would require little to no time investment. The resulting data matrix would be generated containing all the samples of interest and all the mass spectral peak intensities for those samples. If such a tool could be developed, researchers from a variety of backgrounds could focus on the analysis and machine learning questions that interest them, without having to invest their efforts into the data extraction and formatting aspects of the process.

The data formatter we developed is simply called LevR; its approach to processing the MS data and the GUI that controls it are shown in Figures 1 and 2. The mass spectral data are used directly to populate a data matrix, where each column in the matrix is a sample, a single mass spectrometry data file, and each row in the matrix is a feature. Each sample and feature pair contains the sum of the peak abundances that appear in a narrow slice (m/z bin) of the mass spectrometry data. For example, the mass spectrometry data could be binned to include features for each 0.1 Da present in the spectra, as shown in Figure 2-1a. In this case, a portion of the mass spectrum that covers a range of 1.3 Da is represented by 13 bins, and four of the bins are populated with peaks. Figure 2-1b shows the data for the sample in Figure 2-1a, populating the first column in the data table. In this case, since only 4 peaks were present in the spectrum, only four of the features (m/z bins) are populated with numerical data. The tool also has the capacity to remove bins that are not populated by a certain percentage of the samples; this parameter is fully adjustable by the user. Furthermore, while the trivial example in Figure 2-1 shows the processing of just a single spectrum, and the subsequent processing of 4 other samples (spectra not shown), the script additionally processes as many high-resolution scans as the user chooses – either all the scans in the data file or all the scans in a selected elution range.

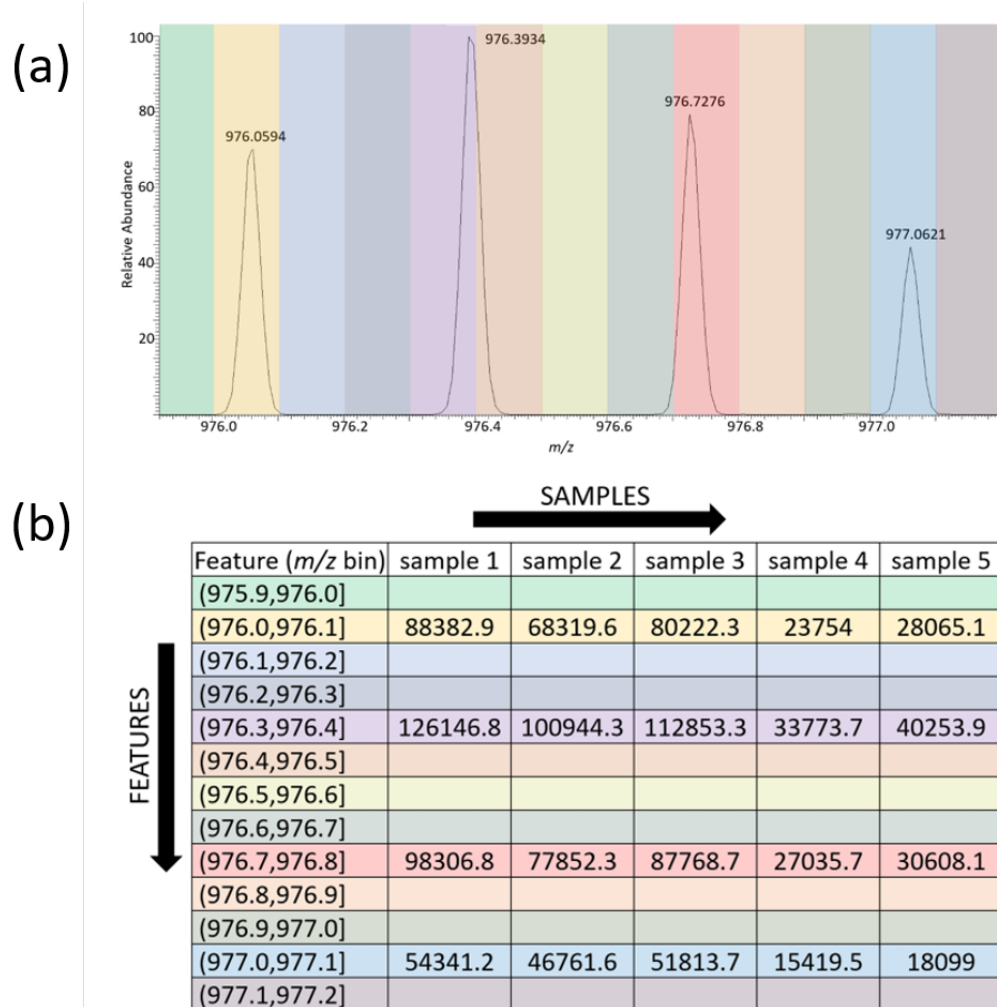


Figure 2-1. Visual depiction of the binning process. (a) 2 Da. range of spectrum from glycopeptide data set, with bin width set to 0.1 Da. Each colored slice of the spectrum represents a bin. (b) Data table depicting how data is arranged by LevR. The m/z value is the experimental m/z value from the spectrum. The feature is the narrow m/z range (bin) assigned to the experimental observation. Each sample occupies a column, and each sample- m/z pair contains the intensity of the m/z peak from the spectrum.

Figure 2-2 shows the interface the users see. The name of the folder with the data present is input, along with the mass range desired, the bin width, and the percent of empty bins allowable. After selecting the desired conditions, the software builds the data matrix of interest.

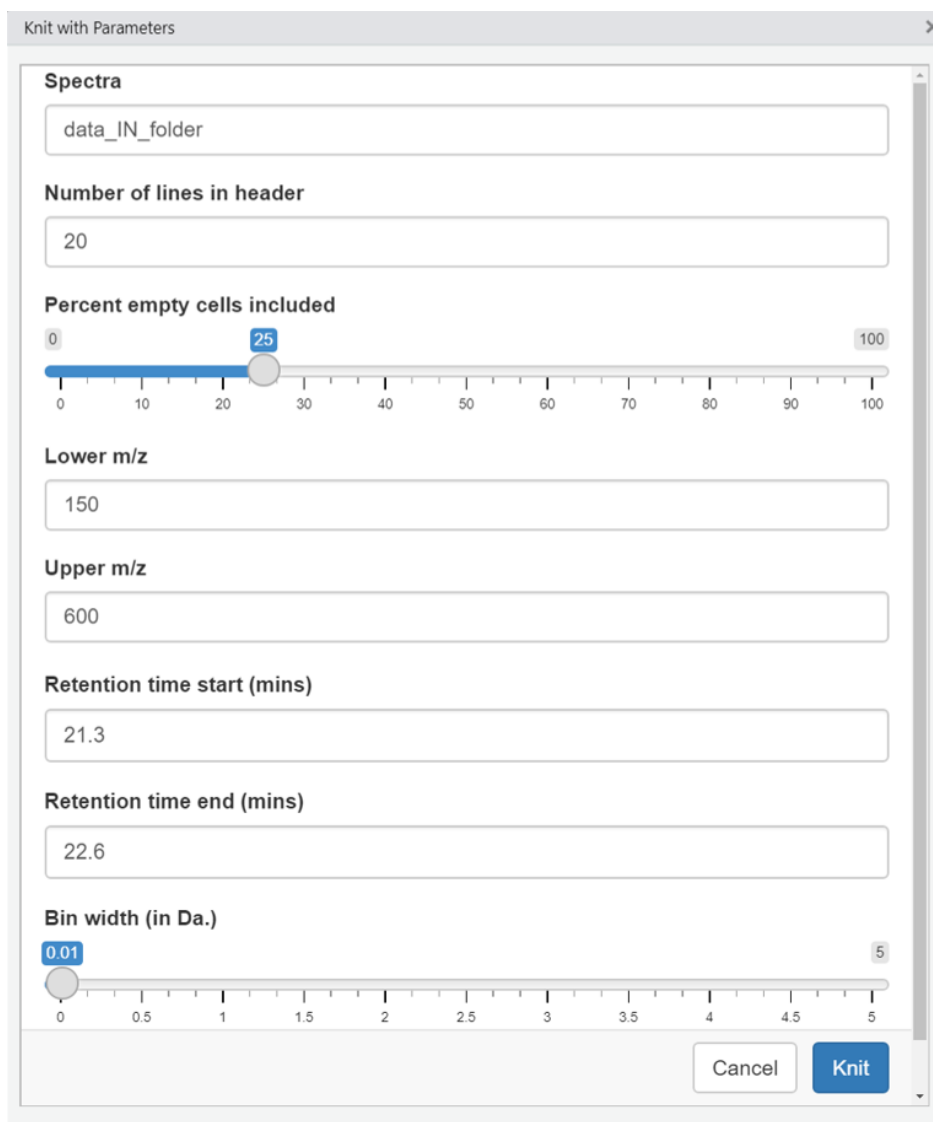


Figure 2-2. Graphical User Interface (GUI) from LevR.

