Accelerating high-throughput proteomics with proteome-selective isolation chromatography (P-SLICY)

By

Darien Bo-Hin Yeung

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

In partial fulfillment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

Department of Biochemistry and Medical Genetics

University of Manitoba

Winnipeg

Thesis Abstract

Proteomics is one of the most effective ways to study protein regulation including characterizing the changes in protein quantity as a result of biological perturbation. Liquid chromatography paired with tandem mass spectrometry (LC-MS/MS) is the most commonly used analytical technology in this field. After two decades of technological advancements, the quantitation of the total human proteome still precludes us. These advancements have allowed for high throughput quantitation of near tens of thousands of proteins in hours of analysis time. However, 30% of the proteome remain elusive in routine proteomic analysis. Current LC-MS/MS methods rely on increasing the number of peptides identified to yield more protein identifications. But it has been shown by Bekker-Jensen et al. that this is a logarithmic relationship with diminishing returns.

This diminishing return is because a majority of the peptides belong to high abundance or large proteins and are highly redundant. Prior peptide-centric approaches have shown promising improvements to proteome coverage by enriching a subset of peptides containing specific amino acids for analysis rather than analyzing the whole digested proteome. By reducing the number of peptides necessary to analyze and the overall protein identification redundancy, these approaches can potentially bypass the logarithmic relationship between peptide and protein identifications. We aim to develop a peptide-centric approach that can be generally applicable for a variety of sample types by enriching a subset of peptides that have a particular physicochemical property via chromatographic approaches. Therefore, our approach is named Proteome-Selective Isolation Chromatography (P-SLICY). In this thesis, we will describe the fundamental topics within proteomics, chromatography, and mass spectrometry and the innovations towards understanding peptide chromatography to develop P-SLICY.

Acknowledgements

I would like to thank my supervisor, Prof. Oleg Krokhin for his unwavering support, guidance, and mentorship throughout my PhD degree. Такого учителя, как Вы, непросто найти. I would also like to thank the members of my committee Prof. John Wilkins, Prof. Barbara Triggs-Raine, and Prof. Pingzhao Hu for all your insights and knowledge. I learned a lot through the process of this scientific journey, and you all have contributed immensely to my development.

Thank you to all the past and present colleagues in the Krokhin lab, and at the Manitoba Centre for Proteomics and Systems Biology. I want to take the opportunity to thank Prof. Rene Zahedi and Prof. Janilyn Arsenio for the opportunities you have given me to contribute to amazing collaborative projects at MCPSB. I would like to thank Ying (Tenny) Lao, Victor Spicer, Dr. Xuehui (Helen) Jiang, Quinn Neale, Dr. Arzu Ozturk, Megan Rodriguez, Dr. Hemshekhar Mahadevappa, Immanuel for all your support as great friends and helping me grow as a scientist.

Furthermore, I would like to thank Dr. Yasushi Ishihama for the opportunity to be a part of your lab and learn amazing techniques. 本当にたくさんのことを教えてくれてありがとうございます. Thank you to Prof. Kosuke Ogata, Ayana Tomioka, Echo Geronca, Moritz Weissmeier, Junqi Liang, and all the lab members of the Ishihama lab.

Most importantly, I want to thank my family for always believing in and supporting me.

Dedication

To my family, friends, and team members

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Biological importance and study of proteins

There are approximately 3 trillion cells that compose the human body and these cells work in unison to support the survival of the organism.[1] A variety of cell types exists, each with their phenotype, role, and purpose.[2] Proteins are a class of molecules that contribute immensely to the behaviour and function of the cell.[3,4] The set of proteins expressed by the organism is the proteome and cells can have different proteome composition depending on their cell type. Fundamentally, proteins are biopolymers that contain specific arrangements of amino acids defined by the organism's genome.[3,5,6] There are twenty naturally occurring amino acids that participate in the construction of proteins and the sequence is critical towards protein function. Furthermore, proteins often operate in conjunction with other proteins through complexes[7–9] or pathways to execute cellular tasks.[10–12] Aberrations to these proteins, their interactions, and the involved pathways can result in disease and hence, proteins are a key research subject for future medical interventions and diagnostics.[13,14]

The research of proteins and their relationship with disease can involve the study of a protein's function, structure, localization, interaction, and regulation. For instance, the disruption of a protein's function can be found in diseases like cystic fibrosis where amino acid mutations in the CFTR protein, a chloride transporter, compromises the transporter's ability to traffic chloride out of the cell.[15–17] In conditions like Parkinson's disease, mis-folding events that alter a protein's structure can contribute to neurodegeneration.[18–20] Alterations to a protein's sequence or structure can lead to improper localization of the protein. The mutation in the protein ATP7B impedes the protein's ability to express at the cell membrane and results in the observed copper accumulation

for Wilson's disease.[21,22] These mark a few examples of many diseases that are associated with changing the protein properties.

Aside from studying proteins in isolation, it is important to contextualize how the proteins interact with one another. Protein-protein interactions (PPI) involve the noncovalent interaction between multiple proteins.[23,24] In the case of human immunodeficiency virus infections, the viral glycoprotein gp120 can interact with the human CCR5 receptor to initiate the viral infection of the cell.[25,26] Hence, pharmacological intervention to disrupt this interaction can minimize the virus's ability to replicate using the human cell's machinery. Beyond PPI, other interactions can manifest as pathways, involving the interaction of a series of molecules, that results in changes to the cell. These pathways can be involved in biological processes such as signaling, apoptosis, metabolism. Ultimately, aberrations to these processes can alter the proteome composition in a cell. Studying the changes in the proteome composition is part of the protein regulation subfield.

Protein regulation studies the generation, modification, and degradation of proteins.[20,27] Each of these processes are complex involving many proteins and pathways. For instance, the process of generating a protein is not trivial and, in a simplified view, includes the (i) construction of the transcriptional protein complex at the promoter region of the gene, (ii) trafficking proteins into the nucleus from the cytosol, (iii) processing the transcriptional product to be viable for protein synthesis, (iv) trafficking the processed transcriptional product to the cytosol, (v) generation of nucleic acid-amino acid complexes as building blocks for synthesis, (vi) recruiting chaperones to fold the protein into a correct structure. There are many regulatory components that control each process in this simplified view of protein generation. Furthermore, each regulatory components can be regulated by other factors as well and it is unclear if there is a unified regulator that provides the context of the cell state or phenotype. Dysregulations that occur during these processes can lead

to changes in the composition of the cellular proteome.[19] Temporal information can be powerful to investigate how protein regulation change over time and how alterations to the regulation can induce disease-like phenotypes in cells, since these are all dynamic processes involving rates of change for each protein. In an ideal scenario, continuous monitoring of the molecular changes can provide temporal information. This is similar to how chemical reactions are monitored via reaction rates to determine the mechanistic involvement of reactants, intermediates, and products.[28] Except biology is complex. Cellularly, there are many reactions that are occurring simultaneously that can involve the same set of proteins. This makes it difficult to delineate which proteins are reactants, intermediates, or products. Therefore, this continuous approach is not yet possible on the whole proteome. Instead, researchers are using taking "snapshots" of the proteome to infer the changes over time.

The study of regulation as a "snapshot" can be done by analyzing the concentration of the proteins in the proteome at a particular time.[29–32] This can be a limitation in some studies, but until future technologies can overcome these limitations, it allows us to understand which proteins are participates in a cellular response or disease. By comparing the proteome responses between disease versus not disease states, we aim to find the proteins involved with a disease. This type of analysis is differential expression.[33,34]

1.1.1 Analytical Methods for Studying Proteins via Differential Expression

There are many biological techniques that are utilizing the concept of differential expression. A recent example is the lateral flow antibody tests used to screen if a particular patient has a coronavirus disease 2019 (COVID-19) infection. An antibody is a type of protein that can recognize a target protein through specific binding and is a useful tool for protein detection. In this case, the

target protein is the spike protein on the virus's surface. By sampling the patient's mucosa and subjecting the sample to an antibody test, the results can classify the patients with viral protein as COVID-19 positive and otherwise, COVID-19 negative. In this simple scenario, we are evaluating the differential expression of a viral protein in two different conditions.

To evaluate higher complexity problems such as degree of infection, the spike protein would not be able to provide sufficient context and a different set of molecules would be needed.[35,36] Currently, there are biological questions that do not have existing molecular assays to help answer and need discovery methodologies to find the molecules that differentially express. A potential approach to bridge this gap for proteins is proteomics, the study of the proteome.

One of the aims of proteomics is to understand changes in protein expression without pre-determining a group of proteins for analysis.[37] This aim is part of discovery proteomics. The challenge in this subdomain is trying to detect all ~20,000 proteins that the human cell can express.[38–41] Using antibody tests will be laborious involving at least ~20,000 tests per biological condition. Furthermore, only ~3000 human proteins have commercially available antibodies. Researchers in proteomics integrate mass spectrometry (MS) as the analytical platform because this technology can analyze any protein. Therefore, practitioners use MS as an unbiased discovery tool. MS is a tool that measures the mass to charge ratio (*m/z*) of molecules. The *m/z* is critical to identify the identity of the protein. Analyzing the intact protein with MS is part of top-down proteomics. There are technical limitations that prevent top-down proteomics from studying a wide range of protein concurrently in high-throughput methods. Therefore, researchers cut proteins into shorter fragments that are peptides. Proteases are enzymes used to cleave proteins into smaller peptides. The MS then analyze the peptides as surrogate to determine the proteome composition of the sample. This latter approach is called bottom-up proteomics and it is the most commonly used

method.[42] The workflow to generate the peptides analyzed in bottom-up proteomics is shown in Figure 1.1. Conventionally, we start with sample which can be cell lines, tissue, or biological fluids. The first step is to isolate and solubilize proteins by using detergents and other chaotropic agents. For cell lines and tissues, this the lysis step is critical to disrupt the cell membrane. Upon isolation of the proteins, we use proteases to digest proteins into peptides.

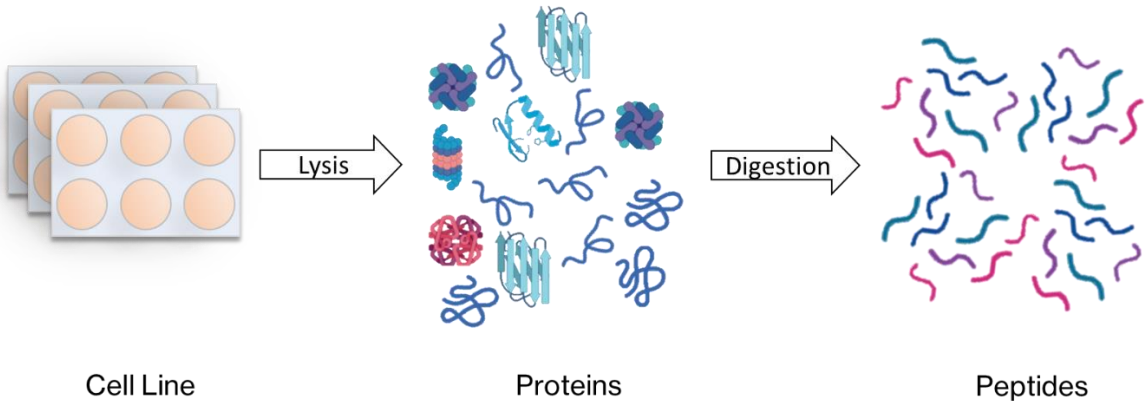


Figure 1.1: Sample preparation workflow to generate peptides for bottom-up proteomics

Peptides have directional sequences. When writing out the sequence of a peptide, the left side of the peptide is the N-terminus, and the right side is the C-terminus. The sequences GGGGT and TGGGG are not the same peptide as the directional sequences are different. However, when using MS directly, these two peptides will have the same *m/z* value; thus, distinguishing them will be difficult. The specialized procedure, tandem MS (often written as MS/MS) can sequence the peptide inside the mass spectrometer. We can determine the sequence by fragmenting the peptide with high-energy ions inside the instrument.[43] During peptide fragmentation, we characterize the amino acids that break off the peptide in sequential order to reconstitute the peptide sequence. An additional liquid chromatography (LC) step can separate peptides to improve (expand time for the analysis) their delivery into the MS and can resolve peptides with similar *m/z* based on the peptides'

chemical properties.[44] By physically connecting the two approaches in series, the resulting method is LC-MS/MS.

At the time of writing, LC-MS/MS is the gold standard for bottom-up proteomics.[41] However, there are still significant challenges to detecting all the proteins the human genome encodes. Proteins express across a large dynamic range and depending on samples can span 8 orders of magnitude or more.[45–47] Current approaches in LC-MS/MS have difficulty detecting the lower abundance proteins due to the peptide signals from high abundance proteins overshadowing lower abundance species. Consider, for an analogy, that LC-MS/MS can only detect 100 total proteins and there are three different proteins in the sample. If protein A has 100 copies in the sample, protein B has 10 copies, and protein C has 1 copy, there is only a 1% chance that protein C will be detected. In this oversimplified analogy, the limited capacity of the LC-MS/MS can undermine the signals relating to the low abundance protein C. Researchers address this problem by expanding the capabilities of the mass spectrometers.[48–56] Despite efforts in this regard, the MS capacity is still limiting the detection of the whole human proteome. Shishkova et al. suggests that in addition to MS enhancements, LC approaches also should improve.[57] We agree with this perspective and from our experiences with peptide chromatography, we conclude that chromatographic advancements can allow deeper interrogation of the expressed proteome. Hence, we will describe the fundamentals of chromatography and MS in the following sections to contextualize how we are approaching improving protein detection and quantitation in this thesis.

1.2 Chromatography

1.2.1 Introduction to Liquid Chromatography

Chromatography is the method for separating complex mixtures into individual compounds.[58] Each compound has their own unique chemical properties that is defined by their structure that varies the compound's interaction strength with other molecules. In chromatography, the material used to separate compounds is called the stationary phase. The stationary phase engages in noncovalent interactions with compounds in a complex mixture and their interaction strength will vary the resolution of separation. A mobile phase solvent carries the compounds through the chromatographic column containing the stationary phase and the balance between the compound-stationary phase and compound-mobile phase interaction allows for individual compounds to be isolated from mixtures. The extraction of the compounds from the stationary phase is done by the elution process. If the compound has stronger interaction with the mobile phase, then the compound will leave the stationary phase material and associate with the mobile phase to elute. This relationship is visualized in Figure 1.2 where the blue peptide has a stronger interaction with the mobile phase therefore eluting from the sorbent. Conversely, the orange peptide has a stronger interaction on the stationary phase to retain on the column. The retention time when the compounds elute from the column depends on the interaction strength between the compound, the stationary phase, and the mobile phase. Modern liquid chromatography approaches often rely on the high-pressure liquid chromatography (HPLC) instruments to facilitate this separation process.[59,60]

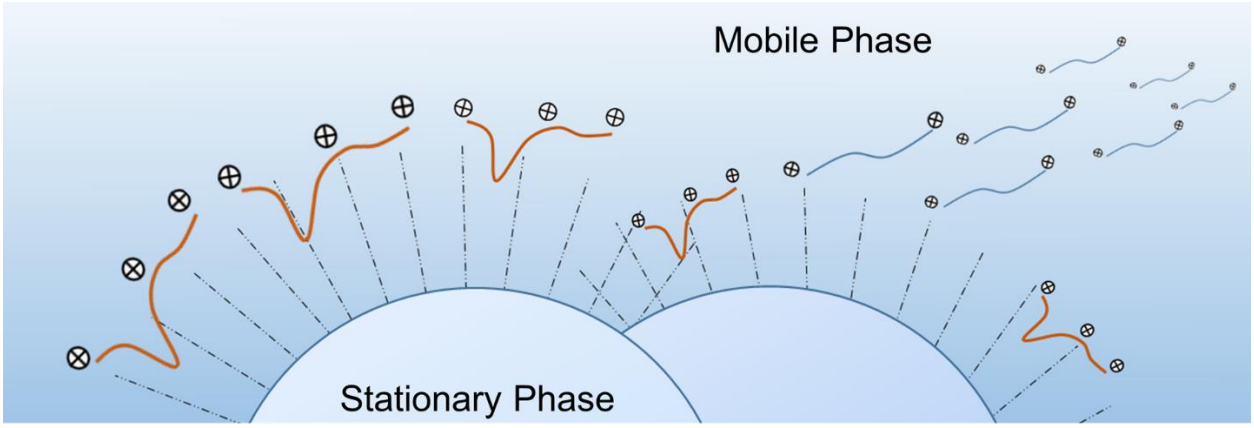


Figure 1.2 Peptide interaction with stationary phase and mobile phase during chromatographic separation.

1.2.2 Chromatographic Interactions

1.2.2.1 Compound-stationary phase interactions

Conceptually, the compound-stationary phase interactions are modelled via the Langmuir isotherm[61] which describes the occupation of compounds on available interaction sites found on the stationary phase, as seen in Figure 1.3. In this model, a compound can adsorb on or desorb from an interaction; however, each interaction site can only accommodate one molecule as the assumption is that only one layer of interaction can occur. In reality, adsorption and desorption processes are more complex and can involve multiple layers of interaction sites. Furthermore, each compound will have different interaction strength with the stationary phase. However, this simplified explanation can aid in conceptualizing most chromatographic topics. When more molecules occupy the stationary phase, it becomes more challenging for another molecule to adsorb onto the stationary phase. Increasing the stationary phase surface area will increase the number of

available interaction sites; thus, in chromatography, it is beneficial to have a higher surface area. This topic will be elaborated in Section 1.2.2.4.

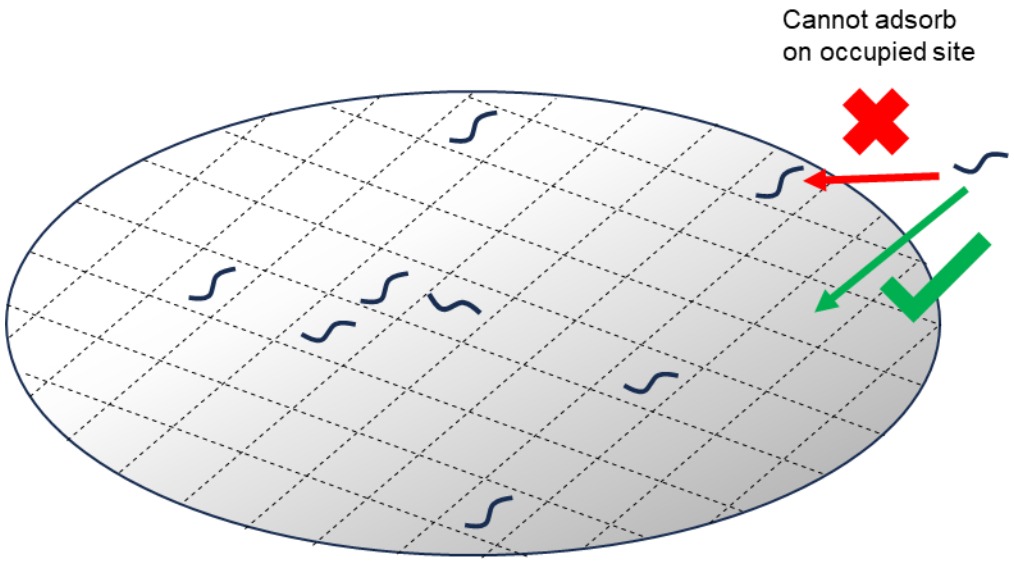


Figure 1.3: Peptide adsorption on the available interaction sites on the stationary phase sorbent. Incoming peptides can access unoccupied sites but cannot adsorb on occupied sites.

The chromatographic interaction with a compound can be tuned based on the interaction strength with the stationary phase. In chromatography, we refer to this interaction strength as affinity. The balance between the affinity for the stationary phase and the mobile phase establishes the basis for achieving good chromatographic separation. Therefore, in the following subsections, we will explain the types of interaction between the compound, stationary phase, and mobile phase to understand how we optimize chromatographic processes for proteomics research.

1.2.2.2 Types of stationary-phase interactions

Stationary phase interactions conventionally refer to the non-covalent interaction such as London dispersion forces, van der Waals interaction, ionic interaction, dipole-dipole moments. In the context of peptide separation via liquid chromatography, we can summarize these interactions into

hydrophobic, hydrophilic, and electrostatic interaction that modulate a majority of peptide-stationary phase interactions. Hydrophobicity is the physical property of a compound to repel water. The solubility ratio of the compound in octanol (a hydrophobic solvent) and water quantifies this physical parameter. Compounds with higher hydrophobicity will have high octanol solubility and low water solubility. Of the 20 amino acids that build proteins and peptides, 9 are hydrophobic which are alanine, valine, leucine, isoleucine, methionine, proline, phenylalanine, tyrosine, tryptophan. Hydrophilicity is the opposite of hydrophobicity. For this physical property, the compounds tend to be polar and have higher solubility in water than octanol. There are 11 amino acids of the 20 that are hydrophilic such as lysine, arginine, histidine, aspartic acid, glutamic acid, cysteine, serine, threonine, asparagine, glutamine, glycine. Electrostatic interaction is the attraction or repulsion of two charged molecules. If the compound and the stationary phase have opposing charge polarity, they electrostatically attract. However, if they both have the same charge polarity, they repel. Only 5 amino acids of the 20 in their ionic forms can participate in electrostatic interactions with a charged stationary phase which are lysine, arginine, histidine, aspartic acid, and glutamic acid. Furthermore, the terminal ends of peptides will be charged which can participate in electrostatic interaction. Stationary phases can interact with the peptide in different ways depending on the sorbent properties and ionization state of individual residues and terminal ends. This can be modulated by altering the pH or incorporating additives in the mobile phase.

With respect to stationary phase materials, used in the initial HPLC development in mid 20th century, they appeared in two main forms.[62] They were classified as the 'normal-phase' and 'reversed-phase' and they were named as such due to historical reasons. Silica-based columns with hydrophilic surfaces in conjunction with more hydrophobic stationary phases were initially used, hence, they are called normal phase. The surface of silica is decorated with hydrophilic hydroxy

(OH) groups. Thus, the hydrophilic surface is appropriate for polar interactions with hydrophiles as seen in Figure 1.4. However, hydrophobes do not interact well in this type of system, therefore, hydrophobic modifications to the silica were made. Thus, since hydrophobicity is the opposite behaviour to hydrophilicity, this mode is called reversed phase (RP).[63] The separation of this class of compound improves with a silica surface that is modified with long hydrophobic chains typically in lengths of 18 carbons. Resulting from this, the RP sorbent is called C18. Aside from hydrophobic molecules, there were demands to separate small ionic compounds which normal phase was not sufficient to retain these analytes. As a result, ion exchange (IEX) was implemented to support electrostatic separation between the compound and the stationary phase.[64]

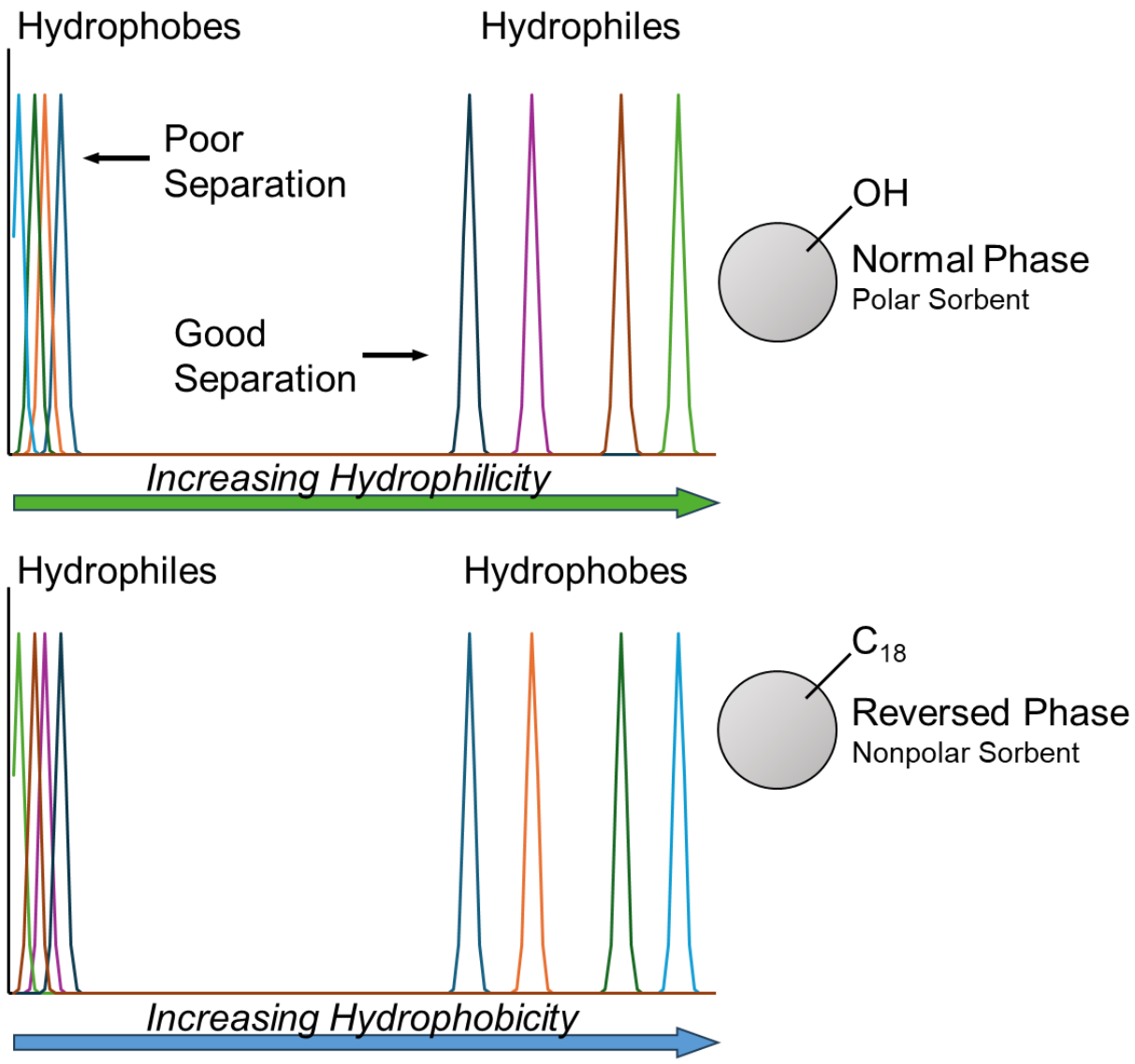


Figure 1.4: Difference in retention properties between polar normal phase sorbent and non-polar reversed phase sorbent.

In proteomics, the stationary phase options are mainly RP, hydrophilic interaction liquid chromatography (HILIC), or ion exchange (IEX).[65,66] Normal phase is less popular due to poor compatibility of non-polar solvents electrospray ionisation technique used in with LC-MS/MS.[67] HILIC is used instead of normal phase separation of hydrophilic peptides as this method utilizes

water mixable organic solvents such as acetonitrile – compatible with electrospray ionization. In practice, RP is the most prevalent separation mechanism in proteomics due to presence of hydrophobic residues in the vast majority of peptides.[66,68] These peptides have a strong interaction between its hydrophobic regions with the RP stationary phases to retain on the column shown in Figure 1.5. Conversely, hydrophilic peptides have low retention on RP and elute early during the separation procedure. HILIC has the opposite characteristics to RP where the hydrophilic peptides have higher retention on the column.[69–71] In HILIC, the peptide interaction is different from other separation mechanisms. Usually, HILIC is considered a variation of normal phase mode as it uses mixtures of water and non-polar solvents as mobile phases. Therefore, water will have strong hydrophilic interaction with the hydrophilic stationary phase surface illustrated in Figure 1.5. The retention of the peptide depends on partitioning with the water layer rather than direct interaction with the stationary phase.[72,73] Finally, IEX is applicable to separate different groups of charged peptides where RP and HILIC cannot achieve this.[74–79] There are two subtypes of IEX. The first subtype is strong cation exchange (SCX), which uses an anionic surface to attract cationic molecules for separation electrostatically. The second subtype is strong anion exchange (SAX) which has a cationic surface to attract anionic molecules.

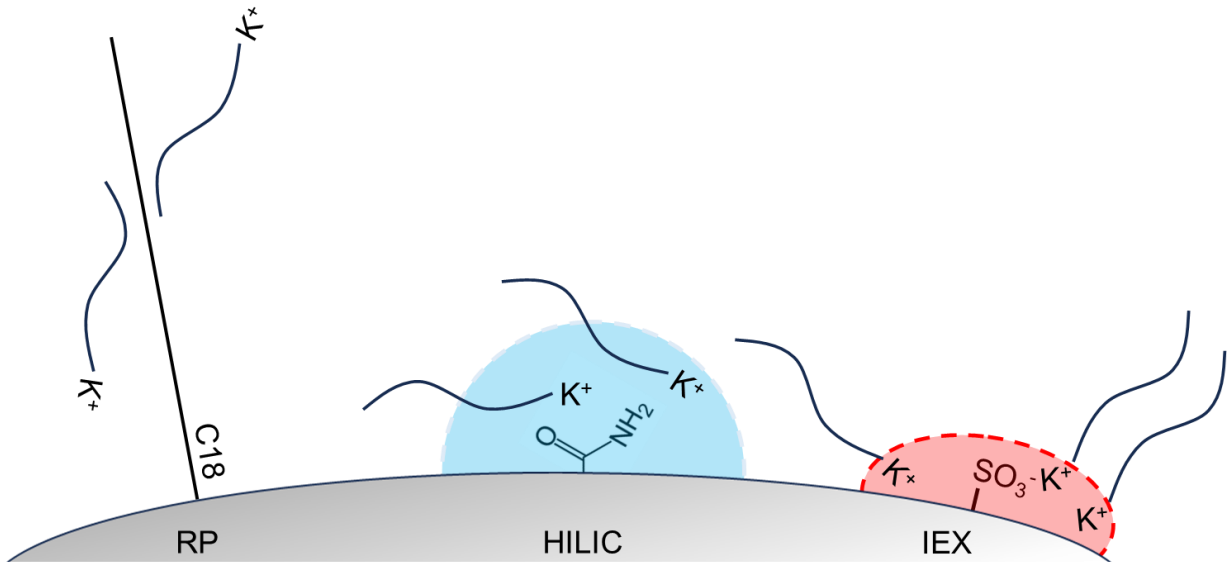


Figure 1.5: Interaction of peptides with RP, HILIC, and IEX stationary phases

1.2.2.3 Localization of chromatographic interactions

With the understanding of the types of stationary phase, we need to establish where the interactions occur, and the location of the interaction sites. The chromatographic sorbent defines the surface area of a stationary phase and the higher the surface area, the more interaction sites are available. The porosity of the particles is an important parameter to consider where smaller pore sizes can accommodate more pores on the sorbent and increase surface area.[80,81] However, the size of the pores should be 4-10 times the size of the compounds being separated to avoid poor chromatographic separation.[82] There will be size exclusion effects that are not representative of the chromatographic interaction if the pore size is too small. In proteomics, 100 Å to 200 Å pore sizes can accommodate the separation of peptides. When deciding the pore size to be used for peptide separation, one must consider the peptide-ion pair size. Formic acid (FA), for instance, will be able

to form larger hydrated ions readily. Hence, the most popular 100 Å pore size may not be large enough for peptides to penetrate into the sorbent and utilize the entire surface area.[83]

If the peptide can enter the pores, one must also consider the structure of the stationary phase. Using RP as an example, the C18 structure will contract or expand depending on the mobile phase composition.[84,85] This effectively changes the accessible sorbent volume of the stationary phase.[86] Since peptides interact directly with the aliphatic structures in C18, the accessible sorbent volume directly relates to the number of interaction sites. Under aqueous conditions, the C18 structure takes a contracted form and decreases the accessible sorbent volume. Therefore, at low concentrations of organic solvent, C18 has fewer interaction sites than at higher organic conditions where the C18 is fully solvated in its extended structure shown in Figure 1.6. Different vendors will have varying degrees of contraction and extension of the C18 structure. Ideally, the C18 should have extended structures regardless of conditions to maximize accessible interaction area.

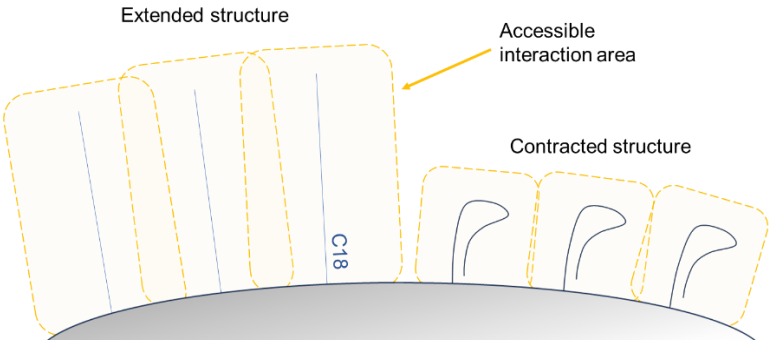


Figure 1.6: Illustration of extended and contracted structure for C18 with the accessible interaction area that differ between the two conformations.

In the extended structure of the stationary phase, there can be residual silanol groups on the silica from incomplete reaction when modifying the surface with C18. The peptides that interact with the C18 will exhibit the expected hydrophobic interaction. However, if the peptides interact

with the silica, the peptides participate in hydrophilic interaction as well.[87] If a peptide has multiple competing mechanisms that simultaneously affect the peptide's separation, this can widen the peak and decrease the efficiency of separation. Therefore, the end-capping procedure was introduced to cap the residual silanols with small aliphatic chains to minimize unintended hydrophilic interactions.[88,89]

1.2.2.4 Types of mobile-phase interactions

In addition to the stationary phase interaction, the mobile phase also plays a critical role in the retention of a peptide. Mobile phases are solvents that elute the compounds from the stationary phase, but they can be a mix of multiple solvents and other eluents' additives. There are two ways to mix solvents. The isocratic method involves mixing the solvents to a final fixed composition, and this mixture is constant throughout the chromatographic separation.[90,91] On the other hand, the gradient method dynamically changes the concentration of the solvents in the mobile phase as the separation occurs. Isocratic separations are simple to use to separate low complexity mixtures. Generally, isocratic separation is ideal for analyses involving less than 10 components in the mixture.[91] However, isocratic methods are not feasible for complex mixtures involving a wide breadth of interaction strengths. The mobile phase composition may not be optimal for all analytes. Therefore, gradient separations are better for the separation of complex mixtures. By varying the solvent composition over time, this method provides the efficient elution of the analytes. As an example, RP separations employ gradual increase of organic solvent (acetonitrile, methanol) concentration in mixtures with water. HILIC separations are opposite: starting with low water content in water/acetonitrile mixture and increasing it concentration gradually.[90] The digested

human proteome is an extremely complex mixture of peptides, and hence, most separations in proteomics used gradient methods.[68]

Additives can be incorporated into mobile phases to enhance the interaction between the compounds and stationary phases, and to maintain constant pH.[92–96] In proteomics, formic acid (FA) is commonly used as it facilitates peptide ionization during ESI. Peptides can contain basic amino acids like lysine (K), arginine (R), and histidine (H). These three amino acids are hydrophilic and therefore, will have minimal interaction with the stationary phase. At the acidic pH, the basic amino acids will ionize to carry a positive charge. Negatively charged formate anions in the mobile phase form ion pairs with the charged basic amino acids to insulate the charge.[94,97] Furthermore, this enhances the hydrophobicity of peptides containing those amino acids and leads to better retention. Increasing the hydrophobicity of the additive will also increase the hydrophobicity of the peptides accordingly as seen when using the more hydrophobic trifluoroacetic acid (TFA). The latter, however, reduces peptide detection sensitivity as it suppresses peptide ionization in ESI.[98]

Additives can also decrease the retention of peptides in other chromatographic scenarios. In IEX, charged peptides form strong interactions with the stationary phase. Therefore, ionic salt additives are necessary to displace peptides from the column by competing for the same interaction sites.[99,100] When the salts occupy the interaction sites, the peptides can no longer bind to the sorbent and will elute off the column. IEX methods often use gradients that change the concentration of the additive salt to elute peptides rather than changing the organic phase.[99,100] Resulting from this type of gradient, IEX modes can separate peptides of different charges.

1.2.2.5 Interpreting the chromatogram

After separating the molecules through liquid chromatography, the instrument reports a chromatogram, which contains the elution profiles of all the separated compounds. The x-axis often represents the elapsed time for the separation process and the y-axis reports the intensity values from the detector, such as UV absorbance at 210-220 in case of peptides. When peptides elute under gradient conditions, the signal intensity for the elution process often follows a gaussian distribution (Figure 1.7A). The corresponding time value for the peak maxima is referred to as retention time and represents qualitative parameter used for peaks' identification. In the overlapping elution shown in Figure 1.7A, peptides 1 and 2 have an RT of 40 minutes and 50 minutes, respectively. However, due to their wide peak distributions, these two peptides are co-eluting. Hence, when we visualize the total intensity, the chromatogram looks like a peak with a shoulder to the right. Figure 1.7B demonstrates the simulated chromatogram for 100 peptides with narrow peak distributions and is highly complex. There are many overlapping or co-eluting peaks, which is common in peptide separation for bottom-up proteomics.[101–103] When separating hundreds of thousands of peptides in real-world samples, the chromatogram increases in complexity, as shown in Figure 1.7C. Clearly, identification of individual peptides based on retention time is impossible in this case and requires detection devices with higher resolution power such as mass spectrometers.

To interpret the chromatogram, we can split the visualization into two regions. In Figure 1.7C, the red region represents the signals collected starting from sample injection to the arrival of the first gradient portion (in case of gradient elution) to the column. This region is known as the gradient delay time, which corresponds to the internal volume of the plumbing the sample and the

gradient travelled through prior to reaching the column. The peptides that are identified in this region do not retain on the column and are considered flowthrough peptides. The yellow region after this dead volume contains the retention information of the peptides eluting from the column with respect to the gradient program. Generally, the earlier the elution of a peptide represents a weaker retention on the column and the later the elution is associated with strongly retaining peptides. Based on the retention time, we can infer the interaction strength between the peptide and the chromatographic column. The intensity of a peak observed by the detector, represented by the y-axis, correlates with the amount of peptide found in the sample. The higher signal, the more of that peptide is present.

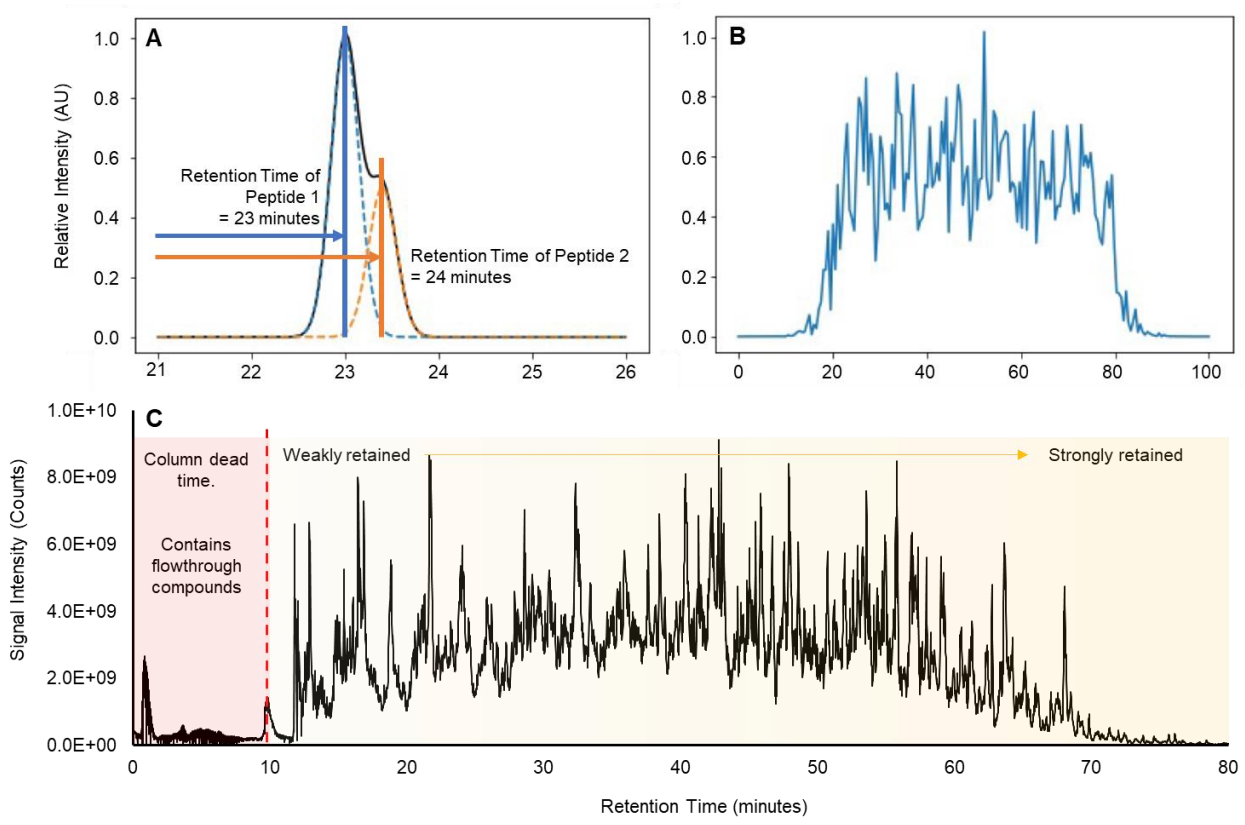


Figure 1.7: (A) Simulated chromatogram for the separation of two co-eluting peptides. (B) Simulated chromatogram for 100 peptides with identical peak widths and height to demonstrate chromatogram complexity with high number of compounds. (C) Experimental LC-MS/MS chromatogram for the separation of peptides from the tryptic digest of a human whole cell extract. The region in red denotes the dead time and the elution of flowthrough peptides. The region in yellow is the area where gradient separation of retained peptides occur.

1.2.3 Applications of Chromatography in Proteomics

In proteomic LC-MS/MS analyses, chromatography is used to separate extremely complex peptide mixtures prior to analysis in MS system. The most basic methodologies consist of direct connection of the separation system to mass spectrometer. These methods can be classified as one dimensional

LC-MS/MS or Single Shot Proteomics (SSP)[39,48–56] as shown in Figure 1.8. The objective is to identify as much proteins as possible in one injection of the digested proteome using LC-MS/MS. Chromatographic system is used in this case to spread peptide population and deliver it into the mass spectrometer in uniform fashion to maximize probability of successful detection. Multidimensional chromatography (MDC) involves two stages of separation and used for extremely complex samples, when a single LC separation is insufficient.[104–110] The first stage is the first dimension in MDC where the peptides separate into multiple fractions using an HPLC system. Subsequently, each fraction is injected into the LC-MS/MS analysis system, which is the second dimension, to perform multiple SSP experiments. The output from these experiments combine to represent the results for one sample; hence, MDC is a laborious approach. Typically, in SSP, the samples are analyzed for 1-2 hours using reversed phase (C18) columns. The shorter run times are more favourable for high-throughput experimentations. MDC is using different (compared to C18) separation chemistry and often requires significantly more time than SSP but allows for a deeper sample interrogation and higher number of protein identification.[110,111]

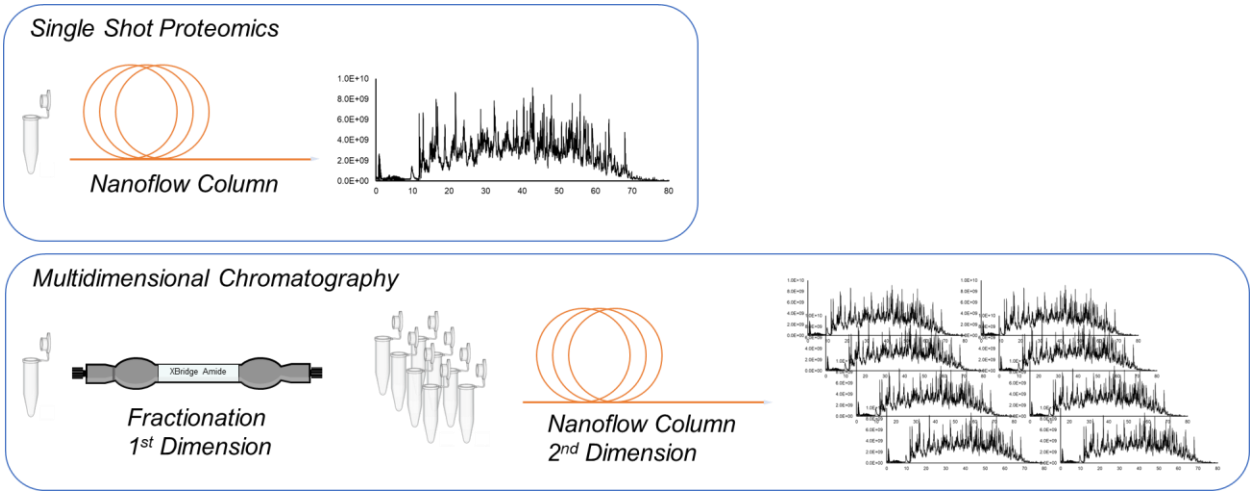


Figure 1.8: The schematic outlining the processes for single shot proteomics and multidimensional chromatography.