2.4.2 Test set one: Fingerprints.

To test the utility of this approach for generating useful machine-learning ready data sets, we developed a challenge data set in-house by acquiring mass spectrometry data of fingerprints that had been subjected to two different storage conditions. While all the fingerprints were deposited onto aluminum foil, half the foil samples were immediately subjected to extraction with organic solvent. The other half of the samples were left to sit for 24 hours prior to extraction. Previous researchers²⁴⁻²⁷ have indicated that some of the lipids in the fingerprints that

are not immediately processed can undergo chemical changes, and this difference causes changes in a few peaks' intensities in the mass spectrometry data. The fingerprint samples for the data set were acquired over numerous days and the MS data was acquired in two separate analyses more than a week apart. No effort was made to control other variables that may impact the lipid distribution, such as the depositor's diet or exercise status or the laboratory conditions (e.g., heat, humidity, light). This was intentional so that the data would be sufficiently variable and challenging to classify. We sought to know whether it would be possible to classify the fingerprints' age by simply extracting the full mass spectral data, binning it, as described above, and conducting machine learning on the output matrix. If the classification were feasible in this paradigm, this outcome would demonstrate that the difficult up-front work of identifying the changing compounds may be eliminated. Furthermore, it would show that LevR could be applicable to a variety of other problems where researchers do not know whether a successful classification would be possible with their samples. This tool would enable screening of data for good classification outcomes prior to going through the laborious process of identifying the features that *might* be useful.

The data in Figure 2-3a clearly show that fingerprint age can be determined with a reasonable degree of accuracy using the data sets generated by LevR and classified by the Aristotle Classifier, a new machine learning tool developed by our group. In Figure 2-3a, the output data from the Aristotle Classifier shows that a total of 70 samples were classified and about 85% were correctly assigned to their group. Using a leave-one-out classification method, so test samples are never included in their training set, most of the (aged) samples, which are the first 35 samples shown, have Results of greater than zero, indicating that they are assigned to the aged group. By contrast, the non-aged samples, which range from Sample number 36 to 70 in

the data set, mostly receive Results of less than zero, indicating that they are part of the non-aged group. A minority of the samples, which appear in red quadrants, were misassigned.

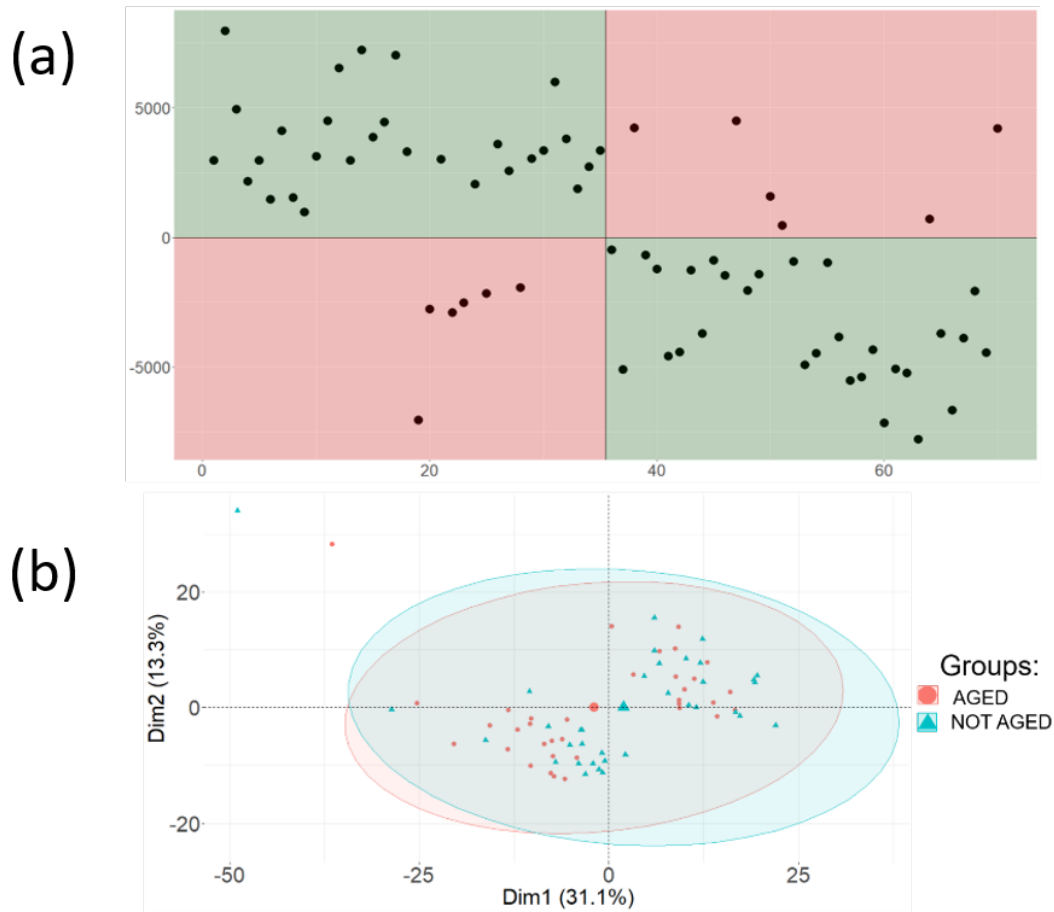


Figure 2-3. Comparison of Aristotle Classifier and PCA results for the fingerprint data set. (a) Results from the Aristotle Classifier for 70 fingerprint samples: 35 of each group (not aged and aged). Correctly classified samples are highlighted in the green quadrants. (b) PCA results of the same 70 fingerprint samples from panel a.

The data in Figure 2-3b show a PCA plot of the same data used for supervised classification in Figure 2-3a. In this case, the two sets of samples, which are colored either orange or blue, are completely intermixed on the PCA plot. This Figure indicates that the difference imparted by leaving the samples out on a benchtop for a day was a small and difficult-to-detect difference, and other attributes contribute significantly more to the variability within the data. The samples would have separated into their respective groups (aged or not aged) had

the difference in the samples due to the aging process been one of the most significant contributors to sample variability. Rather, the first two principal components represent more than 50% of the variability within the samples, and this variability is not attributable to the two different sample types.

In summary then, the simple data processor, LevR, was useful for rapidly rendering a data matrix for 70 different lipidomics samples from deposited fingerprints. By coupling this software with a new machine learning tool, the Aristotle Classifier, the samples could be discriminated as either being aged or immediately processed, with reasonable accuracy (~83%), even though the target differences in the sample were minor compared to other properties that contributed to the samples' variability and features were not pre-selected for classification. This proof of concept, therefore, demonstrates the possibility of performing supervised machine learning directly on the full mass spectrometry data file for samples acquired by direct infusion experiments, without first identifying peaks of interest and quantifying them across a sample set.

2.4.3 Test set two: Glycopeptides.

In a second analysis challenge, LC-MS data of glycopeptides from IgG were interrogated. In this case, the classification challenge was to determine whether the IgG glycoforms matched a native glycosylation profile or a non-native form, which was intentionally generated in the laboratory by modifying IgG with fucosidase, an enzyme that trims fucose off the IgG glycans. More details describing the samples and their preparation are in the Experimental section. Again, the full mass spectral data including the elution window for the IgG glycoforms was used to build the data matrix, but only a small number of peaks within the data set carry the information content necessary to distinguish the two groups: Any bin that did not include peaks corresponding to glycopeptides would be uninformative. The data set contains 12,000 features (each corresponding to a 0.1 Da bin), in which only 120 could be associated with glycopeptides

by our parameters (15 glycopeptides x the first 8 isotopic peaks); thus, the vast majority of the features would not be useful for classification. So, we again sought to determine whether extracting the MS data over the entire elution window for the glycopeptides, without including an identification step where the potentially relevant features were selected first, would lead to a viable data set that could be classified correctly.

The data in Figure 2-4 show the results of supervised (2-4a) and unsupervised (2-4b) classification of this data set. Figure 4a clearly shows that classification with the Aristotle Classifier was successful, and about 90% of the samples are correctly classified as either possessing a native or modified glycosylation profile. Likewise, the data in Figure 4b show that, as expected, the glycosylation difference is not the factor that generates most of the variability within the sample set. A plot of the first two principal components shows no ability to distinguish the native (blue) samples from the non-native (orange) ones. The fact that the samples were not readily separable by their principal components in Figure 2-4b is not surprising because the change in glycosylation was subtle, and the vast majority of the features in the data set did not correspond to glycopeptide masses. Even considering the fact that the glycosylation difference is slight and that only a fraction of the peaks in the data set were impacted by this difference, the combined workflow of first extracting all of the MS data using LevR and then subjecting it to the Aristotle Classifier shows promise for machine learning applications on MS data, even in the case where the differences in the data set are subtle and lurking in a background of many uninformative peaks.