1.2.3.1 Chromatographic Advancements in Single Shot Proteomics

Interfacing chromatography online with MS is available in a variety of sizes. Smaller (nanolitre per minute flowrate format – ie. nano-flow) columns are used in proteomics to increase absolute detection sensitivity, critical for minute proteomic samples.[112–114] Typically, these columns are 75-100 µm in inner diameter with a length varying from 100 mm to 500 mm. These smaller columns typically use a few hundred nanoliters per minute flow rates compared to a few hundred microliters per minute for analytical size columns (e.g. 4.6 x 150 mm) and results in higher sensitivitiy.[112–117]

In addition to using nanoflow columns, researchers are also decreasing the sorbent particle sizes to increase peak sharpness. It is common to see researchers changing their column particles from 5 to 3, to sub-2 µm sorbents.[118–121] However, smaller particle sizes correlate to higher column pressures on an HPLC. This can be a practical problem for LC systems where higher pressures require specialized equipment and reduce reliability of solvent delivery. Incorporating a column heater to increase column temperature can decrease the pressure significantly while maintaining superior peak shape in comparison to 3 µm sorbents.[122–125] The peak shape improvements can lead to an increase in detection sensitivity for peptides. It is important to note that heating the column reduces peptide retention and leads to the loss of poorly retained hydrophilic peptides. Therefore, the optimal particle size depends on many factors, such as type of LC equipment, samples, requirements for optimal analysis time, etc.

Gradients are also important to consider for improving the detection of peptides in SSP. There are two predominant methods for designing gradients: linear and non-linear.[126–128] The linear gradient (for typical reversed-phase C18 separations) increases the concentration of the organic solvent in the mobile phase at a fixed rate over time. Typically, 40% ACN will be sufficient to elute

all peptides. Therefore, if the gradient spans over 90 minutes and ACN concentration increases from 0% to 40%, then the rate of increase is 0.44% ACN/minute. However, linear gradients yield chromatograms that resemble a gaussian distribution of eluting peptides. The earlier and later regions in the chromatogram have a low number of peptides in comparison to the middle region. This is not efficient for peptide detection. Therefore, a non-linear gradient approach addresses this by accelerating the increase of ACN concentration in the less populated areas to have peptides distribute equally along the chromatogram.[126] The non-linear gradient is important chromatographic advancement in proteomics to achieve high quality results that we elaborate in Section 1.5.

1.2.3.2 Chromatographic Advancements in Multidimensional Chromatography

MDC is an extension of SSP where there is an additional pre-fractionation component prior to LC-MS/MS.[105] The chromatographic method used for the pre-fractionation step is called the first dimension separation. The subsequent LC-MS/MS procedure performed for each fraction is the second dimension, and often employ the same chromatography approach of RP HPLC with formic acid ion-pairing modifier as SSP. Conceptually, MDC is designed to improve detection of peptides by distributing them uniformly over multiple fractions. Therefore, each fraction will be a less complex mixture than the unfractionated mixture. An important consideration for MDC is that the two dimensions should separate peptides with different selectivity.[129] The spread of peptides improves with increasing difference between the two chromatographic modes and this topic is explored in Chapter 2. The even distribution of peptides' population across fractions achieved by the differences between the two chromatographic modes maximizes the chances for successful unique peptide identifications by mass spectrometer, which will be further elaborated in Section

1.3. There are a variety of chromatographic method applicable in the first dimension to improve peptide detection such as RP HPLC, HILIC, IEX, mixed-mode separation.[104–107,109,110]

Generally, reversed-phase separations are not suitable as the first dimension techniques prior to RP HPLC (formic acid) due to similarity in retention mechanisms.[104] Drastically changing pH of the eluent allows to alleviate this problem by altering the amino acids' ionization properties. Acidic residues are neutral at acidic pH (with formic acid) and ionized at pH 10. Conversely, basic amino acids are ionized at acidic pH and become neutral at basic pH. Therefore, the same peptide can have different interactions with the RP sorbent at acidic and basic conditions. Taking advantage of this difference in interaction, Gilar et al. demonstrated that separating peptides into fractions using RP at pH 10 is effective to increase peptide detection when acidic pH RP is the second dimension.[130] Development of silica-based RP sorbents that are stable in a wide range of pH values from 2-10 was necessary to make pH 10 RP possible.[131] This first dimension technique reduces the number of peptides that co-elute in acidic pH RP in each fraction. A metric to evaluate this effect is called separation orthogonality[104,132] and used to evaluate the dissimilarity between two chromatographic methods for peptide separation, which we elaborate in Section 1.5. Aside from basic pH RP, C18 stationary phase can be used at any pH depending on pH range stability[133,134] and can involve other ion pairing reagents[93,96] to improve separation orthogonality. The overall goal when optimizing these techniques is to minimize the complexity of each fraction to enhance the overall peptide detection.

Beyond pH 10 RP, SCX is a viable approach to separate peptides in the first dimension of MDC. One of the first MDC methods in proteomics from Raida et al. used cation exchange chromatography.[135] Washburn et al. extends this concept using an automated approach to improve the throughput of MDC.[136] The technique is multidimensional protein identification technology

(MUDPIT). Both methodologies use a salt gradient to elute the peptides; however, other methods are possible to achieve peptide separation with SCX. Establishing a pH gradient on the LC can also be an alternative option to separate peptides with different ionization points (pI) on SCX.[137] Furthermore, using TFA is effective in displacing peptide-stationary phase interaction with peptide-acid interaction via ion-pairing.[138] There are many creative ways to approach MDC with a variety of chromatographic modes. This is one of the reasons why there are on-going efforts utilizing MDC in proteomics. However, be efficient at designing MDC methods, one must understand how peptide separate. In our workflow, we use peptide retention time prediction models to improve this understanding and aid in method development.

1.2.4 Peptide retention time prediction models

Peptide retention time prediction models were proposed to improve method development in chromatography to facilitate selection of eluent conditions for complete separation of peptide mixtures. In the early 1980s, Rose et al.[139–141] and Kyte et al.[142] established the quantitative hydrophobicity scales for understanding how amino acids in the protein sequence influence the protein's three-dimensional structure. These hydrophobicity values were obtained from X-ray crystallography data or reported amino acid physicochemical properties like distribution within water-octanol biphasic systems. Since HPLC is able to separate compounds based on their physicochemical interaction with the sorbent, chromatographers attempted to contribute to the development of representative hydrophobicity scales through chromatographic methodologies. This notion was supported by Molnar et al.[143] where they established the correlation between partition coefficients with the C18 retention time of amino acids. In 1980, Meek attempted a

computational approach to solve for the hydrophobicity values.[144] He developed a model that uses the composition of amino acids to predict the experimental retention time of the 25 peptides chosen for the study. The resulting variables trained in this linear model corresponded to the hydrophobicity values of the amino acids when conjugated as a peptide. Subsequent approaches to solve for amino acid hydrophobicity values was done by Guo et al.[145,146] which involved chromatographic separation synthetic peptides that were substituted at one position with each of the 20 amino acids. Beyond amino acid contribution to retention on C18, Mant et al. later reported that the length of peptide also influenced the interaction with the stationary phase.[147] At the time, the state of the art retention models were additive effects of amino acid composition along with the peptide length. The accuracy of these models expressed as pearson correlation R^2 values for linear correlation of experimental versus predicted retention times has been somewhat misleading as these models were developed for very small peptide collections such that the models were prone to overfitting.

During the early 2000s, MS-based proteomics was advancing and required retention time prediction algorithms to assist with quality control and data validation.[148] Since most peptide MS-based identification strategies have some degree of uncertainty, retention time has been viewed as an independent parameter to increase confidence of peptide sequence assignments. A variety of retention time prediction algorithms were developed to accommodate this demand. Methods involved linear regression,[149–152] statistical learning,[153–159] and artificial neural networks.[160,161] At this time, the goal of proteomic researchers was to generate accurate prediction models for practical use, while chromatographic specialists tried to deepen their insights into actual peptide retention mechanism. These developments have been fuelled by ever-growing size of high-quality retention datasets derived from proteomic LC-MS/MS experiments.

In all these approaches, accuracy was a consistent problem, especially for linear regression methods. Palmblad et al.[151] and Gilar et al.[162] employed linear regression models to model peptide retention under a variety of chromatographic conditions and reported the limitations to obtain accurate predictions. The limitations of all these models were the use of amino acid composition to model interaction, thus leaving behind peptide sequence in which these resides were connected. The sequence of peptide has a large influence on a peptide's physicochemical properties. In 2004, Krokhin et al. published the Sequence-Specific Retention Time Calculator (SSRCalc) algorithm, which was the first algorithm to incorporate sequence-specific peptide features.[97] Its first version considered the alteration of amino acids' hydrophobicity values depending on proximity to peptide termini. This resulted in development of the most accurate model in the field at the time. The algorithm was further refined in later years by incorporating corrections for (i) nearest neighbor effects, (ii) presence of clusters of hydrophobic amino acids, (iii) presence of proline, (iv) isoelectric point, (v) peptide length, and (vi) propensity of helical structure formation.[163]

SSRCalc became the predominant retention prediction engine in proteomics. Although its prediction accuracy supersedes other retention prediction model, there were a class of peptides that consistently has low prediction accuracy. These peptides were speculated to have helical structures when interacting with reversed-phase and behave differently from random-coil peptides. In 2014, Krokhin et al. attempted to characterize the retention for this class of peptides by studying stabilization of peptide helical structures by N-cap motifs.[164] The N-cap motif was initially reported in 1970s from crystallography studies on the alpha-helical portion of proteins. In terms of peptides separation through RP, the peptides capable of forming helices will contort into a helix with a hydrophobic face that maximizes interaction with the C18 phase. Therefore, it is common to observe helical peptides will have experimental retention times that are much higher than the

predicted retention time. In this study, it was found that the N-cap motif was consistent among peptide with high positive retention shift; however, predicting whether the peptide will engage in helical formation remain elusive.

Due to steady increase in the size of retention datasets (up to few hundred thousand by 2016) there was a resurgence of retention time prediction algorithms that involved incorporating deep machine learning-based approaches. Gessulat et al. in 2019 developed Prosit[165] which employed an encoder-decoder model. This scheme tried to learn how peptides interact by extracting the intangible amino acid relationship-of-relationships using sequential layers of bidirectional gated recurrent units (GRU). GRUs were also used for the decoders that resolves the retention time after computing the values through a series of fully connected layers. This large model had a total of 2 million parameters for modeling retention time of peptides on C18. In 2021, Bouwmeester et al. published DeepLC that used deep convolutional neural networks (CNN) predict retention of post-translationally modified peptides.[166] This model tries to extract spatial relationships between amino acids and understand their interaction with the C18. However, these understandings are also intangible due to their use of sequential layers that model relationship-of-relationships. These techniques were able to achieve higher prediction accuracy than SSRCalc. However, these models have all classical traits of black box approaches. The retention features, which drive peptide retention cannot be extracted directly through these models.

Therefore, shallow learning models are the best of both worlds. By combining the physics-driven features from SSRCalc along with the spatial architecture and encoding schemes from Prosit and DeepLC, we should be able to surgically interrogate what each feature does in these shallow models. Yeung et al. reported the use of shallow learning models for this approach and found that GRUs are more effective at modeling peptides sequence with respect to predicting retention

times.[167] Furthermore, they elucidated the underlying parameters that can contribute to a molecule's hydrophobicity. Through their optimization of the model space, they found that by restricting variable space to have only one variable per amino acid – the result variables directly correlate with hydrophobicity values. However, when more variables were allowed for each amino acid, some variables emerged to correlate with other physicochemical properties like cross collision section and accessible surface area. This may implicate that hydrophobicity was driven by cross collision section and the accessible surface area of an amino acid. Further investigation into shallow learning models may elucidate explainable features that were not previously obvious for peptide retention and chromatography.

1.3 Mass Spectrometry

1.3.1 Introduction to Mass Spectrometry

Mass spectrometry is a general term for wide variety of analytical methodologies which allow measuring masses of molecules.[168] Fundamentally any mass spectrometer consists of three main components: ionization, mass filtering (separation), and detection. Advanced hybrid mass analyzers are capable of molecules' fragmentation followed by mass filtering and detection of correspondent fragments. This section outlines general principles of operation of hybrid ESI - quadrupole – orbitrap mass spectrometer capable of ion fragmentation.

Traditionally, MS used harsh ionization methods like electron impact, where an electron beam excites an incoming molecule to become an ion in the gas phase.[168,169] This form of ionization breaks the molecule apart into smaller fragments and is challenging to resolve when multiple molecules enter into the mass spectrometer. Furthermore, electron impact is not readily compatible with LC for peptide analysis.[170] Harsh ionization is not applicable to proteomics. Soft ionization

techniques such as electrospray ionization (ESI)[171,172] and matrix-assisted laser desorption ionization (MALDI)[173] yield intact ions that have *m/z* ratios corresponding to the molecular mass of the target plus a proton(s) for positively charged peptide ions ($[M+H]^+$, ($[M+2H]^{2+}$, etc.). ESI and MALDI differ in typical charge states observed. MALDI yields predominantly singly charged, while ESI results in multiply charged molecular ions. Eventually, ESI found much wider application in proteomics as it can be directly connected to LC for continuous infusion. The intact peptide *m/z* is critical for further studies involving a process known as data-dependent acquisition for peptide identification, which we explain section 1.3.3. As ESI is continuously ionizing compounds that elute from the LC column, there will be a plethora of analytes that enter the MS simultaneously. Ion filtering devices are used to measure *m/z* values and to select specific ions for fragmentation if required.

Quadrupoles are an ion filtering device that stabilizes the flight path of ions with a particular *m/z*. Only the incoming ions with the target *m/z* progress further into the instrument and all other ions are destabilized and not analyzed. There are two main types of acquisition experiments available for quadrupole-orbitrap MS instruments. MS1 is the spectral analysis of all intact ions that are introduced to the MS at a given time point. The quadrupoles will not filter for this type of acquisition. Masses of all ions will be measured in a high resolution orbitrap mass analyser as we want the detector to observe all ions. MS2 is the spectral analysis involving the fragmentation of target ions. In this scenario, the quadrupoles filter to isolate a specific ion with *m/z* value determined in preceding MS1 orbitrap mass measurement. Subsequently, the selected ions are transported to the collision cell where they collide with inert gas molecules generating fragment ions. In the context of peptides, the collision cell will fragment the peptides into smaller peptide fragments or

even amino acids when the high energy ions ablate the peptide. Subsequently fragment ions are transferred to orbitrap for mass measurement.

There are a variety of mass analyzers available including magnetic sector, time-of-flight, linear ion traps, quadrupole ion traps, and orbitraps[168]. The instrument we operate throughout this thesis utilizes orbitrap analyzer. In comparison, the other mass analyzers physically count ions with ion-counting devices;[174–177] however, orbitrap detectors use image current detection to infer the number of ions that are inside the detector.[178,179] If a cluster of ions is near an electrode, the ions can induce a current on the electrode without contact, this is the image current.[180] The smaller the distance between the ions and the electrode, the higher the image current. The orbitrap detectors have an internal rod, and the electrode is the outer barrel of the detector illustrated in Figure 1.9. When ions are inside the orbitrap detector, the ions oscillate around the internal rod. If an electronic excitation is active inside the orbitrap detector, the ions will increase their radius of oscillation that brings the ions closer to the outer electrode. This will result in an increase in image current on the electrode. Subsequently, ions will relax back to their original oscillations at different rates depending on the *m/z.* Therefore, the image current collected from the start of ion excitation to the end of relaxation will compass the different frequencies of ion relaxation shown in Figure 1.9. Upon using the Fourier transform on the image current signal, we can solve which ions are inside the detector for that scan.[181,182] This method has a higher resolution than traditional ion counting detectors.[168,178,179] Therefore, the orbitrap detector is classified as a high-resolution mass analyzer. This resolution is critical to discern the signals belonging to peptide *m/z* apart from confounding signals from adjacent *m/z*, as seen in Figure 1.9.

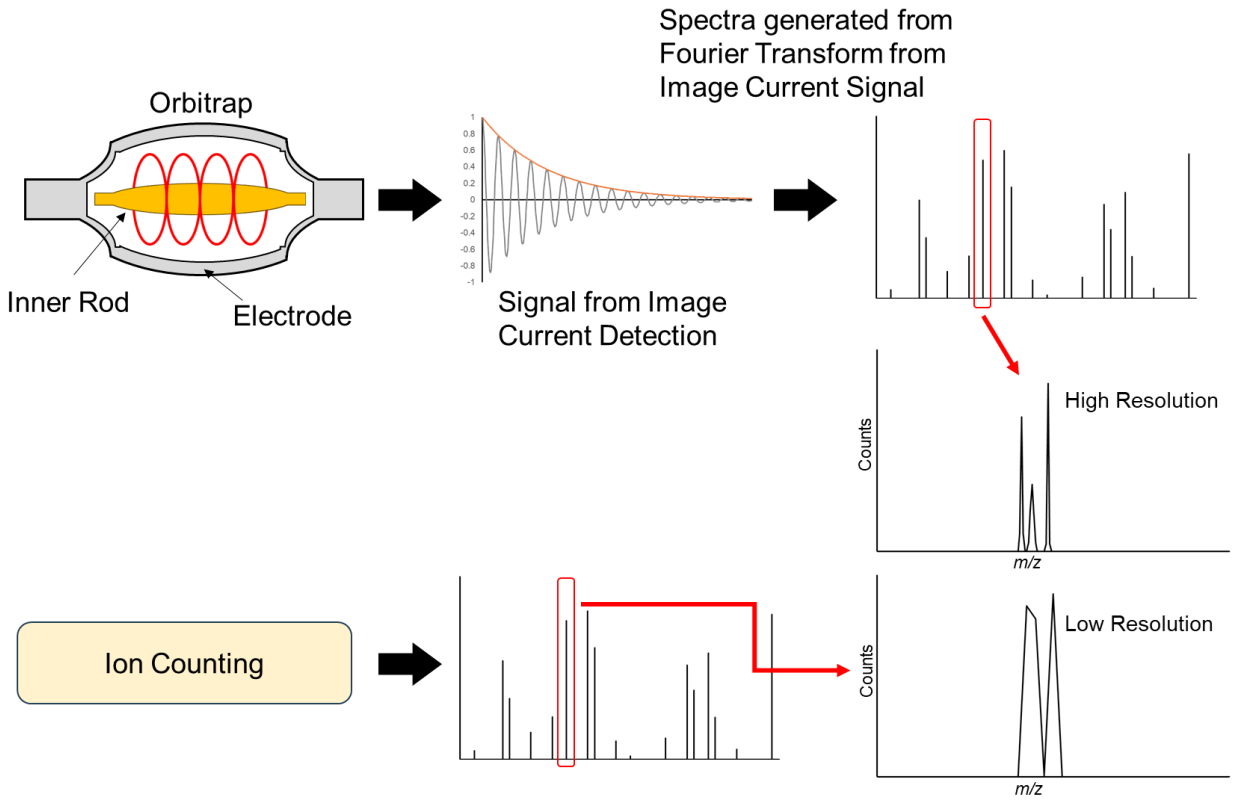


Figure 1.9: Schematic diagram outlining the structure of an orbitrap detector and the comparison data acquisition workflow for an orbitrap acquisition versus the traditional ion counting MS approaches to yield high resolution and low resolution spectra, respective

1.3.2 Interpreting an MS spectra

The data generated from an MS is represented as a spectrum. The spectrum reports the number of ions the MS detected for each *m/z* value. The MS1 spectra are used for surveying all ionized peptides entering the mass spectrometer and subsequent decision on which ion to select for fragmentation. Accurate MS1 measurements are also critical for assigning peptide identity based on calculated *m/z* value. However, the mass of intact molecule alone is not sufficient for confident identification in complex samples. MS2 spectra provided the information needed for confident

peptide identification as they contain information on positioning of individual amino acid residues encoded in m/z values of fragment ions. In the context of proteomics, the fragmentation process is well characterized for peptides when using higher-energy collision dissociation (HCD) in orbitrap-based MS systems.[183,184]

Peptides fragment in HCD along the C-N (peptide) bond between the carbon of the carbonyl and the nitrogen of amino group.[185,186] The fragment containing the N-terminal portion of the peptide is referred to as b-ion. On the other hand, the fragment containing the C-terminal of the peptide is called the y-ion. Adjacent b or y ions differ by the mass corresponding to the amino acid residue. Therefore, finding a consecutive series of fragment ions can unequivocally infer the amino acid sequence.

The b/y ion fragmentation series is unique to a peptide. However, additional difficulties arise from the fact that b and y ions are present on the same spectra and some fragments could be missed due to lower intensity signals that may overlap with other fragments.[187]

1.3.3 Data dependent acquisition and protein identification workflows

Initially, the MS1 spectra surveys all ion species entering the mass spectrometer following LC-separation. From the MS1 spectra, the data dependent acquisition (DDA) algorithm selects ions above a particular signal intensity threshold as seen in Figure 1.10.[188–190] An intensity threshold is put in place because low ion count MS1 usually leads to poor MS2 spectral quality. The vast majority of DDA experiments in this thesis follow the same DDA protocol: an initial MS1 is collected to select the top 20 ions in the spectra for MS2 acquisition. From this type of analysis that typically is a 1-2 hour LC-MS acquisition, approximately ~60000-80000 MS2 spectra will be generated. It is important also to note that not all ions in the MS1 are peptide ions. The ions can

also be electronic noise, chemical noise, or sample contaminants.[168,191] Hence, enhancing peptide signals would decrease the possibility of analyzing non-peptide ions.

Typically, the algorithm starts collecting MS2 spectra from the most intense ion to the least intense.[188–190] After the first round of MS2 acquisition, *m/z* values for fragmented ions are placed into exclusion list to prevent re-selection until a certain amount of time elapses, which the operator defines depending on expected width of chromatographic peak.[121] There is only a limited amount of time to collect the necessary information during the elution of a peak. Hence, operators need to optimize their LC-MS/MS methods to maximize data collection during the elution process.

LC-MS/MS data from DDA experiments contain information on *m/z* values for the intact peptides (precursor ions from MS1) and their correspondent fragments from MS2 measurements. Conventional search algorithms like X!Tandem[192,193] or MSFragger[194] first check in a peptide library (generated from protein sequences database) the potential peptides that have *m/z* values matching the precursor *m/z*. From this list of peptides, an algorithm calculates the b/y ions for each peptide. Subsequently, the algorithm matches each set experimental MS2 spectra with the calculated b/y ions for a potential peptide candidate, shown in Figure 1.10. The matching process attempts to determine the number of matched b/y ions for each peptide candidate. Since the assumption in most basic DDA algorithms is that each spectrum can identify at most one peptide, the peptide with the highest number of fragment matches is the identified peptide. This peptide-spectrum pair is called the peptide-spectrum match (PSM). Each peptide in the library has additional information such as the protein of origin. Based on a PSM, we can associate the signal intensity to a protein along with additional LC-MS/MS information. If a peptide maps to multiple proteins, the operator must decide on the validity of including the peptide for downstream analysis.

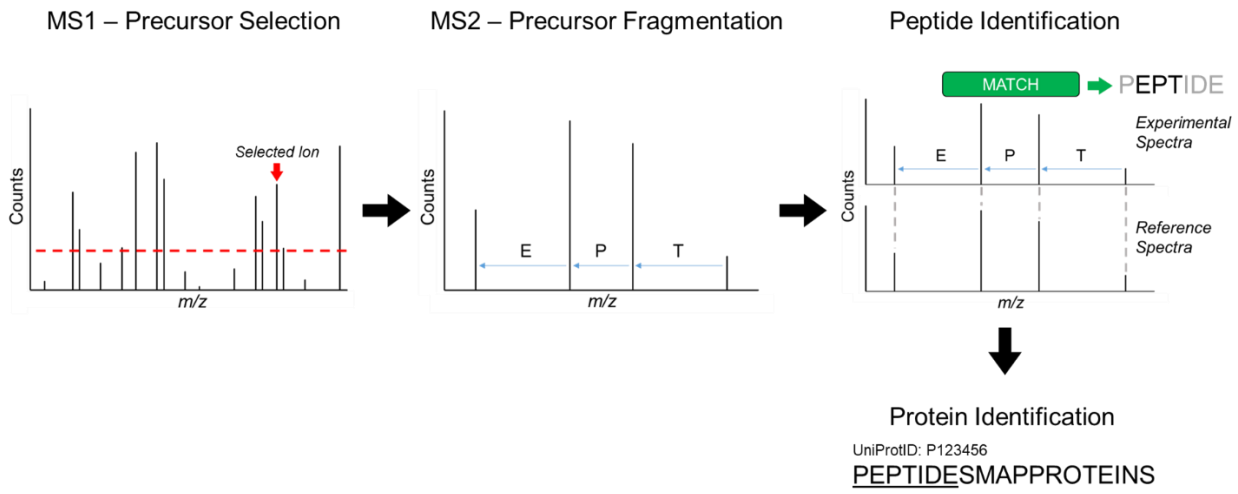


Figure 1.10: Data analysis workflow for the data dependent acquisition outlining the procedure from MS1 to MS2 to peptide identification of a peptide ion.

1.4 Quantitative Proteomics and Current Methodologies

Now that we have explored the topics of proteomics (Section 1.1), chromatography (Section 1.2), and mass spectrometry (Section 1.3), how can we apply LC-MS/MS to investigate changes to a proteome in biological systems? Detecting and identifying proteins is insufficient to contextualize changes in biological perturbations. To apply principles of differential expression, there needs to be a way to extract the protein expression information from the MS data. During the peptide and protein identification process, the protein identification contains the MS information for its constituent peptides. We can extract signal intensities for each protein to compare the relative changes at different biological conditions to perform differential expression. There are two main approaches for quantitation of the proteome involving DDA that are isobaric tag-based quantitation and label free quantitation.[195,196]

1.4.1 Isobaric Tag-based Multiplex Quantitation Methodologies

An approach to quantifying proteins across different experiments is to incorporate isobaric tags that acts as a small molecule barcode.[195,197–199] These tags chemically attach to digested peptides, allowing us to distinguish which biological experiments the peptides are from. These barcode tags are designed to have the same mass because the reporter mass of the tag is compensated by a mass normalizer portion of the molecule by incorporating various stable isotopes, shown in Figure 1.11. Therefore, labeled peptides of the same sequence from different biological experiments have identical mass and other chemical properties. When these peptides are mixed after labeling, they separate together through LC system and simultaneously enter the mass spectrometer. Upon fragmentation, the reporter ion separates from the mass normalizer. Therefore, the corresponding MS2 spectra contain the standard b/y ion fragments and the fragments for the reporter ions in the low *m/z* region corresponding to individual labeled samples as shown in Figure 1.11. Since each isobaric tag reports a different biological condition, the fragment intensity for each tag corresponds to the peptide intensity in the respective experiments. This is a multiplexed approach to quantify multiple experiments simultaneously in one LC-MS/MS experiment. The effect of these isobaric tags on the retention of peptides is explored in Chapter 6.

Gygi et al. demonstrated the first isobaric tag known as isotope-coded affinity tags that were labeling the cysteine through an alkylation step.[200,201] The initial commercial offering of such isobaric tags was known as iTRAQ reagents that utilize N-hydroxysuccinimide (NHS) chemistry to derivatize peptide's primary amine sites.[202,203] The resulting peptide has a mass shift of +144.102 and has higher retention on RP chromatography because the tag is hydrophobic. Due to the chemical structure of iTRAQ, only 4-plex[203] and 8-plex[202] are available; thus, comparisons for

differential expression can only evaluate eight conditions simultaneously. This can be limiting for establishing statistical significance when evaluating differential expression with hypothesis testing approaches. If the experimental design surpasses eight conditions, then more channels are needed. The second type of isobaric tags, tandem mass tags (TMT), also utilize NHS chemistry.[204] With a different chemical structure, this isobaric tag can accommodate up to ten channels. When sample sets start to expand for comparing multiple conditions with many replicates, ten channels can be insufficient. Practitioners may need an additional channel to normalize the signals between LC-MS/MS acquisition.[205] Thus, the eleventh channel is made available to accommodate larger studies. Further expansion of TMT resulted in a new tag structure with the name TMTpro, which allows for up to 18-plex analysis.[206] Novel approaches of mixing multiple isobaric tags in succession gave rise to a new sub-field known as hyperplexing.[207–210] This technology can have up to 45 channels in one LC-MS/MS acquisition.[209] However, these high-plex experiments vary in their consistency and success.

Complex chemical composition of the tags and similarity of reporter ions generated from fragmentation of different precursor ions represents serious problem.[211–213] Channel bleed is the result of ion intensities from other channels that can confound adjacent channels.[212,213] This can be due to impurities during the synthesis of the tags or co-isolation of co-eluting peptides in the MS.[205,212,213] In conventional workflows, utilizing MS2 can result in the isolation of co-eluting peptides with similar *m/z*. Therefore, their reporter intensities amalgamate as if they were one peptide.

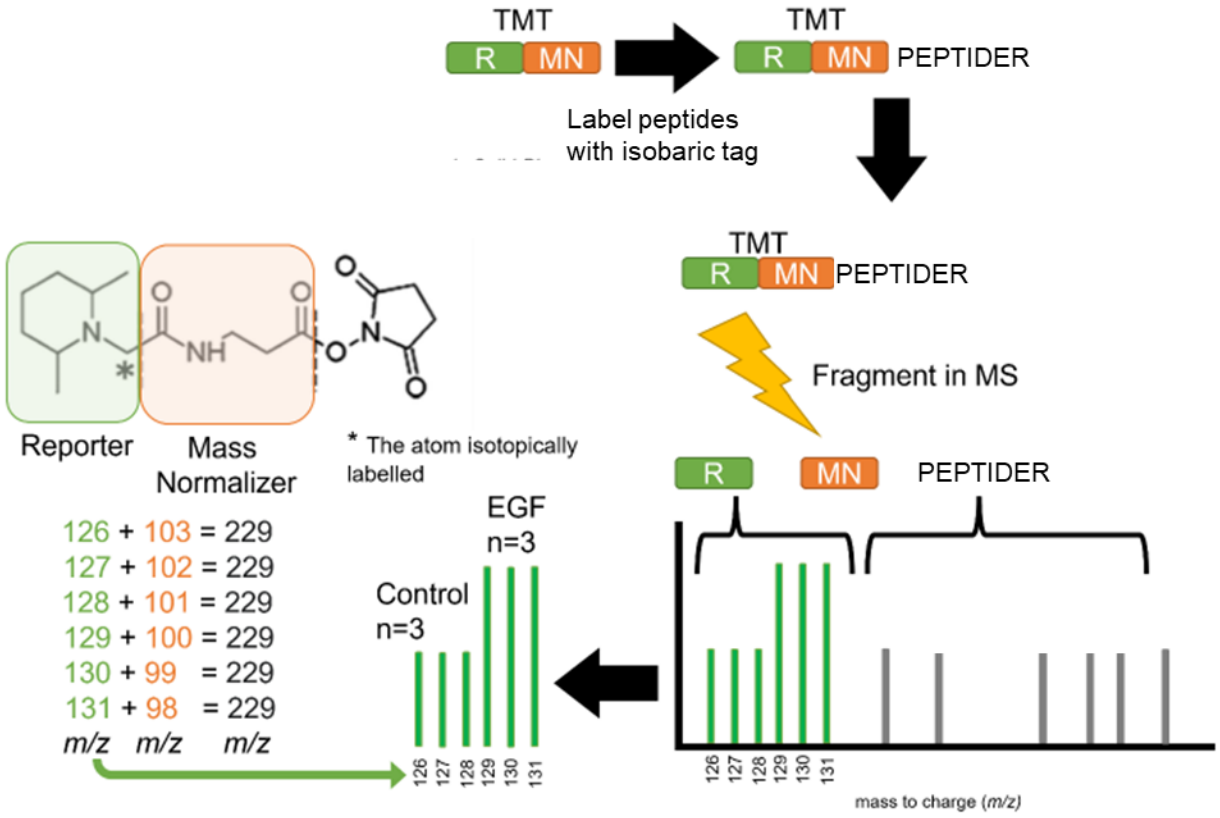


Figure 1.11: Schematic diagram for the analysis of isobaric tags on peptides for multiplex quantitation on MS.

1.4.2 Label-Free Quantitation of Proteins in Bottom-up Proteomics

An alternative approach is label-free quantification where SSP directly analyze the digest peptides without additional labeling.[195,196,214,215] Each experimental condition has an independent SSP experiment. After protein identification, the quantification workflow extracts the peptide signal intensities across the SSP acquisitions to perform differential expression. However, there are different strategies to extract the peptide intensities from the MS data. Since there are two different

types of scans for the MS, peptide quantitation workflows can use on MS1 or MS2 intensities, each having their own advantages and disadvantages.[196,214,216–218]

MS1 quantitation extracts the peptide information from the MS1 spectra based on the *m/z* of the peptide.[214,216,218,219] Since the MS collects these spectra at constant interval, there is a guarantee to collecting multiple data points per each chromatographic peak during the peptide elution. Having more points per peak can enhance the accuracy of quantitation.[220,221] If too little sampling is done during the elution, it is possible to underestimate the true concentration of the peptide. It is important to note that this quantitation method still relies on MS2 spectra to identify the peptides in the sample. The disadvantage to this method is the assumption that the *m/z* is unique to a target peptide. If the target peptide and a contaminant share overlapping *m/z* values, the quantitation is inaccurate as we will be quantifying the peptide plus the contaminants together.[222] Therefore, high resolution MS systems are used for this type of approach to minimize the occurrence of these overlaps.

MS2 quantitation uses the peptide information from the MS2 spectra to report the intensity value for the peptide.[218,223–226] This method is more specific than the MS1 approach as we are quantifying the protein amount based on the signal intensities of the identified fragment ions in the MS2 spectra. However, DDA is a stochastic process and there is no guarantee that the MS2 for a particular peptide will be collected for all samples that are interrogated for quantitation. Therefore, MS2 quantitation is prone to encountering the missing value problem where certain biological conditions will have no quantitative information available. Furthermore, the dynamic exclusion parameter limits the number of data points collected across LC peak. This can result in an individual peptide being fragmented before or after reaching true maximum of chromatographic peak, resulting in inadequate quantitation accuracy.

Protein quantitation is not trivial and is dependent on the operator's training to make informed decision to either proceed with MS1 or MS2. Furthermore, there needs to be a balance between optimizing the LC-MS/MS method to detect more proteins versus detecting proteins more consistently. The general rule of thumb is to have at least two unique peptides per protein to perform quantitation.[227–229] This increases the confidence that the protein of interest is being identified correctly. Peptides that are used for protein quantitation ideally should not carry post-translational modifications or contain amino acids sites that are prone to spontaneous modifications or single amino acid variations.[230,231] The latter is especially critical in clinical proteomics. The measured peptide intensities must be consistent between replicates. Coefficient of variation (CV) metrics is used to evaluate measurements' reproducibility.[232–234] CV estimates the variance of the peptide quantitation in the replicate analysis. Low variability is favourable and increases the significance in hypothesis testing when performing differential expression analysis.

1.5 Current Advancements in DDA-based LC-MS/MS Methods for Proteome Quantitation

We have elaborated in the previous sections about the well-defined technology used in contemporary LC-MS/MS-based proteomics. Conversely, our conceptual understanding as a community of how proteins differentially express and which proteins are involved in multitude of biological processes remains elusive. There is always a notion of *what if the protein that was not detected is a key factor regulating a biological process*? To iteratively improve our understanding of how proteins respond to stimulus, one of the ways is to create LC-MS/MS methods that can quantify the entire proteome.[40,41] Regardless of analysis depth and time spent on LC-MS/MS acquisitions, there is a lack of methods that can provide quantitative information on the entire

expressed proteome. Enhancing peptide detection and protein coverage in proteomic experiments requires further improvements to both, MS and LC technologies.

1.5.1 MS Technologies to Increase Proteome Coverage

In the early stages of proteomics, the main MS systems used ion counting devices for detectors and therefore, were lower resolution. During the 2010s, the Thermo Fisher LTQ Velos was the predominant MS used for proteomics but had a slow MS2 scan speed of 1 MS2 per second or 1 Hz.[235] This means the MS could only acquire 3600 MS2 spectra in a 1-hour MS acquisition.[57] Hence, the proteomic methods involving the LTQ Velos were often slow and could take up to 8 hours to identify 5000 proteins.[236] Subsequently, in 2011, the TripleTOF 5600 MS systems was introduced from AB SCIEX as a higher scan speed instrument achieving a maximum of 50 Hz and this instrument was able to identify 1000 yeast proteins in 1 hour. In 2012, the Thermo Fisher released the newer Orbitrap Q Exactive instrument had better performance to acquire 13 MS2 spectra per second and was the first commercial instrument with a high resolution orbitrap detector. Through this innovation, it was possible to detect ~4000 yeast proteins in 4 hours.[237] Through further optimizations to the acquisition methods, it was possible to achieve ~4800 protein identifications from human cell lysates in 4 hours.[56] Subsequently, in 2014, the Orbitrap Q Exactive was upgraded to the Q Exactive HF model where the sensitivity of the instrument increased. Therefore, ~5000 protein identifications from human cell lysates were attainable in a 90 minute LC-MS/MS acquisition.[55] Following these achievements, the enhanced Q Exactive HF-X and Exploris 480 further improved identification output to result in identifications of ~5900 human protein in 30 minutes LC-MS/MS acquisitions.[53,54] The advanced scan modes available for these two instruments made it possible to detect peptides for a large number of proteins in a short period

of time. Kawashima et al. further pushed the limits using the advanced scan modes to detect 10,000 proteins in 2 hours.[39]

Most recently, the Orbitrap Astral is the newest orbitrap-based MS instrumentation that can yield 12,000 protein identifications in 1 hour.[40] These identification statistics are achieved by simultaneously applying the advanced scan modes, novel instrument architecture, and MDC together. However, the 5-minute LC-MS/MS acquisitions for each of the 24 fractions obtained from the MDC is limited to ~5000 protein identifications. It is likely the improvements to MS instrumentation are reaching a plateau for protein identification and vendors are instead focusing on throughput over depth. With each instrument succession, the MS collects more spectra in less time, but it is unrealistic to rely on a brute-force approach to mine mass spectral information for protein identifications.

Upgrading the MS technology can eventually attain comprehensive proteome coverage. However, that constrains the MS as the bottleneck for increasing the number of protein identifications when designing a method. Furthermore, as a field, depending on MS improvements results in an overall dependency on instrument vendors to drive the state of the art in proteomics. LC is another component to the LC-MS/MS methods that can be improved for enhancing proteome coverage. This notion is of particular importance for labs which cannot afford yearly upgrades of MS equipment. Moreover, any improvement that are made upstream of the MS involving LC technology will provide impact on overall LC-MS/MS performance regardless of MS device used.

1.5.2 LC Technologies to Increase Proteome Coverage

Researchers used a different mentality when designing LC technologies for increasing the proteome coverage. One of the factors that affected protein identification for LC-MS/MS methods

was that the separation processes was too fast for the scan speed of the MS instrument. Initially, researchers used long gradient approaches to increase the time needed for peptide separation and therefore, the MS has more opportunities to detect unique peptides. Iwasaki et al., in 2010, utilized a 350 cm long column to separate the *E. Coli* proteome with a 41-hour long gradient.[238] The team successfully identified ~2700 proteins that encompassed most of the proteins expressed by *E. Coli*. In 2011, Kocher et al. implemented a 10-hour long gradient was implemented to separate the human proteome isolated from the HeLa cell line.[239] This approach also resulted in ~2700 protein identifications but achieved this in shorter time. In the following year, Zhou et al. reported the identification of ~4000 human proteins with a 10 hour long gradient but the LC system operated at lower flow rates.[240]

Beyond increasing gradient lengths for SSP experiments, researchers also used MDC approaches for enhancing proteome coverage in an experiment.[104–107,109,110] Furthermore, there were efforts towards high fraction number MDC experiments to improve the depth of protein identification using three dimensional chromatography and fractionation with isoelectric focusing. Spicer et al. separated 720 µg of human whole cell lysate into 126 fractions through a three-dimensional pairing of (1) acidic pH with heptafluorobutyric acid ion pairing additive on RP, (2) basic pH RP, and (3) acidic pH with FA additive on RP.[129] Resulting from this extensive separation, 14,000 proteins were identified with ~250,000 unique peptides. Despite achieving one of the highest human proteome coverage, it was a low throughput method requiring 8 days to analyze the fractions for one biological sample. Branca et al. used a two-dimensional approach involving isoelectric focusing in the first dimension that was able to achieve similar protein identification statistics.[241] The first dimension was a non-conventional high-resolution isoelectric focusing method that separated peptides based isoelectric point. The peptides were fractionated into 72

fractions that were then analyzed by LC-MS/MS. From this protocol, the analysis detected ~80,000 unique peptides that identified ~13,000 proteins. The latter method had a lower peptide identification likely because the fractions from isoelectric focusing resulted in co-elutions in the second dimension. This may have saturated the DDA schedule despite each fraction contained only a subset of peptides from the proteome. The saturation issue highlights the fundamental limitations of DDA-based approaches towards increasing proteome coverage in LC-MS/MS proteomics.

1.5.3 Current Limitations in DDA-based LC-MS/MS

Despite all the significant achievements in LC and MS technology, one of the prevailing issues with DDA-based LC-MS/MS is associated with the basic principles of DDA approach. If a column effluent at any time point during separation has more components than the DDA algorithm can schedule for MS2, then the MS becomes saturated. During saturation, lower abundance species are less likely to be detected. Therefore, desaturating the DDA schedule will enhance the detection of lower abundance species. The problem is exemplified due to presence of very large (producing many peptides like titin) proteins and very abundant species (like albumin in blood). The latter yields highly abundant peptides, which usually elute from LC column as very wide peaks, which prevents the selection of low abundant species for MS2 in these regions. Simplifying the sample mixture and decreasing the total number of unique peptide sequences in the sample is one way to achieve this desaturation. Selecting which peptides to remove, however, is not trivial. If we remove key peptides that are critical to the identification of a protein, that protein will not be identified. In the past, there has been investigation into sub-sampling the peptide space from the proteome through techniques known as peptide-centric proteomics. The goal is to minimize the number of

peptides needed to represent all proteins in the proteome through selective isolation of peptides with specific physicochemical properties or amino acids.

1.5.4 Prior work on Selective Isolation of Peptides from the Proteome

Peptide-centric proteomics involved methods that select groups of peptides for LC-MS/MS analysis that reduce protein identification redundancy on a peptide level.[242–254] In 2002, a chromatographic enrichment method called Combined Fractional Diagonal Chromatography (COFRADIC) was used to isolate methionine-containing peptides.[253] Close to all of proteins in *E. Coli* have at least one methionine residue, hence, the methionine-containing peptides could potentially be used for identifying the entire proteome. From the enriched peptides analyzed on LC-MS/MS, ~800 *E. Coli* proteins were identified. The enrichment was done with two rounds of RP separation. The first round distributed the peptides into multiple fractions along the chromatogram. The methionine-containing peptides in each fraction were chemically oxidized which altered the retention properties of the peptide. Oxidized methionine (Mso) has a lower hydrophobicity than methionine and therefore, the peptides containing Mso elute earlier than other peptides in the corresponding fraction. Hence, the second round of RP chromatography under identical conditions captured this group of Mso-containing peptides that were subsequently, submitted for LC-MS/MS analysis. Through this labor-intense procedure, the protein identification software identified ~1300 peptides belonging to 800 proteins, which was an achievement at the time. Effectively, this method was able to reduce redundancy for the DDA algorithm to improve protein identification.

Aside from isolating peptides with methionine, Liu et al. demonstrated that thiol-decorated resins were able to selectively enrich cysteine-containing peptides.[254] When cysteine-containing

peptides were incubated with the thiol resins, they formed disulfide bonds with each other. Washing the beads removed the unretained peptides and the cysteine-containing peptides were eluted for LC-MS/MS analysis. The resulting analysis detected ~5900 peptides to yield ~3000 proteins identifications. Without the enrichment step, ~10000 peptides were detected to identify ~3000 proteins. The comparison between samples with and without enrichment illustrated that selecting a group of peptides can reduce the redundancy of protein identifications.