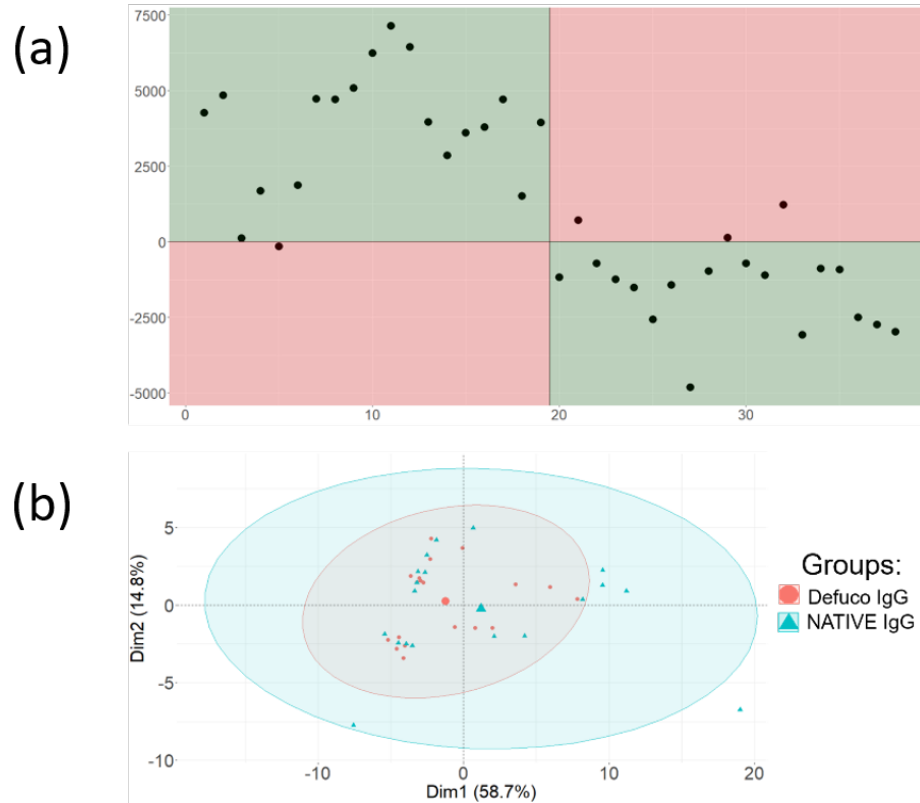


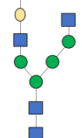
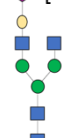
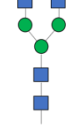
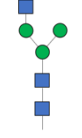
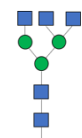
Figure 2-4. Comparison of Aristotle Classifier and PCA results for the full IgG glycopeptide data set. (a) Results from the Aristotle Classifier for 38 IgG glycopeptides: 19 of each group (native and partially defucosylated). Correctly classified samples are highlighted in the green quadrants. (b) PCA results of the same 38 IgG glycopeptide samples from panel a.

The results in Figure 2-4a are exciting, but a logical next question is: could we have done better at classifying the data by limiting the analysis to only the glycopeptide peaks? Supervised learning methods, like the Aristotle Classifier, achieve their enhanced predictive power over unsupervised methods, like PCA, by first determining the features that best discriminate the two states. Those features are then weighted more heavily than others in the final classification, although too many uninformative peaks can negatively impact the model's performance. We wanted to determine whether the classification would have been more successful had those uninformative features been removed in advance. Furthermore, we sought to verify that the

glycopeptide peaks were, in fact, the ones that had been selected by the classifier as the “important features” in the classification shown in Figure 4a.

Determining which m/z regions in the spectra were weighted most heavily in the resulting classification is straightforward: The version of the Aristotle Classifier used for this work, AC.2021,⁴⁷ includes a built-in matrix called FeatureScore, which includes how each feature was weighted for the final result score for each sample. FeatureScores can be positive, if the feature indicates the sample of interest is more like one sample type or negative, if the feature indicates that the sample is more like the alternative sample type. Therefore, to determine which features most impacted the classifier’s weightings overall, the absolute values of the feature scores were summed across the sample set. The resulting data is shown in Table 1- showing 19 of the top 20 features were associated with IgG glycoforms. For each of the IgG-related features, the relevant glycoform, FeatureScore, and bin are included. This result indicates that the embedded feature selection and weighting component of this particular classifier is effective at identifying the relevant features in the presence of many uninformative ones.

Table 1. Top 20 highest scoring features (*m/z* bins) as determined by the Aristotle Classifier. Within each glycoform section, features are ordered from highest to lowest scores. All features but one matched to an expected glycoform. Note, the *m/z* value includes the IgG2 peptide (EEQFNSTFR).

rank	<i>m/z</i> bin	feature score	glycan composition
1	(1024.73,1024.74]	20274	[Hex]5[HexNAc]4[NeuAc]1
2	(1025.07,1025.08]	19408	
4	(1024.4,1024.41]	17168	
8	(1025.4,1025.41]	11986	
10	(1024.74,1024.75]	11128	
3	(970.38,970.39]	17558	[Hex]4[HexNAc]4[NeuAc]1
7	(970.72,970.73]	12620	
16	(971.05,971.06]	6688	
17	(970.71,970.72]	6662	
6	(1229,1229.01]	14866	[Hex]3[HexNAc]4
11	(1228.5,1228.51]	10412	
13	(1229.5,1229.51]	8284	
18	(820.34,820.35]	6456	
19	(819.33,819.34]	5404	
20	(820,820.01]	5226	
9	(1126.96,1126.97]	11620	[Hex]3[HexNAc]3
15	(1127.46,1127.47]	6718	
12	(887.36,887.37]	9556	[Hex]3[HexNAc]5
14	(887.7,887.71]	7684	
5	(900.41,900.42]	16620	unidentified

But could the classification be more successful if only the glycopeptides had been included in the first place? To answer this question, we first identified all the relevant *m/z* bins that would contain glycopeptide peaks, as described in the experimental section, and reclassified the data using only those features. The results appear in Figure 2-5a, for supervised classification, and 2-5b, where the unsupervised PCA plot is provided. The PCA plot clearly shows that removing all the bins that do not contain glycopeptide information reduces the overall

variability in the data, and the two sample types, natively glycosylated or modified, are now somewhat separable using this unsupervised method.

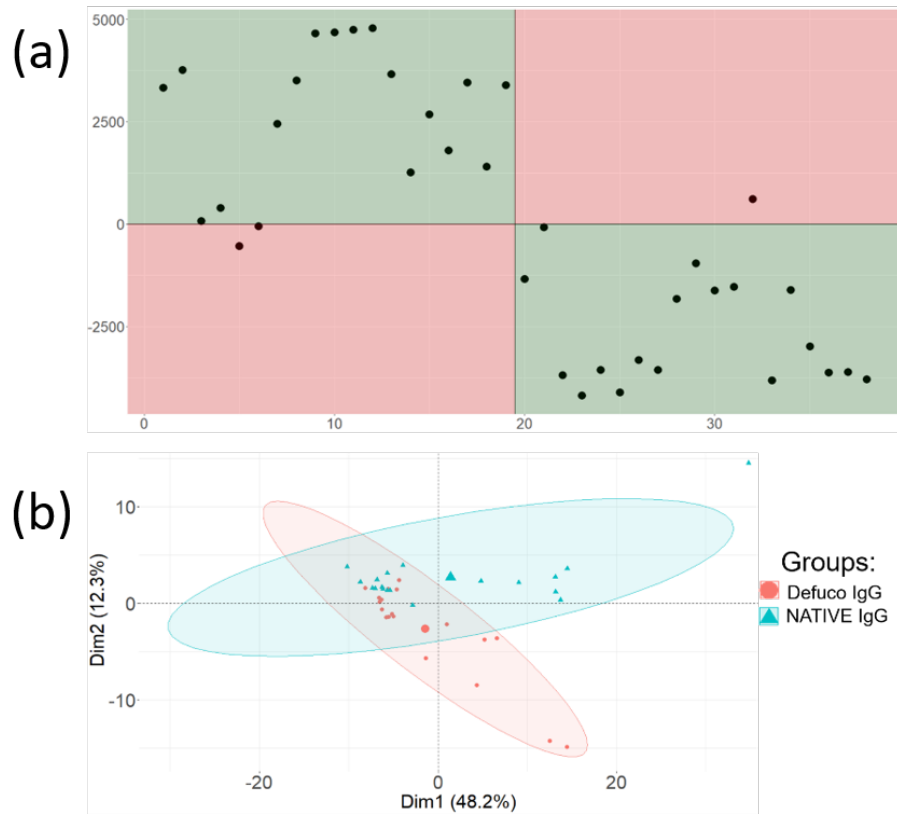


Figure 2-5. Comparison of Aristotle Classifier and PCA results for the refined IgG glycopeptide data set, including only features associated with glycopeptides. (a) Results from the Aristotle Classifier for 38 IgG glycopeptides data after removing all non-glycopeptide associated features. The misclassification rate did not change. The magnitude of the Y axis decreased slightly; this is due to a data set with reduced number of features. (b) PCA results for the refined IgG glycopeptide data set.

This outcome is consistent with the well-known principal that removing uninformative features generally improves one's ability to discriminate the different biological states. Yet, the data in Figure 5a, showing the supervised classification of this data set with reduced features, is essentially identical to the result obtained in Figure 4a, where 12,000 uninformative features were still present in the data set. The fact that Figure 2-4a and Figure 2-5a look so similar is a desirable result: It unequivocally demonstrates that, when using the right kind of classifier, the

data set need not be preprocessed to remove unnecessary, uninformative features. Rather, both the glycopeptide example and the fingerprint example in Figure 2-3, show that when a measurable difference is present in two different sample groups, machine learning and mass spectrometry can be exploited to identify that difference and classify samples into their respective groups, using the straight-forward workflow shown here.