Selective Capture of Peptide (SCAPE) followed a different route to achieve redundancy reduction.[251] This method depletes a group of peptides from the sample rather than performing enrichment. The procedure involves reversibly modifying the primary amines on a peptide where the N-terminal and the lysine side chain would be modified but other basic amino acids would not be perturbed. This resulted in peptides containing histidine and arginine to have charges greater than +1. All other peptides will have a 0 charge. The 0 charge peptides were enriched in the flow through of SCX and histidine/arginine containing peptides were depleted from the sample. Subsequently, the amine modification was removed from the histidine/arginine deficient peptides for LC-MS/MS separation.

The authors of these methods hypothesized that this could be a viable way to enhance protein identification. However, these methods were not widely used in proteomics due to the laborious nature of their workflow. They employ various version of chemical modifications, which rarely result in 100% yields and prone to side reactions.[251,253] Furthermore, it would be difficult to perform fractionation procedures such as COFRADIC consistently to conduct quantitation.

SCAPE and COFRADIC are essentially semi-MDC methods involving multiple rounds of chromatography to achieve redundancy reduction. Enriching a select class of peptides with a specific amino acid can be restrictive since not all organisms share the same amino acid

abundances.[248,251,253] If we enrich a group of peptides based on chemical property, that can be more applicable across various organisms and sample types. We can take inspiration from these semi-MDC methods to pre-fractionate peptides using a first dimension column. We then only analyze a portion of the fractions. It is difficult to select an appropriate chromatographic mode to create such a procedure because there is a lack of systematic evaluations for 2D separations. This information is critical to decide what chromatographic mode in the first dimension can yield a portion of peptides that can achieve the highest protein identification with the least amount of peptides needed.[104] Furthermore, not many chromatographic modes have prediction models available to evaluate where peptides distribute for *in-silico* simulations of enrichment procedures. These are issues we will need to address to selectively isolate peptide populations through chromatography for increasing proteome coverage via the desaturation of the DDA schedule.

1.5.5 Our Approach to Increasing Proteome Coverage

Fundamentally, our goal is to create a chromatographic method that can desaturate the DDA schedule through reducing redundant protein identifications on a peptide-level. There are three critical objectives we need to address to successfully construct this method.

The first objective is to identify a chromatographic mode that has a group of peptides that can be isolated to identify the entire proteome. Furthermore, this group of peptide needs to be evenly distributed in the RP LC-MS/MS acquisition to maximize unique peptide detection to improve protein identification. Orthogonality is a parameter that can evaluate the distribution of peptides between two chromatographic modes. Gilar et al. were the first to report on a metric that can quantify this relationship—based on filling percentage of respective orthogonality plots.[132] Conceptually, the orthogonality metric will allow us to determine which fractions in the first

dimension is able to evenly distribute peptides in the second dimension. To identify a suitable first dimension chromatographic mode, we studied sixteen different chromatographic settings to be paired with acidic pH RP. Within these sixteen modes, they include RP, HILIC, IEX, and mixed mode columns, which contains a blend of functional groups corresponding to two or more separation mechanisms. We also wanted to extend the orthogonality metric reported by Gilar et al. to see if the resulting orthogonality values correlate with protein identification statistics. This portion of my research project is outlined in Chapter 2.

The second objective was to model the selected chromatographic modes to understand the peptide retention behaviour on these columns and also to predict *in-silico* ideal peptide enrichment regions that can minimize redundant protein identifications. We expect to observe an increase in protein identification when we subject this group of chromatographically enriched peptides for LC-MS/MS analysis. Therefore, we used SSRCalc retention time prediction approaches to model the peptide retention times in ZIC-HILIC stationary phase at pH 4.5 (Chapters 3) and pH 3 (Chapter 4) to be potentially used for redundancy reduction in proteomics. Finally, in Chapter 5, we demonstrate the utility of such prediction models towards designing redundancy reduction chromatographic methods using the XBridge Amide HILIC mode, that we previously investigated.

The third objective to develop a fractionation technique that can isolate peptides reproducibly for quantitation. With the previous reporting of peptide isolation techniques yielding high variability, we want the redundancy reduction method to be applicable for quantitation. Prior work involving rapid sample preparation schemes employed packed chromatographic material into pipette tips. The separation of peptides occurs within the tip where the mobile phase passes through the stationary phase via centrifugation. These techniques have been previously used in protein quantitation and is promising in terms of attaining low CVs. Therefore, it would be advantageous

to evaluate if it is possible to create a protocol that can support reproducible enrichment of peptides that reduce the redundancy of protein identification for a sample. In the future, we want to integrate isobaric tag labelling to our redundancy reduction workflows. Thus, in Chapter 6, we modelled the changes in retention behaviour of peptides that are labelled with the commercially available isobaric tags.

1.6 References

(1) Bianconi, E.; Piovesan, A.; Facchin, F.; Beraudi, A.; Casadei, R.; Frabetti, F.; Vitale, L.; Pelleri, M. C.; Tassani, S.; Piva, F.; Perez-Amodio, S.; Strippoli, P.; Canaider, S. An Estimation of the Number of Cells in the Human Body. *Ann. Hum. Biol.* **2013**, *40* (6), 463–471. https://doi.org/10.3109/03014460.2013.807878.

(2) Fleck, J. S.; Camp, J. G.; Treutlein, B. What Is a Cell Type? *Science* **2023**, *381* (6659), 733–734. https://doi.org/10.1126/science.adf6162.

(3) Nussinov, R.; Tsai, C.-J.; Jang, H. Protein Ensembles Link Genotype to Phenotype. *PLoS Comput. Biol.* **2019**, *15* (6), e1006648. https://doi.org/10.1371/journal.pcbi.1006648.

(4) Sul, J.-Y.; Wu, C. K.; Zeng, F.; Jochems, J.; Lee, M. T.; Kim, T. K.; Peritz, T.; Buckley, P.; Cappelleri, D. J.; Maronski, M.; Kim, M.; Kumar, V.; Meaney, D.; Kim, J.; Eberwine, J. Transcriptome Transfer Produces a Predictable Cellular Phenotype. *Proc. Natl. Acad. Sci.* **2009**, *106* (18), 7624–7629. https://doi.org/10.1073/pnas.0902161106.

(5) Stamatoyannopoulos, J. A. What Does Our Genome Encode? *Genome Res.* **2012**, *22* (9), 1602–1611. https://doi.org/10.1101/gr.146506.112.

(6) Lu, S.; Zhang, J.; Lian, X.; Sun, L.; Meng, K.; Chen, Y.; Sun, Z.; Yin, X.; Li, Y.; Zhao, J.; Wang, T.; Zhang, G.; He, Q.-Y. A Hidden Human Proteome Encoded by "non-Coding" Genes. *Nucleic Acids Res.* **2019**, *47* (15), 8111–8125. https://doi.org/10.1093/nar/gkz646.

(7) Tompa, P.; Fuxreiter, M. Fuzzy Complexes: Polymorphism and Structural Disorder in Protein–Protein Interactions. *Trends Biochem. Sci.* **2008**, *33* (1), 2–8. https://doi.org/10.1016/j.tibs.2007.10.003.

(8) Nooren, I. M. A.; Thornton, J. M. Diversity of Protein-Protein Interactions. *EMBO J.* **2003**, *22* (14), 3486–3492. https://doi.org/10.1093/emboj/cdg359.

(9) Pereira-Leal, J. B.; Levy, E. D.; Teichmann, S. A. The Origins and Evolution of Functional Modules: Lessons from Protein Complexes. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **2006**, *361* (1467), 507–517. https://doi.org/10.1098/rstb.2005.1807.

(10) Maghsoudi, Z.; Nguyen, H.; Tavakkoli, A.; Nguyen, T. A Comprehensive Survey of the Approaches for Pathway Analysis Using Multi-Omics Data Integration. *Brief. Bioinform.* **2022**, *23* (6), bbac435. https://doi.org/10.1093/bib/bbac435.

(11) García-Campos, M. A.; Espinal-Enríquez, J.; Hernández-Lemus, E. Pathway Analysis: State of the Art. *Front. Physiol.* **2015**, *6*, 383. https://doi.org/10.3389/fphys.2015.00383.

(12) Wu, X.; Hasan, M. A.; Chen, J. Y. Pathway and Network Analysis in Proteomics. *J. Theor. Biol.* **2014**, *362*, 44–52. https://doi.org/10.1016/j.jtbi.2014.05.031.

(13) Kavallaris, M.; Marshall, G. M. Proteomics and Disease: Opportunities and Challenges. *Med. J. Aust.* **2005**, *182* (11), 575–579. https://doi.org/10.5694/j.1326-5377.2005.tb06817.x.

(14) Zubair, M.; Wang, J.; Yu, Y.; Faisal, M.; Qi, M.; Shah, A. U.; Feng, Z.; Shao, G.; Wang, Y.; Xiong, Q. Proteomics Approaches: A Review Regarding an Importance of Proteome Analyses in Understanding the Pathogens and Diseases. *Front. Vet. Sci.* **2022**, *9*, 1079359. https://doi.org/10.3389/fvets.2022.1079359.

(15) Bradbury, N. A. CFTR and Cystic Fibrosis: A Need for Personalized Medicine. In *Ion Channels and Transporters of Epithelia in Health and Disease*; Hamilton, K. L., Devor, D. C., Eds.; Springer New York: New York, NY, 2016; pp 773–802. https://doi.org/10.1007/978-1-4939-3366-2_24.

(16) Rowe, S. M.; Miller, S.; Sorscher, E. J. Cystic Fibrosis. *N. Engl. J. Med.* **2005**, *352* (19), 1992–2001. https://doi.org/10.1056/NEJMra043184.
(17) Sheppard, D. N.; Welsh, M. J. Structure and Function of the CFTR Chloride Channel. *Physiol. Rev.* **1999**, *79* (1), S23–S45. https://doi.org/10.1152/physrev.1999.79.1.S23.
(18) Chaudhuri, T. K.; Paul, S. Protein-misfolding Diseases and Chaperone-based Therapeutic Approaches. *FEBS J.* **2006**, *273* (7), 1331–1349. https://doi.org/10.1111/j.1742-4658.2006.05181.x.
(19) Le Quesne, J. P.; Spriggs, K. A.; Bushell, M.; Willis, A. E. Dysregulation of Protein Synthesis and Disease. *J. Pathol.* **2010**, *220* (2), 140–151. https://doi.org/10.1002/path.2627.
(20) Vogel, C.; Marcotte, E. M. Insights into the Regulation of Protein Abundance from Proteomic and Transcriptomic Analyses. *Nat. Rev. Genet.* **2012**, *13* (4), 227–232. https://doi.org/10.1038/nrg3185.
(21) Huster, D.; Hoppert, M.; Lutsenko, S.; Zinke, J.; Lehmann, C.; Mössner, J.; Berr, F.; Caca, K. Defective Cellular Localization of Mutant ATP7B in Wilson's Disease Patients and Hepatoma Cell Lines. *Gastroenterology* **2003**, *124* (2), 335–345. https://doi.org/10.1053/gast.2003.50066.
(22) Polishchuk, R. S. Cellular Function of ATP7B (Wilson ATPase). In *Clinical and Translational Perspectives on WILSON DISEASE*; Elsevier, 2019; pp 45–56. https://doi.org/10.1016/B978-0-12-810532-0.00006-9.
(23) Rao, V. S.; Srinivas, K.; Sujini, G. N.; Kumar, G. N. S. Protein-Protein Interaction Detection: Methods and Analysis. *Int. J. Proteomics* **2014**, *2014*, 147648. https://doi.org/10.1155/2014/147648.
(24) Ofran, Y.; Rost, B. Analysing Six Types of Protein–Protein Interfaces. *J. Mol. Biol.* **2003**, *325* (2), 377–387. https://doi.org/10.1016/S0022-2836(02)01223-8.
(25) Olson, W. C.; Rabut, G. E. E.; Nagashima, K. A.; Tran, D. N. H.; Anselma, D. J.; Monard, S. P.; Segal, J. P.; Thompson, D. A. D.; Kajumo, F.; Guo, Y.; Moore, J. P.; Maddon, P. J.; Dragic, T. Differential Inhibition of Human Immunodeficiency Virus Type 1 Fusion, Gp120 Binding, and CC-Chemokine Activity by Monoclonal Antibodies to CCR5. *J. Virol.* **1999**, *73* (5), 4145–4155. https://doi.org/10.1128/JVI.73.5.4145-4155.1999.
(26) Biswas, P.; Tambussi, G.; Lazzarin, A. Access Denied? The Status of Co-Receptor Inhibition to Counter HIV Entry. *Expert Opin. Pharmacother.* **2007**, *8* (7), 923–933. https://doi.org/10.1517/14656566.8.7.923.
(27) Cooper, G. M. *The Cell: A Molecular Approach*, 2. ed.; ASM Press [u.a.]: Washington, DC, 2000.
(28) Yeung, Darien. Automation of Reaction Monitoring, University of Victoria. http://hdl.handle.net/1828/10667.
(29) Mulvey, C. M.; Breckels, L. M.; Crook, O. M.; Sanders, D. J.; Ribeiro, A. L. R.; Geladaki, A.; Christoforou, A.; Britovšek, N. K.; Hurrell, T.; Deery, M. J.; Gatto, L.; Smith, A. M.; Lilley, K. S. Spatiotemporal Proteomic Profiling of the Pro-Inflammatory Response to Lipopolysaccharide in the THP-1 Human Leukaemia Cell Line. *Nat. Commun.* **2021**, *12* (1), 5773. https://doi.org/10.1038/s41467-021-26000-9.
(30) Ordureau, A.; Kraus, F.; Zhang, J.; An, H.; Park, S.; Ahfeldt, T.; Paulo, J. A.; Harper, J. W. Temporal Proteomics during Neurogenesis Reveals Large-Scale Proteome and Organelle Remodeling via Selective Autophagy. *Mol. Cell* **2021**, *81* (24), 5082-5098.e11. https://doi.org/10.1016/j.molcel.2021.10.001.

(31) Khan, M. M.; Chattagul, S.; Tran, B. Q.; Freiberg, J. A.; Nita-Lazar, A.; Shirtliff, M. E.; Sermswan, R. W.; Ernst, R. K.; Goodlett, D. R. Temporal Proteomic Profiling Reveals Changes That Support Burkholderia Biofilms. *Pathog. Dis.* **2019**, *77* (2). https://doi.org/10.1093/femspd/ftz005.
(32) Novo, P.; Dell'Aica, M.; Janasek, D.; Zahedi, R. P. High Spatial and Temporal Resolution Cell Manipulation Techniques in Microchannels. *The Analyst* **2016**, *141* (6), 1888–1905. https://doi.org/10.1039/C6AN00027D.
(33) Zhu, Y.; Orre, L. M.; Zhou Tran, Y.; Mermelekas, G.; Johansson, H. J.; Malyutina, A.; Anders, S.; Lehtiö, J. DEqMS: A Method for Accurate Variance Estimation in Differential Protein Expression Analysis. *Mol. Cell. Proteomics* **2020**, *19* (6), 1047–1057. https://doi.org/10.1074/mcp.TIR119.001646.
(34) Haynes, W. A.; Higdon, R.; Stanberry, L.; Collins, D.; Kolker, E. Differential Expression Analysis for Pathways. *PLoS Comput. Biol.* **2013**, *9* (3), e1002967. https://doi.org/10.1371/journal.pcbi.1002967.
(35) Gisby, J. S.; Buang, N. B.; Papadaki, A.; Clarke, C. L.; Malik, T. H.; Medjeral-Thomas, N.; Pinheiro, D.; Mortimer, P. M.; Lewis, S.; Sandhu, E.; McAdoo, S. P.; Prendecki, M. F.; Willicombe, M.; Pickering, M. C.; Botto, M.; Thomas, D. C.; Peters, J. E. Multi-Omics Identify Falling LRRC15 as a COVID-19 Severity Marker and Persistent pro-Thrombotic Signals in Convalescence. *Nat. Commun.* **2022**, *13* (1), 7775. https://doi.org/10.1038/s41467-022-35454-4.
(36) Babačić, H.; Christ, W.; Araújo, J. E.; Mermelekas, G.; Sharma, N.; Tynell, J.; García, M.; Varnaite, R.; Asgeirsson, H.; Glans, H.; Lehtiö, J.; Gredmark-Russ, S.; Klingström, J.; Pernemalm, M. Comprehensive Proteomics and Meta-Analysis of COVID-19 Host Response. *Nat. Commun.* **2023**, *14* (1), 5921. https://doi.org/10.1038/s41467-023-41159-z.
(37) Shuken, S. R. An Introduction to Mass Spectrometry-Based Proteomics. *J. Proteome Res.* **2023**, *22* (7), 2151–2171. https://doi.org/10.1021/acs.jproteome.2c00838.
(38) Spicer, V.; Ezzati, P.; Neustaeter, H.; Beavis, R. C.; Wilkins, J. A.; Krokhin, O. V. 3D HPLC-MS with Reversed-Phase Separation Functionality in All Three Dimensions for Large-Scale Bottom-Up Proteomics and Peptide Retention Data Collection. *Anal. Chem.* **2016**, *88* (5), 2847–2855. https://doi.org/10.1021/acs.analchem.5b04567.
(39) Kawashima, Y.; Nagai, H.; Konno, R.; Ishikawa, M.; Nakajima, D.; Sato, H.; Nakamura, R.; Furuyashiki, T.; Ohara, O. Single-Shot 10K Proteome Approach: Over 10,000 Protein Identifications by Data-Independent Acquisition-Based Single-Shot Proteomics with Ion Mobility Spectrometry. *J. Proteome Res.* **2022**, *21* (6), 1418–1427. https://doi.org/10.1021/acs.jproteome.2c00023.
(40) Guzman, U. H.; Martinez-Val, A.; Ye, Z.; Damoc, E.; Arrey, T. N.; Pashkova, A.; Renuse, S.; Denisov, E.; Petzoldt, J.; Peterson, A. C.; Harking, F.; Østergaard, O.; Rydbirk, R.; Aznar, S.; Stewart, H.; Xuan, Y.; Hermanson, D.; Horning, S.; Hock, C.; Makarov, A.; Zabrouskov, V.; Olsen, J. V. Ultra-Fast Label-Free Quantification and Comprehensive Proteome Coverage with Narrow-Window Data-Independent Acquisition. *Nat. Biotechnol.* **2024**. https://doi.org/10.1038/s41587-023-02099-7.
(41) Martinez-Val, A.; Guzmán, U. H.; Olsen, J. V. Obtaining Complete Human Proteomes. *Annu. Rev. Genomics Hum. Genet.* **2022**, *23* (1), 99–121. https://doi.org/10.1146/annurev-genom-112921-024948.

(42) Dupree, E. J.; Jayathirtha, M.; Yorkey, H.; Mihasan, M.; Petre, B. A.; Darie, C. C. A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field. *Proteomes* **2020**, *8* (3), 14. https://doi.org/10.3390/proteomes8030014.
(43) Papayannopoulos, I. A. The Interpretation of Collision-induced Dissociation Tandem Mass Spectra of Peptides. *Mass Spectrom. Rev.* **1995**, *14* (1), 49–73. https://doi.org/10.1002/mas.1280140104.
(44) Beavis, R. C.; Bolbach, G.; Ens, W.; Main, D. E.; Schueler, B.; Standing, K. G. Automated Dry Fraction Collection for Microbore High-Performance Liquid Chromatography—Mass Spectrometry. *J. Chromatogr. A* **1986**, *359*, 489–497. https://doi.org/10.1016/0021-9673(86)80101-7.
(45) Schwanhäusser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M. Global Quantification of Mammalian Gene Expression Control. *Nature* **2011**, *473* (7347), 337–342. https://doi.org/10.1038/nature10098.
(46) Zubarev, R. A. The Challenge of the Proteome Dynamic Range and Its Implications for In-Depth Proteomics. *PROTEOMICS* **2013**, *13* (5), 723–726. https://doi.org/10.1002/pmic.201200451.
(47) Hortin, G. L.; Sviridov, D. The Dynamic Range Problem in the Analysis of the Plasma Proteome. *J. Proteomics* **2010**, *73* (3), 629–636. https://doi.org/10.1016/j.jprot.2009.07.001.
(48) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. The One Hour Yeast Proteome. *Mol. Cell. Proteomics* **2014**, *13* (1), 339–347. https://doi.org/10.1074/mcp.M113.034769.
(49) Trujillo, E. A.; Hebert, A. S.; Brademan, D. R.; Coon, J. J. Maximizing Tandem Mass Spectrometry Acquisition Rates for Shotgun Proteomics. *Anal. Chem.* **2019**, *91* (20), 12625–12629. https://doi.org/10.1021/acs.analchem.9b02979.
(50) Meier, F.; Brunner, A.-D.; Koch, S.; Koch, H.; Lubeck, M.; Krause, M.; Goedecke, N.; Decker, J.; Kosinski, T.; Park, M. A.; Bache, N.; Hoerning, O.; Cox, J.; Räther, O.; Mann, M. Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Mol. Cell. Proteomics* **2018**, *17* (12), 2534–2545. https://doi.org/10.1074/mcp.TIR118.000900.
(51) Meier, F.; Beck, S.; Grassl, N.; Lubeck, M.; Park, M. A.; Raether, O.; Mann, M. Parallel Accumulation–Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device. *J. Proteome Res.* **2015**, *14* (12), 5378–5387. https://doi.org/10.1021/acs.jproteome.5b00932.
(52) Messner, C. B.; Demichev, V.; Bloomfield, N.; Yu, J. S. L.; White, M.; Kreidl, M.; Egger, A.-S.; Freiwald, A.; Ivosev, G.; Wasim, F.; Zelezniak, A.; Jürgens, L.; Suttorp, N.; Sander, L. E.; Kurth, F.; Lilley, K. S.; Mülleder, M.; Tate, S.; Ralser, M. Ultra-Fast Proteomics with Scanning SWATH. *Nat. Biotechnol.* **2021**, *39* (7), 846–854. https://doi.org/10.1038/s41587-021-00860-4.
(53) Bekker-Jensen, D. B.; Martínez-Val, A.; Steigerwald, S.; Rüther, P.; Fort, K. L.; Arrey, T. N.; Harder, A.; Makarov, A.; Olsen, J. V. A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients. *Mol. Cell. Proteomics* **2020**, *19* (4), 716–729. https://doi.org/10.1074/mcp.TIR119.001906.
(54) Kelstrup, C. D.; Bekker-Jensen, D. B.; Arrey, T. N.; Hogrebe, A.; Harder, A.; Olsen, J. V. Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *J. Proteome Res.* **2018**, *17* (1), 727–738. https://doi.org/10.1021/acs.jproteome.7b00602.

(55) Scheltema, R. A.; Hauschild, J.-P.; Lange, O.; Hornburg, D.; Denisov, E.; Damoc, E.; Kuehn, A.; Makarov, A.; Mann, M. The Q Exactive HF, a Benchtop Mass Spectrometer with a Pre-Filter, High-Performance Quadrupole and an Ultra-High-Field Orbitrap Analyzer. *Mol. Cell. Proteomics* **2014**, *13* (12), 3698–3708. https://doi.org/10.1074/mcp.M114.043489.
(56) Pirmoradian, M.; Budamgunta, H.; Chingin, K.; Zhang, B.; Astorga-Wells, J.; Zubarev, R. A. Rapid and Deep Human Proteome Analysis by Single-Dimension Shotgun Proteomics. *Mol. Cell. Proteomics* **2013**, *12* (11), 3330–3338. https://doi.org/10.1074/mcp.O113.028787.
(57) Shishkova, E.; Hebert, A. S.; Coon, J. J. Now, More Than Ever, Proteomics Needs Better Chromatography. *Cell Syst.* **2016**, *3* (4), 321–324. https://doi.org/10.1016/j.cels.2016.10.007.
(58) Mokrushin, S. G. Definition of Chromatography. *Nature* **1956**, *178* (4544), 1244–1245. https://doi.org/10.1038/1781244c0.
(59) Abdu Hussen, A. High-Performance Liquid Chromatography (HPLC): A Review. *Ann. Adv. Chem.* **2022**, *6* (1), 010–020. https://doi.org/10.29328/journal.aac.1001026.
(60) Mant, C. T.; Chen, Y.; Yan, Z.; Popa, T. V.; Kovacs, J. M.; Mills, J. B.; Tripet, B. P.; Hodges, R. S. HPLC Analysis and Purification of Peptides. In *Peptide Characterization and Application Protocols*; Fields, G. B., Ed.; Walker, J. M., Series Ed.; Methods in Molecular Biology™; Humana Press: Totowa, NJ, 2007; Vol. 386, pp 3–55. https://doi.org/10.1007/978-1-59745-430-8_1.
(61) Langmuir, I. THE ADSORPTION OF GASES ON PLANE SURFACES OF GLASS, MICA AND PLATINUM. *J. Am. Chem. Soc.* **1918**, *40* (9), 1361–1403. https://doi.org/10.1021/ja02242a004.
(62) Robards, K.; Haddad, P. R.; Jackson, P. E. High-Performance Liquid Chromatography—Separations. In *Principles and Practice of Modern Chromatographic Methods*; Elsevier, 2004; pp 305–380. https://doi.org/10.1016/B978-0-08-057178-2.50009-1.
(63) Molnár, I.; Horváth, C. Reverse-Phase Chromatography of Polar Biological Substances: Separation of Catechol Compounds by High-Performance Liquid Chromatography. *Clin. Chem.* **1976**, *22* (9), 1497–1502. https://doi.org/10.1093/clinchem/22.9.1497.
(64) Fritz, J. S. Early Milestones in the Development of Ion-Exchange Chromatography: A Personal Account. *J. Chromatogr. A* **2004**, *1039* (1–2), 3–12. https://doi.org/10.1016/j.chroma.2003.12.068.
(65) Zhang, X.; Fang, A.; Riley, C. P.; Wang, M.; Regnier, F. E.; Buck, C. Multi-Dimensional Liquid Chromatography in Proteomics—A Review. *Anal. Chim. Acta* **2010**, *664* (2), 101–113. https://doi.org/10.1016/j.aca.2010.02.001.
(66) Shi, Y.; Xiang, R.; Horvath, C.; Wilkins, J. The Role of Liquid Chromatography in Proteomics. *J. Chromatogr. A* **2004**, *1053* (1–2), 27–36. https://doi.org/10.1016/S0021-9673(04)01204-X.
(67) Jiang, P.; Lucy, C. A. Coupling Normal Phase Liquid Chromatography with Electrospray Ionization Mass Spectrometry: Strategies and Applications. *Anal. Methods* **2016**, *8* (35), 6478–6488. https://doi.org/10.1039/C6AY01419D.
(68) Lenčo, J.; Jadeja, S.; Naplekov, D. K.; Krokhin, O. V.; Khalikova, M. A.; Chocholouš, P.; Urban, J.; Broeckhoven, K.; Nováková, L.; Švec, F. Reversed-Phase Liquid Chromatography of Peptides for Bottom-Up Proteomics: A Tutorial. *J. Proteome Res.* **2022**, *21* (12), 2846–2892. https://doi.org/10.1021/acs.jproteome.2c00407.
(69) Bensaddek, D.; Nicolas, A.; Lamond, A. I. Evaluating the Use of HILIC in Large-Scale, Multi Dimensional Proteomics: Horses for Courses? *Int. J. Mass Spectrom.* **2015**, *391*, 105–114. https://doi.org/10.1016/j.ijms.2015.07.029.

(70) Boersema, P. J.; Mohammed, S.; Heck, A. J. R. Hydrophilic Interaction Liquid Chromatography (HILIC) in Proteomics. *Anal. Bioanal. Chem.* **2008**, *391* (1), 151–159. https://doi.org/10.1007/s00216-008-1865-7.
(71) Alpert, A. J. Hydrophilic-Interaction Chromatography for the Separation of Peptides, Nucleic Acids and Other Polar Compounds. *J. Chromatogr. A* **1990**, *499*, 177–196. https://doi.org/10.1016/S0021-9673(00)96972-3.
(72) Dinh, N. P.; Jonsson, T.; Irgum, K. Probing the Interaction Mode in Hydrophilic Interaction Chromatography. *J. Chromatogr. A* **2011**, *1218* (35), 5880–5891. https://doi.org/10.1016/j.chroma.2011.06.037.
(73) Hemström, P.; Irgum, K. Hydrophilic Interaction Chromatography. *J. Sep. Sci.* **2006**, *29* (12), 1784–1821. https://doi.org/10.1002/jssc.200600199.
(74) Venne, A. S.; Solari, F. A.; Faden, F.; Paretti, T.; Dissmeyer, N.; Zahedi, R. P. An Improved Workflow for Quantitative N-terminal Charge-based Fractional Diagonal Chromatography (ChaFRADIC) to Study Proteolytic Events in *Arabidopsis Thaliana*. *PROTEOMICS* **2015**, *15* (14), 2458–2469. https://doi.org/10.1002/pmic.201500014.
(75) Gussakovsky, D.; Neustaeter, H.; Spicer, V.; Krokhin, O. V. Sequence-Specific Model for Peptide Retention Time Prediction in Strong Cation Exchange Chromatography. *Anal. Chem.* **2017**, *89* (21), 11795–11802. https://doi.org/10.1021/acs.analchem.7b03436.
(76) Zhang, T.; Liu, X.; Rossio, V.; Dawson, S. L.; Gygi, S. P.; Paulo, J. A. Enhancing Proteome Coverage by Using Strong Anion-Exchange in Tandem with Basic-pH Reversed-Phase Chromatography for Sample Multiplexing-Based Proteomics. *J. Proteome Res.* **2023**, acs.jproteome.3c00492. https://doi.org/10.1021/acs.jproteome.3c00492.
(77) Ritorto, M. S.; Cook, K.; Tyagi, K.; Pedrioli, P. G. A.; Trost, M. Hydrophilic Strong Anion Exchange (hSAX) Chromatography for Highly Orthogonal Peptide Separation of Complex Proteomes. *J. Proteome Res.* **2013**, *12* (6), 2449–2457. https://doi.org/10.1021/pr301011r.
(78) Betancourt, L. H.; De Bock, P.-J.; Staes, A.; Timmerman, E.; Perez-Riverol, Y.; Sanchez, A.; Besada, V.; Gonzalez, L. J.; Vandekerckhove, J.; Gevaert, K. SCX Charge State Selective Separation of Tryptic Peptides Combined with 2D-RP-HPLC Allows for Detailed Proteome Mapping. *J. Proteomics* **2013**, *91*, 164–171. https://doi.org/10.1016/j.jprot.2013.06.033.
(79) Chan, K. C.; Issaq, H. J. Fractionation of Peptides by Strong Cation-Exchange Liquid Chromatography. In *Proteomics for Biomarker Discovery*; Zhou, M., Veenstra, T., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2013; Vol. 1002, pp 311–315. https://doi.org/10.1007/978-1-62703-360-2_23.
(80) Gritti, F.; Guiochon, G. The Quantitative Impact of the Mesopore Size on the Mass Transfer Mechanism of the New 1.9 Mm Fully Porous Titan-C18 Particles. I: Analysis of Small Molecules. *J. Chromatogr. A* **2015**, *1384*, 76–87. https://doi.org/10.1016/j.chroma.2015.01.047.
(81) Bagge, J.; Enmark, M.; Leśko, M.; Limé, F.; Fornstedt, T.; Samuelsson, J. Impact of Stationary-Phase Pore Size on Chromatographic Performance Using Oligonucleotide Separation as a Model. *J. Chromatogr. A* **2020**, *1634*, 461653. https://doi.org/10.1016/j.chroma.2020.461653.
(82) Godinho, J. M.; Naese, J. A.; Toler, A. E.; Boyes, B. E.; Henry, R. A.; DeStefano, J. J.; Grinias, J. P. Importance of Particle Pore Size in Determining Retention and Selectivity in Reversed Phase Liquid Chromatography. *J. Chromatogr. A* **2020**, *1634*, 461678. https://doi.org/10.1016/j.chroma.2020.461678.
(83) Yeung, D.; Spicer, V.; Zahedi, R. P.; Krokhin, O. V. Paradigm Shift: Major Role of Ion-Pairing-Dependent Size Exclusion Effects in Bottom-Up Proteomics Reversed-Phase Peptide

Separations. *Anal. Chem.* **2024**, *96* (23), 9721–9728. https://doi.org/10.1021/acs.analchem.4c02035.
(84) Nagae, N; Yamamoto, K; Kadota, C. Study of State of Alkyl Chains for Reversed-Phase: Collapse or Brush Up, 2010. http://chromanik.co.jp/technical/pdf/Pittcon10_2.pdf.
(85) Nagae, N; Tsukamoto T. The Retention Behaviour of Reversed-Phase HPLC Columns with 100% Aqueous Mobile Phase. March 2014. https://www.obrnutafaza.hr/pdf/chromanik_korisno/Retention-Behavior-RP-HPLC-Columns-Aqueous.pdf.
(86) Gritti, F.; Walter, T. H. Retention Loss of Reversed-Phase Columns Using Highly Aqueous Mobile Phases: Fundamentals, Mechanism, and Practical Solutions. *LCGC N. Am.* **2021**, 33–40. https://doi.org/10.56530/lcgc.na.zm6986c6.
(87) Chongying, L. Silanol Group Effect in C18 Bonded Phases in HPLC. *Chromatographia* **1992**, *34* (3–4), 182–184. https://doi.org/10.1007/BF02276135.
(88) Sorgen, C. M.; Burge, L. J. Evaluation of Polar-Embedded and Polar-Endcapped C12-C18 Stationary Phase Vulnerabilities Toward Highly Nucleophilic Compounds: A Case Study. *LCGC N. Am.* **2005**, *23* (8), 762–765.
(89) Layne, J. Characterization and Comparison of the Chromatographic Performance of Conventional, Polar-Embedded, and Polar-Endcapped Reversed-Phase Liquid Chromatography Stationary Phases. *J. Chromatogr. A* **2002**, *957* (2), 149–164. https://doi.org/10.1016/S0021-9673(02)00193-0.
(90) Snyder, L. R.; Kirkland, J. J.; Dolan, J. W. *Introduction to Modern Liquid Chromatography*, 1st ed.; Wiley, 2009. https://doi.org/10.1002/9780470508183.
(91) Schellinger, A. P.; Carr, P. W. Isocratic and Gradient Elution Chromatography: A Comparison in Terms of Speed, Retention Reproducibility and Quantitation. *J. Chromatogr. A* **2006**, *1109* (2), 253–266. https://doi.org/10.1016/j.chroma.2006.01.047.
(92) McCalley, D. V. Effect of Mobile Phase Additives on Solute Retention at Low Aqueous pH in Hydrophilic Interaction Liquid Chromatography. *J. Chromatogr. A* **2017**, *1483*, 71–79. https://doi.org/10.1016/j.chroma.2016.12.035.
(93) Pearson, J. D.; McCroskey, M. C. Perfluorinated Acid Alternatives to Trifluoroacetic Acid for Reversed-Phase High-Performance Liquid Chromatography. *J. Chromatogr. A* **1996**, *746* (2), 277–281. https://doi.org/10.1016/0021-9673(96)00377-9.
(94) Shibue, M.; Mant, C. T.; Hodges, R. S. Effect of Anionic Ion-Pairing Reagent Hydrophobicity on Selectivity of Peptide Separations by Reversed-Phase Liquid Chromatography. *J. Chromatogr. A* **2005**, *1080* (1), 68–75. https://doi.org/10.1016/j.chroma.2005.03.035.
(95) Boyes, B. E.; Dong, M. W. Modern Trends and Best Practices in Mobile-Phase Selection in Reversed-Phase Chromatography. *LCGC N. Am.* **2018**, *36* (10), 752–768.
(96) Kobayashi, H.; Sukegawa, M.; Fujimura, K.; Kubo, T.; Otsuka, K. Effect of Acidic Additives on Peak Capacity and Detectivity in Peptide Analysis Using Nano-Flow LC/MS with Low-Density ODS Modified Monolithic Silica Capillary Columns. *CHROMATOGRAPHY* **2016**, *37* (3), 133–139. https://doi.org/10.15583/jpchrom.2016.012.
(97) Krokhin, O. V.; Craig, R.; Spicer, V.; Ens, W.; Standing, K. G.; Beavis, R. C.; Wilkins, J. A. An Improved Model for Prediction of Retention Times of Tryptic Peptides in Ion Pair Reversed-Phase HPLC. *Mol. Cell. Proteomics* **2004**, *3* (9), 908–919. https://doi.org/10.1074/mcp.M400031-MCP200.

(98) Shou, W.; Naidong, W. Simple Means to Alleviate Sensitivity Loss by Trifluoroacetic Acid (TFA) Mobile Phases in the Hydrophilic Interaction Chromatography–Electrospray Tandem Mass Spectrometric (HILIC–ESI/MS/MS) Bioanalysis of Basic Compounds. *J. Chromatogr. B* **2005**, *825* (2), 186–192. https://doi.org/10.1016/j.jchromb.2005.01.011.
(99) Paull, B.; Nesterenko, P. N. Ion Chromatography. In *Liquid Chromatography*; Elsevier, 2013; pp 157–191. https://doi.org/10.1016/B978-0-12-415807-8.00008-0.
(100) Ngere, J. B.; Ebrahimi, K. H.; Williams, R.; Pires, E.; Walsby-Tickle, J.; McCullagh, J. S. O. Ion-Exchange Chromatography Coupled to Mass Spectrometry in Life Science, Environmental, and Medical Research. *Anal. Chem.* **2023**, *95* (1), 152–166. https://doi.org/10.1021/acs.analchem.2c04298.
(101) Alves, G.; Ogurtsov, A. Y.; Kwok, S.; Wu, W. W.; Wang, G.; Shen, R.-F.; Yu, Y.-K. Detection of Co-Eluted Peptides Using Database Search Methods. *Biol. Direct* **2008**, *3* (1), 27. https://doi.org/10.1186/1745-6150-3-27.
(102) Abdul-Khalek, N.; Wimmer, R.; Overgaard, M. T.; Gregersen Echers, S. Insight on Physicochemical Properties Governing Peptide MS1 Response in HPLC-ESI-MS/MS: A Deep Learning Approach. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 3715–3727. https://doi.org/10.1016/j.csbj.2023.07.027.
(103) Schliekelman, P.; Liu, S. Quantifying the Effect of Competition for Detection between Coeluting Peptides on Detection Probabilities in Mass-Spectrometry-Based Proteomics. *J. Proteome Res.* **2014**, *13* (2), 348–361. https://doi.org/10.1021/pr400034z.
(104) Yeung, D.; Mizero, B.; Gussakovsky, D.; Klaassen, N.; Lao, Y.; Spicer, V.; Krokhin, O. V. Separation Orthogonality in Liquid Chromatography–Mass Spectrometry for Proteomic Applications: Comparison of 16 Different Two-Dimensional Combinations. *Anal. Chem.* **2020**, *92* (5), 3904–3912. https://doi.org/10.1021/acs.analchem.9b05407.
(105) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC−MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome. *J. Proteome Res.* **2003**, *2* (1), 43–50. https://doi.org/10.1021/pr025556v.
(106) Tang, J.; Gao, M.; Deng, C.; Zhang, X. Recent Development of Multi-Dimensional Chromatography Strategies in Proteome Research. *J. Chromatogr. B* **2008**, *866* (1–2), 123–132. https://doi.org/10.1016/j.jchromb.2008.01.029.
(107) Zhang, X.; Fang, A.; Riley, C. P.; Wang, M.; Regnier, F. E.; Buck, C. Multi-Dimensional Liquid Chromatography in Proteomics—A Review. *Anal. Chim. Acta* **2010**, *664* (2), 101–113. https://doi.org/10.1016/j.aca.2010.02.001.
(108) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.-C.; Yates, J. R. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chem. Rev.* **2013**, *113* (4), 2343–2394. https://doi.org/10.1021/cr3003533.
(109) Wolters, D. A.; Washburn, M. P.; Yates, J. R. An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics. *Anal. Chem.* **2001**, *73* (23), 5683–5690. https://doi.org/10.1021/ac010617e.
(110) Roca, L. S.; Gargano, A. F. G.; Schoenmakers, P. J. Development of Comprehensive Two-Dimensional Low-Flow Liquid-Chromatography Setup Coupled to High-Resolution Mass Spectrometry for Shotgun Proteomics. *Anal. Chim. Acta* **2021**, *1156*, 338349. https://doi.org/10.1016/j.aca.2021.338349.

(111) Hinzke, T.; Kouris, A.; Hughes, R.-A.; Strous, M.; Kleiner, M. More Is Not Always Better: Evaluation of 1D and 2D-LC-MS/MS Methods for Metaproteomics. *Front. Microbiol.* **2019**, *10*, 238. https://doi.org/10.3389/fmicb.2019.00238.
(112) Gaspari, M.; Cuda, G. Nano LC–MS/MS: A Robust Setup for Proteomic Analysis. In *Nanoproteomics*; Toms, S. A., Weil, R. J., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2011; Vol. 790, pp 115–126. https://doi.org/10.1007/978-1-61779-319-6_9.
(113) Wilson, S. R.; Vehus, T.; Berg, H. S.; Lundanes, E. Nano-LC in Proteomics: Recent Advances and Approaches. *Bioanalysis* **2015**, *7* (14), 1799–1815. https://doi.org/10.4155/bio.15.92.
(114) Cutillas, P. Principles of Nanoflow Liquid Chromatography and Applications to Proteomics. *Curr. Nanosci.* **2005**, *1* (1), 65–71. https://doi.org/10.2174/1573413052953093.
(115) Yi, L.; Piehowski, P. D.; Shi, T.; Smith, R. D.; Qian, W.-J. Advances in Microscale Separations towards Nanoproteomics Applications. *J. Chromatogr. A* **2017**, *1523*, 40–48. https://doi.org/10.1016/j.chroma.2017.07.055.
(116) Wang, H.; Qian, W.-J.; Mottaz, H. M.; Clauss, T. R. W.; Anderson, D. J.; Moore, R. J.; Camp, D. G.; Khan, A. H.; Sforza, D. M.; Pallavicini, M.; Smith, D. J.; Smith, R. D. Development and Evaluation of a Micro- and Nanoscale Proteomic Sample Preparation Method. *J. Proteome Res.* **2005**, *4* (6), 2397–2403. https://doi.org/10.1021/pr050160f.
(117) Sui, X.; Wu, Q.; Cui, X.; Wang, X.; Zhang, L.; Deng, N.; Bian, Y.; Xu, R.; Tian, R. Robust Capillary- and Micro-Flow Liquid Chromatography–Tandem Mass Spectrometry Methods for High-Throughput Proteome Profiling. *J. Proteome Res.* **2022**, *21* (10), 2472–2480. https://doi.org/10.1021/acs.jproteome.2c00405.
(118) Sorensen, M. J.; Anderson, B. G.; Kennedy, R. T. Liquid Chromatography above 20,000 PSI. *TrAC Trends Anal. Chem.* **2020**, *124*, 115810. https://doi.org/10.1016/j.trac.2020.115810.
(119) Mant, C. T.; Hodges, R. S. Separation of Peptides on HALO 2-Micron Particles. *Curr. Protoc. Protein Sci.* **2016**, *85* (1). https://doi.org/10.1002/cpps.12.
(120) Stejskal, K.; Jeff, O. D. B.; Matzinger, M.; Dürnberger, G.; Boychenko, A.; Jacobs, P.; Mechtler, K. Deep Proteome Profiling with Reduced Carryover Using Superficially Porous Microfabricated nanoLC Columns. *Anal. Chem.* **2022**, *94* (46), 15930–15938. https://doi.org/10.1021/acs.analchem.2c01196.
(121) Johnson, D.; Boyes, B.; Fields, T.; Kopkin, R.; Orlando, R. Optimization of Data-Dependent Acquisition Parameters for Coupling High-Speed Separations with LC-MS/MS for Protein Identifications. *J. Biomol. Tech. JBT* **2013**, jbt.13-2402-003. https://doi.org/10.7171/jbt.13-2402-003.
(122) Staes, A.; Boucher, K.; Dufour, S.; Maia, T. M.; Timmerman, E.; Haver, D. V.; Pauwels, J.; Demol, H.; Vandenbussche, J.; Gevaert, K.; Impens, F.; Devos, S. High-Throughput Nanoflow Proteomics Using a Dual-Column Electrospray Source. *Anal. Chem.* **2024**, *96* (17), 6534–6539. https://doi.org/10.1021/acs.analchem.4c00845.
(123) Blackler, A. R.; Speers, A. E.; Wu, C. C. Chromatographic Benefits of Elevated Temperature for the Proteomic Analysis of Membrane Proteins. *PROTEOMICS* **2008**, *8* (19), 3956–3964. https://doi.org/10.1002/pmic.200800210.
(124) Li, B.; Guo, F.; Hu, H.; Liu, P.; Tan, M.; Pan, J.; Zhai, L. The Characterization of Column Heating Effect in Nanoflow Liquid Chromatography Mass Spectrometry (nanoLC-MS)–Based Proteomics. *J. Mass Spectrom.* **2020**, *55* (1), e4441. https://doi.org/10.1002/jms.4441.

(125) Villacrés, C.; Mizero, B.; Spicer, V.; Viner, R.; Saba, J.; Patel, B.; Snovida, S.; Jensen, P.; Huhmer, A.; Krokhin, O. V. Toward an Ultimate Solution for Peptide Retention Time Prediction: The Effect of Column Temperature on Separation Selectivity. *J. Proteome Res.* **2024**, *23* (4), 1488–1494. https://doi.org/10.1021/acs.jproteome.4c00018.
(126) Moruz, L.; Pichler, P.; Stranzl, T.; Mechtler, K.; Käll, L. Optimized Nonlinear Gradients for Reversed-Phase Liquid Chromatography in Shotgun Proteomics. *Anal. Chem.* **2013**, *85* (16), 7777–7785. https://doi.org/10.1021/ac401145q.
(127) Trudgian, D. C.; Fischer, R.; Guo, X.; Kessler, B. M.; Mirzaei, H. GOAT – A Simple LC - MS / MS Gradient Optimization Tool. *PROTEOMICS* **2014**, *14* (12), 1467–1471. https://doi.org/10.1002/pmic.201300524.
(128) Tyteca, E.; De Vos, J.; Vankova, N.; Cesla, P.; Desmet, G.; Eeltink, S. Applicability of Linear and Nonlinear Retention-time Models for Reversed-phase Liquid Chromatography Separations of Small Molecules, Peptides, and Intact Proteins. *J. Sep. Sci.* **2016**, *39* (7), 1249–1257. https://doi.org/10.1002/jssc.201501395.
(129) Spicer, V.; Ezzati, P.; Neustaeter, H.; Beavis, R. C.; Wilkins, J. A.; Krokhin, O. V. 3D HPLC-MS with Reversed-Phase Separation Functionality in All Three Dimensions for Large-Scale Bottom-Up Proteomics and Peptide Retention Data Collection. *Anal. Chem.* **2016**, *88* (5), 2847–2855. https://doi.org/10.1021/acs.analchem.5b04567.
(130) Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C. Two-dimensional Separation of Peptides Using RP-RP-HPLC System with Different pH in First and Second Separation Dimensions. *J. Sep. Sci.* **2005**, *28* (14), 1694–1703. https://doi.org/10.1002/jssc.200500116.
(131) Kirkland, J. J.; Henderson, J. W.; DeStefano, J. J.; Van Straten, M. A.; Claessens, H. A. Stability of Silica-Based, Endcapped Columns with pH 7 and 11 Mobile Phases for Reversed-Phase High-Performance Liquid Chromatography. *J. Chromatogr. A* **1997**, *762* (1–2), 97–112. https://doi.org/10.1016/S0021-9673(96)00945-4.
(132) Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C. Orthogonality of Separation in Two-Dimensional Liquid Chromatography. *Anal. Chem.* **2005**, *77* (19), 6426–6434. https://doi.org/10.1021/ac050923i.
(133) Lee, D. P. Reversed-Phase HPLC from pH 1 to 13. *J. Chromatogr. Sci.* **1982**, *20* (5), 203–208. https://doi.org/10.1093/chromsci/20.5.203.
(134) Schoenmakers, P. J.; Van Molle, S.; Hayes, C. M. G.; Uunk, L. G. M. Effects of pH in Reversed-Phase Liquid Chromatography. *Anal. Chim. Acta* **1991**, *250*, 1–19. https://doi.org/10.1016/0003-2670(91)85058-Z.
(135) Raida, M.; Schulz-Knappe, P.; Heine, G.; Forssmann, W.-G. Liquid Chromatography and Electrospray Mass Spectrometric Mapping of Peptides from Human Plasma Filtrate. *J. Am. Soc. Mass Spectrom.* **1999**, *10* (1), 45–54. https://doi.org/10.1016/S1044-0305(98)00117-2.
(136) Washburn, M. P.; Yates, J. R. New Methods of Proteome Analysis: Multidimensional Chromatography and Mass Spectrometry. *Trends Biotechnol.* **2000**, *18*, 27–30. https://doi.org/10.1016/S0167-7799(00)00005-6.
(137) Zhu, M.-Z.; Li, N.; Wang, Y.-T.; Liu, N.; Guo, M.-Q.; Sun, B.; Zhou, H.; Liu, L.; Wu, J.-L. Acid/Salt/pH Gradient Improved Resolution and Sensitivity in Proteomics Study Using 2D SCX-RP LC–MS. *J. Proteome Res.* **2017**, *16* (9), 3470–3475. https://doi.org/10.1021/acs.jproteome.7b00443.
(138) Adachi, J.; Hashiguchi, K.; Nagano, M.; Sato, M.; Sato, A.; Fukamizu, K.; Ishihama, Y.; Tomonaga, T. Improved Proteome and Phosphoproteome Analysis on a Cation Exchanger by a

Combined Acid and Salt Gradient. *Anal. Chem.* **2016**, *88* (16), 7899–7903. https://doi.org/10.1021/acs.analchem.6b01232.
(139) Rose, G. D.; Roy, S. Hydrophobic Basis of Packing in Globular Proteins. *Proc. Natl. Acad. Sci.* **1980**, *77* (8), 4643–4647. https://doi.org/10.1073/pnas.77.8.4643.
(140) Rose, G. D. Prediction of Chain Turns in Globular Proteins on a Hydrophobic Basis. *Nature* **1978**, *272* (5654), 586–590. https://doi.org/10.1038/272586a0.
(141) Rose, G. D.; Geselowitz, A. R.; Lesser, G. J.; Lee, R. H.; Zehfus, M. H. Hydrophobicity of Amino Acid Residues in Globular Proteins. *Science* **1985**, *229* (4716), 834–838. https://doi.org/10.1126/science.4023714.
(142) Kyte, J.; Doolittle, R. F. A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.* **1982**, *157* (1), 105–132. https://doi.org/10.1016/0022-2836(82)90515-0.
(143) Molnár, I.; Horváth, C. Separation of Amino Acids and Peptides on Non-Polar Stationary Phases by High-Performance Liquid Chromatography. *J. Chromatogr. A* **1977**, *142*, 623–640. https://doi.org/10.1016/S0021-9673(01)92073-4.
(144) Meek, J. L. Prediction of Peptide Retention Times in High-Pressure Liquid Chromatography on the Basis of Amino Acid Composition. *Proc. Natl. Acad. Sci. U. S. A.* **1980**, *77* (3), 1632–1636. https://doi.org/10.1073/pnas.77.3.1632.
(145) Guo, D.; Mant, C. T.; Hodges, R. S. Effects of Ion-Pairing Reagents on the Prediction of Peptide Retention in Reversed-Phase High-Resolution Liquid Chromatography. *J. Chromatogr. A* **1987**, *386*, 205–222. https://doi.org/10.1016/S0021-9673(01)94598-4.
(146) Guo, D.; Mant, C. T.; Taneja, A. K.; Parker, J. M. R.; Rodges, R. S. Prediction of Peptide Retention Times in Reversed-Phase High-Performance Liquid Chromatography I. Determination of Retention Coefficients of Amino Acid Residues of Model Synthetic Peptides. *J. Chromatogr. A* **1986**, *359*, 499–518. https://doi.org/10.1016/0021-9673(86)80102-9.
(147) Mant, C. T.; Burke, T. W. L.; Black, J. A.; Hodges, R. S. Effect of Peptide Chain Length on Peptide Retention Behaviour in Reversed-Phase Chromatogrphy. *J. Chromatogr. A* **1988**, *458*, 193–205. https://doi.org/10.1016/S0021-9673(00)90564-8.
(148) Moruz, L.; Käll, L. Peptide Retention Time Prediction. *Mass Spectrom. Rev.* **2017**, *36* (5), 615–623. https://doi.org/10.1002/mas.21488.
(149) Spicer, V.; Yamchuk, A.; Cortens, J.; Sousa, S.; Ens, W.; Standing, K. G.; Wilkins, J. A.; Krokhin, O. V. Sequence-Specific Retention Calculator. A Family of Peptide Retention Time Prediction Algorithms in Reversed-Phase HPLC: Applicability to Various Chromatographic Conditions and Columns. *Anal. Chem.* **2007**, *79* (22), 8762–8768. https://doi.org/10.1021/ac071474k.
(150) Bączek, T.; Wiczling, P.; Marszałł, M.; Heyden, Y. V.; Kaliszan, R. Prediction of Peptide Retention at Different HPLC Conditions from Multiple Linear Regression Models. *J. Proteome Res.* **2005**, *4* (2), 555–563. https://doi.org/10.1021/pr049780r.
(151) Palmblad, M.; Ramström, M.; Markides, K. E.; Håkansson, P.; Bergquist, J. Prediction of Chromatographic Retention and Protein Identification in Liquid Chromatography/Mass Spectrometry. *Anal. Chem.* **2002**, *74* (22), 5826–5830. https://doi.org/10.1021/ac0256890.
(152) Strittmatter, E. F.; Kangas, L. J.; Petritis, K.; Mottaz, H. M.; Anderson, G. A.; Shen, Y.; Jacobs, J. M.; Camp, D. G.; Smith, R. D. Application of Peptide LC Retention Time Information in a Discriminant Function for Peptide Identification by Tandem Mass Spectrometry. *J. Proteome Res.* **2004**, *3* (4), 760–769. https://doi.org/10.1021/pr049965y.

(153) Pfeifer, N.; Leinenbach, A.; Huber, C. G.; Kohlbacher, O. Statistical Learning of Peptide Retention Behavior in Chromatographic Separations: A New Kernel-Based Approach for Computational Proteomics. *BMC Bioinformatics* **2007**, *8* (1), 468. https://doi.org/10.1186/1471-2105-8-468.
(154) Pfeifer, N.; Leinenbach, A.; Huber, C. G.; Kohlbacher, O. Improving Peptide Identification in Proteome Analysis by a Two-Dimensional Retention Time Filtering Approach. *J. Proteome Res.* **2009**, *8* (8), 4109–4115. https://doi.org/10.1021/pr900064b.
(155) Golmohammadi, H.; Dashtbozorgi, Z.; Vander Heyden, Y. Support Vector Regression Based QSPR for the Prediction of Retention Time of Peptides in Reversed-Phase Liquid Chromatography. *Chromatographia* **2015**, *78* (1–2), 7–19. https://doi.org/10.1007/s10337-014-2819-1.
(156) Palmblad, M.; Ramström, M.; Markides, K. E.; Håkansson, P.; Bergquist, J. Prediction of Chromatographic Retention and Protein Identification in Liquid Chromatography/Mass Spectrometry. *Anal. Chem.* **2002**, *74* (22), 5826–5830. https://doi.org/10.1021/ac0256890.
(157) Maboudi Afkham, H.; Qiu, X.; The, M.; Käll, L. Uncertainty Estimation of Predictions of Peptides' Chromatographic Retention Times in Shotgun Proteomics. *Bioinformatics* **2017**, *33* (4), 508–513. https://doi.org/10.1093/bioinformatics/btw619.
(158) Žuvela, P.; Macur, K.; Jay Liu, J.; Bączek, T. Exploiting Non-Linear Relationships between Retention Time and Molecular Structure of Peptides Originating from Proteomes and Comparing Three Multivariate Approaches. *J. Pharm. Biomed. Anal.* **2016**, *127*, 94–100. https://doi.org/10.1016/j.jpba.2016.01.055.
(159) Moruz, L.; Tomazela, D.; Käll, L. Training, Selection, and Robust Calibration of Retention Time Models for Targeted Proteomics. *J. Proteome Res.* **2010**, *9* (10), 5209–5216. https://doi.org/10.1021/pr1005058.
(160) Petritis, K.; Kangas, L. J.; Ferguson, P. L.; Anderson, G. A.; Paša-Tolić, L.; Lipton, M. S.; Auberry, K. J.; Strittmatter, E. F.; Shen, Y.; Zhao, R.; Smith, R. D. Use of Artificial Neural Networks for the Accurate Prediction of Peptide Liquid Chromatography Elution Times in Proteome Analyses. *Anal. Chem.* **2003**, *75* (5), 1039–1048. https://doi.org/10.1021/ac0205154.
(161) Petritis, K.; Kangas, L. J.; Yan, B.; Monroe, M. E.; Strittmatter, E. F.; Qian, W.-J.; Adkins, J. N.; Moore, R. J.; Xu, Y.; Lipton, M. S.; Camp, D. G.; Smith, R. D. Improved Peptide Elution Time Prediction for Reversed-Phase Liquid Chromatography-MS by Incorporating Peptide Sequence Information. *Anal. Chem.* **2006**, *78* (14), 5026–5039. https://doi.org/10.1021/ac060143p.
(162) Gilar, M.; Jaworski, A.; Olivova, P.; Gebler, J. C. Peptide Retention Prediction Applied to Proteomic Data Analysis. *Rapid Commun. Mass Spectrom.* **2007**, *21* (17), 2813–2821. https://doi.org/10.1002/rcm.3150.
(163) Krokhin, O. V. Sequence-Specific Retention Calculator. Algorithm for Peptide Retention Prediction in Ion-Pair RP-HPLC: Application to 300- and 100-Å Pore Size C18 Sorbents. *Anal. Chem.* **2006**, *78* (22), 7785–7795. https://doi.org/10.1021/ac060777w.
(164) Spicer, V.; Lao, Y. W.; Shamshurin, D.; Ezzati, P.; Wilkins, J. A.; Krokhin, O. V. N-Capping Motifs Promote Interaction of Amphipathic Helical Peptides with Hydrophobic Surfaces and Drastically Alter Hydrophobicity Values of Individual Amino Acids. *Anal. Chem.* **2014**, *86* (23), 11498–11502. https://doi.org/10.1021/ac503352h.
(165) Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; Reimer, U.; Ehrlich, H.-C.; Aiche, S.; Kuster, B.;

Wilhelm, M. Prosit: Proteome-Wide Prediction of Peptide Tandem Mass Spectra by Deep Learning. *Nat. Methods* **2019**, *16* (6), 509–518. https://doi.org/10.1038/s41592-019-0426-7.
(166) Bouwmeester, R.; Gabriels, R.; Hulstaert, N.; Martens, L.; Degroeve, S. DeepLC Can Predict Retention Times for Peptides That Carry As-yet Unseen Modifications. *Nat. Methods* **2021**, *18* (11), 1363–1369. https://doi.org/10.1038/s41592-021-01301-5.
(167) Yeung, D.; Spicer, V.; Zahedi, R. P.; Krokhin, O. Exploring the Variable Space of Shallow Machine Learning Models for Reversed-Phase Retention Time Prediction. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 2446–2453. https://doi.org/10.1016/j.csbj.2023.02.047.
(168) *Mass Spectrometry: A Textbook*, 3rd edition.; Springer Berlin Heidelberg: New York, NY, 2017.
(169) Märk, T. D.; Dunn, G. H. *Electron Impact Ionization*; Springer-Verlag: Wien, 1985.
(170) Cappiello, A.; Famiglini, G.; Palma, P.; Siviero, A. Liquid Chromatography-electron Ionization Mass Spectrometry: Fields of Application and Evaluation of the Performance of a Direct-EI Interface. *Mass Spectrom. Rev.* **2005**, *24* (6), 978–989. https://doi.org/10.1002/mas.20054.
(171) Edmonds, C. G.; Smith, R. D. [22] Electrospray Ionization Mass Spectrometry. In *Methods in Enzymology*; Elsevier, 1990; Vol. 193, pp 412–431. https://doi.org/10.1016/0076-6879(90)93430-S.
(172) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science* **1989**, *246* (4926), 64–71. https://doi.org/10.1126/science.2675315.
(173) Kaufmann, R. Matrix-Assisted Laser Desorption Ionization (MALDI) Mass Spectrometry: A Novel Analytical Tool in Molecular Biology and Biotechnology. *J. Biotechnol.* **1995**, *41* (2–3), 155–175. https://doi.org/10.1016/0168-1656(95)00009-F.
(174) Dubois, F.; Knochenmuss, R.; Zenobi, R. An Ion-to-Photon Conversion Detector for Mass Spectrometry. *Int. J. Mass Spectrom. Ion Process.* **1997**, *169–170*, 89–98. https://doi.org/10.1016/S0168-1176(97)00210-3.
(175) Brown, K. L.; Tautfest, G. W. Faraday-Cup Monitors for High-Energy Electron Beams. *Rev. Sci. Instrum.* **1956**, *27* (9), 696–702. https://doi.org/10.1063/1.1715674.
(176) Tuithof, H. H.; Boerboom, A. J. H.; Meuzelaar, H. L. C. Simultaneous Detection of a Mass Spectrum Using a Channeltron Electron Multiplier Array. *Int. J. Mass Spectrom. Ion Phys.* **1975**, *17* (3), 299–307. https://doi.org/10.1016/0020-7381(75)87040-9.
(177) Koppenaal, D. W.; Barinaga, C. J.; Denton, M. B.; Sperline, R. P.; Hieftje, G. M.; Schilling, G. D.; Andrade, F. J.; Barnes, J. H.; Iv, I. MS Detectors. *Anal. Chem.* **2005**, *77* (21), 418 A-427 A. https://doi.org/10.1021/ac053495p.
(178) Hu, Q.; Noll, R. J.; Li, H.; Makarov, A.; Hardman, M.; Graham Cooks, R. The Orbitrap: A New Mass Spectrometer. *J. Mass Spectrom.* **2005**, *40* (4), 430–443. https://doi.org/10.1002/jms.856.
(179) Makarov, A. Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Anal. Chem.* **2000**, *72* (6), 1156–1162. https://doi.org/10.1021/ac991131p.
(180) Griffiths, D. J. *Introduction to Electrodynamics*, Fourth edition.; Pearson: Boston, 2013.
(181) Steigerwald, S.; Sinha, A.; Fort, K. L.; Zeng, W.-F.; Niu, L.; Wichmann, C.; Kreutzmann, A.; Mourad, D.; Aizikov, K.; Grinfeld, D.; Makarov, A.; Mann, M.; Meier, F. Full Mass Range

ΦSDM Orbitrap Mass Spectrometry for DIA Proteome Analysis. *Mol. Cell. Proteomics* **2024**, *23* (2), 100713. https://doi.org/10.1016/j.mcpro.2024.100713.
(182) Lange, O.; Damoc, E.; Wieghaus, A.; Makarov, A. Enhanced Fourier Transform for Orbitrap Mass Spectrometry. *Int. J. Mass Spectrom.* **2014**, *369*, 16–22. https://doi.org/10.1016/j.ijms.2014.05.019.
(183) Jedrychowski, M. P.; Huttlin, E. L.; Haas, W.; Sowa, M. E.; Rad, R.; Gygi, S. P. Evaluation of HCD- and CID-Type Fragmentation Within Their Respective Detection Platforms For Murine Phosphoproteomics. *Mol. Cell. Proteomics* **2011**, *10* (12), M111.009910. https://doi.org/10.1074/mcp.M111.009910.
(184) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-Energy C-Trap Dissociation for Peptide Modification Analysis. *Nat. Methods* **2007**, *4* (9), 709–712. https://doi.org/10.1038/nmeth1060.
(185) Brodbelt, J. S. Ion Activation Methods for Peptides and Proteins. *Anal. Chem.* **2016**, *88* (1), 30–51. https://doi.org/10.1021/acs.analchem.5b04563.
(186) Cleveland, J. P.; Rose, J. R. Identification of B-/y-Ions in MS/MS Spectra Using a Two Stage Neural Network. *Proteome Sci.* **2013**, *11* (Suppl 1), S4. https://doi.org/10.1186/1477-5956-11-S1-S4.
(187) Sandberg, A.; Branca, R. M. M.; Lehtiö, J.; Forshed, J. Quantitative Accuracy in Mass Spectrometry Based Proteomics of Complex Samples: The Impact of Labeling and Precursor Interference. *J. Proteomics* **2014**, *96*, 133–144. https://doi.org/10.1016/j.jprot.2013.10.035.
(188) Kreimer, S.; Belov, M. E.; Danielson, W. F.; Levitsky, L. I.; Gorshkov, M. V.; Karger, B. L.; Ivanov, A. R. Advanced Precursor Ion Selection Algorithms for Increased Depth of Bottom-Up Proteomic Profiling. *J. Proteome Res.* **2016**, *15* (10), 3563–3573. https://doi.org/10.1021/acs.jproteome.6b00312.
(189) Guo, J.; Huan, T. Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography–Mass Spectrometry Based Untargeted Metabolomics. *Anal. Chem.* **2020**, *92* (12), 8072–8080. https://doi.org/10.1021/acs.analchem.9b05135.
(190) Goldfarb, D.; Wang, W.; Major, M. B. MSAcquisitionSimulator: Data-Dependent Acquisition Simulator for LC-MS Shotgun Proteomics. *Bioinformatics* **2016**, *32* (8), 1269–1271. https://doi.org/10.1093/bioinformatics/btv745.
(191) Du, P.; Stolovitzky, G.; Horvatovich, P.; Bischoff, R.; Lim, J.; Suits, F. A Noise Model for Mass Spectrometry Based Proteomics. *Bioinformatics* **2008**, *24* (8), 1070–1077. https://doi.org/10.1093/bioinformatics/btn078.
(192) Craig, R.; Beavis, R. C. A Method for Reducing the Time Required to Match Protein Sequences with Tandem Mass Spectra. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2310–2316. https://doi.org/10.1002/rcm.1198.
(193) Craig, R.; Beavis, R. C. TANDEM: Matching Proteins with Tandem Mass Spectra. *Bioinformatics* **2004**, *20* (9), 1466–1467. https://doi.org/10.1093/bioinformatics/bth092.
(194) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry–Based Proteomics. *Nat. Methods* **2017**, *14* (5), 513–520. https://doi.org/10.1038/nmeth.4256.
(195) Bantscheff, M.; Lemeer, S.; Savitski, M. M.; Kuster, B. Quantitative Mass Spectrometry in Proteomics: Critical Review Update from 2007 to the Present. *Anal. Bioanal. Chem.* **2012**, *404* (4), 939–965. https://doi.org/10.1007/s00216-012-6203-4.

(196) Matthiesen, R.; Carvalho, A. S. Methods and Algorithms for Quantitative Proteomics by Mass Spectrometry. In *Mass Spectrometry Data Analysis in Proteomics*; Matthiesen, R., Ed.; Methods in Molecular Biology; Springer New York: New York, NY, 2020; Vol. 2051, pp 161–197. https://doi.org/10.1007/978-1-4939-9744-2_7.
(197) Tao, W. A.; Aebersold, R. Advances in Quantitative Proteomics via Stable Isotope Tagging and Mass Spectrometry. *Curr. Opin. Biotechnol.* **2003**, *14* (1), 110–118. https://doi.org/10.1016/S0958-1669(02)00018-6.
(198) Goshe, M. B.; Smith, R. D. Stable Isotope-Coded Proteomic Mass Spectrometry. *Curr. Opin. Biotechnol.* **2003**, *14* (1), 101–109. https://doi.org/10.1016/S0958-1669(02)00014-9.
(199) Rauniyar, N.; Yates, J. R. Isobaric Labeling-Based Relative Quantification in Shotgun Proteomics. *J. Proteome Res.* **2014**, *13* (12), 5293–5309. https://doi.org/10.1021/pr500880b.
(200) Yi, E. C.; Li, X.; Cooke, K.; Lee, H.; Raught, B.; Page, A.; Aneliunas, V.; Hieter, P.; Goodlett, D. R.; Aebersold, R. Increased Quantitative Proteome Coverage with13C/12C-Based, Acid-Cleavable Isotope-Coded Affinity Tag Reagent and Modified Data Acquisition Scheme. *PROTEOMICS* **2005**, *5* (2), 380–387. https://doi.org/10.1002/pmic.200400970.
(201) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. Quantitative Analysis of Complex Protein Mixtures Using Isotope-Coded Affinity Tags. *Nat. Biotechnol.* **1999**, *17* (10), 994–999. https://doi.org/10.1038/13690.
(202) Choe, L.; D'Ascenzo, M.; Relkin, N. R.; Pappin, D.; Ross, P.; Williamson, B.; Guertin, S.; Pribil, P.; Lee, K. H. 8-Plex Quantitation of Changes in Cerebrospinal Fluid Protein Expression in Subjects Undergoing Intravenous Immunoglobulin Treatment for Alzheimer's Disease. *PROTEOMICS* **2007**, *7* (20), 3651–3660. https://doi.org/10.1002/pmic.200700316.
(203) Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J. Multiplexed Protein Quantitation in Saccharomyces Cerevisiae Using Amine-Reactive Isobaric Tagging Reagents. *Mol. Cell. Proteomics* **2004**, *3* (12), 1154–1169. https://doi.org/10.1074/mcp.M400129-MCP200.
(204) Thompson, A.; Schäfer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Hamon, C. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Anal. Chem.* **2003**, *75* (8), 1895–1904. https://doi.org/10.1021/ac0262560.
(205) Brenes, A.; Hukelmann, J.; Bensaddek, D.; Lamond, A. I. Multibatch TMT Reveals False Positives, Batch Effects and Missing Values. *Mol. Cell. Proteomics* **2019**, *18* (10), 1967–1980. https://doi.org/10.1074/mcp.RA119.001472.
(206) Li, J.; Cai, Z.; Bomgarden, R. D.; Pike, I.; Kuhn, K.; Rogers, J. C.; Roberts, T. M.; Gygi, S. P.; Paulo, J. A. TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing. *J. Proteome Res.* **2021**, *20* (5), 2964–2972. https://doi.org/10.1021/acs.jproteome.1c00168.
(207) Sun, H.; Poudel, S.; Vanderwall, D.; Lee, D. G.; Li, Y.; Peng, J. 29-Plex Tandem Mass Tag Mass Spectrometry Enabling Accurate Quantification by Interference Correction. *PROTEOMICS* **2022**, *22* (19–20), 2100243. https://doi.org/10.1002/pmic.202100243.
(208) Wang, Z.; Yu, K.; Tan, H.; Wu, Z.; Cho, J.-H.; Han, X.; Sun, H.; Beach, T. G.; Peng, J. 27-Plex Tandem Mass Tag Mass Spectrometry for Profiling Brain Proteome in Alzheimer's Disease. *Anal. Chem.* **2020**, *92* (10), 7162–7170. https://doi.org/10.1021/acs.analchem.0c00655.

(209) Wu, Z.; Xiang, W.; Huang, L.; Li, S.; Zhang, X. Hyperplexing Approaches for up to 45-Plex Quantitative Proteomic Analysis. *Anal. Chem.* **2023**, *95* (12), 5169–5175. https://doi.org/10.1021/acs.analchem.3c00237.
(210) Dephoure, N.; Gygi, S. P. Hyperplexing: A Method for Higher-Order Multiplexed Quantitative Proteomics Provides a Map of the Dynamic Response to Rapamycin in Yeast. *Sci. Signal.* **2012**, *5* (217). https://doi.org/10.1126/scisignal.2002548.
(211) McGann, C. D.; Barshop, W. D.; Canterbury, J. D.; Lin, C.; Gabriel, W.; Huang, J.; Bergen, D.; Zabrouskov, V.; Melani, R. D.; Wilhelm, M.; McAlister, G. C.; Schweppe, D. K. Real-Time Spectral Library Matching for Sample Multiplexed Quantitative Proteomics. *J. Proteome Res.* **2023**, *22* (9), 2836–2846. https://doi.org/10.1021/acs.jproteome.3c00085.
(212) Hughes, C. S.; Zhu, C.; Spicer, V.; Krokhin, O. V.; Morin, G. B. Evaluating the Characteristics of Reporter Ion Signal Acquired in the Orbitrap Analyzer for Isobaric Mass Tag Proteome Quantification Experiments. *J. Proteome Res.* **2017**, *16* (5), 1831–1838. https://doi.org/10.1021/acs.jproteome.7b00092.
(213) Werner, T.; Sweetman, G.; Savitski, M. F.; Mathieson, T.; Bantscheff, M.; Savitski, M. M. Ion Coalescence of Neutron Encoded TMT 10-Plex Reporter Ions. *Anal. Chem.* **2014**, *86* (7), 3594–3601. https://doi.org/10.1021/ac500140s.
(214) Asara, J. M.; Christofk, H. R.; Freimark, L. M.; Cantley, L. C. A Label-free Quantification Method by MS/MS TIC Compared to SILAC and Spectral Counting in a Proteomics Screen. *PROTEOMICS* **2008**, *8* (5), 994–999. https://doi.org/10.1002/pmic.200700426.
(215) Zhu, W.; Smith, J. W.; Huang, C.-M. Mass Spectrometry-Based Label-Free Quantitative Proteomics. *J. Biomed. Biotechnol.* **2010**, *2010*, 1–6. https://doi.org/10.1155/2010/840518.
(216) Palomba, A.; Abbondio, M.; Fiorito, G.; Uzzau, S.; Pagnozzi, D.; Tanca, A. Comparative Evaluation of MaxQuant and Proteome Discoverer MS1-Based Protein Quantification Tools. *J. Proteome Res.* **2021**, *20* (7), 3497–3507. https://doi.org/10.1021/acs.jproteome.1c00143.
(217) Zhao, L.; Cong, X.; Zhai, L.; Hu, H.; Xu, J.-Y.; Zhao, W.; Zhu, M.; Tan, M.; Ye, B.-C. Comparative Evaluation of Label-Free Quantification Strategies. *J. Proteomics* **2020**, *215*, 103669. https://doi.org/10.1016/j.jprot.2020.103669.
(218) Wang, X.; Shen, S.; Rasam, S. S.; Qu, J. MS1 Ion Current-based Quantitative Proteomics: A Promising Solution for Reliable Analysis of Large Biological Cohorts. *Mass Spectrom. Rev.* **2019**, *38* (6), 461–482. https://doi.org/10.1002/mas.21595.
(219) Yu, F.; Haynes, S. E.; Nesvizhskii, A. I. IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. *Mol. Cell. Proteomics* **2021**, *20*, 100077. https://doi.org/10.1016/j.mcpro.2021.100077.
(220) Doellinger, J.; Blumenscheit, C.; Schneider, A.; Lasch, P. Increasing Proteome Depth While Maintaining Quantitative Precision in Short-Gradient Data-Independent Acquisition Proteomics. *J. Proteome Res.* **2023**, *22* (6), 2131–2140. https://doi.org/10.1021/acs.jproteome.3c00078.
(221) Pino, L. K.; Just, S. C.; MacCoss, M. J.; Searle, B. C. Acquiring and Analyzing Data Independent Acquisition Proteomics Experiments without Spectrum Libraries. *Mol. Cell. Proteomics MCP* **2020**, *19* (7), 1088–1103. https://doi.org/10.1074/mcp.P119.001913.
(222) Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N. G.; Old, W. M. Quantifying the Impact of Chimera MS/MS Spectra on Peptide Identification in Large-Scale Proteomics Studies. *J. Proteome Res.* **2010**, *9* (8), 4152–4160. https://doi.org/10.1021/pr1003856.
(223) Vaudel, M. MS2-Based Quantitation. In *Proteome Informatics*; Bessant, C., Ed.; The Royal Society of Chemistry, 2016; pp 155–177. https://doi.org/10.1039/9781782626732-00155.

(224) Hoopmann, M. R.; Winget, J. M.; Mendoza, L.; Moritz, R. L. StPeter: Seamless Label-Free Quantification with the Trans-Proteomic Pipeline. *J. Proteome Res.* **2018**, *17* (3), 1314–1320. https://doi.org/10.1021/acs.jproteome.7b00786.
(225) Ishihama, Y.; Oda, Y.; Tabata, T.; Sato, T.; Nagasu, T.; Rappsilber, J.; Mann, M. Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein. *Mol. Cell. Proteomics* **2005**, *4* (9), 1265–1272. https://doi.org/10.1074/mcp.M500061-MCP200.
(226) Rappsilber, J.; Ryder, U.; Lamond, A. I.; Mann, M. Large-Scale Proteomic Analysis of the Human Spliceosome. *Genome Res.* **2002**, *12* (8), 1231–1245. https://doi.org/10.1101/gr.473902.
(227) Gupta, N.; Pevzner, P. A. False Discovery Rates of Protein Identifications: A Strike against the Two-Peptide Rule. *J. Proteome Res.* **2009**, *8* (9), 4173–4181. https://doi.org/10.1021/pr9004794.
(228) Nurul Hidayah, S.; Gaikwad, M.; Heikaus, L.; Schlüter, H. Preparing Proteoforms of Therapeutic Proteins for Top-Down Mass Spectrometry. In *Proteoforms - Concept and Applications in Medical Sciences*; Zhan, X., Ed.; IntechOpen, 2020. https://doi.org/10.5772/intechopen.89644.
(229) Higdon, R.; Kolker, E. A Predictive Model for Identifying Proteins by a Single Peptide Match. *Bioinformatics* **2007**, *23* (3), 277–280. https://doi.org/10.1093/bioinformatics/btl595.
(230) Rozanova, S.; Uszkoreit, J.; Schork, K.; Serschnitzki, B.; Eisenacher, M.; Tönges, L.; Barkovits-Boeddinghaus, K.; Marcus, K. Quality Control—A Stepchild in Quantitative Proteomics: A Case Study for the Human CSF Proteome. *Biomolecules* **2023**, *13* (3), 491. https://doi.org/10.3390/biom13030491.
(231) Nakayasu, E. S.; Gritsenko, M.; Piehowski, P. D.; Gao, Y.; Orton, D. J.; Schepmoes, A. A.; Fillmore, T. L.; Frohnert, B. I.; Rewers, M.; Krischer, J. P.; Ansong, C.; Suchy-Dicey, A. M.; Evans-Molina, C.; Qian, W.-J.; Webb-Robertson, B.-J. M.; Metz, T. O. Tutorial: Best Practices and Considerations for Mass-Spectrometry-Based Protein Biomarker Discovery and Validation. *Nat. Protoc.* **2021**, *16* (8), 3737–3760. https://doi.org/10.1038/s41596-021-00566-6.
(232) Ivanov, M. V.; Garibova, L. A.; Postoenko, V. I.; Levitsky, L. I.; Gorshkov, M. V. On the Excessive Use of Coefficient of Variation as a Metric of Quantitation Quality in Proteomics. *PROTEOMICS* **2024**, *24* (1–2), 2300090. https://doi.org/10.1002/pmic.202300090.
(233) Piehowski, P. D.; Petyuk, V. A.; Orton, D. J.; Xie, F.; Moore, R. J.; Ramirez-Restrepo, M.; Engel, A.; Lieberman, A. P.; Albin, R. L.; Camp, D. G.; Smith, R. D.; Myers, A. J. Sources of Technical Variability in Quantitative LC–MS Proteomics: Human Brain Tissue Sample Analysis. *J. Proteome Res.* **2013**, *12* (5), 2128–2137. https://doi.org/10.1021/pr301146m.
(234) Li, X.; Lee, L. W.; Hayward, C.; Brusniak, M.-Y.; Fong, P.-Y.; McLean, M.; Mulligan, J.; Spicer, D.; Fang, K. C.; Hunsucker, S. W.; Kearney, P. An Integrated Quantification Method to Increase the Precision, Robustness, and Resolution of Protein Measurement in Human Plasma Samples. *Clin. Proteomics* **2015**, *12* (1), 3. https://doi.org/10.1186/1559-0275-12-3.
(235) Pekar Second, T.; Blethrow, J. D.; Schwartz, J. C.; Merrihew, G. E.; MacCoss, M. J.; Swaney, D. L.; Russell, J. D.; Coon, J. J.; Zabrouskov, V. Dual-Pressure Linear Ion Trap Mass Spectrometer Improving the Analysis of Complex Protein Mixtures. *Anal. Chem.* **2009**, *81* (18), 7757–7765. https://doi.org/10.1021/ac901278y.
(236) Thakur, S. S.; Geiger, T.; Chatterjee, B.; Bandilla, P.; Fröhlich, F.; Cox, J.; Mann, M. Deep and Highly Sensitive Proteome Coverage by LC-MS/MS Without Prefractionation. *Mol. Cell. Proteomics* **2011**, *10* (8), M110.003699. https://doi.org/10.1074/mcp.M110.003699.

(237) Nagaraj, N.; Alexander Kulak, N.; Cox, J.; Neuhauser, N.; Mayr, K.; Hoerning, O.; Vorm, O.; Mann, M. System-Wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-Shot Ultra HPLC Runs on a Bench Top Orbitrap. *Mol. Cell. Proteomics* **2012**, *11* (3), M111.013722. https://doi.org/10.1074/mcp.M111.013722.
(238) Iwasaki, M.; Miwa, S.; Ikegami, T.; Tomita, M.; Tanaka, N.; Ishihama, Y. One-Dimensional Capillary Liquid Chromatographic Separation Coupled with Tandem Mass Spectrometry Unveils the *Escherichia Coli* Proteome on a Microarray Scale. *Anal. Chem.* **2010**, *82* (7), 2616–2620. https://doi.org/10.1021/ac100343q.
(239) Köcher, T.; Swart, R.; Mechtler, K. Ultra-High-Pressure RPLC Hyphenated to an LTQ-Orbitrap Velos Reveals a Linear Relation between Peak Capacity and Number of Identified Peptides. *Anal. Chem.* **2011**, *83* (7), 2699–2704. https://doi.org/10.1021/ac103243t.
(240) Zhou, F.; Lu, Y.; Ficarro, S. B.; Webber, J. T.; Marto, J. A. Nanoflow Low Pressure High Peak Capacity Single Dimension LC-MS/MS Platform for High-Throughput, In-Depth Analysis of Mammalian Proteomes. *Anal. Chem.* **2012**, *84* (11), 5133–5139. https://doi.org/10.1021/ac2031404.
(241) Branca, R. M. M.; Orre, L. M.; Johansson, H. J.; Granholm, V.; Huss, M.; Pérez-Bercoff, Å.; Forshed, J.; Käll, L.; Lehtiö, J. HiRIEF LC-MS Enables Deep Proteome Coverage and Unbiased Proteogenomics. *Nat. Methods* **2014**, *11* (1), 59–62. https://doi.org/10.1038/nmeth.2732.
(242) Chen, L.; Shan, Y.; Weng, Y.; Sui, Z.; Zhang, X.; Liang, Z.; Zhang, L.; Zhang, Y. Hydrophobic Tagging-Assisted N-Termini Enrichment for In-Depth N-Terminome Analysis. *Anal. Chem.* **2016**, *88* (17), 8390–8395. https://doi.org/10.1021/acs.analchem.6b02453.
(243) Giron, P.; Dayon, L.; David, F.; Sanchez, J.-C.; Rose, K. Enrichment of N-Terminal Cysteinyl-Peptides by Covalent Capture. *J. Proteomics* **2009**, *71* (6), 647–661. https://doi.org/10.1016/j.jprot.2008.11.009.
(244) Tambor, V.; Hunter, C. L.; Seymour, S. L.; Kacerovsky, M.; Stulik, J.; Lenco, J. CysTRAQ — A Combination of iTRAQ and Enrichment of Cysteinyl Peptides for Uncovering and Quantifying Hidden Proteomes. *J. Proteomics* **2012**, *75* (3), 857–867. https://doi.org/10.1016/j.jprot.2011.09.027.
(245) Sanchez, A.; Sun, W.; Ma, J.; Betancourt, L.; Perez-Riverol, Y.; de-Cossio, J. F.; Padron, G.; Jiang, Y.; He, F.; Gonzalez, L. J.; Besada, V. Selective Isolation of Multiply Charged Peptides: A Confident Strategy for Protein Identification Using a Linear Trap Quadrupole Mass Spectrometer. *Eur. J. Mass Spectrom.* **2012**, *18* (6), 505–508. https://doi.org/10.1255/ejms.1204.
(246) Betancourt, L. H.; De Bock, P.-J.; Staes, A.; Timmerman, E.; Perez-Riverol, Y.; Sanchez, A.; Besada, V.; Gonzalez, L. J.; Vandekerckhove, J.; Gevaert, K. SCX Charge State Selective Separation of Tryptic Peptides Combined with 2D-RP-HPLC Allows for Detailed Proteome Mapping. *J. Proteomics* **2013**, *91*, 164–171. https://doi.org/10.1016/j.jprot.2013.06.033.
(247) Sánchez, A.; González, L. J.; Ramos, Y.; Betancourt, L.; Gil, J.; Besada, V.; Fernández-de-Cossio, J.; Alvarez, F.; Padrón, G. Selective Isolation of Lysine-Free Tryptic Peptides Delimited by Arginine Residues: A New Tool for Proteome Analysis. *J. Proteome Res.* **2006**, *5* (5), 1204–1213. https://doi.org/10.1021/pr060003w.
(248) Foettinger, A.; Leitner, A.; Lindner, W. Selective Enrichment of Tryptophan-Containing Peptides from Protein Digests Employing a Reversible Derivatization with Malondialdehyde and Solid-Phase Capture on Hydrazide Beads. *J. Proteome Res.* **2007**, *6* (9), 3827–3834. https://doi.org/10.1021/pr0702767.

(249) Ramos, Y.; Gutierrez, E.; Machado, Y.; Sánchez, A.; Castellanos-Serra, L.; González, L. J.; Fernández-de-Cossio, J.; Pérez-Riverol, Y.; Betancourt, L.; Gil, J.; Padrón, G.; Besada, V. Proteomics Based on Peptide Fractionation by SDS-Free PAGE. *J. Proteome Res.* **2008**, *7* (6), 2427–2434. https://doi.org/10.1021/pr700840y.
(250) Tuytten, R.; Ruttens, B.; Gheysen, K.; Sandra, K.; De Cremer, K.; Vlieghe, D.; Van Landuyt, N.; Thomas, G.; Martins, J. C.; Sandra, P.; Kas, K.; Verleysen, K. Application of a Combined Weak Cation-Exchange/Crown Ether Column: First Demonstrations of a Versatile Tool for Proteome Subselection. *Anal. Chem.* **2009**, *81* (7), 2456–2469. https://doi.org/10.1021/ac801975b.
(251) Betancourt, L.; Gil, J.; Besada, V.; González, L. J.; Fernández-de-Cossio, J.; García, L.; Pajón, R.; Sanchez, A.; Alvarez, F.; Padrón, G. SCAPE: A New Tool for the Selective CApture of PEptides in Protein Identification. *J. Proteome Res.* **2005**, *4* (2), 491–496. https://doi.org/10.1021/pr049794x.
(252) Van Damme, P.; Van Damme, J.; Demol, H.; Staes, A.; Vandekerckhove, J.; Gevaert, K. A Review of COFRADIC Techniques Targeting Protein N-Terminal Acetylation. *BMC Proc.* **2009**, *3* (S6), S6. https://doi.org/10.1186/1753-6561-3-S6-S6.
(253) Gevaert, K.; Van Damme, J.; Goethals, M.; Thomas, G. R.; Hoorelbeke, B.; Demol, H.; Martens, L.; Puype, M.; Staes, A.; Vandekerckhove, J. Chromatographic Isolation of Methionine-Containing Peptides for Gel-Free Proteome Analysis. *Mol. Cell. Proteomics* **2002**, *1* (11), 896–903. https://doi.org/10.1074/mcp.M200061-MCP200.
(254) Liu, T.; Qian, W.-J.; Strittmatter, E. F.; Camp, D. G.; Anderson, G. A.; Thrall, B. D.; Smith, R. D. High-Throughput Comparative Proteome Analysis Using a Quantitative Cysteinyl-Peptide Enrichment Technology. *Anal. Chem.* **2004**, *76* (18), 5345–5353. https://doi.org/10.1021/ac049485q.

Chapter 2 Separation orthogonality in liquid chromatography-mass spectrometry for proteomic applications: comparison of 16 different two-dimensional combinations

This section contains text and figures from a paper published in *Analytical Chemistry* in 2020, 92, 5, 3904-3912. **Darien Yeung**, Benilde Mizero, Danny Gussakovsky, Nicole Klaassen, Ying Lao, Victor Spicer, and Oleg V. Krokhin.

DY performed data analysis, writing, editing of the manuscript. BM, DG, NK, YL performed the experiments, VS performed data analysis, editing of the manuscript. OVK conceived and supervised the project, performed data analysis, writing, and editing of the manuscript.