2.5 Conclusion

The combined functionality of LevR and the Aristotle Classifier yields exciting results for mass spectrometrists and researchers studying biomarkers. LevR is a plain, yet effective, solution for formatting large amounts of mass spectrometry data. Its coupling to the Aristotle Classifier, a new machine learning tool, results in a powerful workflow that can be accessed by all researchers regardless of coding experience. We imagine its application for biomarker discovery, in which biological samples can be analyzed by mass spectrometry, the data is formatted automatically, and the classifier renders results to indicate if there are detectable differences between the healthy and disease state biological samples. Further, the classifier's results can be leveraged to identify which features contribute most to the difference between sample types. We anticipate LevR will be useful in advancing biomarker discovery to the point of implementation in a clinical setting.

2.6 Acknowledgements

This work was supported by NIH Grant R35GM130354 to HD. The IgG2 data set was prepared and acquired by Dr. Milani Wijeweera Patabandige.

2.7 References

1. Huang, Y.-C.; Chung, H.-H.; Dutkiewicz, E. P.; Chen, C.-L.; Hsieh, H.-Y.; Chen, B.-R.; Wang, M.-Y.; Hsu, C.-C., Predicting Breast Cancer by Paper Spray Ion Mobility Spectrometry Mass Spectrometry and Machine Learning. *Analytical chemistry (Washington)* **2020**, *92* (2), 1653-1657.
2. Sho, K.; Kentaro, Y.; Junichi, A.; Takashi, K.; Hiroyuki, H.; Meguri, T.; Takeaki, I.; Nobuhisa, A.; Junichi, K.; Sen, T.; Kiyoshi, H., A new rapid diagnostic system with ambient mass spectrometry and machine learning for colorectal liver metastasis. *BMC cancer* **2021**, *21* (1), 1-9.
3. Manzi, M.; Palazzo, M. n.; Knott, M. a. E.; Beausery, P.; Yankilevich, P.; Giménez, M. a. I.; Monge, M. a. E., Coupled Mass-Spectrometry-Based Lipidomics Machine Learning Approach for Early Detection of Clear Cell Renal Cell Carcinoma. *Journal of proteome research* **2021**, *20* (1), 841-857.
4. Mészáros, B.; Járvas, G.; Kun, R.; Szabó, M.; Csánky, E.; Abonyi, J.; Guttman, A., Machine Learning Based Analysis of Human Serum N-glycome Alterations to Follow up Lung Tumor Surgery. *Cancers* **2020**, *12* (12).
5. Acharjee, A.; Prentice, P.; Acerini, C.; Smith, J.; Hughes, I. A.; Ong, K.; Griffin, J. L.; Dunger, D.; Koulman, A., The translation of lipid profiles to nutritional biomarkers in the study of infant metabolism. *Metabolomics* **2017**, *13* (3), 25.
6. Zhang, L.; Ma, F.; Qi, A.; Liu, L.; Zhang, J.; Xu, S.; Zhong, Q.; Chen, Y.; Zhang, C.-Y.; Cai, C., Integration of ultra-high-pressure liquid chromatography tandem mass spectrometry with machine learning for identifying fatty acid metabolite biomarkers of ischemic stroke. *Chem. Commun.* **2020**, *56* (49), 6656-6659.
7. Hua, D.; Liu, X.; Go, E. P.; Wang, Y.; Hummon, A. B.; Desaire, H., How to Apply Supervised Machine Learning Tools to MS Imaging Files: Case Study with Cancer Spheroids Undergoing Treatment with the Monoclonal Antibody Cetuximab. *Journal of the American Society for Mass Spectrometry* **2020**, *31* (7), 1350-1357.
8. Zhang, J.; Du, Q.; Song, X.; Gao, S.; Pang, X.; Li, Y.; Zhang, R.; Abliz, Z.; He, J., Evaluation of the tumor-targeting efficiency and intratumor heterogeneity of anticancer drugs using quantitative mass spectrometry imaging. *Theranostics* **2020**, *10* (6), 2621-2630.
9. Barthélemy, M.; Guérineau, V.; Genta-Jouve, G.; Roy, M.; Chave, J.; Guillot, R.; Pellissier, L.; Wolfender, J.-L.; Stien, D.; Eparvier, V.; Touboul, D., Identification and dereplication of endophytic Colletotrichum strains by MALDI TOF mass spectrometry and molecular networking. *Scientific reports* **2020**, *10* (1), 19788.
10. van Oosten, L. N.; Klein, C. D., Machine Learning in Mass Spectrometry: A MALDI-TOF MS Approach to Phenotypic Antibacterial Screening. *Journal of medicinal chemistry* **2020**, *63* (16), 8849-8856.
11. Weis, C. V.; Jutzeler, C. R.; Borgwardt, K., Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review. *Clinical Microbiology and Infection* **2020**, *26* (10), 1310-1317.
12. Xie, Y. R.; Castro, D. C.; Bell, S. E.; Rubakhin, S. S.; Sweedler, J. V., Single-Cell Classification Using Mass Spectrometry through Interpretable Machine Learning. *Analytical chemistry (Washington)* **2020**, *92* (13), 9338-9347.

13. Liebal, U. W.; Phan, A. N. T.; Sudhakar, M.; Raman, K.; Blank, L. M., Machine Learning Applications for Mass Spectrometry-Based Metabolomics. *Metabolites* **2020**, *10* (6), 243.
14. Stanstrup, J.; Broeckling, C. D.; Helmus, R.; Hoffmann, N.; Mathé, E.; Naake, T.; Nicolotti, L.; Peters, K.; Rainer, J.; Salek, R. M.; Schulze, T.; Schymanski, E. L.; Stravs, M. A.; Thévenot, E. A.; Treutler, H.; Weber, R. J. M.; Willighagen, E.; Witting, M.; Neumann, S., The metaRbolomics Toolbox in Bioconductor and beyond. *Metabolites* **2019**, *9* (10).
15. Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G., XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry* **2006**, *78* (3), 779-787.
16. Atherton, T.; Croxton, R.; Baron, M.; Gonzalez-Rodriguez, J.; Gamiz-Gracia, L.; Garcia-Campana, A. M., Analysis of amino acids in latent fingerprint residue by capillary electrophoresis-mass spectrometry. *J. Sep. Sci.* **2012**, *35* (21), 2994-2999.
17. Ifa, D. R.; Manicke, N. E.; Dill, A. L.; Cooks, R. G., Latent Fingerprint Chemical Imaging by Mass Spectrometry. *Science (Washington, DC, U. S.)* **2008**, *321* (5890), 805.
18. Mirabelli, M. F.; Chramow, A.; Cabral, E. C.; Ifa, D. R., Analysis of sexual assault evidence by desorption electrospray ionization mass spectrometry. *J. Mass Spectrom.* **2013**, *48* (7), 774-778.
19. O'Neill, K. C.; Lee, Y. J., Effect of Aging and Surface Interactions on the Diffusion of Endogenous Compounds in Latent Fingerprints Studied by Mass Spectrometry Imaging. *J. Forensic Sci.* **2018**, *63* (3), 708-713.
20. Tang, X.; Huang, L.; Zhang, W.; Zhong, H., Chemical Imaging of Latent Fingerprints by Mass Spectrometry Based on Laser Activated Electron Tunneling. *Anal. Chem. (Washington, DC, U. S.)* **2015**, *87* (5), 2693-2701.
21. Tang, H.-W.; Lu, W.; Che, C.-M.; Ng, K.-M., Gold Nanoparticles and Imaging Mass Spectrometry: Double Imaging of Latent Fingerprints. *Anal. Chem. (Washington, DC, U. S.)* **2010**, *82* (5), 1589-1593.
22. Yagnik, G. B.; Korte, A. R.; Lee, Y. J., Multiplex mass spectrometry imaging for latent fingerprints. *J. Mass Spectrom.* **2013**, *48* (1), 100-104.
23. Zhou, Z.; Zare, R. N., Personal Information from Latent Fingerprints Using Desorption Electrospray Ionization Mass Spectrometry and Machine Learning. *Analytical Chemistry* **2017**, *89* (2), 1369-1372.
24. Pleik, S.; Spengler, B.; Schäfer, T.; Urbach, D.; Luhn, S.; Kirsch, D., Fatty Acid Structure and Degradation Analysis in Fingerprint Residues. *Journal of the American Society for Mass Spectrometry* **2016**, *27* (9), 1565-1574.
25. Hinners, P.; Thomas, M.; Lee, Y. J., Determining Fingerprint Age with Mass Spectrometry Imaging via Ozonolysis of Triacylglycerols. *Analytical Chemistry* **2020**, *92* (4), 3125-3132.
26. Pleik, S.; Spengler, B.; Ram Bhandari, D.; Luhn, S.; Schäfer, T.; Urbach, D.; Kirsch, D., Ambient-air ozonolysis of triglycerides in aged fingerprint residues. *The Analyst* **2018**, *143* (5), 1197-1209.
27. Archer, N. E.; Charles, Y.; Elliott, J. A.; Jickells, S., Changes in the lipid composition of latent fingerprint residue with time after deposition on a surface. *Forensic Science International* **2005**, *154* (2), 224-239.