Abstract

Peptide separation orthogonality for 16 different 2D LC-ESI MS systems has been evaluated using the same mass-spectrometry platform (Triple TOF 5600), the same complex tryptic *S. cerevisiae* digest and identical chemistry (reversed-phase C18, 0.1% formic acid) in the second dimension. To compare and contrast the behavior of the first dimension columns, a large proteomic retention dataset of ~30,000 tryptic peptides was collected for each 2D pairing. The selection of the first dimension system was made to cover the most popular peptide separation modes applied in proteomics: reversed-phase (RP) separations with different pH and ion-pairing modifiers, hydrophilic interaction liquid chromatography (HILIC), strong cation- and anion exchange (SCX, SAX), and mixed mode separations. The separation orthogonality generally increases in the order RP < SCX < HILIC < SAX, with the exception of high pH RP – low pH RP system, which showed the second best orthogonality value (68%), just behind PolySAX LP column (74%). The identification output of 2D LC-MS/MS system is driven by both separation orthogonality and efficiency, making high pH RP the best choice for the first dimension separation. Its performance in combination with a standard C18 at acidic pH can be increased further through the application

of pairwise fraction concatenation. The effect of the latter has been evaluated using *in-silico* fraction concatenation, which proven to show improvement only for RP separations in the first dimension. Concatenation of 2, 3 and 4-5 fractions into one is shown to be the most effective for high pH RP, HFBA- and TFA-based C18 separations, respectively. We also suggest simple guidelines for the unbiased determination of dissimilarity for two separation dimensions and evaluate separation orthogonality in 3D LC-LC-MS separation space for all systems under investigation.

2.1 Introduction

Reversed-phase (RP) HPLC is a leading tool for on-line combination with mass spectrometry in bottom-up proteomics. This combination is optimal due to robustness, MS-compatible separation speed and high detection sensitivity when formic/acetic acids are used as ion-pairing additives.[1] While acquisition speed and mass accuracy[2-4] have been increasing steadily for mass spectrometers, the developments of the LC-component of the system have been incremental as most of the separations in proteomics are still performed on fully porous silica-based C18 phases. The introduction of efficient core-shell UPLC packing materials, the wide applications of splitless nano-flow systems operating at high pressures and multi-dimensional separation techniques are among the recent advancements of separation technology for applications in proteomics.

Multi-dimensional separation methods[5,6] have been developed to increase MS/MS acquisition space (time) when analysing extremely complex peptide mixtures. Such complex digests may contain hundreds of thousands to millions of peptides. While the most sophisticated mass spectrometers can produce high quality MS/MS spectra at a rate of ~40 Hz, this is still not enough for acquiring all possible tandem spectra within the few hours of a typical one-dimensional

run. Another reason for using deeper fractionation is the sample loading limits for 1D nano-flow LC-MS and resulting problems with detection of low abundance species. Due to its simplicity and productivity 2D LC-MS has surpassed gel-based proteomics[7] approaches and now represents the leading method for deep shotgun proteomic analyses.

There are numerous 2D combinations used in proteomics, discussed in the comprehensive review by Di Palma et al.[6] The SCX-RP combination was introduced in 1999 and has been a method of choice for number of years[8,9] and implemented in a convenient on-line format, known as MudPIT. Other combinations are based on RP-RP,[10,11] HILIC-RP,[12] SAX-RP,[13] and mixed mode separation-RP,[14] essentially utilizing all available peptide separation chemistries. Recent publications on benchmarking the productivity of new mass spectrometers use high pH – low pH RP combination, indicating its wide acceptance and growing popularity.[4,15] This method was introduced for silica-[10] and polymer-based[11] RP columns and modified by introducing fraction concatenation strategy to further improve separation orthogonality.[16] Not surprisingly, the vast majority of techniques utilize reversed-phase separation in the last separation dimension, giving good compatibility with ESI. The selection of the first dimension chemistry is dictated by the basic understanding of separation principles, when two independent dimensions have to provide good peak capacity to separate as many components as possible, i.e. have complementary selectivity. Thus, SCX/SAX separation is driven by peptide charge and RP by hydrophobic interactions, promoting separation orthogonality. Separation efficiency, compatibility of the eluents, and robustness of the first dimension separation represents other criteria used during the selection process.[6]

There are several separation modes that are sufficiently orthogonal compared to reversed-phase separation at acidic pH.[5] Systematic studies on the selection of optimal 2D or 3D LC

combinations for proteomics, however, are very rare. The selection of better multidimensional protocols is accomplished by comparing the number of identified peptides/proteins (often relative to previously published results), rather than chromatographic selectivity metrics. Studies like this are being performed on different MS equipment complicating a fair comparison. Moreover, comparison within the same lab using identical equipment may be difficult due to continuous alteration of the mass-spectrometers performance over time due to source contamination and other factors. Overall, there is a need for unbiased evaluation of separation orthogonality using chromatographic criteria.

Pioneering work in this area by Gilar et al.[5] elucidated peak capacity and separation orthogonality for SCX-RP, high pH - low pH RP, low pH –low pH RP (Phenyl-C18, Perfluorophenyl-C18), HILIC-RP, SEC-RP pairings using a dataset of 196 peptides, which allowed for establishment of general trends and the crowning of HILIC at pH 4.5 as "orthogonality champion". The authors implemented a geometric approach based on the distribution statistics of the data points within two-dimensional orthogonality plots. The random distribution of 196 data points in 14x14 matrix will result in filling ~63% of the bins, representing a perfect 100% orthogonality. Identical separation conditions will result in diagonal positioning of all data points, filling ~ 7% of the bins (14 out of 196) and representing 0% orthogonality. All intermediate cases of experimental populating of 2D separation matrices were scaled in relation to these limits. Surprisingly, this approach is rarely used in proteomics studies. Most researchers,[17,18] including our laboratory,[19] employ R^2-value correlation statistics to assess separation orthogonality. A systematic comparison of these two methods was one of the motivation points for this study.

Another motivation was the fact that since 2005 two dimensional separations have become extremely powerful and ubiquitous: dozens of 2D methods have been reported, many generating

up to 100,000 unique peptide identifications in a matter of hours.[4,15] Our laboratory alone used more than a dozen of different 2D LC-MS/MS approaches to study peptide separation selectivity through the development of Sequence-Specific Retention Calculator model in different separation modes: RP (high – low pH),[16] SCX,[20] HILIC,[19,21] SAX, mixed-mode. Embarking on this project in 2016, we decided to use a standard BY4741 yeast tryptic digests for all experiments. This standardization simplifies filtering false positive identifications (critical for prediction models' development) and the comparison between the separation modes as retention of identical molecules are monitored across all these runs. All these experiments have been performed using 2D LC with various columns (conditions) in the first, identical Luna C18(2) columns with formic acid as ion-pairing modifier in the second dimension separation and the same Triple TOF 5600 mass spectrometer. The resulting unique collection of datasets allows not only for the systematic comparison of separation orthogonality, but helps to elucidate its effect on the identification output in otherwise nearly identical 2D-LC-MS/MS settings. This work summarizes our finding on separation orthogonality observed for sixteen different 2D LC combinations, suggests general guidelines for systematic comparison of 2D LC systems, including taking into account mass-spectrometry separation space, and provides overview and limitations of popular fraction concatenation strategy.

2.2 Experimental

2.2.1 Materials, digestion, chromatographic columns.

Tryptic digestions of yeast cell lysate, bovine serum albumin (BSA) and chromatographic alignment have been performed as described elsewhere.[20] Digestion included a standard Cys protection procedure using alkylation with iodoacetamide, followed by proteolysis with trypsin.

Designed standard peptides P1-P6,[22] N2-N10,[23] peptides with different charge[20] were synthesized by BioSynthesis Inc. (Lewisville, TX). These peptides were used for the initial selection of chromatographic conditions and for the retention data alignment in the second dimension by spiking each fraction with ~200 fmol of each peptide of P1-P6 mixture,[22] prior to the separation. The list of first dimension separation columns included (Table 2.1): Luna HILIC 3 µm 200Å (3x50 mm), Luna (2) Silica 3 µm 100Å (3x50mm), Luna C18(2) 5 µm 100Å (1x100 mm, used for TFA and HFBA RP separations) (Phenomenex, Torrance, CA), XBridge Amide 3.5µm 100Å (3x50), Atlantis Silica 3 µm 100Å (3x50), XTerra C18 5 µm 100Å (1x100 used for high pH RP separations) (Waters, Milford, MA); IonPac AS24 (2.1x250mm), IonPac AS32 (2.1x250mm) (Thermo Fisher, Sunnyvale, CA); Polyhydroxyethyl A 5 µm 100Å (2.1x100 mm), Polysulfoethyl A 5 µm 200Å (2.1x100 mm), PolySAX LP 3 µm 100Å (2.1x100 mm) (Poly LC, Columbia, MD), Scherzo SM-C18 3 µm 130Å (3x50 mm) (Imtakt, Kyoto, Japan); Daicel DCpack PTZ 3 µm (3x50 mm) (Daicel, Osaka, Japan). HPLC-grade acetonitrile, de-ionized and LC-MS water grade chemicals were used for the preparation of eluents.

2.2.2 First dimension separation conditions.

Agilent 1100 series HPLC system with UV detector (214 nm) with 100 µL injection loop and 150-300 µL/min flow rate was used for the separations. The separation conditions were selected to spread entire separation space over 40 fractions or more. Initial selection of separation conditions has been accomplished through separation of standard BSA digest and peptide mixtures with varying hydrophobicity (RPLC, P1-P6),[22] hydrophilicity (HILIC, N2-N10),[23] peptide charge for SCX[20] and SAX. 150-200 ug of yeast digest was separated in each case, fractions were lyophilized, spiked with standard peptides P1-P6 for the retention time alignment in the second dimension.[22]

Fractions were dissolved in 40-50 µl of buffer A (0.1% formic acid in water) to provide ~1 ug injections in the second dimensions to avoid column overloading.

Table 2.1: Chromatographic conditions for the first dimension separations used in this study.

Column	Eluent composition	Gradient slope	# of 1 min fractions
HILIC (gradient: % water increase starting from 10:90 water: acetonitrile, 0.3 mL/min)			
Luna HILIC 3 µm 200Å (3x50 mm)	A: 10 mM ammonium formate pH 4.5 B: 10 mM ammonium formate in 90:10 ACN:water	0.5%	39
XBridge Amide 3.5µm 100Å (3x50)		0.7%	38
Luna (2) Silica 3 µm 100Å (3x50)		1%	39
Atlantis Silica 3 µm 100Å (3x50)		1%	38
Polyhydroxyethyl A 3 µm 100Å (2.1x100)		1%	44
ZIC-HILIC 3.5 µm 100Å (2.1x100)		0.7%	41
ZIC-cHILIC ZIC 3 µm 100Å (2.1x100)		0.7%	41
Daicel DCpack PTZ		1%	47
SCX (gradient: salt concentration starting from 0 mM, 0.3 mL/min)			
PolySulfoethyl A 5 µm (2.1x100)	A: 80:20 water: Acetonitrile, 0.1% formic acid. B: same as A plus 500 mM of KCl	8.5 mM KCl per minute	46
SAX (gradient: salt concentration starting from 0 mM, 0.3 mL/min)			
SAX1 IonPac AS24 (2.1x250)	A: 20 mM Tris-HCl pH 8 B: same plus 1M NaCl	8 mM NaCl per minute	43
SAX2 IonPac AS32 (2.1x250)			46 (0.5 min)
SAX 3 PolySAX LP 3 µm 100Å (2.1x100)			40
RPLC (gradient: % acetonitrile increase starting from 100% water, identical ion-pairing in both buffers A and B, 0.15 mL/min)			
XTerra 5 µm 100Å (1x100)	20 mM ammonium formate pH 10	1%	70 (0.5 min)
Luna C18(2) 5 µm 100Å (1x100)	0.1% TFA	1%	70 (0.5 min)
Luna C18(2) 5 µm 100Å (1x100)	0.1% HFBA	1%	70 (0.5 min)
Mixed-mode (gradient: % acetonitrile increase starting from 100% water 0.2 mL/min)			
Scherzo SM-C18	A: 10 mM ammonium formate pH 4.5 B: 10 mM ammonium formate in 90:10 ACN:water	2%	39

XBridge Amide, Atlantis Silica, XTerra C18 (Waters, Milford, MA); Luna HILIC, Luna (2) Silica, Luna C18(2) (Phenomenex, Torrance, CA); IonPac AS24, IonPac AS32 (Thermo Fisher, Sunnywale, CA); Polyhydroxyethyl A, PolySulfoethyl A, PolySAX LP (Poly LC, Columbia, MD), Scherzo (Imtakt, Kyoto, Japan); Daicel DCpack PTZ (Daicel, Osaka, Japan)

2.2.3 Second dimension LC-MS/MS.

LC-MS/MS in the second dimension has been done using a standard data-dependent acquisition protocol using a TripleTOF 5600 mass spectrometer (Sciex, Concord, ON) and 2D LC Ultra system (Eksigent, Dublin, CA) and as described elsewhere.[19] Similar acquisition parameters were used for all experiments. LC settings featured 100µm x 200mm analytical column packed with 3 µm Luna C18(2) (Phenomenex, Torrance, CA), 300 µm x 5 mm PepMap 100 trap-column (ThermoFisher), 500nL/min flow rate, linear ~0.3-0.65% acetonitrile per minute gradient. Both eluents A (water) and B (acetonitrile) contained 0.1% formic acid. The gradient programs consisted of following steps: a linear increase from 0.4 to 31% buffer B (acetonitrile) in 47, 77 or 107 minutes, 5 minutes at 80% B and then 8 minutes at 0.4% B for column equilibration (60, 90 or 120 min total analysis time).

2.2.4 Data Analysis and retention time assignment.

The X!Tandem search engine[24] was used for peptide identifications with following settings: 20 ppm and 50 ppm mass tolerance for parent and daughter ions, respectively; alkylation of Cys with iodoacetamide; the list of potential modifications included Met oxidation, Asn and Gln deamidation, N-terminal cyclization at Cys and Gln. All identified tryptic non-modified peptides (Log (e) < -1) were additionally filtered using retention time prediction in both dimensions. We used existing SSRCalc RP model for formic acid conditions and tabulated retention (Hydrophobicity Index, HI)[22] values for yeast peptides from SSRCalc database for retention time filtering. Only low scoring peptides (-1 > Log (e) > -3) have been subjected to this procedure; no peptides with confident identifications (Log (e) < -3) were excluded. The same retention data

collections of non-modified tryptic peptides have been used for models' development and assessment of separation orthogonality.

Retention times in the first dimension were assigned as the fraction number in which this peptide was found. When peptide signal was distributed between two or more fractions, the intensity weighted average fraction number was used. Compared to retention time assignment using maximum of chromatographic peak (UV or MS detection), this procedure leaves some uncertainty in determining retention value. However, it does not interfere with retention prediction model development and estimating separation orthogonality. Experimental retention values in the second dimension (C18, formic acid) have been expressed in HI (acetonitrile %) units using tabulated HI values for P1-P6 peptides.[22] Therefore all orthogonality plots were done using measured HI vs. fraction number axis' assignment.

2.2.5 Calculation of orthogonality value.

All retention data collections contained ~ 30,000 peptides, except for Scherzo – RP combination. To eliminate uncertainty with precise assigning of the borders for orthogonality plots, we removed ~1% of data points from each of four edges of the plot (Figure 2.1 A). Resulting area is divided into a 100x100 matrix, populated by 10,000 data points randomly selected from each dataset. To smooth out random selection biases, this is repeated 25 times for each 2D combination. The filling percentage has been assigned for each case, and averaged for the 25 replicate simulations. We computed 2D separation orthogonality as follows:

$$\text{ORTHO(2)} = 100 * \left(\frac{\left(\frac{\sum bins}{10000} - 0.01 \right)}{0.631 - 0.01} \right);$$

where ∑bins/10000 is a fractional bins filling, 0.01 and 0.631 are fractional filling for identical separations and random distribution of data points.

Calculating 3D orthogonality, which takes into account two LC dimensions and the MS scale, was done in a similar fashion: the 3D separation space was divided into a 22x22x22 matrix and populated by 10,648 randomly selected data points, replicated 25 times and averaged:

$$\text{ORTHO(3)} = 100 * \frac{\left(\frac{\sum bins}{10648} - 0.002\right)}{0.631 - 0.002}.$$

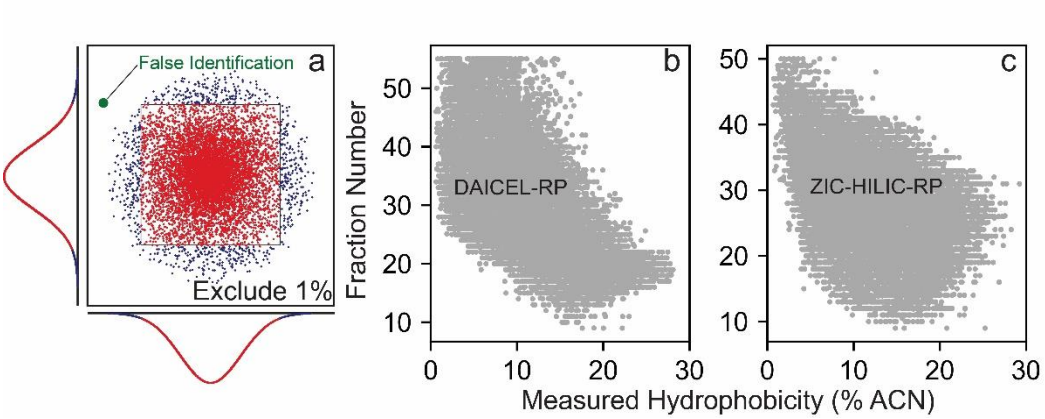


Figure 2.1: Determining peptide separation orthogonality in 2D LC systems. (A) schematic representation of normal peptide distribution in both dimensions. The presence of false positive identification datapoints (shown in green) will severely affect the location of the borders and, as a consequence, the entire procedure of orthogonality calculation; (B) orthogonality plot for Daicel DCpack PTZ –RP system exhibiting high retention (outside of the fraction collection range) of peptides with multiple positive charges on Daicel column; (C) orthogonality plot for fully developed chromatograms in both dimensions of ZIC-HILIC – RP system.

Alongside the geometric approach, the use of R^2-values, the square of Pearson's correlation coefficient, have been investigated for the purposes of quantifying orthogonality. The R^2-values

have been calculated for linear correlation between retention values in two dimensions using the standard Excel function. The optimal orthogonal setups will have low R^2-values on a scale of 0 to 1, a R^2-value of 1 would indicate the system is non-orthogonal.

2.3 Results and Discussion

2.3.1 Determining separation orthogonality.

The ability of a 2D system to uniformly deliver components of the mixture into the mass spectrometer depends on the peptide retention values distributed in two-dimensional space. This in turn is determined by the peptide distribution in the separation range of both columns; these distributions are drastically affected by the type of the gradient used: linear or non-linear, regardless of the eluent and gradient selected. Many separation systems will have non-retained components eluting in the flow-through or highly retained peptides, which will remain on the column under selected conditions. All of these make the determination of orthogonality ambiguous, requiring establishment of general guidelines for an unbiased comparison.

We suggest the following standard steps for determination of separation orthogonality in proteomic 2D runs:

i) Both separations should use linear gradients of the eluent components, which drive the separation: acetonitrile/water percentage or salt concentration for RP/HILIC or SCX/SAX, respectively. Indeed, each separation mode has a characteristic distribution of eluting species across the separation range, usually with lower density at the edges (Figure 2.1 A). The application of non-linear gradients is the most common approach to normalizing (equalizing) the distribution.[14, 25] This will significantly affect separation orthogonality.

ii) Separations in both dimensions should utilize the entire elution range, i.e. there should be no analyte, which can be eluted, remaining on the column. Figure 2.1 B shows an instance where this rule is violated. We employed linear water increasing gradient of 0.5-1% per minute for all HILIC separations, including Daicel DCpack PTZ. However, this column showed additional cation exchange retention mechanism resulting in high retention of positively charged peptides, which remained on the column after 50 min of fraction collection when 1% per minute (10 to 60% water in 50 minutes) was applied. This resulted in abrupt ending of the plot (Figure 2.1 B), compared to fully developed gradients in ZIC-HILIC – RP system as an example (Figure 2.1 C). Consequently, all orthogonality measurements for Daicel DCpack PTZ – RP systems reported here will be biased.

iii) Preferably, comparisons of separation orthogonality should be done using identical numbers of fractions. This condition, however, is very hard to satisfy given continuous alteration of the elution range between stationary phases, as it was shown for HILIC.[19]

iv) One percent of outermost datapoints at each of four sides of orthogonality plot should be excluded to mitigate the potential impact of false positive identifications. A geometric approach is driven by determining coverage of the entire area of orthogonality plot which could be affected by possible inclusion of false positive identifications when proteomic techniques are used for data collection. When such a data point is taken into account, the location of the plots' border will be shifted, severely affecting resulting orthogonality value (shown as hypothetical false hit in Figure 2.1 A). Given the high quality of the data used for this study we did not anticipate significant deviations, however this might not be always the case.

v) Separation orthogonality should be determined using a sufficiently large dataset. It is reasonable to suggest that using larger sets should increase the accuracy of the method. Gilar et al.[5] used 196 data points and 14x14 matrix, which may result in lower accuracy as some of the plots

shown in the original publication had data points with larger deviations from the general population. We chose 100x100 matrix size, which was 63.1% filled by 10,000 purely randomly distributed data points - in exact agreement with previously published data,[5] which corresponds to 100% orthogonality. Using identical separation conditions should result in occupation of a diagonal of 100 cells: 1% area occupation corresponding to 0% orthogonality. All remaining intermediate instances have been scaled accordingly.

Each separation mode is characterized by its unique peptide distribution, and it is reasonable to suggest that the highest possible orthogonality in a real case situation will correspond to identical normal distributions in both dimensions as shown in Figure 2.1 A. 54.5 % of available area will be covered in this case, corresponding to 86% orthogonality level, which will unlikely be reached by any combination of two separation protocols with linear gradients. Table 2.2 confirms that, showing experimental values determined by our approach ranging between 73.4% (PolySAX LP) and 29.3% (C18 TFA).

Table 2.2: Separation orthogonality and its impact on identification output of 2D LC-MS/MS.

First dimension	R^2	Orthogonality ORTHO(2), %	ORTHO(3), %	# of fractions	Instrument time, hrs	# of MS/MS	Peptide IDs	# of non-redundant peptide IDs	# of protein IDs
RP TFA	0.940	29.3	31.3	65	65	489768	256510	35453	3874
RP HFBA	0.750	46.3	47	65	65	526595	257924	44191	4238
RP pH10	0.380	67.7	64.8	68	68	687306	331257	61639	4686
Luna HILIC*	0.036	59.8	63	39	58.5	408721	222856	40160	4066
XBridge Amide	0.130	52	60.5	38	57	389917	207357	44489	4218
ZIC-HILIC	0.200	49.7	57.1	41	61.5	311879	206256	38037	4104
ZIC-cHILIC	0.140	53.6	60.1	41	61.5	352314	225762	39791	4194
Atlantis Silica	0.080	47.8	65	38	57	342317	187556	42715	4165
Luna Silica	0.120	48.5	65	39	58.5	350471	194985	40457	4030
Polyhydroxyethyl A	0.224	60	60.2	44	66	353788	227292	31940	3609
Daicel DCpack PTZ**	0.380	47	52.9	47	47	276146	170995	29801	3923
SCX Polysulfoethyl A	0.130	45.5	59.7	46	69	552954	196470	34454	4185

SAX1 IonPac AS24	0.004	55.5	71.7	43	56.5	418028	266460	38483	4238
SAX2 IonPac AS32	0.044	62.7	71.5	46	46	258732	166601	31886	4018
SAX3 PolySAX LP	0.019	74.3	70.8	40	55	579620	347700	34985	4025
Scherzo SM-C18	0.390	34.3	48.8	39	67.5	336718	158070	22591	3403

* - HILIC columns are listed in the order of increasing hydrophilicity;[23] **- orthogonality values for this column have been affected by incomplete development of chromatogram due to additional cation-exchange interactions as shown in Figure 2.1 B.

2.3.2 Separation orthogonality and its effect on the output of 2D LC-MS/MS analysis.

Since all 2D LC-MS/MS acquisitions in this study were performed using identical LC-MS/MS in the second dimension it was of interest to investigate the effect of separation orthogonality (if any) on the analysis output. This information is not readily available in the literature due to significant differences in the performance of various mass spectrometers used across laboratories. The yeast genome encodes ~4,500 proteins and we expected to see 80-90% of its coverage using 50-70 hours analysis time on TripleTOF 5600. Most of the methods (Table 2.2) showed similar protein identification numbers, suggesting that identification plateau has been reached. Therefore, we used the number of identified unique peptides per hour of instrument time as productivity criteria. The total number of MS/MS or total number of identified spectra can be used as well; however due to differences in separation efficiency and variable distribution of peaks between fractions, these numbers could be higher for low efficiency separation in the first dimension. Figure 2.2 A shows the dependence of analysis output on separation orthogonality expressed as R^2-value. There is no obvious trend in this plot except lower output values for less orthogonal (TFA, Scherzo) and higher output for 2D systems with very low R^2 for SAX separations.

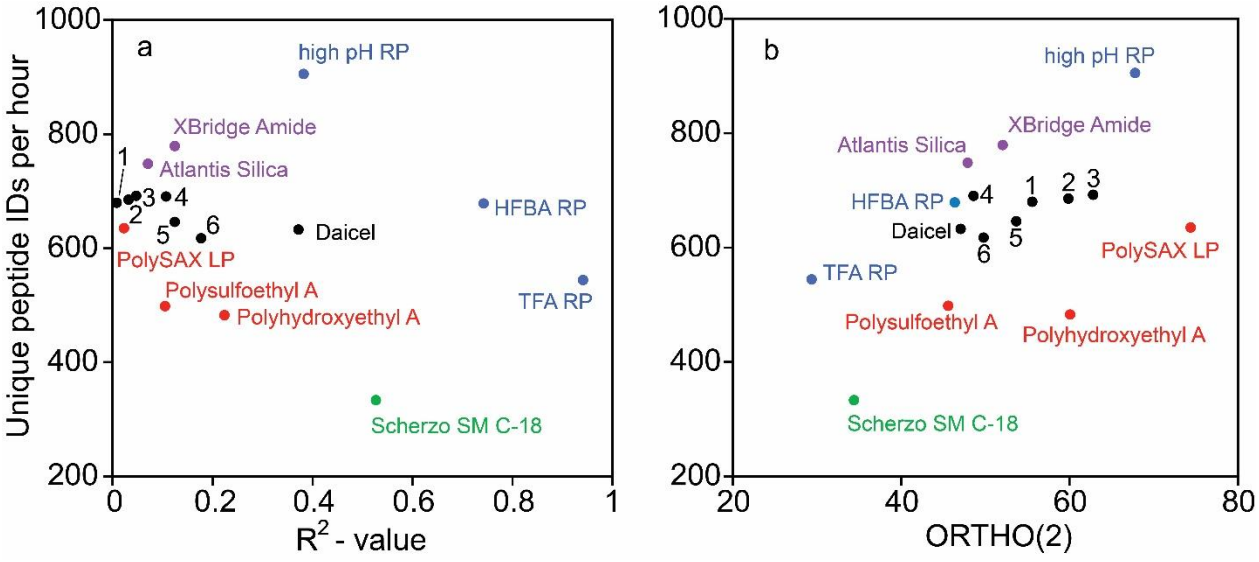


Figure 2.2: The effect of separation orthogonality on the analysis output of 2D LC-MS/MS. (A) correlation between the number of unique peptide IDs per hour and R^2-value of orthogonality plots; (B) similar plot for the separation orthogonality determined by geometric approach. 1 - IonPac AS24, 2 – Luna HILIC, 3 - IonPac AS32, 4 – Luna Silica, 5 – ZIC-cHILIC, 6 – ZIC-HILIC.

Plotting the number of unique peptides identified per hour of instrument time versus separation orthogonality showed definitive linear trend (Figure 2.2 B): analysis output increases with a systems' orthogonality. Yet, there are deviations, which are consistent with the separation efficiency of the column in the first dimension. The best output for given orthogonality value was observed for all RP separations and the HILIC columns from Waters (XBridge Amide and Atlantis Silica) – all showing the best chromatographic peak shape. Lower outputs are characteristic for PolyLC columns and mixed mode Scherzo. These columns exhibited relatively poor peak shape, which could be a consequence of larger particle size for Polyhydroxyethyl A and Polysulfoethyl A columns and mixed mode character of separation for Scherzo. Moreover, it is known that the peptide/protein peak shape on 100 Å PolyLC columns is affected by steric effects, which leads to

diffused peak front. It is obvious, that poor peak shape results in peaks' distribution across larger number of fractions, thus reducing the number of unique identifications. This situation was the most apparent for PolySAX LP. 2D LC-MS/MS using this column resulted in highest number of acquired and identified MS/MS per hour. However, the number of unique identifications was very similar to other modes due to elution of individual components in 2-3 or more fractions.

These results confirm that using geometric approach to determine separation orthogonality is more accurate compared to R^2-value measurement. Indeed there is no obvious correlation between these two values for the systems under investigation, except for the extreme cases of RP TFA, RP HFBA, PolySAX LP separations (Table 2.2) with very low and high orthogonality, respectively.

2.3.3 Separation orthogonality in 3D (LC-LC-MS) separation space.

The MS system itself represents a separation technique similar to the chromatographic dimensions. The separation occurs in *m/z* space in this case and the results of peak assignment of the detectable features will be affected if peptides with similar *m/z* will be clustered together at the same retention time. To visualize this effect, we used heat map orthogonality plots where the color of a data point depends on the peptide *m/z* value (Figure 2.3 A-I). To quantitatively estimate 3D orthogonality values we used a similar geometric approach by splitting the separation space in 22x22x22 matrix and populating it with 10,648 randomly selected data points from datasets. The results of these simulations are shown in Table 2.2. The effect of adding an MS dimension depends on the presence/absence of correlation between *m/z* and retention values in both dimensions. Since all 2D separation systems include identical RP (formic acid) – MS separation space with some correlation between RP retention values and *m/z*, the effect will depend on whether or not there is a correlation

between retention values in the first dimensions and mass-to-charge ratios. As expected, the biggest winners in this case were HILIC separation on bare silicas with low correlations between retention and m/z. At the same time RP-based first dimension separations did not gain much due to some correlation between RP retention and *m/z*.

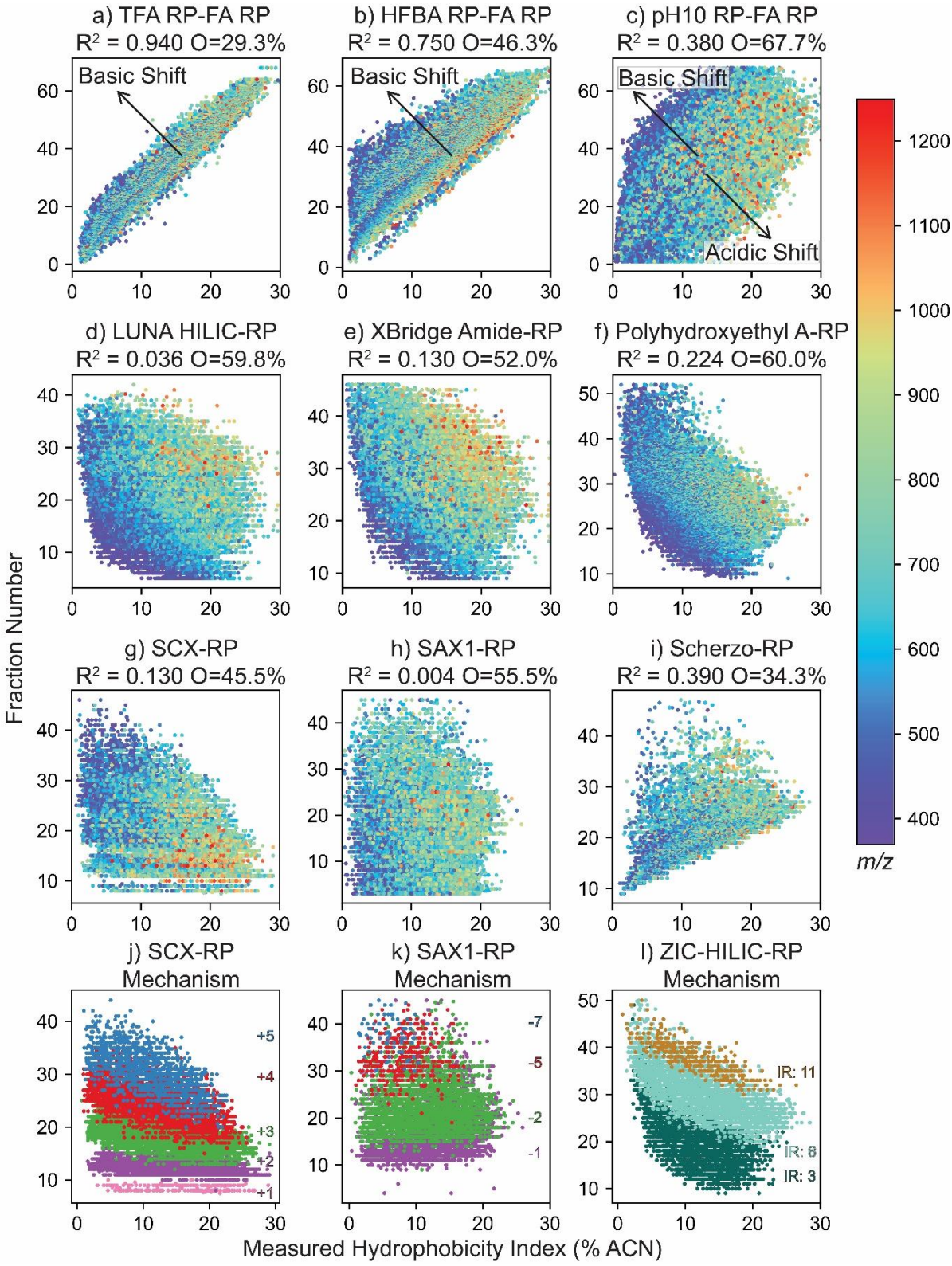

a) TFA RP-FA RP
R² = 0.940 O=29.3%
Basic Shift
b) HFBA RP-FA RP
R² = 0.750 O=46.3%
Basic Shift
c) pH10 RP-FA RP
R² = 0.380 O=67.7%
Basic Shift
Acidic Shift
d) LUNA HILIC-RP
R² = 0.036 O=59.8%
e) XBridge Amide-RP
R² = 0.130 O=52.0%
f) Polyhydroxyethyl A-RP
R² = 0.224 O=60.0%
g) SCX-RP
R² = 0.130 O=45.5%
h) SAX1-RP
R² = 0.004 O=55.5%
i) Scherzo-RP
R² = 0.390 O=34.3%
j) SCX-RP
Mechanism
+5
+4
+3
+2
+1
k) SAX1-RP
Mechanism
-7
-5
-2
-1
l) ZIC-HILIC-RP
Mechanism
IR: 11
IR: 6
IR: 3
Fraction Number
Measured Hydrophobicity Index (% ACN)
m/z
1200
1100
1000
900
800
700
600
500
400

Figure 2.3: Orthogonality plots and factors determining peptide separation dissimilarity in 2D LC peptide separation. (A-I) Orthogonality plots for selected 2D-LC systems using *m/z* heat map scale shown on the right; (J-K) orthogonality in ion-exchange-RP separations is driven by peptide charge; (L) HILIC-RP orthogonality is determined by the number of ionized residues in peptide sequence.

2.3.4 Separation orthogonality in 2D LC systems: general observations and features of separation mechanism, which determine separation orthogonality.

Table 2.2 summarizes separation orthogonality determined for all 2D combinations tested in this study, while Figure 2.3 shows selected orthogonality plots. Several types of the plots are observed based on general retention trends. RP-RP systems (Figure 2.3 A-C) are characterized by diagonal localization of the data points since all of them are driven by hydrophobic interactions. HILIC-RP (Figure 2.3 D-F) systems show anti-correlation because opposite retention mechanisms are used: hydrophilic vs. hydrophobic interactions. In the SCX-RP (Figure 2.3 G) plot, points mostly occupy the middle and bottom-left quadrant, somewhat similar to HILIC. SAX-RP (Figure 2.3 H) visually shows the most uniform distribution. Scherzo-RP plot (Figure 2.3 I) resembles RP-RP combinations with some peptides exhibiting higher retention in mixed-mode, thus occupying the upper-left quadrant.

All of these features are driven by the differences in the separation mechanisms of the various separation modes. Knowing these rules will enhance our knowledge regarding peptide 2D separations and facilitate method development.

RP-RP. Reversed-phase separations using different ion-pairing modifiers and pH represent a good example for illustrating the driving forces of separation orthogonality. The distribution of the peptide population widens with increased hydrophobicity of ion-pairing modifier and changes in eluent pH from 2.8 to 10 (Figure 2.3 A-C). Peptides carrying multiple internal cationic residues are the most affected upon changing hydrophobicity of the ion-pairing additive at acidic pH.[26] For example, when being separated with HFBA, the internally cationic peptides exhibit relatively high retention compared to FA system. This results in their later elution in HFBA dimension (shift to top left quadrant) and early elution when transferred into the second dimension RP-formic acid – shown as a "basic shift" in Figure 2.3 B. The same mechanism is driving orthogonality in TFA-FA system, with a smaller basic shift due to the lower hydrophobicity of TFA compared to HFBA.

Drastic changes in pH from acidic to basic conditions affect the hydrophobic contribution of basic and acidic residues: Asp, Glu become negatively charged and more hydrophilic, whereas Arg, Lys, and His lose protons and become more hydrophobic.[10, 11, 16] This behavior explains deviations from the diagonal distribution of relatively (relative to general population of peptides) basic and relatively acidic peptides (top left quadrant and bottom right quadrant respectively in Figure 2.3 C) – depicted as basic and acidic shifts.

SCX-RP. Electrostatic interactions depend on both, peptide charge and size. Smaller peptides, which on average have a lower number of hydrophobic residues, i.e. lower hydrophobicity, are characterized by stronger electrostatic interactions compared to larger species of the same charge.[20] Because of this, the SCX-RP plot consists of several dependences for peptides of different charges with negative slopes (Figure 2.3 G, J). +1 peptides can be separated from the remaining population, while other groups (+2, +3, +4, etc.) show significant overlap. It should be noted that the orthogonality plot shown here was obtained for a SCX eluent system with 20% acetonitrile,

resulting in suppressed hydrophobic interactions for the ion-exchange dimension.[20] The MudPIT variation of SCX-RP utilizes low acetonitrile eluents for salt elution steps,[8,9] and would exhibit a different orthogonality.

SAX-RP. Both visually and according to the R^2-values, this combination exhibits the highest orthogonality among all LC-LC pairings and shows characteristics similar to SCX-RP due to the interplay between peptide charge and size (Figure 2.3 H,K). However, it is also characterized by a sizeable population of non-retained peptides that are neutral or cationic at pH 8. Clustering of these peptides results in the lower orthogonality values determined by the geometric approach.

HILIC-RP orthogonality is driven by interactions of opposing characteristics: hydrophilic versus hydrophobic. One could expect that, these separation modes should show anti-correlation with a negative slope on the orthogonality plot. However, differences in the chemistry of the interacting phases, pH, and ion-pairing environment results in the drop-shaped orthogonality plot with a negative slope (Figure 2.3 D-F). Electrostatic interactions in HILIC systems provide additional alteration in separation selectivity, as was shown for negatively charged silanol groups on bare silica columns and distal groups of zwitterionic ZIC-HILIC and ZIC-cHILIC. Boersema et al.[17] have suggested that HILIC (pH 3) - RP resembles SCX-RP due to the fact that retention for both HILIC and SCX at acidic pH is driven by the presence of charged residues. Increasing the pH of the eluent to 4.5 adds to this list Asp, Glu and C-terminal carboxyl group, as HILIC retention depends on the number of all ionized residues (IR). Figure 2.3 L shows retention trends in HILIC-RP 2D system to highlight this similarity between SCX-RP and HILIC-RP.

As we have data spanning many HILIC phases, it was of interest to evaluate orthogonality values in this group of 2D LC systems relative to overall hydrophilicity of the phases. Table 2.2 lists HILIC columns in the order of their increasing hydrophilicity from Luna HILIC to

Polyhydroxyethyl A. One should expect increased R^2-value correlations and decreased orthogonality in the same order. Indeed Polyhydroxyethyl A showed by far the highest R^2 correlation (with negative slope) as the most hydrophilic, i.e. opposite to RP. At the same time, unexpectedly, it exhibits almost the same geometric orthogonality as the least hydrophilic Luna HILIC, with a value of ~60%.

Mixed-mode (RP-WAX-WCX) – RP in our study is represented by one example of Scherzo SM-C18. This column is prepared by mixing stationary phases with RP, weak-anion and weak-cation functionality and exhibits stronger retention of highly charged peptides, which occupies the upper left quadrant with significant deviation from diagonal RP-RP driven trend line (Figure 2.3 I).

Overall, the separation orthogonality increases in the following order: RP < SCX < HILIC < SAX. Some of these findings are to be expected given: i) similarity between RP (formic acid) and other RP systems; ii) drastic differences in separation mechanisms between RP and SCX/SAX. At the same time, the sufficiently high orthogonality values of HILIC-RP was somewhat surprising, given the expected anti-correlation between RP and HILIC retention values. The geometric separation orthogonality for high pH RP – low pH RP showed the second best value of 67.7% despite having a 0.38 R^2, characteristic of non-orthogonal separations. Together with superior separation efficiency of reversed-phase mode, robustness, and availability of fraction concatenation options[16] this establishes (Figure 2.2 B) this LC-LC pairing as the most attractive 2D option.

2.3.5 Fraction concatenation for improved separation orthogonality.

Fraction concatenation in 2D LC allows reducing the MS instrument utilization time by combining fractions, components of which will not overlap in the second dimension separation. Examples of its use can be found in COFRADIC applications when identical LC systems are used in both dimensions.[27] The first application of orthogonal separations in bottom-up proteomics was demonstrated by Dwivedi et al.[16] The authors suggested pair-wise concatenation of fractions leading to reduced MS analysis time (2-times) with minimal loss of identification output in high pH RP – low pH RP. The schematic example of concatenation procedure is shown in Figure 2.4. Fractions collected in the first dimension are split in half. First fraction in the first half of the fractions (blue) is then combined with first fraction in the second half (red), second fraction with the second, and so on. The improvement in separation orthogonality can be seen by lower R^2 and higher ORTHO(2) values. One could expect more drastic increases in orthogonality based on visual representation of the concatenated orthogonality plot. However components of two concatenated fractions begin to overlap (Figure 2.4 B), leading to a lower than expected efficacy.

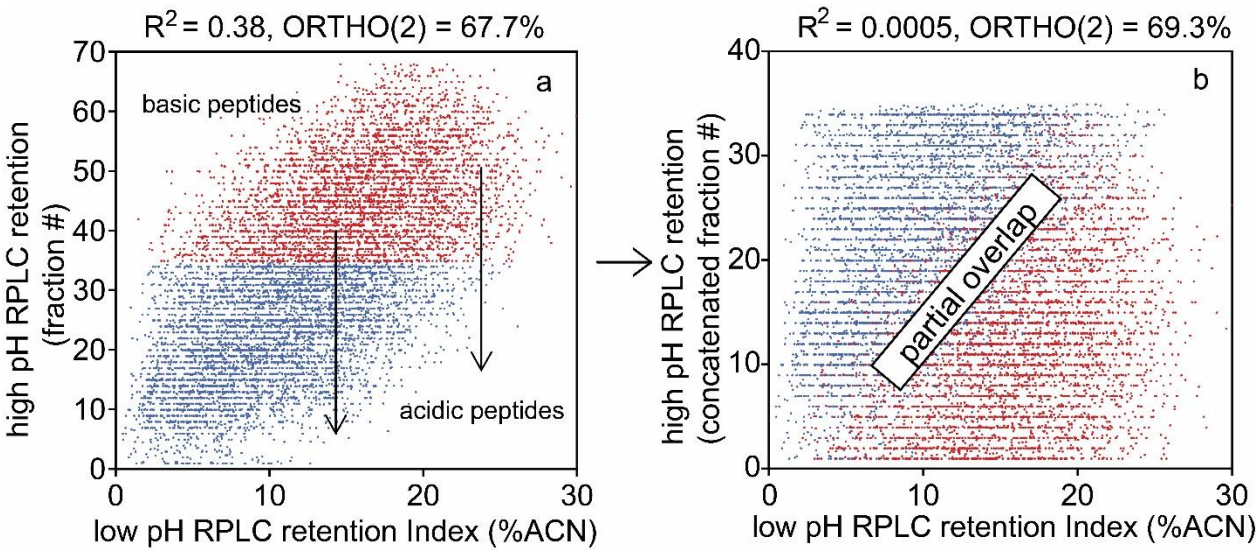


Figure 2.4: Pairwise fraction concatenation in high pH RP – low pH RP leads to increased separation orthogonality. Experimental orthogonality plot (A) has been converted into (B) using *in-silico* simulation (concatenation).