28. Hinners, P.; O'Neill, K. C.; Lee, Y. J., Revealing Individual Lifestyles through Mass Spectrometry Imaging of Chemical Compounds in Fingerprints. *Scientific reports* **2018**, *8* (1), 5149.
29. O'Neill, K. C.; Hinners, P.; Lee, Y. J., Potential of triacylglycerol profiles in latent fingerprints to reveal individual diet, exercise, or health information for forensic evidence. *Analytical Methods* **2020**, *12* (6), 792-798.
30. Bouslimani, A.; Melnik, A. V.; Xu, Z.; Amir, A.; da Silva, R. R.; Wang, M.; Bandeira, N.; Alexandrov, T.; Knight, R.; Dorrestein, P. C., Lifestyle chemistries from phones for individual profiling. *Proceedings of the National Academy of Sciences* **2016**, *113* (48), E7645.
31. van Helmond, W.; van Herwijnen, A. W.; van Riemsdijk, J. J. H.; van Bochove, M. A.; de Poot, C. J.; de Puit, M., Chemical profiling of fingerprints using mass spectrometry. *Forensic chemistry* **2019**, *16*.
32. Ferguson, L. S.; Wulfert, F.; Wolstenholme, R.; Fonville, J. M.; Clench, M. R.; Carolan, V. A.; Francese, S., Direct detection of peptides and small proteins in fingermarks and determination of sex by MALDI mass spectrometry profiling. *The Analyst* **2012**, *137* (20), 4686-4692.
33. Shetage, S. S.; Traynor, M. J.; Brown, M. B.; Chilcott, R. P., Sebomic identification of sex- and ethnicity-specific variations in residual skin surface components (RSSC) for bio-monitoring or forensic applications. *Lipids Health Dis* **2018**, *17* (1), 194-194.
34. Hyde, J.; Runyon, J. R., LCMS Measurement of Steroid Biomarkers Collected from Palmar Sweat. *ChemRxiv* **2020**, 10.26434/chemrxiv.12931769.v1.
35. O'Neill, K. C.; Hinners, P.; Lee, Y. J., Potential of triacylglycerol profiles in latent fingerprints to reveal individual diet, exercise, or health information for forensic evidence. *Analytical Methods* **2020**, *12* (6), 792-798.
36. Desaire, H.; Hua, D., Adaption of the Aristotle Classifier for Accurately Identifying Highly Similar Bacteria Analyzed by MALDI-TOF MS. *Analytical Chemistry* **2020**, *92* (1), 1050-1057.
37. Hua, D.; Patabandige, M. W.; Go, E. P.; Desaire, H., The Aristotle Classifier: Using the Whole Glycomic Profile To Indicate a Disease State. *Analytical Chemistry* **2019**, *91* (17), 11070-11077.
38. Desaire, H.; Patabandige, M. W.; Hua, D., The local-balanced model for improved machine learning outcomes on mass spectrometry data sets and other instrumental data. *Analytical and Bioanalytical Chemistry* **2021**, *413* (6), 1583-1593.
39. He, L.; Diedrich, J.; Chu, Y. Y.; Yates, J. R., 3rd, Extracting Accurate Precursor Information for Tandem Mass Spectra by RawConverter. *Anal Chem* **2015**, *87* (22), 11361-7.
40. R Core Team *R: A language and environment for statistical computing*, R Foundation for Statistical Computing: Vienna, Austria, 2020.
41. Müller, K. *here: A Simpler Way to Find Your Files*, version 1.0.1. ; 2020.
42. Wickham, H., et al., Welcome to the tidyverse. *Journal of Open Source Software* **2019**, *4* (43).
43. Wickham, H., Hester, J. *readr: Read Rectangular Text Data*, 1.4.0; 2020.
44. Wickham, H., Francois, R., Henry, L., Muller, K. *dplyr: A Grammar of Data Manipulation*, 1.0.5; 2021.
45. Dowle, M., Srinivasan, A. *data.table: Extension of 'data.frame'*, 1.14.0; 2021.

46. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag: New York, 2016.
47. Hua, D.; Desaire, H., Improved Discrimination of Disease States Using Proteomics Data with the Updated Aristotle Classifier. *Journal of Proteome Research* **2021**.
48. Kassambara, A., Mundt, F. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 1.0.7; 2020.
49. Ishii, H.; Sakamoto, K.; Ashizawa, K.; Masuyama, K.; Saitoh, M.; Sakamoto, K.; Saigusa, D.; Kasai, H.; Miyazawa, K.; Takeda, S.; Yoshimura, K., Lipidome-based rapid diagnosis with machine learning for detection of TGF- β signalling activated area in head and neck cancer. *Br J Cancer* **2020**, *122* (7), 995-1004.

2.8 Appendix A: LevR for ESI-MS data

All text below, including dashed lines: paste into a new RMarkdown document.

title: "ESI_MS_Title"

output: html_document

params:

spectra:

value: data_IN_folder ## data should be housed in a single folder within the working directory

head:

label: "Number of lines in header"

value: 20

input: numeric

perc_include:

label: "Percent empty cells included" ## this is the percent of empty observations to allow in the matrix

value: 25

input: slider

min: 0

max: 100

p_spec_low:

label: "Lower m/z"

value: 150

input: numeric

p_spec_hi:

label: "Upper m/z"

value: 600

input: numeric

bin_width:

label: "Bin width (in Da.)" ## bin width can correspond with peak width within the spectra, or as defined by user

value: 0.0125

input: slider

min: 0

max: 5

data should be housed in the working directory in a folder named "data_IN_folder"

```
`` {r eval=FALSE, include=FALSE}
```

```
install.packages("here")
```

```
install.packages("tidyverse")
```

```
``
```

```
`` {r include=FALSE, warning = FALSE, message=FALSE}
```

```
library(tidyverse)
```

```
library(readr)
```

```
library(here)
```

```
library(dplyr)
```

```
library(data.table)
```

```
library(knitr)
```

```
library(stringr)
```

```
library(gapminder)
```

```
``
```

```
`` {r include=FALSE}
```

```
here()
```

```
file_path_spec <- as.vector(list.files(here(params$spectra)))
```

```
bins <- seq(from=params$p_spec_low, to=params$p_spec_hi, by=params$bin_width)
```

```
num_of_samps <- length(file_path_spec)
```

```

myfunction <- function(path) {
  mymsdata <- fread((here(params$spectra, path)), skip=params$head, fill=TRUE)
  ms_df <- cbind(mymsdata[,c(1:2)]) %>%
  filter(str_detect(V1, "Instrument", negate = TRUE)) %>%
  filter(str_detect(V1, "Ion", negate=TRUE))%>%
  filter(str_detect(V1, "S", negate=TRUE)) %>%
  as.data.frame()
}
bigdata <- lapply(file_path_spec, myfunction)
names(bigdata) <- file_path_spec
dt <- rbindlist(bigdata, use.names = TRUE, fill=TRUE, idcol="bigdata")

```

```{r include=FALSE, warning = FALSE, message=FALSE}

my_binned_data <- dt %>%
  mutate(mz_bins=cut(as.numeric(V1), breaks=bins)) %>%
  group_by(bigdata, mz_bins) %>%
  summarise(grand_inten=sum(as.numeric(V2)))

my_df <- as.data.frame(my_binned_data)%>%
  pivot_wider(names_from = bigdata, values_from= grand_inten) %>%
  as.data.frame()

my_df$count_NA <- apply(is.na(my_df), 1, sum)
my_df$percent_NA <-(my_df$count_NA)/num_of_samps*100

```

```
drop_NA <- my_df[my_df$percent_NA <= params$perc_include,]
drop_NA[is.na(drop_NA)] <- 0
my_df_zero <- drop_NA
```

```
names(my_df) <- gsub("\\.ms1.*", "", colnames(my_df))
```

```
my_df_mz <- as.data.frame(na.omit(my_df)) ## my_df_mz displays the mz bins for each
feature, along with the data
```

```
final_df <- my_df_zero[-1] ## final_df only has the data. The mz bins column is removed.
```

```
``
```

2.9 Appendix B: LevR for LC-MS data

All text below, including dashed lines: paste into a new RMarkdown document.

title: "LC_MS_Title"

output:

html_document: default

params:

spectra:

label: Spectra

value: LC_data_IN ##folder where the files are located within the working directory

input: text

p_spec_low:

label: Lower m/z

value: 800

input: numeric

p_spec_hi:

label: Upper m/z

value: 2000

input: numeric

bin_width:

label: "Bin width (in Da.)" ## bin width can correspond with peak width within the spectra, or as defined by user

value: 0.0125

input: slider

min: 0

max: 5

perc_include:

label: "Percent empty cells included" ## this is the percent of empty observations to allow in the matrix

value: 25

input: slider

min: 0

max: 100

RTstart:

label: Retention time start (mins)

value: 21.3

input: numeric

RTend:

label: Retention time end (mins)

value: 22.6

input: numeric

```
`` {r eval=FALSE, include=FALSE}
```

```
install.packages("here")
```

```
install.packages("gapminder")
```

```
install.packages("tidyverse")
```

```
install.packages("knitr")
```

```
install.packages("data.table")
```

```
``
```

```
`` {r include=FALSE}
```

```
library(here)
```

```
library(gapminder)
```

```
library(tidyverse)
```

```
library(readr)
```

```
library(dplyr)
```

```
library(data.table)
```



```

library(stringr)
library(knitr)
```



```

## files should be housed in "LC_data_IN" folder within the working directory
```{r include=FALSE}
here()
file_path_spec <- as.vector(list.files(here(params$spectra)))
bins <- seq(from=params$sp_spec_low, to=params$sp_spec_hi, by=(params$bin_width))
num_of_samps <- length(file_path_spec)
```