It was interesting to note that similar to the Dviwedi et al.[16] work, a concatenation strategy based on combining four fractions was proposed later by Wang et al.[28] and used by Mertins et al.[29] for deep fractionation of and detection of PTMs-carrying peptides. The latter report used 96 high pH RP fractions and combined together 4 or 12 fractions giving 24 or 8 concatenated fractions depending on the desired depth of analysis. To address the efficiency of these approaches we performed *in-silico* concatenation of 2-6 fractions into one, and determined the ORTHO(2) values for these datasets. The results shown in Table 2.3 illustrate that pair-wise concatenation provides superior orthogonality for high pH RP and dramatically lowers the R^2 correlation value. Increasing depth of concatenation further reduces performance of 2D LC-MS: R^2-values remain the same, with decreasing orthogonality (Table 2.3). In other words, concatenation of three or more fractions mixed peptides just separated in the first dimension leading to their increased overlap in the second.

Similar simulations have been performed for other first dimension columns. Table 2.3 shows that concatenation strategy is efficient only for 2D pairings with relatively high R^2-values: all RPs, Scherzo and the most hydrophilic HILIC columns Polyhydroxyethyl A and Daicel.

Table 2.3: The impact of multiple fraction concatenation strategy on separation orthogonality.

First dimension separation	no concatenation	2x	3x	4x	5x	6x
RP pH10 (ORTH(2))	67.7	69.3*	62.6	57.9	55.9	53.9
RP pH10 (R^2)	0.38	0.005	0.005	0.004	0.004	0.004
RP TFA	29.3	38.9	44.2	46.5*	46.5*	45.0
RP HFBA	46.4	54.7	56.7*	54.0	50.7	48.7
Luna HILIC	59.8*	55.8	49.7	47.3	44.9	43.5
XBridge Amide	52.0*	50.0	43.6	39.2	36.5	34.6
ZIC-HILIC	49.8*	49	42.7	39.6	37.5	35.6
ZIC-cHILIC	53.7*	49.5	43.4	39.7	38	36
Atlantis silica	47.9*	46.2	40.5	37.9	35.9	34.8
Luna Silica	48.4*	46.5	41.3	38.5	36.6	35.4
Polyhydroxyethyl A	60.0	63.9*	59.8	55.5	54.3	53.5
Daicel DCpack PTZ	46.9	49.6*	45.2	41.8	40.0	38.1
SCX Polysulfoethyl A	45.5*	45.5*	39.2	35.6	32.7	31.3
SAX1 AS24	55.5*	48.9	41.6	36.5	34.3	32.8

SAX2 AS32	62.8*	54.6	48.8	44.4	41.8	39.3
SAX3 PolySAX LP	74.3*	72.0	68.9	65.1	63.8	62.8
Scherzo SM-C18	34.3	36.4	37.2*	36.7	35.4	34.3

* - concatenation multiplier for optimal separation orthogonality; Note that fraction concatenation has been performed *in-silico* using experimental orthogonality data.

2.4 Conclusions

High-throughput proteomics has revolutionized peptide separation science in the past fifteen years providing access to virtually unlimited retention data derived from bottom-up proteomic experiments. Not long ago, retention modeling and studies of separation orthogonality using hundreds of peptides were the state-of-the art. Nowadays, 2D LC-MS/MS approach provides access to massive retention data in any peptide separation mode, which can be utilized in a column format. The final dimension's RPLC-MS serves in this case as universal detection tool, while the chemistry in the first dimension separation varies. We used this advantage of high throughput proteomics to investigate separation orthogonality in 2D LC-MS systems.

Our findings confirmed that using R^2-value correlation between retention times of two columns is not an adequate measure of separation dissimilarity. A geometric approach based on calculation of the filling percentage of 2D separation plot is more accurate, correlating with the identification output of 2D LC-MS/MS. Overall, separation orthogonality increases in the following order: RP < SCX < HILIC < SAX, with the exception of high pH RP – low pH RP system. We suggest that all separation modes except low pH RP can be useful as a first separation dimension without fraction concatenation. Separation efficiency in the first dimension is another critical parameter governing the selection of first dimension conditions. Sharper peaks lead to a

narrower distribution of analytes between fractions and higher efficiency of entire 2D LC-MS/MS. Taking into account both resolution and orthogonality, we concluded that high pH – low pH RP-RP pairing is the most efficient as it provides second-best orthogonality of 67.7%, and is characterized by superior separation efficiency and robustness, in combination yielding the highest analysis output.

The efficiency of the analysis can be further improved by fraction concatenation, which should be, however, used with caution. *In-silico* interrogation of the data showed that multiple fraction concatenation may lead to the loss in orthogonality if it results in excessive overlap of peptide peaks in the second dimension. Concatenation was found to improve separation orthogonality only for 2D combinations with some correlations between retention values in both dimensions: reversed-phase separation, mixed-mode Scherzo and the most hydrophilic HILIC phases. There are, however, optimal concatenation multipliers where moving beyond them will actually reduce the separation orthogonality. We found that 2, 3 and 4-5 fractions concatenation is the most efficient for high pH RP, HFBA- and TFA-based C18 separations, respectively.

Finally, for the first time we considered MS range as an additional separation dimension in overall LC-LC-MS combinations. Its addition increases separation orthogonality further, but only in cases of independent behavior of *m/z* and retention values in the first dimension. Because of that, all reversed-phase separations in the first dimension (TFA, HFBA, pH 10 based eluents), exhibit no gain upon adding the MS dimension.

2.5 References

(1) Bensimon, A.; Heck, A. J.; Aebersold, R. *Annu. Rev. Biochem.* **2012**, *81*, 379-405.
(2) Hu, Q.; Noll, R. J.; Li, H.; Makarov, A.; Hardman, M.; Graham Cooks, R. *J Mass Spectrom* **2005,** *40*, 430-443.
(3) Andrews, G. L.; Simons, B. L.; Young, J. B.; Hawkridge, A. M.; Muddiman, D. C. *Anal Chem* **2011,** *83*, 5442-5446.

(4) Kelstrup, C. D.; Bekker-Jensen, D. B.; Arrey, T. N.; Hogrebe, A.; Harder, A.; Olsen, J. V. *J Proteome Res* **2018**, *17*, 727-738.
(5) Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C. *Anal. Chem.* **2005**, *77*, 6426-6434.
(6) Di Palma, S.; Hennrich, M. L.; Heck, A. J.; Mohammed, S. *J Proteomics* **2012**, *75*, 3791-3813.
(7) Lambert, J. P.; Ethier, M.; Smith, J. C.; Figeys, D. *Anal Chem* **2005**, *77*, 3771-3787.
(8) Wolters, D. A.; Washburn, M. P.; Yates, J. R., 3rd *Anal Chem* **2001**, *73*, 5683-5690.
(9) Washburn, M. P.; Wolters, D. A.; Yates, J. R., 3rd *Nat Biotechnol* **2001**, 19, 242-247.
(10) Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C. *J Sep Sci* **2005**, *28*, 1694-1703.
(11) Toll, H.; Oberacher, H.; Swart, R.; Huber, C. G. *J Chromatogr A* **2005**, *1079*, 274-286.
(12) Boersema, P. J.; Mohammed, S.; Heck, A. J. *Anal Bioanal Chem* **2008**, *391*, 151-159.
(13) Ritorto, M. S.; Cook, K.; Tyagi, K.; Pedrioli, P. G.; Trost, M. *J Proteome Res* **2013**, *12*, 2449-2457.
(14) Yu, P.; Petzoldt, S.; Wilhelm, M.; Zolg, D. P.; Zheng, R.; Sun, X.; Liu, X.; Schneider, G.; Huhmer, A.; Kuster, B. *Anal Chem* **2017**, *89*, 8884-8891.
(15) Bekker-Jensen, D. B.; Kelstrup, C. D.; Batth, T. S.; Larsen, S. C.; Haldrup, C.; Bramsen, J. B.; Sorensen, K. D.; Hoyer, S.; Orntoft, T. F.; Andersen, C. L.; Nielsen, M. L.; Olsen, J. V. *Cell Syst* **2017**, *4*, 587-599 e584.
(16) Dwivedi, R. C.; Spicer, V.; Harder, M.; Antonovici, M.; Ens, W.; Standing, K. G.; Wilkins, J. A.; Krokhin, O. V. *Anal. Chem*. **2008**, *80*, 7036-7042.
(17) Boersema, P. J.; Divecha, N.; Heck, A. J.; Mohammed, S. *J Proteome Res* **2007**, *6*, 937-946.
(18) Zhou, F.; Lu, Y.; Ficarro, S. B.; Adelmant, G.; Jiang, W.; Luckey, C. J.; Marto, J. A. *Nat Commun* **2013**, *4*, 2171.
(19) Spicer, V.; Krokhin, O. V. *J Chromatogr A* **2018**, *1534*, 75-84.
(20) Gussakovsky, D.; Neustaeter, H.; Spicer, V.; Krokhin, O. V. *Anal Chem* **2017**, *89*, 11795-11802.
(21) Krokhin, O. V.; Ezzati, P.; Spicer, V. *Anal Chem* **2017**, *89*, 5526-5533.
(22) Krokhin, O. V.; Spicer, V. *Anal Chem* **2009**, *81*, 9522-9530.
(23) Klaassen, N.; Spicer, V.; Krokhin, O. V. *J Chromatogr A* **2019**, *1588*, 163-168.
(24) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466-1467.
(25) Moruz, L.; Pichler, P.; Stranzl, T.; Mechtler, K.; Kall, L. *Anal Chem* **2013**, *85*, 7777-7785.
(26) Shibue, M.; Mant, C. T.; Hodges, R. S. *J Chromatogr A* **2005**, *1080*, 68-75.
(27) Gevaert, K.; Van Damme, J.; Goethals, M.; Thomas, G. R.; Hoorelbeke, B.; Demol, H.; Martens, L.; Puype, M.; Staes, A.; Vandekerckhove, J. *Mol Cell Proteomics* **2002**, *1*, 896-903.
(28) Wang, Y.; Yang, F.; Gritsenko, M. A.; Wang, Y.; Clauss, T.; Liu, T.; Shen, Y.; Monroe, M. E.; Lopez-Ferrer, D.; Reno, T.; Moore, R. J.; Klemke, R. L.; Camp, D. G., 2nd; Smith, R. D. *Proteomics* **2011**, *11*, 2019-2026.
(29) Mertins, P.; Qiao, J. W.; Patel, J.; Udeshi, N. D.; Clauser, K. R.; Mani, D. R.; Burgess, M. W.; Gillette, M. A.; Jaffe, J. D.; Carr, S. A. *Nat Methods* **2013**, *10*, 634-637.

Chapter 3 Peptide retention time prediction in hydrophilic interaction liquid chromatography: zwitterionic ZIC-HILIC and ZIC-cHILIC stationary phases

This section contains text and figures from a paper published in *Journal of Chromatography A* in 2022, 1679, 463391. **Darien Yeung**, Nicole Klaassen, Benilde Mizero, Victor Spicer, and Oleg V. Krokhin.

DY performed data analysis, writing, editing of the manuscript. NK and BM performed the experiments, VS performed data analysis, editing of the manuscript. OVK conceived and supervised the project, performed data analysis, writing, and editing of the manuscript.

Abstract

Peptide retention time prediction models have been developed for zwitterionic ZIC-HILIC and ZIC-cHILIC stationary phases (pH 4.5 eluents) using proteomics-derived retention datasets of ~30 thousand tryptic peptides each. Overall, hydrophilicity of these stationary phases was found to be similar to the previously studied Amide HILIC phase, but lower compared to bare silicas. Peptide retention is driven by interactions of all charged (hydrophilic) residues at pH 4.5 (Asp, Glu, Arg, Lys, His), but shows specificity according to orientation of functional groups in zwitterionic pair. Thus, ZIC-cHILIC exhibits an increased contribution of negatively charged Asp and Glu due to the distal positioning of positively charged quaternary amines on the stationary phase. These findings confirm that HILIC interactions are driven by both peptide distribution between water layer adsorbed on the stationary phase and by interactions specific to functional groups of the packing material. Sequence-Specific Retention Calculator HILIC models were optimized for these columns showing 0.967-0.976 R^2-values between experimental and predicted retention values. ZIC-HILIC separations represent a good choice as a first dimension in 2D LC-MS of peptide mixtures with correlations between retention values of ZIC-HILIC against RPLC found at 0.197 (ZIC-HILIC) and 0.137 (ZIC-cHILIC) R^2-values, confirming a good orthogonality.

3.1 Introduction

Hydrophilic interaction liquid chromatography (HILIC) is one of the most dynamic fields in separation science due to unmatched diversity of available stationary phases. It was introduced to address the specific need of separating hydrophilic compounds – an option not available for reversed-phase separations. HILIC employs a polar stationary phase and less-polar mobile phases. This modality was first introduced in 1975[1] and in 1990 was named HILIC by Alpert.[2] HILIC exhibits characteristics of three different separation modes: reversed phase liquid chromatography (RPLC), normal phase liquid chromatography (NPLC) and ion-exchange[3] was applied to the separation of a wide range of compounds with polar characteristics, including peptides.[2,4] The latter gained significant attention in recent years due to developments in proteomics.[5,6]

Separation of very hydrophilic peptides,[7,8] selective glycopeptide enrichment[9] and utilization of unique HILIC selectivity in 2D separation schemes[6] remain the major applications of HILIC in bottom-up proteomics. Similar to RPLC,[10,11] to better understand separation in HILIC and develop better analysis protocols, separation scientists aimed at modeling peptide retention. In 1998, Yoshida[12] introduced the first HILIC model, which was based on additive contributions of retention coefficients for individual amino acids. Not surprisingly, these retention contributions showed anti-correlation with RPLC separation mode.[12] Wide application of proteomics/peptidomics prompted significant developments in this area in the past 8 years.[13-16] Most of the models used the same additive approach and 150 or less peptides as optimization datasets. Our laboratory was first to employ 2D LC-MS/MS technique to collect retention data for tens of thousands peptides and use it for model development. HILIC,[17,18] strong cation exchange (SCX),[19] and strong anion exchange (SAX, manuscript in preparation) have been used as first

dimensions of peptide separation and a standard RPLC with formic acid as an ion-pairing modifier, in the second. Sequence Specific Retention Calculator was adopted to the separation on XBridge Amide, Luna HILIC, and two bare silicas: Atlantis Silica and Luna HILIC Silica. Abundance of retention data allowed us to create the first HILIC model with sequence dependent features, influence of the residue position relative to peptide termini, and propensity to form helical structures.

Zwitterionic stationary phases represent a separate class of HILIC columns due to presence of both strongly acidic and basic groups in 1:1 ratio. Sulfoalkylbetaine and phosphorylcholine HILIC stationary phases were introduced by Irgum and co-workers.[20,21] ZIC-cHILIC carries densely bonded, 1:1 charge-balanced phosphorylcholine functional group, while ZIC-HILIC has 1:1 charge-balanced sulfoalkylbetaine functionality of opposite charge orientation. Hydrophilic character of the functional groups promotes strong adsorption of water in the presence of both negatively and positively charged functional groups which allowed simultaneous separation of cation and anions. Commercially available under the trademarks ZIC-HILIC and ZIC-cHILIC, these stationary phases represent extremely attractive alternatives for application in bottom-up proteomics. Mohammed and co-workers provided initial insights into selectivity of ZIC-HILIC and ZIC-cHILIC stationary phases using proteomics derived data.[22,23] First, the authors have assessed the potential of ZIC-HILIC-RP combination for application in 2D LC-MS analysis.[22] They found that separation is pH dependent in pH 3-8 range due to ionization of Asp/Glu and concluded on a mixed-mode character of interaction: polar and electrostatic. Orthogonality of 2D ZIC-HILIC-RP has been evaluated using standard mixture of tryptic peptides from bovine serum albumin, α and β-caseins. Di Palma et al.[23] expanded this approach into both ZIC- and ZIC-cHILIC phases using on-line and offline coupling of the column to mass spectrometer.

The goal of this study is to develop first sequence-dependent peptide retention prediction models derived using high-throughput data for zwitterionic ZIC-HILIC and ZIC-cHILIC columns and compare contributions of individual amino acids into retention on these phases against other HILIC sorbents. The unique feature of these zwitterionic phases is the presence of functional groups with opposing charges and alternative spatial orientations. Will this affect the separation selectivity of peptides and to what degree are the major questions we address in this study. Similar to SSRCalc modeling of all other non-RPLC techniques, we employ 2D LC-MS/MS, which represents a perfect platform to study any separation chemistry regardless of compatibility of eluent systems with MS detection.

3.2 Materials and methods

3.2.1. Materials, chromatographic columns, digests preparation

All chemicals were sourced from Sigma Chemicals (St. Louis, MO) unless noted otherwise. Two different HILIC stationary phases (2.1 mm x 100 mm) were used: SeQuant ZIC-HILIC 3.5 µm 100Å and ZIC-cHILIC 3.0 µm 100Å. Deionized water and HPLC-grade acetonitrile were used for preparation of the eluents. Sequencing grade modified trypsin (Promega, Madison, WI) was used for the digestion. Designed standard peptides P1-P6 and N2-N10[24,25] were synthesized by BioSynthesis Inc. (Lewisville, TX) and used for data alignment in the second dimension and testing separation space in the first dimension, respectively.

Tryptic digests of *S. cerevisiae* and bovine serum albumin (BSA) were prepared as described elsewhere.[17] In both cases Cys residues were reduced and alkylated with iodoacetamide. Mixtures of tryptic peptide mixtures were purified using reversed-phase HPLC and lyophilized.

3.2.2. First dimension separation conditions

Agilent 1100 series HPLC system with UV detector (214 nm), manual injector with 100 uL loop and 300 µL/min flow rate was used for separations. Both eluents: A (water) and B (9:1 acetonitrile:water) contained 10 mM ammonium formate (pH 4.5). Gradients starting from 10% water (100% eluent B) were used. Initial testing using standard N2-N10 peptides was done using 1% per minute water increase gradient. BSA digest was then used for optimization of separation conditions to provide separation window of ~40 min (fractions). 0.7% per minute water increase was chosen – similar to separation conditions on XBridge Amide column described previously.[17] One minute fractions were collected for the separation of the *S. cerevisiae* digest (~100 µg) within expected interval of peptide elution, lyophilized and subjected to the second dimension LC-MS/MS analysis. Lyophilized fractions were re-dissolved in 30 µL of buffer A for the second dimension and spiked with standard peptides P1-P6[24] for the retention time alignment purposes.

3.2.3 Second dimension LC-MS/MS

LC-MS/MS in the second dimension has been done using 2D LC Ultra system (Eksigent, Dublin, CA) and a TripleTOF5600 mass spectrometer (Sciex, Concord, ON) as described elsewhere.[19] 100µm x 200mm analytical column packed with 3 µm Luna C18(2) (Phenomenex, Torrance, CA), was used with 500 nL/min flow rate and 0.4% acetonitrile per minute gradient. Both eluents A (water) and B (acetonitrile) contained 0.1% formic acid. 300 µm x 5 mm PepMap 100 trap-column (ThermoFisher) was used for sample loading with 15 µL per minute flow rate. 10 µL (1/3 of each fraction) was injected. The linear gradient program included: a linear increase from 0.4 to 31%

buffer B in 77 minutes, 5 minutes column wash at 80% B and then 8 minutes at 0.4% B for column equilibration.

3.2.4. Data analysis

X!Tandem search algorithm[26] was applied with following settings: 20 ppm and 50 ppm mass tolerance for parent and daughter ions, respectively; constant modification of Cys with iodoacetamide; standard set of expected PTMs (Met/Trp oxidation, Asn/Gln deamidation, Cys/Gln N-terminal cyclization); one allowed missed cleavage. Identified peptides were assigned with retention values expressed in HI units (% acetonitrile) using tabulated HI values for standard P1-P6 peptides [24]. All identified tryptic non-modified peptides with low identification score ($-1 > \log(E) > -3$) were additionally filtered using retention time prediction in the second dimension using formic acid SSRCalc model and retention database values for *S. Cerevisiae* tryptic peptides.

Retention times in the first dimension were assigned as equal to the fraction number in which this peptide was found, or to the weighted average of the intensity distribution if peptide was found in 2 or more fractions.

3.3 Results and Discussion

3.3.1. Determining hydrophilicity of ZIC stationary phase, selection of gradient slope suitable for 2D LC-MS/MS proteomic analysis

To assess the hydrophilicity of both ZIC columns the mixture of five standard peptides N2-N10[25] was separated over a 1%/min gradient. These sequences were designed to have nearly identical molecular weight and propagating hydrophilicity, which was achieved by substitution of Arg for

Gly-Val sequence. Resulting sequences have decreasing hydrophilicity in the order GVGVRRRDDDPFR (N2), GVGVGVRRDDDPFR (N4), GVGVGVGVRDDDPFR (N6), GVGVGVGVGVDDDPFR (N8), GVGVGVGVGVDDDPFGV (N10) and carry +1, 0, -1, -2, and -3 charge at pH 4.5, respectively. The resulting chromatograms for both columns had the same retention order (Figure 3.1 A, B); however, the N10 - N2 retention window was different between the two zwitterionic stationary phases. ZIC-HILIC carries a negatively charged group on the distal interface,[21,23] where the peptide-stationary phase interaction occurs. Cationic peptides have stronger electrostatic attraction to the ZIC-HILIC stationary phase leading to higher retention times compared to ZIC-cHILIC, as demonstrated by peptide N2. The opposite is true for negatively charged peptides N6, N8, N10, which have a lower retention times compared to ZIC-cHILIC. Anionic peptides engage in electrostatic repulsion with the ZIC-HILIC stationary phase thus exhibiting lower retention times compared to opposite arrangement of charged groups on ZIC-cHILIC.

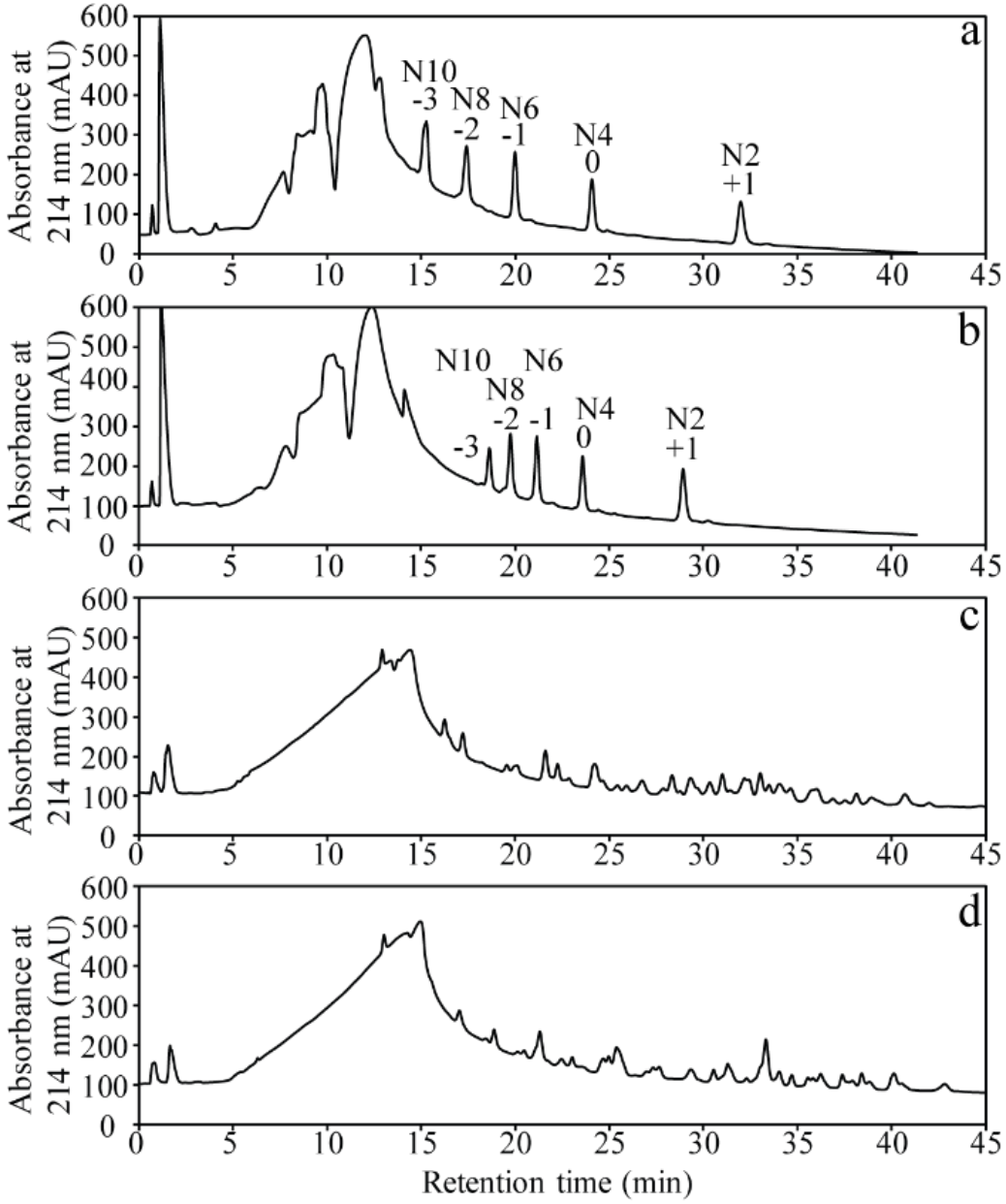


Figure 3.1: Separation of standard peptide mixture (a, b) and tryptic digest of bovine serum albumin (c, d) using ZIC-HILIC (a, c) and ZIC-cHILIC (b, d). (a, b) used 1% and (c, d) 0.7% increase of water per minute gradients.

It is important to understand that the separation mechanism is not solely dependent on the distal charged groups. The proximity between the charged distal and proximal ends of the zwitterion can exhibit weaker electrostatic interactions relative to ionic exchange columns.[23] Based

on the chromatogram of both ZIC-HILIC and ZIC-cHILIC, we can understand that electrostatic interaction has a role in the separation of the peptides. However, since the retention order is not reversed and neutral peptide N4 has identical retention times on both columns, this indicates that the hydrophilic tether in the zwitterionic stationary phase has a dominating influence on the separation mechanism. It is clear that retention of charged analytes on zwitterionic HILIC phases is affected by electrostatic attraction/repulsion.

Comparison of different HILIC columns[25] showed that XBridge Amide and both ZIC columns exhibit virtually identical retention for neutral N4 peptide and very similar average retention values for peptides in the standard mixture. This suggests similar hydrophilicity for these phases. Therefore, we selected to use 0.7%/min gradient to ensure wide enough distribution of peptides throughout the elution range of typical tryptic digests such as bovine serum albumin (Figure 3.1 C,D) and whole cell lysate of *S. Cerevisiae*. One-minute fractions were collected in 0-55 min interval and submitted for LC-MS/MS analysis in the second dimension.

3.3.2 Identification output for two 2D LC-MS/MS analyses and peptide retention prediction filtering

The 2D runs for both ZIC/ZICc-HILIC led to peptide identification in 41 out of 55 fraction collected. Each fraction was analyzed using identical dilution (~1µg or less of peptides). The identification rate for all the acquired MS/MS was ~65% and each unique peptide was identified by approximately on average 8.5 MS/MS spectra. Table 3.1 illustrates the acquisition statistics for the 2D LC-MS/MS analysis. The higher identification output for ZIC-cHILIC may serve as an

indication of more favorable separation orthogonality between HILIC and RP dimensions as discussed further in section 3.4.

Table 3.1: Acquisition statistics of 2D (ZIC/ZICc-HILIC-RP) LC-MS/MS analysis for zwitterionic HILIC separation of whole cell yeast tryptic digest.

HILIC separation mode	Number of fractions	Total LC-MS time (hr)	Amount injected (µg)	# of MS/MS	# of identified peptides	# of non-redundant peptide IDs	# of protein IDs
ZIC-HILIC	41	61.5	~40	311879	206256	38037	4235
ZIC-cHILIC	41	61.5	~40	352314	225762	39791	4324

Retention values for 34878 and 36728 unique non-modified peptides have been measured for ZIC-HILIC and ZIC-cHILIC, respectively. The average length of the peptides was ~14 residues with ~20000 overlapping peptides found in both runs. Average net charge at pH of the eluents used was negative 0.65 spanning a range of +4 to -14. Datasets were further refined by removing low confidence peptides of -1 > Log(e) > -3 with large retention prediction errors (±2% acetonitrile) in the second separation dimension. Careful inspection of the dataset, supported by existing database of retention values of *S. Cerevisiae* helped to maintain a representative distribution of length of peptides. This is important as short peptides usually suffer from low confidence scores, and can be mistakenly removed in the refinement process. Only 0.5-0.7% of peptides were removed, consistent with the False Positive Rate of 0.4-0.5% for these runs estimated by X!tandem. The remaining 33755 (ZIC-HILIC) and 35392 (ZIC-cHILIC) peptides were used for training the retention time prediction model. As expected, the vast majority of sequences were Arg or Lys

terminated. However, both datasets contained sufficient number of protein C-terminal sequences (454 and 430, respectively) to ensure confident assignment of C-terminal retention coefficients for the residues other than arginine or lysine.

3.3.3 ZIC/ZICc-HILIC models' optimization

Predictive models for ZIC/ZICc-HILIC have been developed using an approach described in our previous publications.[17,18] It starts with additive model (20 parameters, one for each amino acid) with peptide length correction, followed by introduction of position-dependent retention coefficients (*Rc*, five on each termini plus internal *Rc* - 220 parameters). Final modifications included taking into account possible effects of peptide helicity, clustering of hydrophobic residues, peptide charge and conversion of the model output into gradient % units - % water in case of HILIC. The algorithms require a large data density to prevent possible overfitting of the model. Typically, the additive models for random collection of peptides are built using 5-6 data points per variable.[27,28] However, SSRCalc models based on high-throughput proteomic data consistently exceed ~100:1 ratio.[17,18]

Both ZIC models used a dataset of 33000+ tryptic peptides and resulted in excellent R^2-value correlations between predicted and experimental values: 0.967 for ZIC-HILIC and 0.976 for ZIC-cHILIC (Figure 3.2 A, B). When the output of the model is converted into water % units, the slope of the plot retention time (fraction number) vs. SSRCalc hydrophilicity is reciprocal to gradient slope used (1/1.42 ~0.7). The same feature helps in determining experimental conditions for the separation of individual peptides or complex mixtures in proteomics. For example, Figure 3.2 C shows frequency distribution of tryptic peptides from *S. Cerevisiae* for different HILIC

phases. It is interesting to note that bare silicas exhibit highest retentivity, likely due to additional retention of cationic peptides via interaction with charged silanol groups.

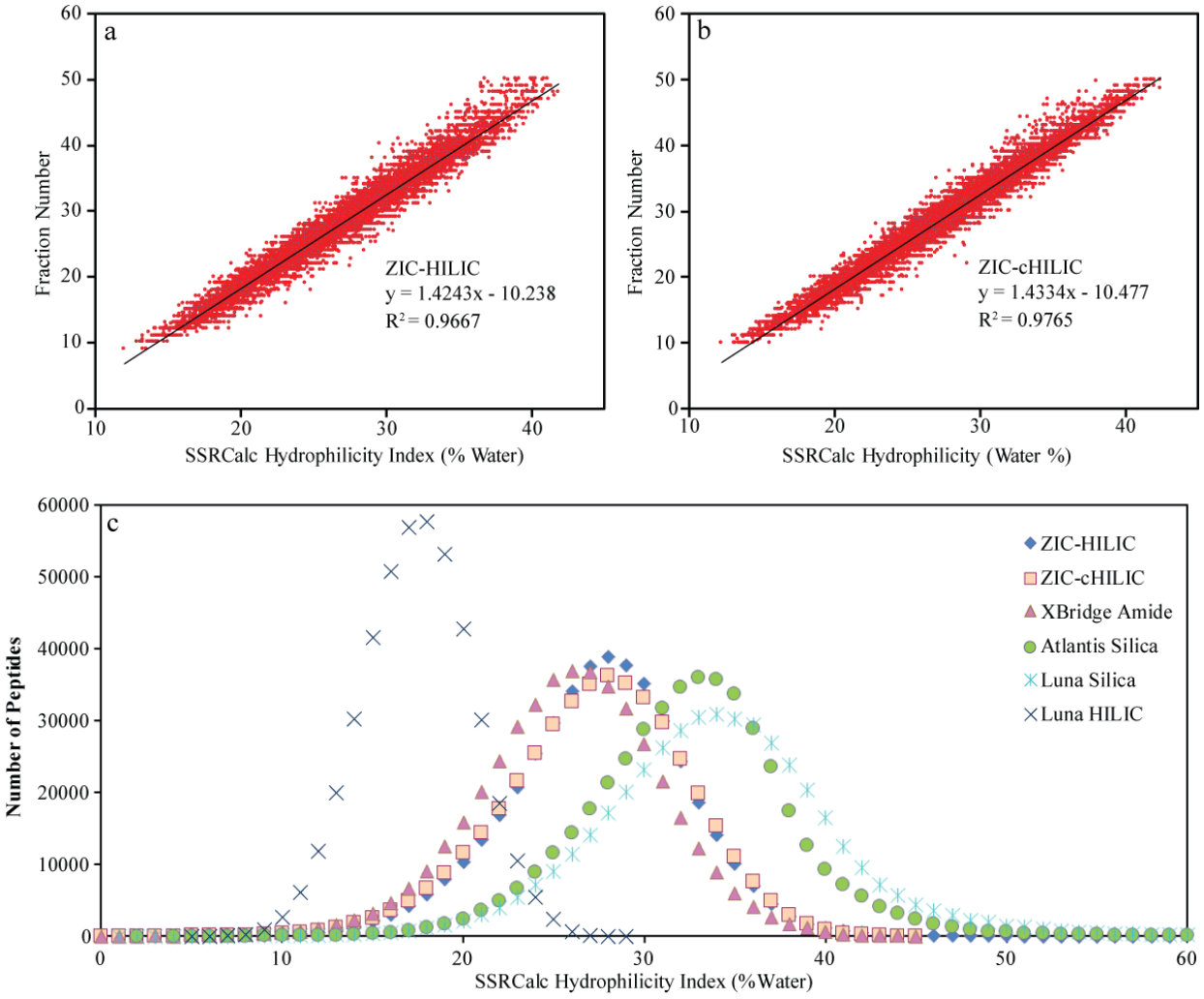


Figure 3.2: SSRCalc peptide retention time prediction for (A) ZIC-HILIC or (B) ZIC-cHILIC. (C) Predicted distribution of tryptic peptides from *S. Cerevisiae* across separation space for 6 different HILIC phases (XBridge Amide, Atlantis Silica, Luna Silica, and Luna HILIC); data from previous publications[17,18] and current work

We also found that peptide retention on different HILIC stationary phases varies significantly, making developing individual retention prediction models necessary. Thus,

application of ZIC-cHILIC model provided R^2-value correlations of 0.94, 0.96, 0.68, 0.63 and 0.87 when applied to Luna HILIC, XBridge amide, Atlantis Silica, Luna Silica[18] and ZIC-HILIC data, respectively.

3.3.3.1 Composition-specific features of ZIC separation: retention coefficients

Each amino acid in a peptide contributes to the retention on the column and the individual contributions can be quantified as retention coefficients.[27-29] In the position dependent models such as SSRCalc, the internal *Rc* represents the most valuable information on retention mechanism and usually used for comparison. Figure 3.3 A shows normalized retention coefficients for ZIC-HILIC and ZIC-cHILIC in comparison to neutral XBridge amide and bare silica Atlantis Silica columns. As with the other HILIC columns, the retention on ZIC-HILIC phases is driven by charged hydrophilic residues. K, R, H, D, and E as shown by their high positive retention coefficients. The ZIC-HILIC coefficients for the basic amino acids are also higher than the ZIC-cHILIC ones as the cationic functional groups of the amino acids have electrostatic attraction with ZIC-HILIC stationary phase increasing retention and electrostatic repulsion with ZIC-cHILIC stationary phase decreasing retention. The opposite situation is observed for negatively charged Asp and Glu, their retention contributions are higher on ZIC-cHILIC. These findings are in complete agreement with separation trends observed for the standard peptide mixture (Figure 3.1 A, B). It was interesting to compare the variation of retention coefficients relative to neutral HILIC phase and bare silica with negatively charged silanol groups. ZIC-HILIC with distal positioning of negatively charged sulfonate[21,23] shows behavior seemingly intermediary between neutral amide phase (XBridge Amide) and silica (Atlantis Silica) as seen in Figure 3.3. It is observed to have higher retention values for Arg, Lys, His and lower retention for Asp and Glu compared to Amide phases.

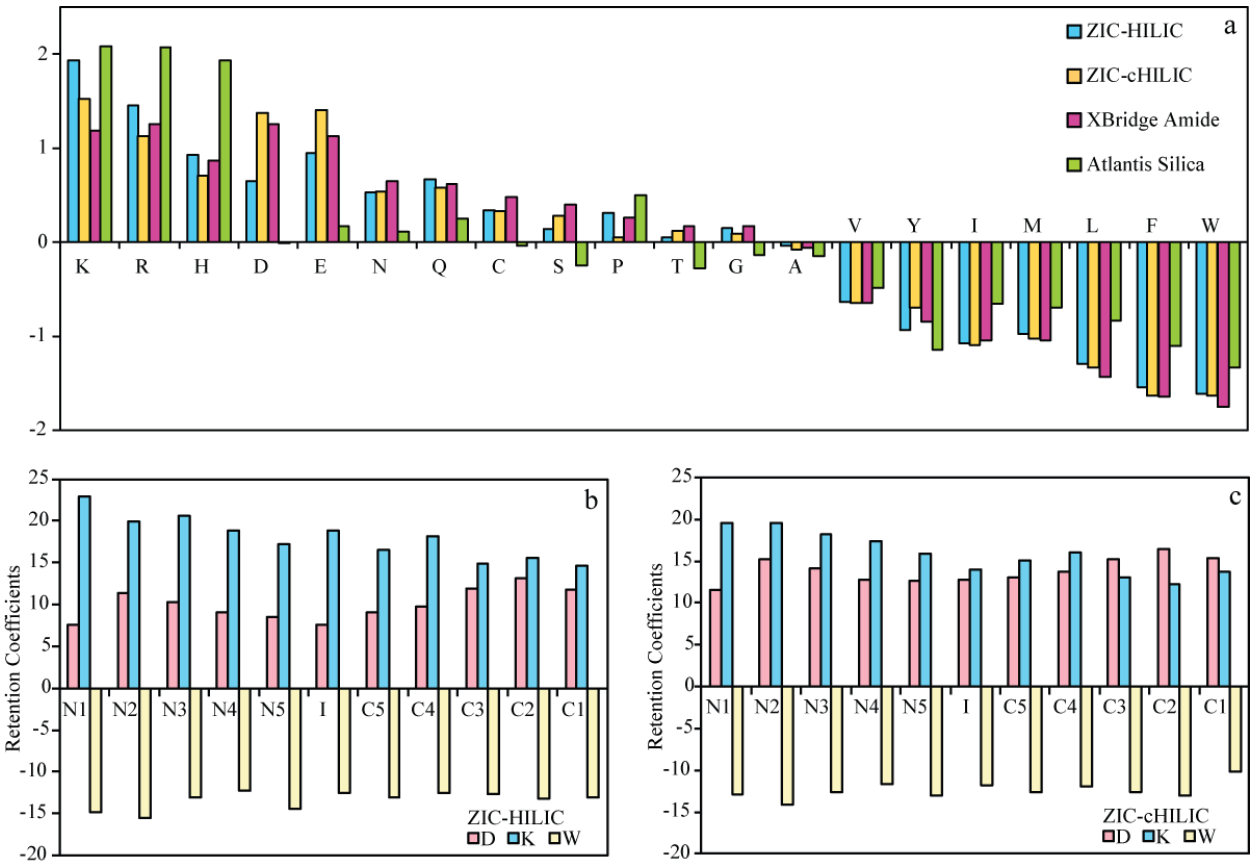


Figure 3.3: Individual amino acid contribution (normalized retention coefficients, Rc) to retention time for ZIC and ZICc-HILIC compared to (A) XBridge Amide and Atlantis Silica columns. Positional dependence of positively (Lys), negatively (Asp) and hydrophobic (Trp) residues' retention contributions for (B) ZIC-HILIC and (C) ZIC-cHILIC.

3.3.3.2. Optimization of position dependent (relative to peptide N- and C-termini) retention coefficients.

Our previous work[10,17,18] has demonstrated the utility of expanding retention coefficients to include position-dependent effects from the terminal ends of the N and C termini. In the expansion of the algorithm for position-dependent contributions, amino acids are categorized into 11 positions. The

positions are, starting from the N-terminus (N1), N2, N3, N4, N5, internal, C5, C4, C3, C2, and C-terminus C1. Any amino acids located more than five positions inside of the peptide sequence are assigned as internal. Position-dependent retention coefficients were introduced to account for unique properties of peptide termini. Termini effects are observed due to specific peptide orientation[30] and ion-pairing interactions at N-terminal amino and C-terminal carboxy groups.[31] Figure 3.3 B, C shows comparison of position dependent coefficients for some positively/negatively charged and hydrophobic residues. Variation of the *Rc* indicates significant effect of this residue position, which improves overall model accuracy upon introduction of separate retention values. Basic and acidic resides are affected the most. In the instance of K, R and H, the retention contribution decreases systematically with the distance from N-termini. The operating pH in this separation is 4.5, a region where basic amino acids side chains and peptide N-termini are protonated. In cation-exchange separations[19,30] and HILIC on bare silicas,[18] this leads to the N-terminal being a primary point of contact, thus exerting an "orientation effect". Positively charged side chains of Lys, Arg, and His will have stronger interactions when located at N-terminus. Surprisingly, this was the case for both ZIC phases (Figure 3.3 B, C). While this is expected for ZIC-HILIC with the distal cation exchange group, ZIC-cHILIC showed similar behavior, albeit in a lesser degree. Presence of residual silanol groups may result in such effect and represent the most plausible explanation of identical peptide orientation effect on both columns.

Contribution of negatively charged hydrophilic Asp is larger at both termini compared to internal positions due to better accessibility, with exception of N1. Being located at the peptide N-termini Asp affects basicity of terminal amino group and reduces its hydrophilic contribution.[32] The same effect was observed for all previously studied HILIC phases. Contributions of

hydrophobic residues remain basically unchanged depending on the position as shown for Trp (Figure 3.3 B, C).

3.3.3.3. Additional model corrections: peptide helicity, clustering of hydrophobic amino acids, peptide charge.

The effect of peptide helicity in HILIC is less profound compared to reversed-phase separations. While amphipathic helicity should result in increased HILIC retention due to preferential interaction of the hydrophilic face of the helix, the blocking of hydrogen-bonds between the backbone amino and carbonyl groups in the helix leads to decreased retention. These two effects largely compensate each other. Therefore, our approach for helical correction in HILIC models are based on simple additive approaches. We optimized weighted variables for select sequence features of: clusters of hydrophobic amino acids, N-cap sequences (NP, GP, SP, TP, DP),[33] and multiple Gly/Pro in the sequence (all features increase HILIC retention).

Final inspection of prediction errors showed significant positive deviations from predicted values for peptides carrying +2 to +4 charges under the separation conditions used, suggesting possible involvement of cation-exchange interactions. The number of these basic peptides in our dataset was relatively low (~5%) as most of the tryptic fragments carry two positively charged groups and likely to have one or more acidic Asp or Glu. The N-terminus and C-terminal side chain of Arg or Lys is being offset by the negative charge of N-terminal amino and Asp/Glu side chains. This effect of increased retention of basic peptides was observed for both ZIC-HILIC and to a lesser degree, ZIC-cHILIC. Similar to peptide orientation effects observed for position dependent retention coefficients, this finding contradicts the spatial arrangement of functional groups on ZIC-cHILIC [21, 23] and can be caused by the presence of remaining unreacted silanol

groups. Meanwhile, peptides carrying multiple negatively charged groups (up to -15 net charge) did not exhibit significant deviations.

To address this increased retention for +2 to +4 peptides, the final correction ($\boldsymbol{\Delta}$ expressed in water % units) was introduced: $\boldsymbol{\Delta} = \frac{A*z}{ln(N)}$, where z and N is peptide charge and length, respectively. Scaling coefficients A was optimized as follows: A = 2 for +3/+4 peptides and 1 for +2 peptides for ZIC-HILIC. As for ZIC-cHILIC, A = 1.5 for +3/+4 peptides and 0.75 for +2 peptides. We can see that the scaling for A is greater for ZIC-HILIC and this confirms that the charge effect is stronger for ZIC-HILIC.

3.3.4 Separation orthogonality between ZIC/ZICc and RP separation modes

Plotting retention values correlation between two separation modes is a standard procedure to assess the 2D separation orthogonality. Compared to previous reports on HILIC-RP systems orthogonality,[22,34] we had access to significantly larger datasets acquired for both ZIC columns, which provide deeper insight into features determining orthogonality in these systems. HILIC and reversed-phase separations are driven by interactions of opposite character: hydrophilic vs. hydrophobic. These separation modes should show anti-correlation: higher HILIC retention should result in lower retention in RP. However, differences in chemistry of interacting phases, pH and ion-pairing environment results in much wider distributions, such orthogonal plots are teardrop-shaped (Figure 3.4 A, B) with R^2 correlations of 0.197 and 0.137 for ZIC-HILIC and ZIC-cHILIC, respectively, confirming a good orthogonality for both modes but favoring slightly the latter. Similar to Gilar et al.,[34] we measured separation orthogonality using a geometric approach based on calculating filling percentage of a 100x100 matrix with 10,000 randomly selected data points.

Resulting values also slightly favour ZIC-cHILIC (53.5% orthogonal) as opposed to ZIC-HILIC (49.7% orthogonal).

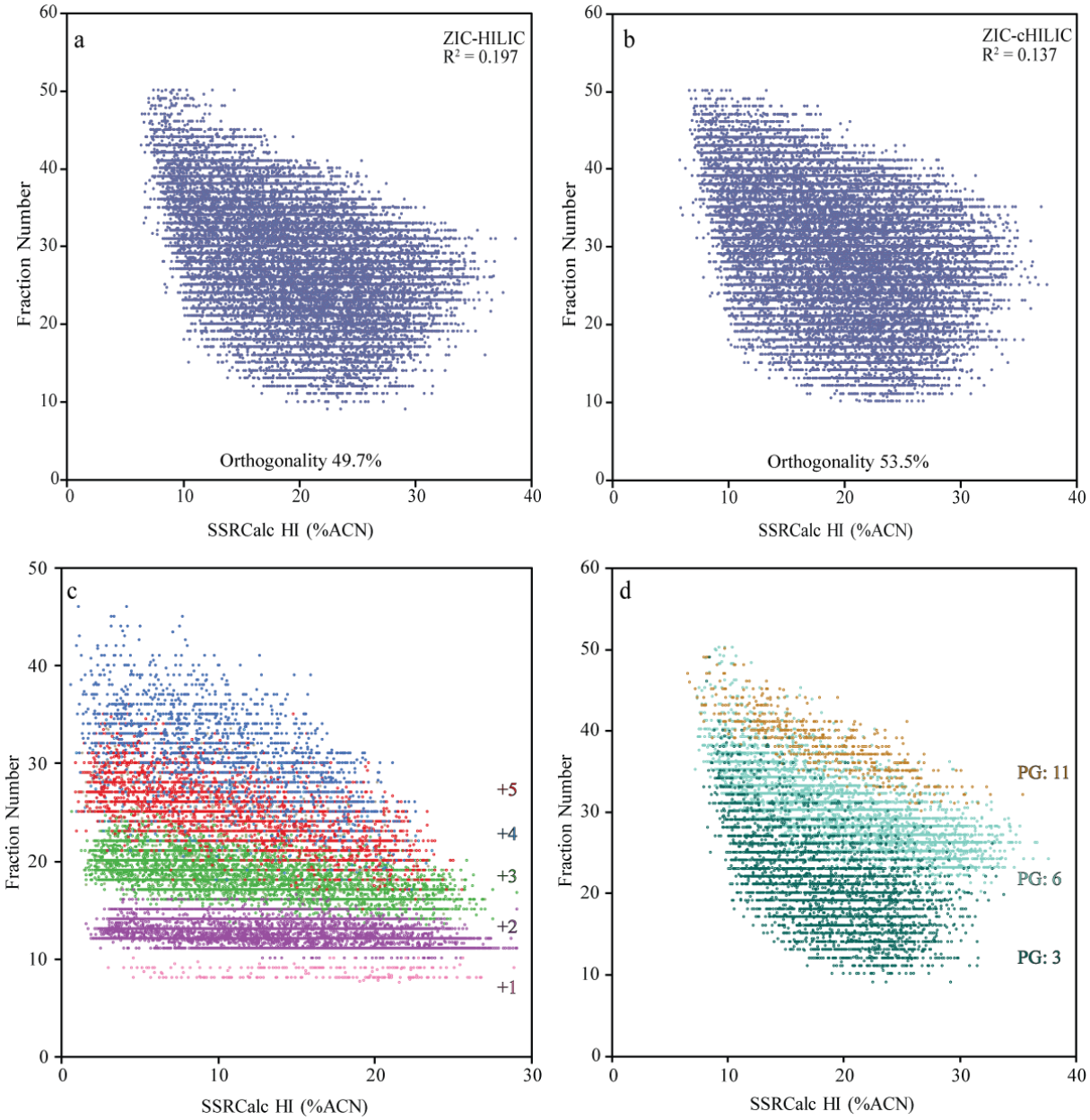


Figure 3.4: Orthogonality plots for ZIC/ZIC-cHILIC – RP and its comparison to SCX-RP. A, B – retention values correlations for ZIC-HILIC – RP and ZIC-cHILIC – RP, respectively. The effects of peptide charge in (C) SCX-RP and (D) number of charged residues in ZIC-HILIC (pH 4.5) – RP.

Boersema et al.[22] suggested that separation orthogonality in ZIC-HILIC (pH 3) – RP is similar to SCX-RP mode due to the fact that retention is determined by residues charged at pH 3: Arg, Lys, His. Figure 3.4 C shows the orthogonality plot for SCX-RP system (data taken from previous publication[19]), which consists of a series of linear bands for peptides with different net charge. Anti-correlation (negative slope) is observed due to lower electrostatic interaction of large molecules, which tend to be more hydrophobic. Similar plot for ZIC-HILIC (pH 4.5) – RP shows less definitive trends with wider distribution of peptides of the same charge, considering protonated Arg, Lys, His, N-termini, and fully ionized side chains of Asp, Glu and C-terminal carboxyl group. Indeed, peptides with -1 net charge at pH 4.5 could carry a different number of positive and negative charges: e.g. +4 and -5 or +2 and -3. The total number of hydrophilic groups strongly contributing in HILIC retention would be different in these two cases. Therefore, the total number of charged residues should be used to assess retention trends in HILIC – RP separation system. Figure 3.4 D shows orthogonality plot for peptides with 3, 6, and 11 polar groups (PG), which fit to 3 trend lines showing higher RP retention for peptides with lower HILIC retention. The negative slope of the trends in this case is caused by opposing retention mechanism (hydrophilic vs. hydrophobic) rather than the influence of peptide size in SCX-RP system.

3.4. Conclusions

Using high-throughput retention data collection based on proteomic methodology, we were able for the first time to acquire extended retention datasets and train high quality retention time prediction models to compare the separation mechanism on zwitterionic HILIC phases. One of the key points is maintaining the quality of the data through careful selection of chromatographic conditions and refining the dataset to capture peptides identified with high confidence based on

both mass-spectrometry analysis and their chromatographic properties. Using high quality data in the standard framework for Sequence-Specific Retention Calculator model optimization allowed for rapid *in-silico* analysis of separation orthogonality between the zwitterionic HILIC and reversed-phase separations in 2D LC-MS/MS, in-depth research into separation mechanisms, and comparing of the difference between the separation modality based on the difference in charge surface of the zwitterionic stationary phase. The results illustrate that the two columns exhibit similar separation mechanism based on the interaction of the most hydrophilic (charged) residues at pH 4.5: Arg, Lys, His, Asp and Glu. The difference arise from alternative orientation of charged groups between ZIC-HILIC and ZIC-cHILIC, which explains additional ion-exchange features of the separation mechanism. ZIC-HILIC carries distal negatively charged groups and favor retention of basic residues. ZIC-cHILIC has positively charged quaternary ammonium group closer to the surface and increases retention contributions of negatively charged Asp and Glu. Apart from that difference, both phases have hydrophilicity similar to neutral XBridge Amide column. Generally, all peptides neutral at pH 4.5 have almost identical retention times on these three columns. Position dependent effects have been established for both ZIC columns and showed similar increase of retention contributions for positively charged residues closer to N-terminus of a peptide. While this behavior is expected for ZIC-HILIC with distal cation-exchange functionality, observing it for ZIC-cHILIC was somewhat surprising. This feature could be a consequence of residual silanol groups on the surface of the sorbent, which was supported by finding predominantly positive prediction errors for cationic peptides following optimization of an intermediate version of the model based solely on position-dependent retention coefficients.

Both ZIC phases represent viable candidates for the selection of first separation dimension in 2D LC-MS/MS. R^2-values for correlations between HILIC and RP dimensions slightly favor

orthogonality for ZIC-cHILIC: 0.197 vs. 0.137. The driving force behind separation orthogonality is somewhat similar to SCX-RP system with one important difference: SCX retention at acidic pH is determined by the number of positively charged residues, while retention in HILIC – by the number of all ionized at pH 4.5 residues. Through retention time modeling, it is now possible to quantitatively ascertain the similarities and differences between the peptides and stationary phase interactions from a physical and chromatographic chemistry standpoint. Having the insight from variety of HILIC columns would in turn help the understanding other chromatographic modalities in our pursuit to cover all possible peptide separation modes RP, HILIC, SCX, SAX, and mixed-mode separations.

3.5. References

(1) Linden, J.C.; Lawhead, C.L. Liquid chromatography of saccharides. *J. Chromatogr.*, **1975**, 105, 125-133.
(2) Alpert, A.J. Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. *J. Chromatogr.,* **1990**, 499, 177-196.
(3) Buszewski, B.; Noga, S. Hydrophilic interaction liquid chromatography (HILIC)--a powerful separation technique. *Anal. Bioanal. Chem.*, **2012**, 402, 231-247.
(4) Oyler, A.R.; Armstrong, B.L.; Cha, J.Y.; Zhou, M.X.; Yang, Q.; Robinson, R.I.; Dunphy, R.; Burinsky, D.J. Hydrophilic interaction chromatography on amino-silica phases complements reversed-phase high-performance liquid chromatography and capillary electrophoresis for peptide analysis. *J. Chromatogr. A.*, **1996**, 724, 378–383.
(5) Boersema, P.J.; Mohammed, S.; Heck, A.J. Hydrophilic interaction liquid chromatography (HILIC) in proteomics. *Anal. Bioanal. Chem.*, **2008**, 391, 151-159.
(6) Di Palma, S.; Hennrich, M.L.; Heck, A.J.; Mohammed, S. Recent advances in peptide separation by multidimensional liquid chromatography for proteome analysis., *J. Proteomics*, **2012**, 75, 3791-3813.
(7) Yang, Y.; Boysen, R.I.; Hearn, M.T. Hydrophilic interaction chromatography coupled to electrospray mass spectrometry for the separation of peptides and protein digests. *J. Chromatogr. A,* **2009**, 1216, 5518-5524.
(8) Simon, R.; Enjalbert, Q.; Biarc, J.; Lemoine, J.; Salvador, A. Evaluation of hydrophilic interaction chromatography (HILIC) versus C(18) reversed-phase chromatography for targeted quantification of peptides by mass spectrometry. *J. Chromatogr. A*, **2012**, 1264, 31-39.
(9) Hagglund, P.; Bunkenborg, J.; Elortza, F.; Jensen, O.N.; Roepstorff, P. A new strategy for identification of N-glycosylated proteins and unambiguous assignment of their glycosylation sites using HILIC enrichment and partial deglycosylation. *J. Proteome Res.*, **2004**, 3, 556-566.

(10) Krokhin, O.V. Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-A pore size C18 sorbents. *Anal. Chem.,* **2006**, 78, 7785-7795.

(11) Petritis, K.; Kangas, L.J.; Yan, B.; Monroe, M.E.; Strittmatter, E.F.; Qian, W.J.; Adkins, J.N.; Moore, R.J.; Xu, Y.; Lipton, M.S.; Camp, D.G.; Smith, R.D. Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Anal. Chem.*, **2006**, 78, 5026-5039.
(12) Yoshida, T.Y. Calculation of peptide retention coefficients in normal-phase liquid chromatography. *J. Chromatogr. A*, **1998**, 808, 105-112.
(13) Gilar, M.; Jaworski, A. Retention behavior of peptides in hydrophilic-interaction chromatography. *J. Chromatogr. A*, **2011**, 1218, 8890-8896.

(14) Harscoat-Schiavo, C.; Nioi, C.; Ronat-Heit, E.; Paris, C.; Vanderesse, R.; Fournier, F.; Marc, I. Hydrophilic properties as a new contribution for computer-aided identification of short peptides in complex mixtures. *Anal. Bioanal. Chem.*, **2012**, 403, 1939-1949.
(15) Le Maux, S.; Nongonierma, A.B.; FitzGerald, R.J. Improved short peptide identification using HILIC-MS/MS: retention time prediction model based on the impact of amino acid position in the peptide sequence. *Food Chem*., 2015, 173, 847-854.
(16) Badgett, M.J.; Boyes, B.; Orlando, R. Prediction of Peptide Retention Times in Hydrophilic Interaction Liquid Chromatography (HILIC) Based on Amino Acid Composition. In Chromatography Today; International Labmate Limited: St. Albans, Hertfordshire, 2015; (Nov-Dec), pp 39−42.
(17) Krokhin, O.V.; Ezzati, P.; Spicer, V. Peptide Retention Time Prediction in Hydrophilic Interaction Liquid Chromatography: Data Collection Methods and Features of Additive and Sequence-Specific Models. *Anal. Chem.*, **2017**, 89, 5526-5533.
(18) Spicer, V.; Krokhin, O.V. Peptide retention time prediction in hydrophilic interaction liquid chromatography. Comparison of separation selectivity between bare silica and bonded stationary phases. *J. Chromatogr. A*, **2018**, 1534, 75-84.
(19) Gussakovsky, D.; Neustaeter, H.; Spicer, V.; Krokhin, O.V. Sequence-Specific Model for Peptide Retention Time Prediction in Strong Cation Exchange Chromatography. *Anal. Chem.*, 2017, 89, 11795-11802.
(20) Jiang, W.; Irgum, K. Tentacle-type zwitterionic stationary phase prepared by surface-initiated graft polymerization of 3-[N,N-dimethyl-N-(methacryloyloxyethyl)-ammonium] propanesulfonate through peroxide groups tethered on porous silica. *Anal. Chem.*, 2002, 74, 4682-4687.
(21) Jiang, W.; Fischer, G.; Girmay, Y.; Irgum, K. Zwitterionic stationary phase with covalently bonded phosphorylcholine type polymer grafts and its applicability to separation of peptides in the hydrophilic interaction liquid chromatography mode, *J. Chromatogr. A*, **2006**, 1127, 82-91.
(22) Boersema, P.J.; Divecha, N.; Heck, A.J.; Mohammed, S. Evaluation and optimization of ZIC-HILIC-RP as an alternative MudPIT strategy. *J. Proteome Res.*, **2007**, 6, 937-946.
(23) Di Palma, S.; Boersema, P.J.; Heck, A.J.; Mohammed, S. Zwitterionic hydrophilic interaction liquid chromatography (ZIC-HILIC and ZIC-cHILIC) provide high resolution separation and increase sensitivity in proteome analysis. *Anal. Chem*., **2011**, 83, 3440-3447.
(24) Krokhin, O.V.; Spicer, V. Peptide retention standards and hydrophobicity indexes in reversed-phase high-performance liquid chromatography of peptides. *Anal. Chem*., **2009**, 81, 9522-9530.

(25) Klaassen, N.; Spicer, V.; Krokhin, O.V. Universal retention standard for peptide separations using various modes of high-performance liquid chromatography. *J. Chromatogr. A*, **2019**, 1588, 163-168.
(26) Craig, R.; Beavis, R.C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **2004**, 20, 1466-1467.
(27) Gilar, M.; Xie, H.; Jaworski, A. Utility of retention prediction model for investigation of peptide separation selectivity in reversed-phase liquid chromatography: impact of concentration of trifluoroacetic acid, column temperature, gradient slope and type of stationary phase. *Anal. Chem.*, **2010**, 82, 265-275.
(28) Shamshurin, D.; Spicer, V.; Krokhin, O.V. Defining intrinsic hydrophobicity of amino acids' side chains in random coil conformation. Reversed-phase liquid chromatography of designed synthetic peptides vs. random peptide data sets. *J. Chromatogr. A*, **2011**, 1218, 6348-6355.

(29) Guo, D.; Mant, C.T.; Taneja, A.K.; Parker, J.M.R.; Hodges, R.S. Prediction of peptide retention times in reversed-phase high-performance liquid chromatography I. Determination of retention coefficients of amino acid residues of model synthetic peptides. *J. Chromatogr*. **1986**, 359, 499.
(30) Alpert, A.J.; Petritis, K.; Kangas, L.; Smith, R.D.; Mechtler, K.; Mitulovic, G.; Mohammed, S.; Heck, A.J. Peptide orientation affects selectivity in ion-exchange chromatography, *Anal. Chem.*, **2010**, 82, 5253-5259.
(31) Krokhin, O.V.; Craig, R.; Spicer, V.; Ens, W.; Standing, K.G.; Beavis, R.C.; Wilkins, J.A. An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: its application to protein peptide mapping by off-line HPLC-MALDI MS. *Mol. Cell. Proteomics*, **2004**, 3, 908-919.
(32) Krokhin, O.V.; Anderson, G.; Spicer, V.; Sun, L.; Dovichi, N.J. Predicting Electrophoretic Mobility of Tryptic Peptides for High-Throughput CZE-MS Analysis. *Anal. Chem.*, **2017**, 89, 2000-2008.
(33) Spicer, V.; Lao, Y.W.; Shamshurin, D.; Ezzati, P.; Wilkins, J.A.; Krokhin, O.V. N-capping motifs promote interaction of amphipathic helical peptides with hydrophobic surfaces and drastically alter hydrophobicity values of individual amino acids. *Anal. Chem.*, **2014**, 86, 11498-11502.
(34) Gilar, M.; Olivova, P.; Daly, A.E.; Gebler, J.C. Orthogonality of separation in two-dimensional liquid chromatography. *Anal. Chem.*, **2005**, 77, 6426-6434.

Chapter 4 Peptide retention time prediction for hydrophilic interaction liquid chromatography at acidic pH in formic-acid based eluents

This section contains text and figures from a paper submitted to *Journal of Chromatography A* in 2024. **Darien Yeung**, Victor Spicer, and Oleg V. Krokhin.

DY designed and performed experiments, performed data analysis, writing, editing of the manuscript. VS performed data analysis, editing of the manuscript. OVK conceived and supervised the project, performed data analysis, writing, and editing of the manuscript.

Abstract

Peptide separation selectivity was evaluated for hydrophilic interaction liquid chromatography (HILIC) ZIC-HILIC, ZIC-cHILIC, and XBridge Amide sorbents using formic acid as eluent additive (pH 2.7). Sequence-specific retention prediction algorithms were trained using retention datasets of ~30,000 peptides for each column. Our retention models were able to attain ~0.98 R2-value and yielded retention coefficients that can be probed to understand peptide-stationary phase interaction. Overall, the hydrophilicity for these columns decreased when the mobile phase changed pH from 4.5 to 2.7, when using 0.1% formic acid in the mobile phase. The acidic residues became protonated, and the resultant hydrophilic interaction is dampened at the lower pH, leaving only the basic residues as the primary hydrophilic interactors. Hence, peptides of increasing charge have higher retention. In this comparison between the three columns, ZIC-HILIC has the highest chromatographic resolution between groups of peptides of different charge. From the position-dependent retention coefficients for ZIC-HILIC at pH 2.7, we found that the amino acids at the terminal positions of the peptide modulate the basicity of the N-terminal amino group or the C-terminal Arg/Lys for tryptic peptides. With respect to the separation orthogonality between HILIC and acidic pH RP for two dimensional separations, the orthogonality values were lower at pH 2.7

than operating HILIC at pH 4.5 for the first dimension. We also demonstrate that ZIC-HILIC was able to distinguish citrullinated and deamidated peptides based on predicted retention values.

4.1. Introduction

Hydrophilic interaction liquid chromatography (HILIC) is a separation method leveraging the interaction of hydrophilic analytes with an aqueous layer adsorbed on to a polar stationary phase.[1,2] This chromatographic mode was characterized in 1990 by Alpert, initially, demonstrating that cation-exchange and hydrophilic neutral materials can retain hydrophilic compounds at high organic solvent concentrations. Subsequently, various hydrophilic sorbents were introduced for HILIC separation.[1,3-5] This type of sorbents showed the ability to retain very hydrophilic compounds that are commonly found in the flow through in reversed-phase (RP) separations.[6-8] Furthermore, pairing HILIC online with mass spectrometry (MS) results in a higher detection sensitivity than RP with electrospray ionization MS.[9,10] This is because the mobile phase for HILIC contains a higher concentration of organic solvents which is more favorable for efficient electrospray ionization. There are a variety of functional groups that are available as stationary phases for HILIC now that this method has been popularized for small-molecules separations.[8,11] Despite these advantages for HILIC, RP is still the dominant chromatographic mode in bottom-up proteomics for peptide separation due to its versatility and ability to achieve higher chromatographic resolution.[12,13]

Large scale HILIC peptide separation with MS detection had been evaluated by Horie et al. where the authors utilized monolithic columns with ureidopropyl function groups to serve as the polar stationary phase.[14,15] Extending the column length from 30 cm to 400 cm and increasing the gradient time to 4 hours were necessary to achieve comparable chromatographic resolution to

RP. When the authors compared HILIC and RP separations while maintaining the same column dimensions and gradient time, the resulting LC-MS/MS analyses of HeLa cell digests showed similar protein identification outputs. Both analytical approaches were able to detect ~10,000 unique peptides to identify ~2600 proteins, but the signal intensities were much higher when HILIC was interfaced with the MS. The improvements in sensitivity were beneficial for targeted quantitation of peptides where Simon et al. reported up to a 14-fold increase in signal intensity.[10] Beyond monolithic columns, particle-based zwitterionic HILIC (ZIC-HILIC) stationary phases were introduced to increase the retention of highly polar small molecules, which would be appropriate for the separation of basic tryptic peptides.[16-18] The zwitterions on the stationary phase have a proximal charge that is closer to the silica and a distal charge further away that acts as the surface of the stationary phase. ZIC-HILIC features a negatively charged group on the distal surface and a positively charged one on the proximal site, whereas ZIC-cHILIC is the opposite. These separation modes had been used directly with MS for peptide separation but suffered from poor chromatographic resolution.[16] Despite all these efforts to use HILIC online with MS, the problems of low resolution or impractical analysis times due to longer re-equilibration were too much to accept for the potential sensitivity improvements afforded by the mobile phase composition.

Aside from direct online separation of peptides, multidimensional chromatography (MDC) represents a popular alternative to reduce sample complexity prior to the MS analysis. During the early years of MS-based proteomics, HILIC became a promising chromatographic mode to use in the first dimension for multi-dimensional chromatography.[8,19] Boersma et al.[20] and Di Palma et al.[21] demonstrated that the separation of peptides using ZIC-HILIC/ZIC-cHILIC columns paired with RP were able to obtain improved protein identification outputs compared to the SCX-RP

approaches.[22] Peptide separation on ZIC-HILIC was further evaluated for a variety of pH conditions and the authors reported that pH 3 had the highest orthogonality when paired with RP at acidic pH. Longworth et al.[23] and Intoh et al.[24] also observed protein identification improvements when they used the ZIC-HILIC-RP pairings for isobaric tag-based quantitation. The authors attributed these improvements to the orthogonal chromatographic properties of HILIC with RP. In a comparison involving six chromatographic modes, Gilar et al. demonstrated that silica HILIC at pH 4.5 was the most orthogonal with RP, surpassing both pH 10 RP and strong cation exchange, which were the most common first dimensions for MDC.[25] Our recent orthogonality evaluation of sixteen chromatographic modes agreed that HILIC(pH 4.5)-RP are highly orthogonal pairings for a variety of hydrophilic sorbents.[26] However, in a comparison between HILIC-RP and strong anion exchange-RP pairings, Bensaddek et al.[27] observed lower orthogonality for the former when they separated peptides through the TSKgel Amide 80 column using trifluoroacetic acid as the mobile phase additive. Although both ZIC-HILIC and TSKgel Amide 80 columns were previously evaluated at pH $\leq$ 3, their opposing conclusions may suggest that stationary phase features play an important role for HILIC separation at acidic condition.

To better understand the chromatographic behavior of peptides in HILIC, a series of separation modeling studies have been undertaken. Yoshida et al. presented the first additive model that used the retention data for 121 synthetic peptides separated on the TSKgel Amide 80 under acidic conditions with trifluororacetic acid and achieved an R^2-value of 0.94.[28] Gilar et al. expanded on this concept and developed models for three different HILIC sorbents.[29] The authors used 150 synthetic peptides to understand the differences in separation selectivity through the retention coefficients obtained in their linear regression models. The accuracy of the resulting models varied from 0.92 to 0.97. Hoarscoat-Schiavo et al. improved the modeling accuracy of

peptide separation in TSKgel Amide 80 to an R^2-value of 0.97 for a collection of 58 relatively short peptides.[30] However, a common problem between all these prediction models is their limited size of datasets that do not appropriately account for various sequence-specific effects observed in HILIC separations.

Our laboratory modeled HILIC separations using Sequence-Specific Retention Calculator (SSRCalc) approach. The peptide HILIC retention information has been collected through 2D LC-MS/MS analyses of complex yeast tryptic digests with various HILIC stationary phases employed in the first separation dimension. This allowed for the detection of +30,000 unique peptides for each of seven HILIC columns we previously evaluated. We used this workflow for modeling peptide separation on amide,[31] zwitterionic,[32] diol, and silica-based[33] HILIC sorbents. Generally, HILIC columns have sequence-specific effects involving the N-terminal positions of amino acid residues, where the most variations in retention coefficients occur compared with the internal position. Upon comparing the peptide separation on various HILIC columns, we find that the overall peptide retention (sorbent hydrophilicity) increases in the following order diol < zwitterionic = amide < silica. Similar to the modeling and orthogonality comparison efforts from Gilar et al.,[25] we employed the pH 4.5 ammonium formate buffers.

In prior studies, Boersma et al.[20] suggested that acidic conditions improve the orthogonality of ZIC-HILIC with RP. Furthermore, acidifying the mobile phases protonates the residual silanol groups leading to improved separation efficiency. Hence, we decided to investigate peptide separation selectivity of HILIC under acidic (formic acid) conditions using large scale proteomics-derived datasets. We selected ZIC-HILIC, ZIC-cHILIC, and XBridge Amide columns since these stationary phases are very similar in hydrophilicity, while the zwitterionic functionalities may result in substantial changes in separation selectivity. Our approach to studying the peptide

retention behaviour and selectivity differences is based on the 2D LC-MS/MS (HILIC-RP) analysis of complex mixtures of tryptic peptides and developing SSRCalc models that can provide quantitative insights into contributions of individual amino acid residues.

4.2 Experimental Section

4.2.1 Chemicals, and Sample Preparation

Unless otherwise noted, all chemicals were sourced from Fisher Scientific (Waltham, MA). Mass-spectrometry compatible yeast protein extract was acquired from Promega (Madison, WI). Proteins were reduced, alkylated with iodoacetamide and digested using 1:50 trypsin:protein ratio for 14 h at 37°C. Peptides were subsequently acidified with 0.5% trifluoroacetic acid and desalted using SOLA-hRP solid phase extraction cartridges (Fisher Scientific) according to manufacturer's recommendations. The trapped peptides were eluted at 65% acetonitrile in 0.1% trifluoroacetic acid. The digests for bovine serum albumin (Sigma Aldrich, St. Louis, MI) were prepared using a standard in-solution digestion with trypsin. Samples were then resuspended in respective buffers A for LC-MS/MS or LC-UV analyses. HPLC-grade water and acetonitrile (both 0.1% formic acid, Honeywell, Charlotte, NC) were used for eluents preparation. The six synthetic peptides were synthesized by Bio-Synthesis Inc (Lewisville, TX) with sequences designed elsewhere[34] P1: LGGGGGGDGSR, P2: LGGGGGGDFR, P3: LLGGGGDFR, P4: LLLGGDFR, P5: LLLLDFR, P6: LLLLLDFR.

4.2.2 First dimension separation conditions

Agilent 1100 series HPLC system (Agilent, Santa Clara, CA, USA) with UV detector (214 nm), manual injector with 100 µL loop and 200 µL/min flow rate were used for separations. Both

eluents: A (water) and B (acetonitrile) contained 0.1% formic acid. The columns used in this work were the Waters XBridge Amide 3.5 µm, 130 Å, 50 x 3.0 mm (Waters, Milford, MA), Merck SeQuant ZIC-HILIC 3.5 µm, 100 Å, 100 x 2.1 mm, and ZIC-cHILIC 3 µm, 100 Å, 100 x 2.1 mm (both Merck, Darmstadt, Germany). For all three columns, the same gradient was used to separate standard peptide mixtures, BSA and yeast digests. The gradient program included a linear increase of water contents from 10 to 70% in 60 min, followed by column wash at 90% water.

All samples were resuspended in the starting conditions' buffer 10% water/90% acetonitrile in 0.1% formic acid for first dimension HILIC separation. 100 µg of *S. Cerevisiae* digest was injected into the column, and fractions were collected every minute from the sample's injection until the end of the chromatogram (based on UV profile). Fractions were then lyophilized and resuspended in 30 µL of eluant A to be analyzed using LC-MS/MS.

4.2.3 Second dimension separation conditions

LC-MS/MS in the second dimension used the 2D LC Ultra system (Eksigent, Dublin, CA) with TripleTOF 5600 mass spectrometer (Sciex, Concord, ON, Canada) operating in data dependent analysis mode. Self-packed analytical column Luna C18(2) 3 µm, 100 Å, 100 µm x 200 mm (Phenomenex, Torrance, CA) was used at 500 nL/min flow rate. The gradient program included a linear increase from 0.4 to 31% B for 77 minutes and to 80% B for 5 minutes as a column wash, followed by 0.4% B for column equilibration. Both eluents A (water) and B (acetonitrile) contained 0.1% formic acid. The trap column was the PepMap 100 0.3 x 5 mm (Fisher Scientific) used for sample loading with a 15 µL/min flow rate, and each sample had an injection volume of 10 µL. Each fraction was spiked with the standard P1-P6 peptides[34] for the retention time alignment purposes.

4.2.4 In-vitro modification of tryptic peptides

This set of experiments to evaluate peptide retention with citrullinated Arg, carbamylated Lys and deamidated Asn residues followed experimental protocol described elsewhere.[35] An equimolar mixture of 11 commercially available proteins, all sources from Sigma Aldrich, (chicken, human, bovine, mouse, and porcine) albumin, (human and bovine) apotransferrin, (human and bovine) lactoferrin, human fibrinogen, and human fibronectin) was used for *in-vitro* modification to generate citrullinated, carbamylated, and deamidated peptides. Deamidated peptides were readily observed for Asn-Gly containing peptides during the overnight tryptic digestion at 37°C. Carbamylated peptides were generated by incubating 100 µg of the equimolar protein mixture in 2M Urea at 60°C for 32 hours prior to trypsin digestion using the same procedure as above. Citrullinated peptides were generated by incubating 100 µg of the equimolar protein mixture with human peptidyl-arginase deiminase type 2 (PAD-2, Sigma Aldrich) at 55°C, pH 7.2 for 32 hours with 10 mM $CaCl_2$ as a reaction buffer. Following the citrullination, the proteins were in-solution digested with trypsin as described previously.

4.2.5 Data Analysis

X!Tandem search engine was used with the following settings: 20 ppm and 50 ppm mass tolerance for parent and daughter ions, respectively; constant modification of Cys with iodoacetamide; standard set of expected post-translational modifications (Met/Trp oxidation, Asn/Gln deamidation, Cys/Gln N-terminal cyclization); two allowed missed cleavage. The resulting identified peptides had retention values expressed in HI units (Hydrophobicity Index, %

acetonitrile) converted using the known HI values for our retention standard P1-P6 peptides.[34] Only tryptic peptides with expectation scores less than -3 were selected for the optimization dataset. The retention values of the first dimension were assigned using the fraction number corresponding to where the peptide was detected. If the peptide is observed in multiple adjacent fractions, a weighted average of the intensity values was used to establish the decimal fraction number in which the peptide eluted. Fraction numbers were also converted into the corresponding water percentage based on the gradient program and measured gradient delay time.

4.3 Results and Discussion

4.3.1 Comparison of peptide retentivity and major selectivity features for XBridge Amide, ZIC-HILIC, ZIC-cHILIC columns at acidic pH

Figure 4.1 shows the separation of six synthetic peptides, P1-P6 for all three columns using a 1% ACN per minute gradient. Throughout this study, the mobile phases contained 0.1% formic acid, and therefore referred hereafter as pH 2.7, corresponding to pH of water solution. Overall, the chromatograms for XBridge Amide and ZIC-HILIC are similar in pattern and have wider peaks than observed in ZIC-cHILIC. Peptide retention increases with respect to hydrophilicity where the most hydrophobic P6 eluted first and the most hydrophilic P1 eluted last. Among the three columns, peptide retention increases in order of: ZIC-cHILIC < XBridge Amide < ZIC-HILIC. Similar retention trends have been observed for the separation of the standard BSA digest. We find that these differences in peptide retentivity are driven by the nature of the functional groups. The zwitterionic group on ZIC-cHILIC exposes the cationic quaternary ammonium on the distal surface which electrostatically repels positively charged peptides. Therefore, ZIC-cHILIC has the lowest peptide retention out of the three HILIC columns at pH 2.7. Whereas, ZIC-HILIC has the

highest retention because this sorbent has a negative charge on the distal surface to electrostatically attract basic peptides.

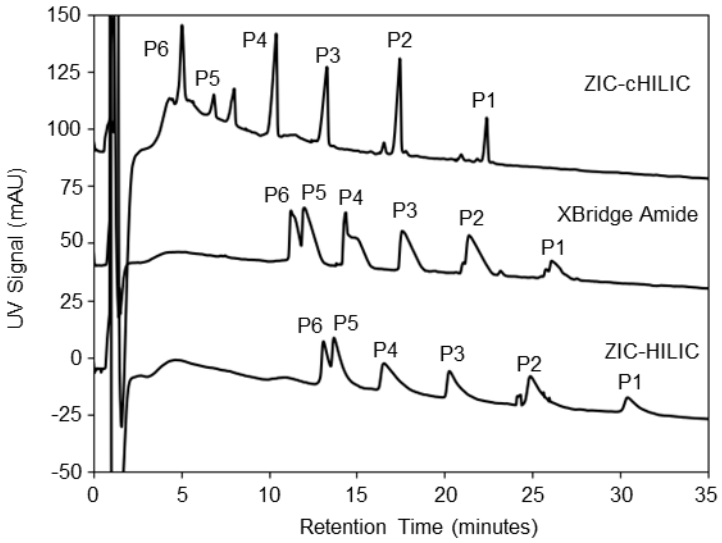


Figure 4.1: LC-UV chromatogram of six synthetic peptides for the three HILIC columns at pH 2.7.

Table 4.1 shows the identification output for the 2D LC-MS/MS experiments for the three columns. Approximately, 1 µg of peptides were analyzed by the second dimension LC-MS/MS for each fraction. On average, there are ~30,000 non-redundant peptide identifications for each 2D experiment. ZIC-cHILIC had the narrowest peptide elution range resulting in 35 fractions collected. Conversely, peptide separation in ZIC-HILIC spanned a wider elution range giving 47 fractions. Based on the data we collected here and previous study using pH 4.5 eluents,[32] the HILIC separation of peptides with each column at pH 2.7 demonstrated a lower retention than at pH 4.5. This indicates the columns exhibit a lower degree of hydrophilicity at pH 2.7. The acidic residues in the peptides lose their negative charge at pH 2.7 and therefore, have reduced hydrophilic contribution to peptide retention on HILIC. To compare the general peptide retention trends between the different column conditions, we converted the peptide retention data from fraction numbers to the corresponding eluant water percentage using the first-dimension gradient slope.

The difference in hydrophilicity increases in order of: ZIC-HILIC < XBridge Amide < ZIC-cHILIC.

Table 4.1: Peptide and protein identification statistics from the 2D-LC/MS/MS runs for the three HILIC first dimension columns at acidic conditions.

Column	# of peptides	# of non-redundant peptides	# of proteins	# of fractions
ZIC-cHILIC	106442	26556	3711	35
XBridge Amide	124784	28218	3504	38
ZIC-HILIC	153320	35154	3952	47

Figure 4.2 demonstrates that peptide retention increases with peptide charge for all three columns operating at pH 2.7. This is consistent with the previous report from Boersma et al. [20] that demonstrated a similar trend in peptide separation when using ZIC-HILIC with 0.1% formic acid in the mobile phase. ZIC-HILIC has the highest resolution between the peptides of different charges because of its anionic surface strongly retaining basic peptides. XBridge Amide and ZIC-cHILIC cannot replicate this charge separation effect and have large regions of overlaps between charges. The propensity for ZIC-HILIC to participate in charge-based separation for peptides is similar to the pattern observed in strong cation exchange, albeit without the use of salt gradients to elute components of the mixture.

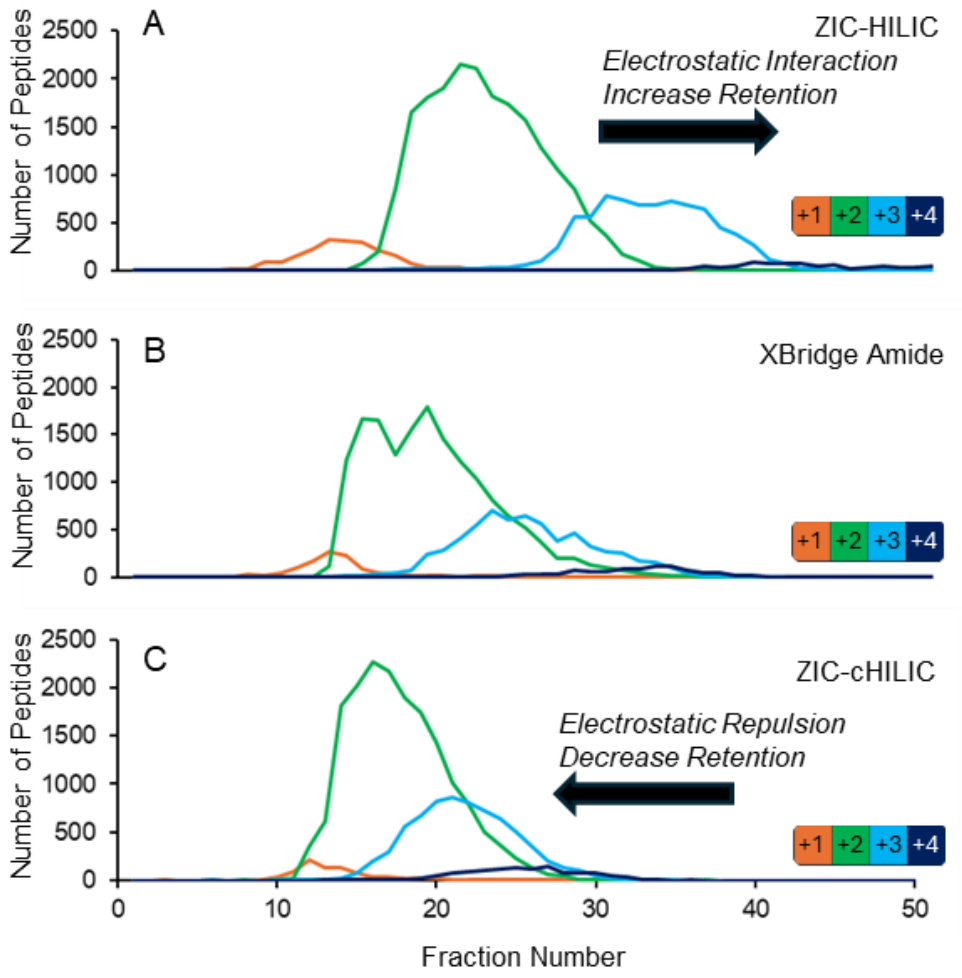


Figure 4.2: Charge separation of +1 (orange), +2 (green), +3 (light blue), +4 (dark blue) peptides for the three HILIC columns at pH 2.7 with formic acid modifier in the mobile phase.

4.3.2 Peptide retention time prediction models for HILIC separations at pH 2.7

4.3.2.1 Composition dependent features driving peptide retention.

Collecting large-scale retention datasets allowed us to develop sequence-specific models for different HILIC separations at pH 2.7. The models use the same optimization scheme and variable structure as previously reported SSRCalc algorithm for the separation of peptides on ZIC-HILIC and ZIC-cHILIC at pH 4.5.[32] In brief, the amino acid contributions were modelled by assigning

eleven position-based coefficients to each of the twenty amino acids encoded via one-hot encoding. Within the eleven coefficients, they correspond to the first five positions on the N-terminus and the last five positions on the C-terminus of a peptide. The eleventh coefficient corresponds to all remaining positions that are classified as internal. We then use the peptide sequences and their retention times to optimize the SSRCalc models for each column to maximize the resulting R^2-values for the predicted versus experimental retention plots. Resulting correlations reaching ~0.98 R^2-values are shown in Figure 4.3.

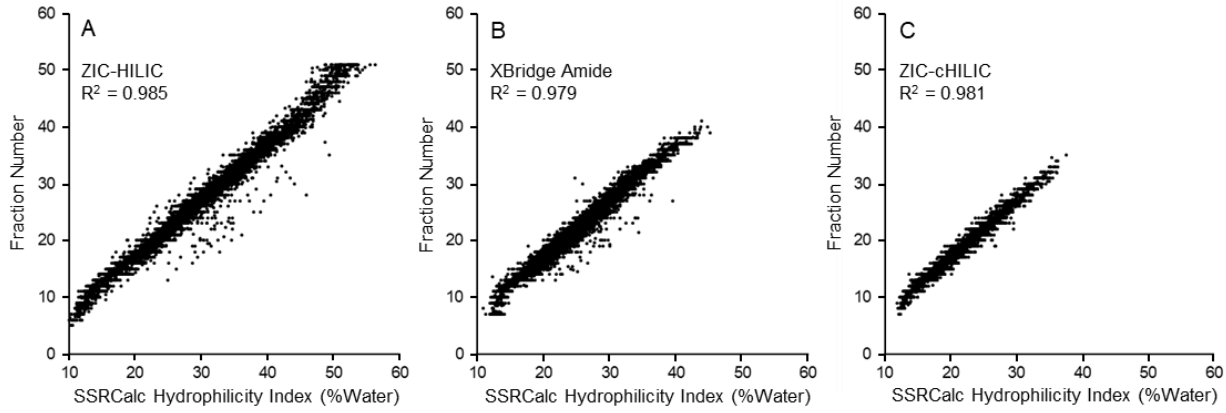


Figure 4.3: SSRCalc peptide retention time prediction models for (A) ZIC-HILIC, (B) XBridge Amide, (C) ZIC-cHILIC at pH 2.7 using 0.1% formic acid in the mobile phase.

Figure 4.4 A shows the normalized internal retention coefficients for the three columns under investigation. All three columns share similar trends where hydrophilic residues increase peptide retention while hydrophobic residues decrease retention. The major selectivity differences between the columns come from the contributions of basic and acidic amino acids. At pH 2.7, Lys/Arg/His are the dominant contributors to peptide retention across all three columns as these are only residues carrying positive charge. However, they have the strongest effect on ZIC-HILIC and the weakest on ZIC-cHILIC. Meanwhile, acidic residues, Asp/Glu, have the opposite relationship where ZIC-cHILIC has the highest normalized retention coefficients for these acidic

amino acids and ZIC-HILIC has the lowest. The peptide interactions in HILIC can involve electrostatic interaction from the stationary phase in addition to the partitioning of peptides into the aqueous layer. Therefore, due to the anionic distal charge on the surface of the ZIC-HILIC sorbent, the electrostatic attraction with Lys/Arg/His increases retention while the repulsion of Asp/Glu decreases retention. Conversely, ZIC-cHILIC has cationic charges on the distal surface and demonstrates the opposing retention characteristics of ZIC-HILIC. It was interesting to note that some hydrophilic amino acids Asn/carbamidomethyl-Cys/Ser/Thr have lower retention contributions in ZIC-HILIC compared to the other two columns. Overall, the retention coefficients extracted for ZIC-HILIC share similar trends with Strong Cation Exchange (SCX) separation[36] where basic residues are the primary retention contributors, and all other hydrophilic/hydrophobic residues have minimal and very uniform contributions. The only difference between HILIC and SCX at acidic pH in this regard is the negative contribution of hydrophobes into the overall retention for the former separation mode.