```

```{r include=FALSE, eval=FALSE}
## this takes about 5 sec per LC file

myfunction <- function(path) {
  mymsdata <- fread((here(params$spectra, path)), skip=22, fill=TRUE)
  ms_df <- cbind(mymsdata[,c(1:3)])
  clean_1_ms_df <- ms_df %>%
  filter(str_detect(V1, "Instrument", negate = TRUE)) %>%
  filter(str_detect(V1, "Ion", negate=TRUE))%>%
  filter(str_detect(V1, "S", negate=TRUE))
  samp1 <- ms_df %>%
  filter(str_detect(V1, "RetTime"))%>%
  mutate(rett=readr::parse_number(as.character(V1)))
  samp_tab <- left_join(clean_1_ms_df, samp1) %>% fill(rett) %>% na.omit()
  samp_df <- as.data.frame(samp_tab)
  GP_rettime <- samp_df[between(samp_df$rett, params$RTstart, params$RTend),]
}

bigdata <- lapply(file_path_spec, myfunction)

```


```


```

```

names(bigdata) <- file_path_spec
dt <- rbindlist(bigdata, use.names = TRUE, fill=TRUE, idcol="bigdata")

...

`` {r include=FALSE, warning = FALSE, message=FALSE}

my_binned_data <- dt %>%
  mutate(mz_bins=cut(as.numeric(V1), breaks=bins)) %>%
group_by(bigdata, mz_bins) %>%
  summarise(grand_inten=sum(as.numeric(V2), na.rm = FALSE)) %>% na.omit()

my_df <- as.data.frame(my_binned_data)%>%
  pivot_wider(names_from = bigdata, values_from= grand_inten) %>%
  as.data.frame()

my_df$count_NA <- apply(is.na(my_df), 1, sum)
my_df$percent_NA <-(my_df$count_NA)/num_of_samps*100

drop_NA <- my_df[my_df$percent_NA <= params$perc_include,]
drop_NA[is.na(drop_NA)] <- 0
my_df_zero <- drop_NA

names(my_df) <- gsub("\\.ms1.*", "", colnames(my_df))

my_df_mz <- as.data.frame(na.omit(my_df))
final_df <- my_df_zero[-1]
...

## feature_matrix is ready for Aristotle Classifier

```

```
```{r include=FALSE}  
here()
AC_df <- rbind(final_df, 1)
feature_matrix <- as.matrix(AC_df[1:num_of_samps])
```
```

Chapter 3 Future directions

3.1 Summary

From the research conducted in Chapter 2, there are two main tasks that relate to the continuation of research related to fingerprints, mass spectrometry, and machine learning: 1) the use of fingerprints as a biological sample that could be used in a clinical setting, and 2) the optimization of data acquisition and processing prior to doing machine learning.

3.2 Studying metabolic health by fingerprint samples

Following the initial fingerprint studies used for generating a model data set for mass spectrometry and machine learning applications, we recognized the potential of using fingerprints as a non-invasive biological sample that could be used as a health status readout. To test this hypothesis, we designed an experiment in which the fingerprint donor practiced a fasting mimicking diet (FMD), and fingerprints were collected over the course of 11 days total. Based on existing literature, we anticipated to see changes in the fingerprint composition as a function of which day in the fasting cycle it was collected. The fingerprints were prepared following the protocol described in Chapter 2, except all fingerprints were prepared immediately, such that the only intentional variability was the fingerprint donor's diet. Again, other sources of variability were not attempted to be controlled. We hypothesized that changes in an individual's diet could be monitored via changes in their fingerprint composition, by combining the strengths of the direct infusion ESI-MS method and the Aristotle Classifier. If changes were detectable between the two sample groups, we imagined applications where this approach might be useful.

3.2.1 Fingerprint Applications for monitoring nourishment

One area of human health that fingerprints could potentially revolutionize is monitoring nourishment of newborn babies. Adequate nourishment is critical for metabolic and

physiological function as well as growth, development, and energy supply. An added benefit for mom and baby who breastfeed is a strengthened immune system and decreased risk for various conditions, including metabolic disorders and chronic inflammatory disease.¹ For breastfeeding mothers, concerns about exclusive breastfeeding and whether their baby was adequately nourished ranked as one of the top self-reported reasons they discontinued breastfeeding.² The fact that there is no easy way to monitor a newborn's nourishment results in delayed intervention; recording growth measurements like weight and length can be useful, but this method is not sensitive enough to detect changes over the course of a few days. Rather, it typically requires at least a week of measurements before the clinician can make an informed decision for course of action; this is certainly anxiety-inducing for new parents, whose baby is passing developmental milestones on a daily basis. This often leads the mother to discontinue breastfeeding and switch to formula feeding, potentially prematurely. While a fed baby is always best, having the ability to easily monitor their nourishment could lead to an increase in sustained breastfeeding practices, which would benefit both mom and baby, far beyond infancy.

Some studies have attempted to determine what biomarkers might be useful for monitoring infant nourishment; indeed, there are distinct changes in the lipidomic profile that occur depending on the nourishment status and whether they are breast-fed or formula-fed.³⁻⁵ Although these studies have been useful in identifying potential biomarkers, the methods require blood samples which are challenging to obtain; drawing blood from a newborn every day is not realistic from a logistical standpoint and not desirable for a new parent. We wondered if the fingerprint workflow developed in Chapter 2 could be used to monitor breastfed newborns' nourishment, given that there is a known difference observed via other biological samples like blood. Since collecting fingerprints and analyzing them by mass spectrometry is non-invasive

and relatively simple, we pursued studies to explore how fingerprints could be used to address this common concern of new parents.

3.2.2 Modeling changes in nourishment via fingerprints

As a model of significant diet changes, which would mimic healthy and malnourished states, we designed a study in which the fingerprint donor- whose baseline is a plant-based diet- started with a high calorie diet and then completed a fasting mimicking diet program. The goal was to induce a significant biological change and to determine whether the method applied to aged fingerprints, in Chapter 2 of this thesis, could also be applied to this data set; collect samples, extract lipids, analysis by direct infusion ESI-MS, format the resultant data using LevR, and do machine learning using the Aristotle Classifier. If these changes were detectable and the classifier had sustained performance, it would be feasible to pursue more in-depth studies with the aim of translation to a clinical setting.

For proof of concept, we followed a similar protocol as described in the experimental section of Chapter 2, but rather than aging being the variable, the fingerprint donor followed a strict diet regimen, as described in more detail in the next section. Fingerprints were collected every day for 11 days. Overall, 63 fingerprint samples were collected and analyzed. There were two instrument days used to acquire the data and the sample set was split, such that the sample order was identical for both days (i.e., if two samples were collected each on Days 1, 2, and 3, the first day of data acquisition included the first sample from all three days, and likewise with the second day).

3.3 Experimental Methods

3.3.1 Diet Regimen

The fingerprint donor followed ProLon from L-Nutra, which is a fasting mimicking diet used for its supposed effects on metabolism and subsequent enhancement of healthy aging.⁶ The

fasting mimicking diet (FMD) has been studied via the collection of anthropometric data and blood samples for its potential anti-aging, anti-cancer, and decreasing the occurrence of aging related disease, like multiple sclerosis or cognitive decline, and metabolic disorders like cardiovascular disease and diabetes.⁷⁻⁹ The FMD consists of a 5-day cycle, during which the participant only consumes the calories provided in the diet box. For our studies, baseline status fingerprints were collected, along with high calorie diet status prior to the FMD, for the duration of the FMD cycle, and for 4 days after the cycle was completed.

3.3.2 Fingerprint Collection and Preparation

The collection and preparation of fingerprint samples was performed by adapting previously described methods,¹⁰⁻¹⁵ and varies slightly from the method described in Chapter 2. A single donor was used, and prior to fingerprint deposition, the donor swiped her fingertips over regions of the face that typically have high sebum secretion prior to depositing the fingerprints onto aluminum foil. These groomed fingerprints were collected over a series of days, limited to 6 fingerprint deposits (3 from each hand) per collection period.

Immediately after fingerprint deposition, the aluminum foil squares containing the fingerprints were rolled loosely using clean tweezers and placed into individual 2 mL screw thread sample vials with PTFE closure. 200 μ L dichloromethane was added to each, and the vials were vortexed for 1 minute, followed by 1 minute of rest, and removal of the foil. Then, to each vial, 200 μ L deionized water was added, vortexed for 1 minute, followed by 1 minute of rest, prior to liquid-liquid extraction. The aqueous layer was removed, and the organic layer was kept in the vial with an additional 200 μ L dichloromethane. All samples were stored -20 °C until analysis, such that only one thaw cycle occurred. Gas-tight Hamilton syringes were used throughout the experiment. For analysis, an aliquot of 44 μ L of the fingerprint sample solution

described above was diluted with 500 μL dichloromethane and 400 μL NH_4OAc in MeOH to achieve 1 mM ammonium acetate in the final solution.

3.3.3 ESI-MS conditions

Direct infusion ESI-MS analysis of the extracted fingerprint lipid samples was performed using an Orbitrap Fusion Tribrid mass spectrometer (ThermoScientific, San Jose, CA). The mass spectrometer was operated in negative ion mode with a sample injection flow rate of 3 $\mu\text{L}/\text{min}$. The heated-electrospray source was held at 2.3 kV while the ion transfer tube temperature, sweep, aux, and sheath gas flow rates were set at 300 $^\circ\text{C}$, 2, 5, and 10 Arb units, respectively. The full MS scans for the m/z range of (150-700) were acquired in the Orbitrap with a resolution of 60k. The AGC target value for the full MS scan was 5×10^4 , and the maximum injection time was 100 ms. For each sample, 30 scans were averaged for each file. Between analysis of every sample, a methanol/dichloromethane mixture was injected at 10 $\mu\text{L}/\text{min}$ for approximately 10 minutes or until the total ion count had returned to its baseline, established at the beginning of the experiment.

3.3.4 Specific settings used for fingerprint samples

The settings used for the analyses in this manuscript were as follows: 25% empty cells allowed, 20 lines in header, Lower m/z : 150, Upper m/z : 700, Bin width: 0.001 Da.

3.3.5 Aristotle Classifier settings and submission to the Aristotle Classifier

The output matrix generated by LevR in the previous section was modified by the addition of a row of 1's to last row of the matrix, as required by the Aristotle Classifier.¹⁶ K (repeats) value was set to 1000, and X value was set to 8.

3.4 Preliminary results

After sample processing, mass spectral data collection, and application of LevR and the Aristotle Classifier, the classification result in Figure 3-1 was generated, where each point

corresponds to a sample. In Figure 3-1a, the samples 1 through 63 are arranged in ascending order of sample collection, such that the left most observation is sample number 1 and the right most is sample 63. Samples 58 through 63 were misclassified; this seems to indicate that these last samples collected (1 day and 3 days post-FMD cycle) are more similar to the regular diet state rather than the fasting state. Another data set collected under the same conditions described above will be useful in determining to what extent this is repeatable.

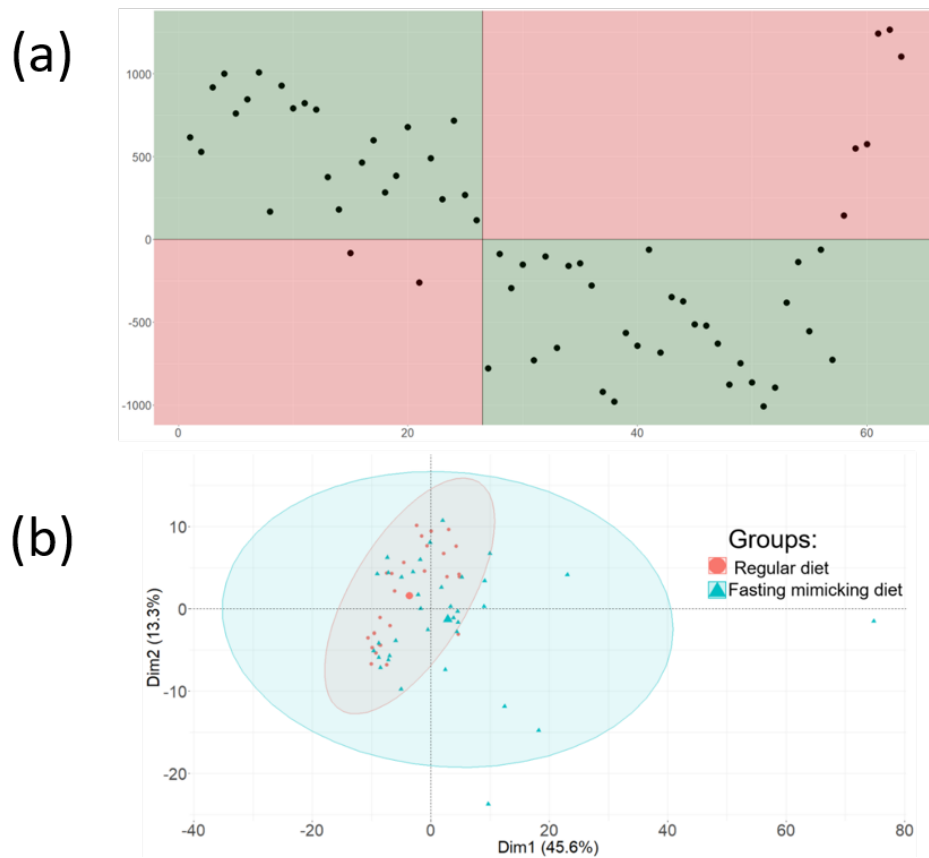


Figure 3-1. Comparison of Aristotle Classifier and PCA results for the full FMD data set. (a) Results from the Aristotle Classifier for 63 samples total. The correctly classified samples are highlighted in the green quadrants. The x-axis is ordered such that sample 1 is on the far left, and sample 63 is on the far right. (b) PCA results of the same 63 samples in panel a.

3.5 Discussion

The next experiments will aim to validate of the findings above and perform additional experiments to unequivocally assign compositions to the features scored by the Aristotle Classifier as being the most significant for distinguishing between samples. To do this, the fingerprint donor will complete the ProLon FMD diet for a second time and fingerprint samples will be collected. The same protocol will be followed to prepare the fingerprint samples for mass spectrometry analysis. During the mass spectrometry experiment, additional tandem MS experiments will be done on a shortened list of high scoring features determined from the first set of experiments so that the identity of the peaks is determined. To generate a peak list to use for this purpose, a similar score extraction method can be used as was described in Chapter 2. Table 2 shows the top 20 features and their predicted identity, but to unequivocally assign compositions to these mass spectral peaks, tandem MS experiments will be necessary. The resultant MS_n spectra acquired will be compared to spectra collected on standard lipids from a library.

Table 2: Top 20 features as scored by the Aristotle Classifier and their possible composition based on HRMS and mass error.

| Feature score | Composition match | m/z bin average | theoretical m/z (-H) | mass error (ppm) |
|---------------|-------------------|-----------------|----------------------|------------------|
| 43206 | C8H14O3 | 157.0865 | 157.0870 | -3 |
| 16742 | C9H16O3 | 171.1025 | 171.1027 | -1 |
| 12524 | C10H17O3N | 198.1135 | 198.1136 | 0 |
| 8634 | C12H21O3N | 226.1445 | 226.1449 | -2 |
| 4168 | C15H27O3N | 268.1915 | 268.1918 | -1 |
| 3878 | C12H21O4N | 242.1395 | 242.1398 | -1 |
| 1690 | C8H15O2N | 156.1025 | 156.1030 | -3 |
| 1532 | C10H17O4N | 214.1085 | 214.1085 | 0 |
| 1142 | C9H14O3 | 169.0865 | 169.0870 | -3 |
| 1132 | C13H23O4N | 256.1555 | 256.1554 | 0 |
| 1078 | C10H20O4 | 203.1285 | 203.1289 | -2 |
| 1034 | C8H10O4 | 169.0505 | 169.0506 | -1 |
| 896 | C11H19O3N | 212.1295 | 212.1292 | 1 |
| 824 | C10H20O3 | 187.1335 | 187.1340 | -3 |
| 760 | C9H15O2N | 168.1025 | 168.1030 | -3 |
| 720 | C10H16O3 | 183.1025 | 183.1027 | -1 |
| 698 | C11H16O3 | 195.1025 | 195.1027 | -1 |
| 694 | C14H28O3 | 243.1965 | 243.1966 | 0 |
| 650 | C9H15O4N1 | 200.0925 | 200.0928 | -2 |
| 630 | C13H23O3N | 240.1605 | 240.1605 | 0 |

3.6 Normalization of MS data to improve outcomes from merged data sets

From our initial studies using fingerprints as a sample type, we identified trends in the data that need to be investigated and addressed in future work with fingerprints. The first trend we identified is that instrument day was a significant contributor to the variability within a sample set with both groups. This became apparent when employing unsupervised classification methods to the fingerprint data. The PCA results were being used as a benchmark to show that the differences between the two sample types were subtle and numerous, and that unsupervised methods were not able to distinguish between sample types. What we found, however, is that there was significant clustering that correlated with the day the data was acquired. Figure 3-2a shows the original PCA generated, where each color is associated with a sample type- aged or not aged. This plot clearly shows that PCA does not separate the two sample types, which was

expected. There was, however, clustering related to other variables. These two clusters are circled on Figure 3-2a. The PCA groups were rearranged to color by data acquisition and the results are shown in Figure 3-2b. Note that the only change was what groups were used to color the plot. Neither the data, nor the clustering, has changed from Figure 3-2a to 3-2b. This indicates that a significant contributor to the variability within the data set is related to the day the data was acquired.

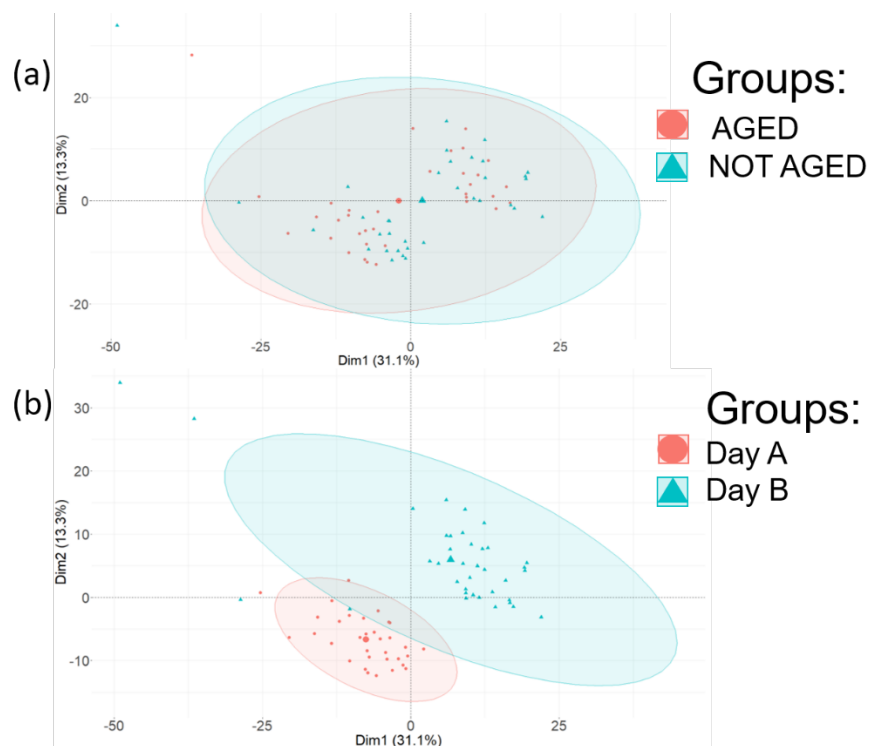


Figure 3-2. PCA results for fingerprint data set, colored by (a) sample type or (b) day of data acquisition.

As more data was acquired, this hypothesis was tested. In Figure 3-3, the PCA plot was generated with 4 days' worth of data and colored by day. Again, the PCA results indicate that a significant contributor of variability between the data sets is related to the day the data was acquired. This is problematic, particularly with the generation of large mass spectrometry data sets depending on the ability to combine data acquired on different days. To address this

significant limitation, which would limit the applications possible with the method developed herein, a future research goal is to identify a normalization strategy that could be applied to all data sets to eliminate the variability contributed by instrument day. If this could be achieved, the acquisition of MS data on different days, months, or even years, would not be a problem. This would enable generation of large data sets with virtually no limitation. This is the ideal scenario for those interested in developing robust machine learning tools.

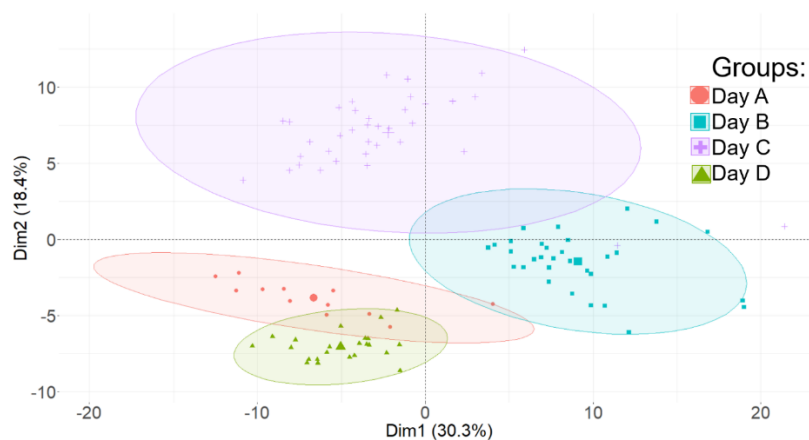


Figure 3-3. PCA results for fingerprint data set colored by day of data acquisition, with 4 days of data total.

3.7 Conclusion

In conclusion, the specific aims of future research involving fingerprints, mass spectrometry, and machine learning should be to establish a normalization strategy that can be applied to data sets and to pursue the possibility of using fingerprints as a clinical readout for health status.

3.8 Acknowledgements

This work was supported by NIH Grant R35GM130354 to HD.

3.9 References

1. Ma, J.; Li, Z.; Zhang, W.; Zhang, C.; Zhang, Y.; Mei, H.; Zhuo, N.; Wang, H.; Wang, L.; Wu, D., Comparison of gut microbiota in exclusively breast-fed and formula-fed babies: a study of 91 term infants. *Scientific reports* **2020**, *10* (1), 15792.
2. Li, R.; Fein, S. B.; Chen, J.; Grummer-Strawn, L. M., Why Mothers Stop Breastfeeding: Mothers' Self-reported Reasons for Stopping During the First Year. *Pediatrics* **2008**, *122* (Supplement 2), S69.
3. Suchdev, P. S., What Pediatricians Can Do to Address Malnutrition Globally and at Home. *Pediatrics* **2017**, *139* (2), e20161666.
4. Prentice, P.; Koulman, A.; Matthews, L.; Acerini, C. L.; Ong, K. K.; Dunger, D. B., Lipidomic analyses, breast- and formula-feeding, and growth in infants. *J Pediatr* **2015**, *166* (2), 276-81.e6.
5. Acharjee, A.; Prentice, P.; Acerini, C.; Smith, J.; Hughes, I. A.; Ong, K.; Griffin, J. L.; Dunger, D.; Koulman, A., The translation of lipid profiles to nutritional biomarkers in the study of infant metabolism. *Metabolomics* **2017**, *13* (3), 25.
6. Wei, M.; Brandhorst, S.; Shelehchi, M.; Mirzaei, H.; Cheng, C. W.; Budniak, J.; Groshen, S.; Mack, W. J.; Guen, E.; Di Biase, S.; Cohen, P.; Morgan, T. E.; Dorff, T.; Hong, K.; Michalsen, A.; Laviano, A.; Longo, V. D., Fasting-mimicking diet and markers/risk factors for aging, diabetes, cancer, and cardiovascular disease. *Science translational medicine* **2017**, *9* (377), eaai8700.
7. Galetti, V.; Brnic, M.; Lotin, B.; Frigeri, M., Observational Study of Lipid Profile and C-Reactive Protein after a Seven-Day Fast. *Nutrients* **2021**, *13* (1).
8. Deligiorgi, M. V.; Liapi, C.; Trafalis, D. T., How Far Are We from Prescribing Fasting as Anticancer Medicine? *Int J Mol Sci* **2020**, *21* (23), 9175.
9. Cheng, C.-W.; Villani, V.; Buono, R.; Wei, M.; Kumar, S.; Yilmaz, O. H.; Cohen, P.; Sneddon, J. B.; Perin, L.; Longo, V. D., Fasting-Mimicking Diet Promotes Ngn3-Driven β -Cell Regeneration to Reverse Diabetes. *Cell* **2017**, *168* (5), 775-788.e12.
10. Archer, N. E.; Charles, Y.; Elliott, J. A.; Jickells, S., Changes in the lipid composition of latent fingerprint residue with time after deposition on a surface. *Forensic Science International* **2005**, *154* (2), 224-239.
11. Hinnners, P.; O'Neill, K. C.; Lee, Y. J., Revealing Individual Lifestyles through Mass Spectrometry Imaging of Chemical Compounds in Fingerprints. *Scientific reports* **2018**, *8* (1), 5149.
12. Hinnners, P.; Thomas, M.; Lee, Y. J., Determining Fingerprint Age with Mass Spectrometry Imaging via Ozonolysis of Triacylglycerols. *Analytical Chemistry* **2020**, *92* (4), 3125-3132.
13. O'Neill, K. C.; Hinnners, P.; Lee, Y. J., Potential of triacylglycerol profiles in latent fingerprints to reveal individual diet, exercise, or health information for forensic evidence. *Analytical Methods* **2020**, *12* (6), 792-798.
14. Pleik, S.; Spengler, B.; Ram Bhandari, D.; Luhn, S.; Schäfer, T.; Urbach, D.; Kirsch, D., Ambient-air ozonolysis of triglycerides in aged fingerprint residues. *The Analyst* **2018**, *143* (5), 1197-1209.
15. Pleik, S.; Spengler, B.; Schäfer, T.; Urbach, D.; Luhn, S.; Kirsch, D., Fatty Acid Structure and Degradation Analysis in Fingerprint Residues. *Journal of the American Society for Mass Spectrometry* **2016**, *27* (9), 1565-1574.

16. Hua, D.; Patabandige, M. W.; Go, E. P.; Desaire, H., The Aristotle Classifier: Using the Whole Glycomic Profile To Indicate a Disease State. *Analytical Chemistry* **2019**, *91* (17), 11070-11077.