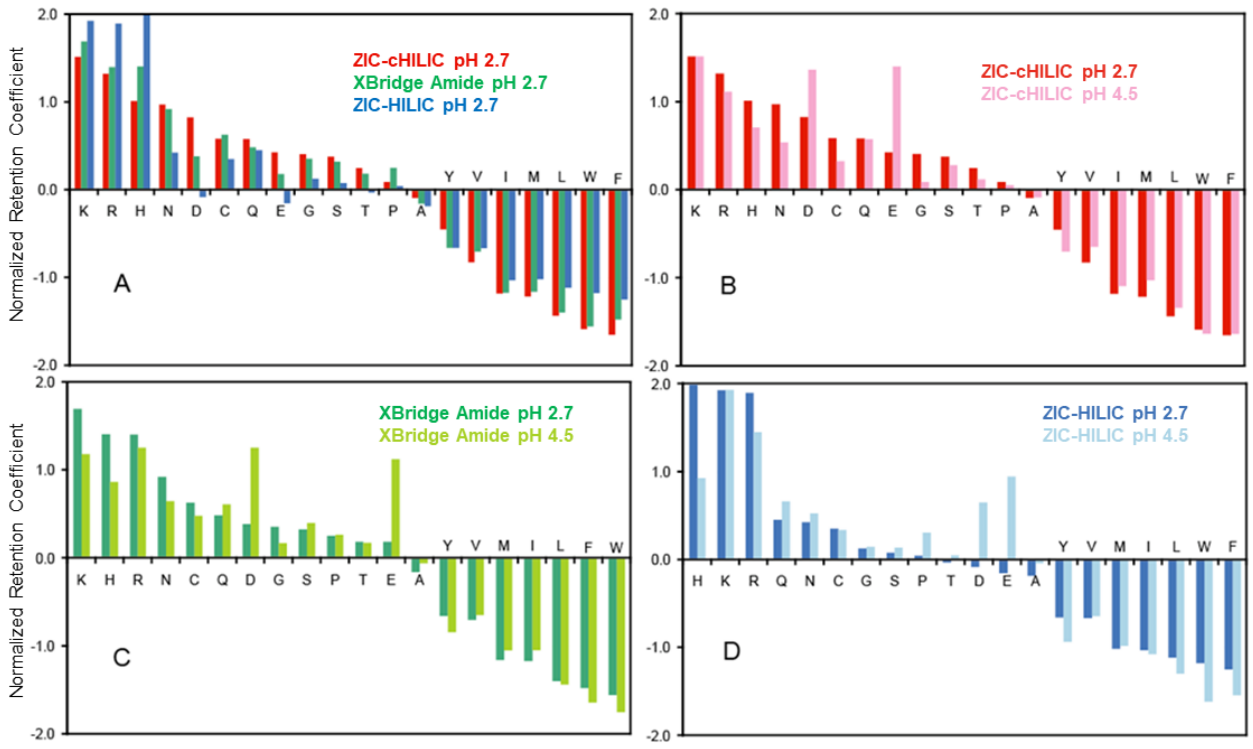


Figure 4.4: Normalized amino acid retention coefficients obtained from SSRCalc models of (A) the three HILIC phases at acidic conditions, (B) ZIC-cHILIC at pH 2.7 and pH 4.5, (C) XBridge Amide at pH 2.7 and pH 4.5, and (D) ZIC-HILIC at pH 2.7 and pH 4.5.

Hydrophobic residues decrease the retention of peptides during separation in HILIC at pH 2.7. The general trends between the three columns are consistent; however, there are difference in specificities where the negative retention contributions from ZIC-cHILIC and XBridge Amide for hydrophobes are greater than ZIC-HILIC. In ZIC-HILIC, the basic peptides partition readily into the water layer on the sorbent and have electrostatic attraction between the peptide and ZIC-HILIC stationary phase. This attraction can pull the peptides closer to the stationary phase to support hydrophobic interactions with the polymeric graft that may dampen the hydrophobic repulsion. Among the HILIC columns, aromatic residues retain the least except for Tyr, due to its polar hydroxyl group increasing retention. The retention coefficients for aliphatic residues decrease in order of Ala > Val > Ile > Leu, which trends negatively with their accessible surface area (ASA).

Having a higher ASA for the aliphatic hydrophobes in HILIC can increase their hydrophobic repulsion with the water layer to exhibit lower retention on column.

In comparison to the same three columns at pH 4.5 (Figure 4.4 B-D), the retention factors are similar for basic and hydrophobic amino acids across all the columns. Generally, our comparison of HILIC at the two pH conditions reveal that the general stationary phase characteristics are maintained with minor differences in specificity. The major difference in retention contribution is from acidic residues Asp/Glu since they change charge states between pH 4.5 to 2.7, which changes the hydrophilicity of the peptides. The loss of the negative charge on the acidic residues causes a decrease in retention for all columns evaluated. XBridge Amide exhibited the greatest loss in normalized retention coefficient as a result of acidifying the mobile phase to pH 2.7. Interestingly, when formic acid is used in the mobile phase ZIC-cHILIC demonstrated an increase in retention coefficient for Asn/carbamidomethylated Cys that is higher than the other two columns. For the hydrophobes, the repulsive effects are similar between the two pH conditions for ZIC-cHILIC and XBridge Amide. However, it is noted that the retention contribution of hydrophobes decreases for ZIC-HILIC at pH 2.7.

4.3.2.2 Position dependent features affecting peptide retention.

The contributions of individual residues into peptide retention are different for terminal positions compared to internal ones as shown in Figure 4.5. Since the vast majority of tryptic peptides carry Arg or Lys at their C-termini, for this comparison we considered N-terminal (N1) and the residues preceding Arg/Lys (C2). Similarly, plots in Figure 4.5 are lacking internal Arg and Lys due to their low abundance inside tryptic peptides. One important conclusion can be drawn based on the behavior of the hydrophilic neutral residues Gln/Asn/Ser/Thr and the hydrophobic Trp/Phe. Their

respective contributions positive and negative, are higher for both terminal positions. The acidic and basic residues also show unique behavior, which scales in accordance with sorbent chemistry. N-terminal Asp is considerably more hydrophobic and more hydrophilic in N1 and C2 positions compared to internal one, respectively. Asp at the N-terminal position significantly decreases the basicity of N-terminal amino group,[36] therefore, reducing the peptide charge and hydrophilicity. This effect is more pronounced for the zwitterionic sorbents than XBridge Amide. When placed in the C2 position, Asp reduces the ion-pairing interaction between formic acid and the adjacent Arg/Lys residue. Hence, the positive charge is more hydrophilic in the absence of ion-pairing to increase retention. This is also observed for Glu at both the N1 and C2 positions. Conversely, His at the C2 position recruits more formate counterions to engage in ion-pairing to dampen the basicity of the C-terminal Arg/Lys.

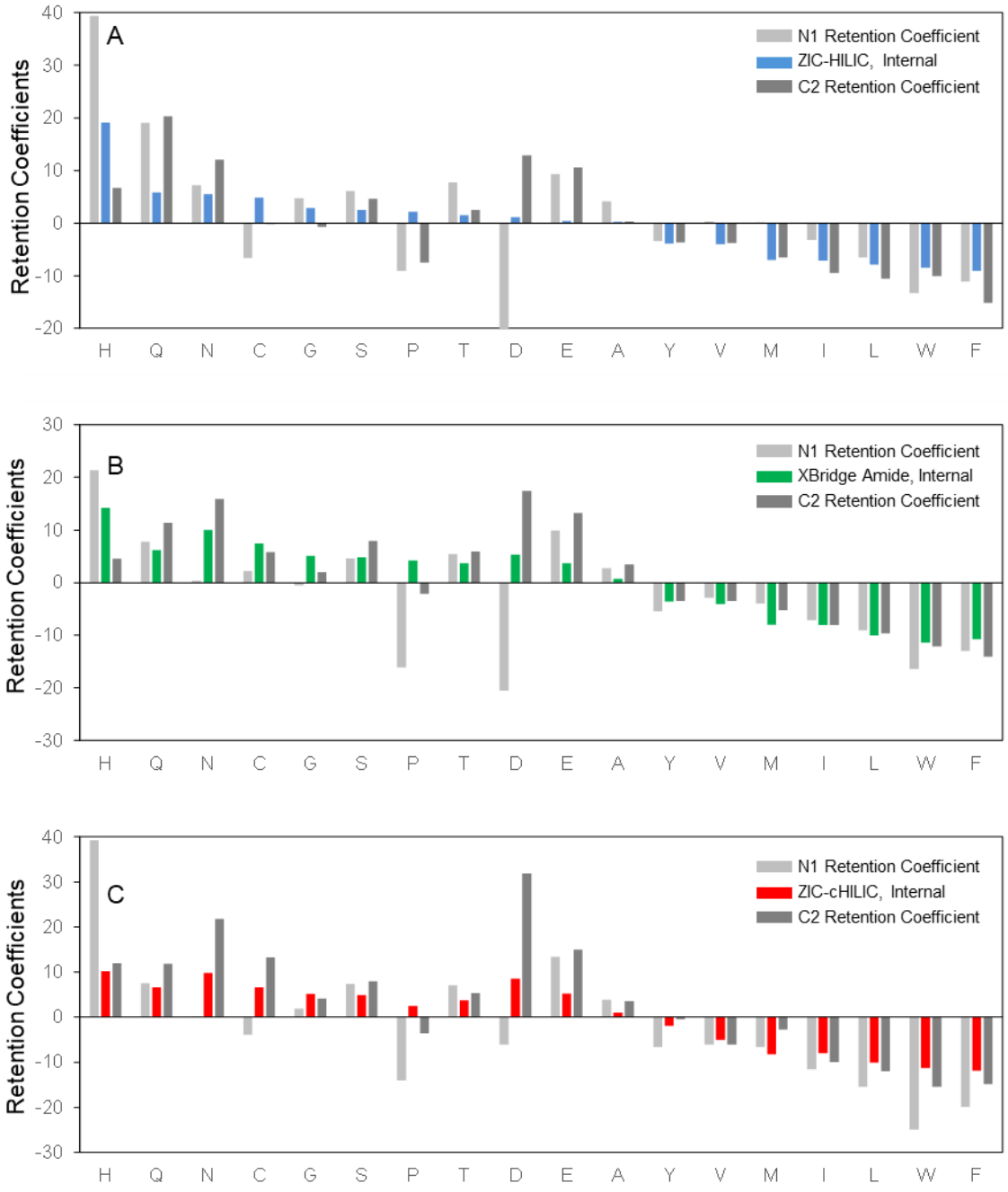


Figure 4.5: Position-dependent retention coefficients extracted from the SSRCalc model for (A) ZIC-HILIC, (B) XBridge Amide, (C) ZIC-cHILIC at pH 2.7 at position N1, internal, and C2 of the peptide.

4.3.3 Separation Orthogonality of HILIC separations at acidic pH vs. RP HPLC (formic acid)

Figure 4.6 shows our use of the geometric filling approach outlined by Gilar et al. [25] and Yeung et al. [26], to demonstrate the orthogonality for the three columns, with respect to acidic pH RP in the second dimension. The resulting orthogonality values are 69.0%, 73.2%, and 72.6% for ZIC-HILIC, XBridge Amide, and ZIC-cHILIC columns at pH 2.7, respectively. Currently, basic pH 10 RP paired with acidic pH RP is the gold standard for MDC and this pairing has a higher orthogonality value of 74.5%[26] than HILIC separation at pH 2.7. Furthermore, comparing the latter to their pH 4.5 counterparts, the higher pH conditions (4.5) had increased orthogonality values of 77.2%, 80.92%, 79.91% for ZIC-HILIC, XBridge Amide, and ZIC-cHILIC, respectively. Boersma et al.[20] previously concluded that using formic acid additives in the mobile phase for ZIC-HILIC separations resulted in higher orthogonality for MDC separation. However, this was likely due to their limited size in peptide dataset and inaccurate orthogonality evaluation. The authors used the R^2 values to determine orthogonality which we previously reported[26] was not a metric that accurately reflects the distribution of peptides across two dimensions of separation.

Fundamentally, separation orthogonality relate the differences in peptide retention between two chromatographic modes. These differences can involve changing the stationary phase properties (separation mechanism) or changing peptides' physicochemical properties (often through alteration of pH). When we pair HILIC at pH 2.7 with acidic pH RP that is also at pH 2.7, only the stationary phase is different. On the other hand, when we pair basic pH 10 RP with acidic pH RP, peptides' charge states change dramatically. The combination of changing the separation mechanism and charge results in higher orthogonality values than changing the individual parameters by themselves. Hence, HILIC at pH 4.5 with acidic pH RP is more orthogonal than

using HILIC at pH 2.7 in the first dimension. Furthermore, this is consistent with the report from Bendassek et al.[27] that HILIC-RP had lower orthogonality than conventional MDC pairings when they used in the first dimension the TSKgel Amide 80 HILIC column with trifluoroacetic acid additives in the mobile phase. Therefore, for MDC experiments involving HILIC-RP, we suggest to perform the HILIC separation at a different pH than RP to allow the change in chromatographic mode and peptide charge to increase the orthogonality.

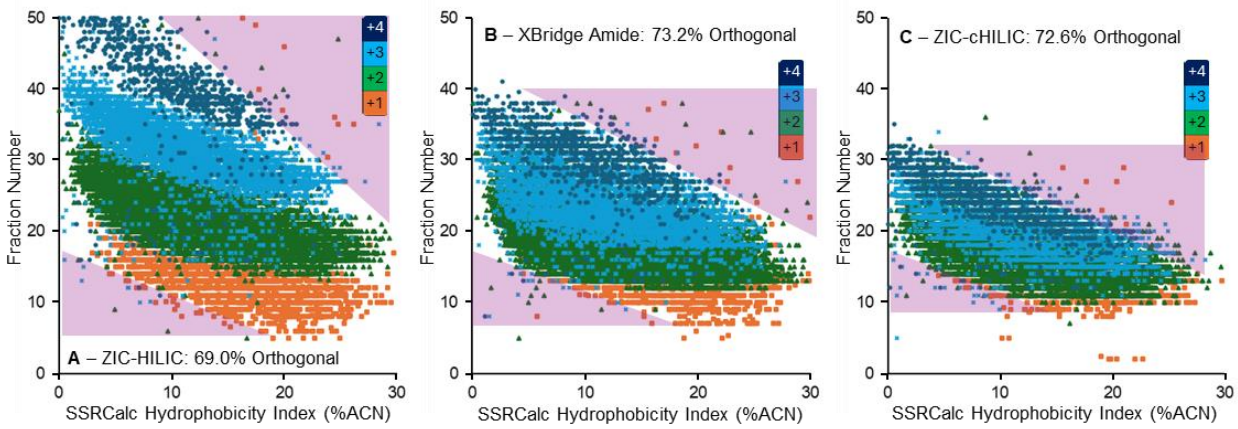


Figure 4.6: Orthogonality plots for (A) ZIC-HILIC, (B) XBridge Amide, and (C) ZIC-cHILIC using formic acid in the mobile phase compared with RP HPLC (formic acid). Peptides with different charge states are labeled +1 (orange), +2 (green), +3 (light blue), and +4 (dark blue). Regions that are not occupied by peptides in all three orthogonality plots are highlighted in purple. Larger proportion of these unoccupied areas explains gradual increase in separation orthogonality from ZIC-HILC to ZIC-cHILIC.

Figure 4.6 A is showing the ZIC-HILIC and RP peptide orthogonality plot with each charge group being well separated. The electrostatic attraction between the basic peptides and the stationary phase shifts the charge distribution vertically along the y-axis. The effect was less pronounced in the XBridge Amide and ZIC-cHILIC orthogonality plots. Furthermore, the slopes for each charge group is similar indicating that the hydrophilic interaction between peptides of

different charges are consistent. This is opposite to what we observed for SCX, which had different slopes for each charge group when visualized on an SCX-RP orthogonality plot.[26] Upon further investigation of the peptide retention in these charge groups, ZIC-cHILIC and XBridge Amide appear to have a hydrophilic focussing effect. This effect is seen where hydrophilic peptides elute over a narrower range in the first dimension than peptide of the same charge that are more hydrophobic, shown in Figure 4.6B. In the +3 peptide distribution of XBridge Amide, at >5% ACN, the peptide distribution spanned over 10+ fractions. However, at 1% ACN, the hydrophilic peptide distribution constricted to span only 5 fractions.

4.3.4 Separation of Citrullinated/Carbamylated/Deamidated Peptides

The charge separation in acidic ZIC-HILIC at pH 2.7 was found similar to SCX. Therefore, specific separation tasks utilizing SCX should also be addressable by acidic pH ZIC-HILIC. Post-translational modifications (PTMs) like citrullination, carbamylation (charge reducing) have SCX retention shifts which are drastically different compared to deamidation.[35] Villacres et al. have suggested this feature for confident identification of citrullinated peptides, which feature the same mass shift as the abundant deamidation.[35] In this work we applied ZIC-HILIC at pH 2.7 to a similar workflow. Through a two-dimensional separation of non-modified and citrullinated/carbamylated/deamidated peptides, in Figure 4.7 A, we observed that ZIC-HILIC exhibit significant decrease in retention of citrullinated/carbamylated peptides compared to deamidated products. Removal of positive charge from arginine's and lysine's side chain upon citrullination and carbamylation significantly decreases peptide retention compared to deamidation, which features no change in charge status at acidic pH. Using peptide retention prediction by SSRCalc ZIC-HILIC allows for confident distinguishing of citrullinated vs. deamidated peptides using prediction errors as shown in Figure 4.7 B. Moreover, the application

of ZIC-HILIC is more convenient compared to SCX, where subsequent LC-MS/MS analysis requires additional desalting for the latter.

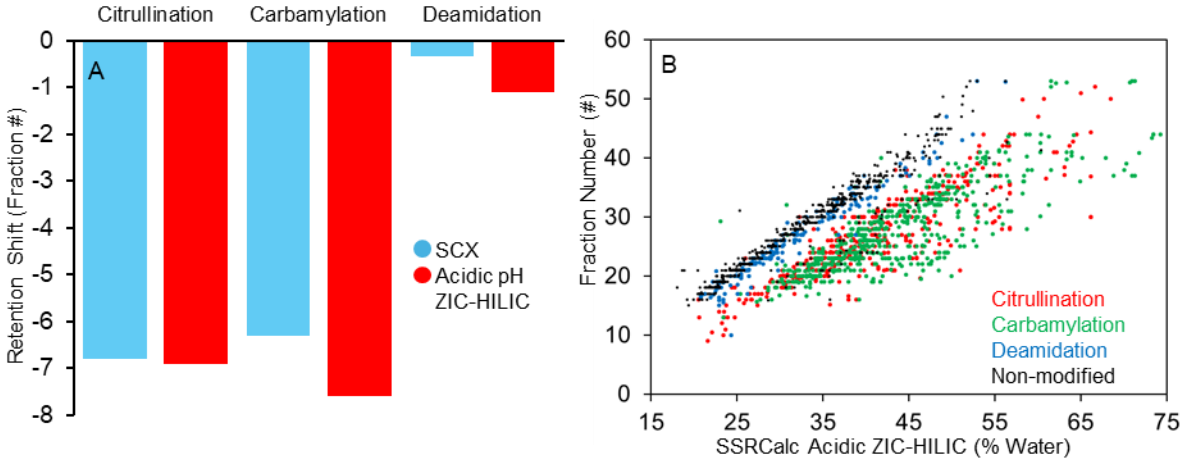


Figure 4.7: (A) Retention shifts of citrullinated, carbamylated, and deamidated peptides in ZIC-HILIC at pH 2.7. (B) SSRCalc ZIC-HILIC retention prediction for non-modified and modified (citrullination/carbamylation/deamidation) peptides observed in the 2D LC-MS/MS analysis of model protein mixture.

4.4 Conclusion

We employed 2D LC-MS/MS methodology for generating large scale retention data collections to study peptide separation selectivity for HILIC separations at acidic pH (2.7). Using column of interest as the first separation dimension allows studying any separation system regardless of its compatibility with downstream MS detection. ZIC-HILIC, ZIC-cHILIC and XBridge Amide stationary phases were selected due to similar overall hydrophilicity, resulting in ~30,000 unique peptide identifications each. For all three columns there was a distinct decrease in peptide retentivity relative to the HILIC separations at pH 4.5 studied before.[32] Overall, peptide retention at pH 2.7 increased in the order ZIC-cHILIC < XBridge Amide < ZIC-HILIC – in agreement with

chemistry of stationary phases. ZIC-HILIC has negatively charged distal functional group inducing electrostatic interactions with positively charged peptides. ZIC-cHILIC features the opposite zwitterion orientation, while X-Bridge Amide is neutral. Peptides' positively charged groups are the major contributors in retention on all three stationary phases. ZIC-HILIC has the strongest interaction strength towards positive charge, allowing almost complete separation of peptides from differently charged populations.

Sequence-specific retention prediction models have been developed for three separation settings, reaching R^2-values ~0.98. The retention coefficients derived from these models confirmed that the main residues contributing to retention are Lys/Arg/His. We also found that Asp/Glu had much lower retention contributions at pH 2.7 than at pH 4.5 By studying the position dependent retention coefficients, we found that the variability between internal and terminal positions is driven by improving accessibility of the residues located near termini and acidic/basic properties of individual amino acids. The separation of peptides by HILIC at pH 2.7 paired with the standard acidic pH RP resulted in having lower orthogonality values than using HILIC at pH 4.5 as the first dimension. Increasing the pH allowed for the alteration of peptide charge (through deprotonation of Asn and Glu) resulting in enhanced separation orthogonality. Furthermore, ZIC-HILIC has potential for peptide charge separation and can be used for confident identification of citrullinated versus deamidated peptides based solely on their retention properties.

This study of HILIC separation at acidic pH was done using formic acid-based mobile phases to cover the most popular eluent conditions compatible with ESI MS. It would be beneficial to expand our understanding of ion-pairing effects that may modulate hydrophilic interactions and partitioning of peptides with the water layer on the stationary phase by varying eluent modifiers (acetic, trifluoracetic, heptafluorobutyric acids). The peptide-ion pair interaction had been well

studied for RP HPLC and used for modulating peptide retentivity and separation selectivity. This provided a degree orthogonality between RP separations using different ion-pairing reagents. Moving beyond formic acid for HILIC separation at acidic pH may further enhance the orthogonality for peptide separation in MDC analyses. Furthermore, it would be very attractive to explore higher pH HILIC separations where negatively charged residues would provide a major contribution into peptide retention. These efforts would lead to expanding chromatographic toolbox for peptide separation in analytical scale HPLC and might improve uptake of HILIC methods in bottom-up proteomics.

4.5 References

(1) Alpert, A.J. Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. *J. Chrom. A.*, **1990**, 499, 177-196.
(2) Hemstrom, P.; Irgum, K. Hydrophilic interaction chromatography. *J. Sep. Sci.*, **2006**, 29, 12, 1784-1821.
(3) McCalley, D.V. Understanding and manipulating the separation in hydrophilic interaction liquid chromatography. *J. Chrom. A.*, **2017**, 1523, 49-71.
(4) Jandera, P. Stationary and mobile phases in hydrophilic interaction chromatography: a review. *Anal. Chim. Acta*, **2011**, 692, 1-25.
(5) Yoshida, T. Peptide Separation in Normal Phase Liquid Chromatography. Anal. Chem. 1997, 69, 3038-3043.
(6) Dinh, N.P.; Jonsson, T.; Irgum, K. Water uptake on polar stationary phases under conditions for hydrophilic interaction chromatography and its relation to solute retention. *J. Chrom. A.*, **2013**, 1320, 33-47.
(7) Dinh, N.P.; Jonsson, T.; Irgum, K. Probing the interaction mode in hydrophilic interaction chromatography. *J. Chrom. A.*, **2011**, 1218, 35, 5880-5891.
(8) Kohler, I.; Verhoeven, M.; Haselberg, R.; Gargano, A.F.G. Hydrophilic interaction chromatography – mass spectrometry for metabolomics and proteomics: state-of-the-art and current trends. *Microchem. J.* **2022**, 175, 106986.
(9) Naidong, W. Bioanalytical liquid chromatography tandem mass spectrometry methods on underivatized silica columns with aqueous/organic mobile phases. *J. Chrom. B.*, **2003**, 796,2, 209-224.
(10) Simon, R.; Enjalbert, Q.; Biarc, J.; Lemoine, J. Salvador, A. Evaluation of hydrophilic interaction chromatography (HILIC) versus C18 reversed-phase chromatography for target quantification of peptides by mass spectrometry. *J. Chrom. A.*, **2012**, 1264, 31-39.
(11) Gama, M.R.; Bottoli, C.B.G. Hydrophilic Interaction Liquid Chromatography of Small Molecules, in: Meyers, R.A. (Eds), Encyclopedia of Analytical Chemistry: Applications, Theory,

and Instrumentation, Wiley, New Jersey, 2018, https://doi.org/10.1002/9780470027318.a9385.pub2.
(12) Lenco, J.; Jadeja, S.; Naplekov, D.K.; Krokhin, O.V.; Khalikova, M.A.; Chocholous, P.; Urban, J.; Broeckhoven, K.; Novakova, L.; Svec, F. Reversed-Phase Liquid Chromatography of Peptides for Bottom-Up Proteomics: A Tutorial, *J. Proteome Res.*, **2022**, 21, 12, 2846-2892.
(13) Alpert, A.J. HILIC at 21: Reflections and perspective. *J. Chrom. A.*, **2011**, 1218, 35, 5879.
(14) Horie, K.; Ikegami, T.; Hosoya, K.; Saad, N.; Fiehn, O.; Tanaka, N. Highly efficient monolithic silica capillary columns modified with poly(acrylic acid) for hydrophilic interaction chromatography. *J. Chrom. A.*, **2007**, 1164, 198-205.
(15) Horie, K.; Kamakura, T.; Ikegami, T.; Wakabayashi, M.; Kato, T.; Tanaka, N.; Ishihama, Y. Hydrophilic Interaction Chromatography Using a Meter-Scale Monolithic Silica Capillary Column for Proteomics LC-MS. *Anal. Chem.*, **2014**, 86, 8, 3817-3824.
(16) Di Palma, S.; Boersma, P.J.; Heck, A.J.R.; Mohammed, S. Zwitterionic hydrophilic interaction liquid chromatography (ZIC-HILIC and ZIC-cHILIC) provide high resolution separation and increase sensitivity in proteome analysis. *Anal. Chem.* **2011**, 83, 9, 3440-3447.
(17) Yu, L.W.; Floyd, T.R.; Hartwick, R.A. The synthesis and characterization of chemically bonded silica-based zwitterion-exchangers for HPLC. *J. Chromatogr. Sci.* **1986**, 24, 5, 177-182.
(18) Jiang, W.; Irgum, K. Synthesis and Evaluation of Polymer-Based Zwitterionic Stationary Phases for Separation of Ionic Species. *Anal. Chem.* **2001**, 73, 9, 1993-2003.
(19) Boersma, P.J.; Mohammed, S.; Heck, A.J.R. Hydrophilic interaction liquid chromatography (HILIC) in proteomics. *Anal. Bioanal. Chem.*, **2008**, 391, 151-159.
(20) Boersma, P.J.; Divecha, N.; Heck, A.J.R.; Mohammed, S. Evaluation and Optimization of ZIC-HILIC-RP as an Alternative MudPIT Strategy. *J. Proteome Res.* **2007**, 6, 3, 937-946.MDC for proteomics
(21) Di Palma, S.; Mohammed, S.; Heck, A.J.R. ZIC-cHILIC as a fractionation method for sensitive and powerful shotgun proteomics. *Nat. Protoc.* **2012**, 7, 2041-2055.
(22) Wolters, D.A.; Washburn, M.P., Yates, J.R. An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics. Anal. Chem. 2001, 73, 23, 5683-5690.
(23) Longworth, J; Noirel, J.; Pandhal, J.; Wright, P.C.; Vaidyanathan, S. HILIC- and SCX-based quantitative proteomics of Chlamydomonas reinhardtii during nitrogen starvation induced lipid and carbohydrate accumulation. *J. Proteome Res.*, **2012**, 11, 12, 5959-5971.
(24) Intoh, A.; Kurisaki, A.; Fukuda, H.; Asashima, M. Separation with zwitterionic hydrophilic interaction liquid chromatography improves protein identification by matrix-assisted laser desorption/ionization-based proteomic analysis. *Biomed. Chromatogr.*, **2009**, 23, 6, 607-614.
(25) Gilar, M.; Olivova, P.; Daly, A.E.; Gebler, J.C. Orthogonality of separation in two-dimensional liquid chromatography. *Anal. Chem.*, **2005**, 77, 19, 6426-6434.
(26) Yeung, D.; Mizero, B.; Gussakovsky, D.; Klaassen, N.; Lao, Y.; Spicer, V; Krokhin, O.V. Separation Orthogonality in Liquid Chromatography-Mass Spectrometry for Proteomic Applications: Comparison of 16 Different Two-Dimensional Combinations. *Anal. Chem.*, **2020**, 92, 5, 3904-3912.
(27) Bensaddek, D.; Nicolas, A.; Lamond, A.I. Evaluating the use of HILIC in large-scale, multi dimensional proteomics: Horses for courses? *Int. J. Mass Spectrom.* **2015**, 30, 391, 105-114.
(28) Yoshida, T. Calculation of peptide retention coefficients in normal-phase liquid chromatograph. J. Chrom. A. 1998, 808, 105-112.
(29) Gilar, M.; Xie. H.; Jaworski, A.; Utility of Retention Prediction Model for Investigation of Peptide Separation Selectivity in Reversed-Phase Liquid Chromatography: Impact of

Concentration of Trifluoroacetic Acid, Column Temperature, Gradient Slope and Type of Stationary Phase. *Anal. Chem.*, **2010**, 82, 1, 265-275.
(30) Hoarscoat-Schiavo, C.; Nioi, C.; Ronat-Heit, E.; Paris, C.; Vanderesse, R.; Fournier, F.; Marc, I. Hydrophilic properties as a new contribution for computer-aided identification of short peptides in complex mixtures. *Anal. Bioanal. Chem.*, **2012**, 403, 1939-1949.
(31) Krokhin, O.V.; Ezzati, P.; Spicer, V. Peptide Retention Time Prediction in Hydrophilic Interaction Liquid Chromatography: Data Collection Methods and Features of Additive and Sequence-Specific Models. *Anal. Chem.* **2017**, 89, 5526-5533.
(32) Yeung, D.; Klaassen, N.; Mizero, B.; Spicer, V.; Krokhin, O.V. Peptide retention time prediction in hydrophilic interaction liquid chromatography: Zwitterionic sulfoalkylbetaine and phosphorylcholine stationary phases. *J. Chrom. A.*, **2020**, 1619, 460909.
(33) Spicer, V.; Krokhin, O.V. Peptide retention time prediction in hydrophilic interaction liquid chromatography. Comparison of separation selectivity between bare silica and bonded stationary phases. *J. Chrom. A.*, **2018**, 1534, 75-84.
(34) Krokhin, O.V.; Spicer, V. Peptide Retention Standards and Hydrophobicity Indexes in Reversed-Phase High-Performance Liquid Chromatography of Peptides. *Anal. Chem.*, **2009**, 81, 22, 9522-9530.
(35) Villacres, C.; Spicer, V.; Krokhin, O.V. Confident Identification of Citrullination and Carbamylation Assisted by Peptide Retention Time Prediction. J. Proteome Res., 2021, 20, 3, 1571-1581.
(36) Krokhin, O.V.; Anderson, G.; Spicer, V.; Sun, L.; Dovichi, N.J. Predicting Electrophoretic Mobility of Tryptic Peptides for High-Throughput CZE-MS Analysis. *Anal. Chem.*, **2017**, 89, 3, 2000-2008.

Chapter 5 One-step proteome coverage enhancement for single-shot proteomics using Proteome-Selective Isolation Chromatography (P-SLICY)

This section contains text and figures from a paper submitted to *Analytical Chemistry* in 2024. **Darien Yeung**, Victor Spicer, and Oleg V. Krokhin.

DY designed and performed experiments, performed data analysis, writing, editing of the manuscript. VS performed data analysis, editing of the manuscript. OVK conceived and supervised the project, performed data analysis, writing, and editing of the manuscript.

Abstract

Improvements to the performance of mass spectrometers have increased the number of proteins identified and quantified in single shot proteomics (SSP). Approaches utilizing data-independent acquisition have shown promising results in detecting near-comprehensive proteomes. However, data dependent acquisition (DDA) can potentially be more sensitive with longer ion accumulation times for a precursor, allowing the detection of lower abundance peptides in complex samples. One of the problems with DDA is the inefficient ion selection method for MS2, favoring the detection of large and high abundance peptides that are likely to result in redundant identification of high abundance proteins. We developed a chromatographic pre-fractionation approach called Proteome Selective Isolation Chromatography to reduce redundant protein identification on a peptide level before SSP. By isolating the hydrophobic peptide population in the sample using the XBridge Amide HILIC column, we reduced the peptide level redundancy from a unique peptide per protein ratio of ~8 to ~2. This additional one-step pre-fractionation procedure improved identification by 20% and also allowed more proteins to be quantified. Furthermore, adapting the method to a pipette-tip based approach enhanced the consistency to a median coefficient of variation of 11%. Incorporating P-SLICY into an SSP workflow can enhance the detection and quantification of proteins with lower abundance.

5.1 Introduction

Constructing methods to optimize the detection of peptides and proteins has always been a critical area of study for proteomics. The overarching goal is to identify and quantify the complete proteome of a sample and ideally, have it done in the quickest way possible. Currently, liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) is the gold standard in proteomics for quantitation.[1] However, it is challenging for any approach to quantify the comprehensive human proteome from a single sample.[2] At the time of writing this article, no method has achieved this milestone. This is a challenging subject because potentially ~20,000 different proteins express over a concentration range along ≥8 orders of magnitude.[3] Such a large range surpasses the detectable range of MS systems operating with data dependent acquisition (DDA). Zubarev discussed that MS is limiting the detection of proteins to the abundant candidates.[4] Therefore, biologically relevant low abundance proteins like cytokines and chemokines are at a constant disadvantage for detection in modern LC-MS/MS methods.[5-7] Contemporary method development in bottom-up proteomics tries to increase the number of proteins identified through two main approaches: single shot proteomics (SSP) and multidimensional chromatography (MDC).

The concept of SSP is to identify and quantify the proteome using only one LC-MS/MS run. The protein identification statistics can be improved by increasing the number of MS/MS acquired that yield peptide identification. This is measured by the peptide-spectrum match (PSM) rate which is the ratio between the number of PSMs versus the total number of spectra collected. The distribution of peptide hydrophobicity values reaches a maximum in the middle of reversed-

phase LC (RP LC) separations. Therefore, using non-linear gradients can increase the total number of acquired MS/MS and enhance the PSM rate compared to a linear one with the same total acquisition time. Due to the efficacy of non-linear gradients, they are prevalent in SSP[8,9] and nanoflow LC systems like the EvoSep One use pre-programmed non-linear gradients for peptide separations.[10] Alternatively, the number of proteins identified positively correlates with an increasing total number of MS/MS acquired. This is a practical approach especially in the past with lower scan speed MS instrumentation. The easiest way to obtain more MS/MS in SSP is to extend the LC gradient to longer durations. In 2010, Iwasaki et al. demonstrated the whole proteome detection of ~2000 proteins in E. Coli using a 41 hour long gradient.[11] In 2011, Thakur et al. used an 8 hour long gradient for higher complexity mammalian samples, which resulted in ~5000 protein identifications.[12] Researchers in subsequent years used 10 hour gradients to achieve similar performances.[13,14] In 2019, Hebert et al.[15] paired long gradients with ion mobility filtration to detect 7900 proteins in 6 hours.

MS parameters optimization for SSP can increase the number of MS/MS and improve the quality of MS/MS spectra. Often, this involves parameters like automatic gain control and maximum ion injection times. Higher quality spectra can enhance the likelihood of peptide identification to increase the number of protein identifications. This type of MS optimization was applied to analyze complex mammalian samples where Pirmoradian et al.[16] in 2013 identified 4825 proteins in 4 hours. Recently in 2019, using newer instrumentation, Trujillo et al.[17] further revealed that narrowing the scan range for DDA can achieve ~5100 protein identifications in 1 hour. Considering which detector to use can also be critical. Some operators prefer ion-trap-based detection from Tribrid instruments to perform high scan speed acquisitions to boost the number of MS/MS collected.[15,17] Conversely, others use orbitrap-based detection to increase MS/MS quality

to improve protein identification.[16,18,19] The past five years have shown the popularization of data independent acquisition (DIA) where dia-PASEF readily detects ~7000 proteins in 1.5 hours[20] and detection of 10,000+ proteins in SSP has been achieved in less than 2 hours.[21,22]

Multidimensional chromatography (MDC) approaches are historically the general option to attain more protein identifications than SSP. This is done by dividing the peptides in a sample over multiple sub-fractions using chromatography. Therefore, the MS analysis of each sub-fraction will increase the total number of MS/MS acquired for the sample. Commonly, MDC uses two different types of chromatographic modes to separate peptides where the second mode is acidic pH RP online with the MS. Wolters et al. reported the use of strong cation exchange (SCX) to separate peptides in the first dimension in their online MUDPIT method.[23] Subsequently, Gilar et al. introduced the basic pH RP as an orthogonal first dimension to the second dimension acidic pH RP.[24] The use of basic pH RP further improved the number of proteins identified due to its superior orthogonality compared to SCX and is a common methodology used in MDC.[25] Our previous evaluation showed that using higher orthogonality first dimensions will increase the number of identified proteins. Deeper fractionations have also been explored by leveraging orthogonal chromatography[26-28] and isoelectric focusing[29] to achieve 10,000+ protein identifications from one sample. Although these methods are sensitive, they require 30-100+ hours of instrument time and large amounts of input samples.

Both SSP and MDC have been important to the development of proteomics, however, both methods cannot obtain the complete human proteome in one sample thus far. Based on the prior work from Picotti et al., the yeast peptides that cannot be observed from standard untargeted approaches were observed when using targeted MS approaches.[30] The sensitivity is further enhanced in orbitrap-based instruments that utilize prolonged C-trap isolation. When discussing

modern untargeted approaches, DIA and DDA are often compared against each other. Although DDA currently identifies fewer proteins than DIA in the same sample, there is a potential for DDA to observe deeper into the proteome because it can preferentially extend the C-trap isolation time for lower abundance peptides. A critical problem in DDA is the inefficient ion-selection algorithm that favors the detection of larger and higher abundance peptides. Resulting from this, it is not surprising to see protein entries in an SSP experiment that have >10 unique peptide observations. This redundancy is preventing the detection of peptides associated with lower abundance proteins. Therefore, a chromatographic pre-fractionation procedure prior to SSP could filter out some of the peptides that will contribute to redundant protein identification and reduce protein identification redundancies. In this work, we aim to determine if the redundancy reduction can yield an increase in the number of protein identifications. Here we outline the development of a one-step peptide-level redundancy reduction approach we call Proteome-Selective Isolation Chromatography (P-SLICY). Hydrophilic Interaction Liquid Chromatography (HILIC) was used for the method development – one of the several available options. This method is critical to serve as a proof-of-concept to understand how protein identification redundancies correlate with the number of proteins identified.

5.2 Experimental

5.2.1 Materials and Protein Digestions.

All chemicals and solvents for sample and mobile phase preparation were purchased from Thermo Fisher (Waltham, MA). Sequencing-grade modified trypsin was obtained from Promega (Madison,

WI) and used to digest proteins. The human whole cell digests were prepared from cultured human Jurkat cells using the SP3 procedure.

5.2.2 HILIC Fractionation and Pooling.

HPLC-based fractionation for HILIC was done using the Waters XBridge Amide, 3.5 µm, 200 Å, 3x50 mm column (Milford, MA). An Agilent 1100 series HPLC system was used for peptide separations and was equipped with a UV detector (214 nm). Samples were injected manually using a 100µL loop. Buffer A contained 10 mM ammonium formate in H_2O at pH 4.5 and buffer B contained 10 mM ammonium formate in 90% ACN/10% H_2O at pH 4.5. Column equilibration was done using 90% B at 300 µL/min, and the gradient used to separate the peptides was 1% water/minute increase starting from 90% B at 300 µL/min. A digest containing 100 µg of peptides was resuspended in 90% B and injected into the column for the 2D fractionation. Subsequently, the fractions were collected every minute. The fractions were then lyophilized, resuspended in 0.1% FA, and submitted to LC-MS/MS analysis.

5.2.3 HILIC P-SLICY Step Gradient and Peptide Enrichment.

Step-gradient pre-fractionation was also done using the same Water XBridge Amide column, buffer composition, and flow rate. The column was equilibrated using the target water concentration for the specific slicing point. A sample containing 10 µg of peptides was resuspended in the same target water concentration. Upon injection of the peptides into the column, the flow through peptides were collected. To prevent the isocratic elution of captured peptides, we decreased the concentration of water to 10% five minutes after injection. After the complete

collection of flow through peptides, the water concentration was increased to 91% water to elute the captured peptides.

5.2.4 HILIC P-SLICY Tips Packing and Loading Condition.

To obtain XBridge Amide 3.5 μm 200 Å stationary phase material, we unpacked a 3x50 mm column to isolate the loose sorbent to pack into tips. Leveraging the consistent pressure between Evosep Pure Tips (Evosep, Odense, Denmark), we loaded 1.26 mg of beads using LC-MS Optima Grade water as the medium. As the sample elution buffer has a composition of (23% water/77% ACN), flow through peptides were not retained on the C18; thus, the C18 pad acts as a frit. As the flow rate on HPLC for XBridge Amide fractionation is 300 μL/minute, we empirically determined that 1.26 mg of sorbent packed onto an EvoSep Pure Tip will require 9000 RCF to achieve a similar flow rate. The loading buffer was 10% water/90% ACN with 10 mM ammonium formate at pH 4.5, and the elution buffer was 23% water/77% ACN with 10 mM ammonium formate at pH 4.5, both made fresh on the day of P-SLICY isolation. To equilibrate the tip, 150 μL of loading buffer was passed through at 9000 RCF for 30 seconds at room temperature and an expected 5 μL of residual buffer should remain to keep the sorbent wet. Lyophilized sample peptides were resuspended with loading buffer to a concentration of 0.5 μg/μL and 20 μL were loaded on the tip at 9000 RCF and the flow-through peptides were eluted with 10 μL of elution buffer and these peptides are the isolated P-SLICY portion of peptides used for analysis.

5.2.5 LC-MS/MS analysis.

Nanoflow LC-MS/MS analysis is done with an EASY-nLC 1200 coupled to an Orbitrap Exploris 480 (Thermo Fisher Scientific, Bremen, Germany). 1 µg of peptide (determined by nanodrop) was loaded on to a self-pack Luna C18(2) analytical column (3µm particle, 100µm ID x 300 mm). Buffer A was 0.1% formic acid in H_2O and buffer B - 0.1% formic acid in 80% ACN. A linear gradient at 300nL/min starting from 1%B was ramped to 45%B in 74 minutes followed by a 1 minute increase to 90% and was held at high ACN for 15 minutes. MS analysis was conducted in data dependent acquisition mode (DDA) with a cycle time of 1 second and dynamic exclusion of 15 seconds. MS1 scans were acquired at 60K resolution from 375-1500 *m/z* (AGC 300%, 50ms maximum ion injection time). Precursor ions with charges between +2 and +5 were selected for MS2 isolating at a *m/z* 1.6 window. MS2 spectra were collected at 30K resolution (AGC 100%, automated maximum ion injection time), and the precursor was fragmented at a normalized collision energy of 30%.

5.2.6 Database Search and Data Analysis.

The resulting instrument data files were searched using FragPipe v21 and the human SwissProt database (downloaded Oct 16, 2022; 20,328 target sequences) with decoys and contaminants added via the FragPipe utility. The MSFragger[31] search was done with precursor mass tolerances set to 20 ppm and fragment mass tolerances to 25 ppm with recalibration and parameter optimization selected. Tryptic cleavage was used with a peptide length range of 7-35 amino acids and a peptide mass of 500-5000. Up to two variable modifications were allowed per peptide, including Oxidation of Met (+15.9949 Da), Gln > pyro-Glu (-17.0265 Da), and Glu > pyro-Glu (-18.0106 Da).

Carbamidomethylation of Cys (+57.02146) as fixed modification. All other unspecified parameters are set to default. MSBooster[32] was used along with Percolator[33] and ProteinProphet[34] for identification validations and reporting values were filtered at a 1% false discovery rate (FDR). IonQuant[35] was used for MS1 based quantitation with match-between-runs enabled at 1% FDR and with the intensities normalized.

5.3 Results and Discussion

5.3.1 Individual fractions from MDC may identify more proteins than SSP.

When performing a standard 2D separation pairing basic pH RP with acidic pH RP with fraction concatenation,[36] we observed that individual fractions had more proteins identified than SSP. However, the number of MS/MS collected, PSMs obtained, and unique peptides identified in individual fractions are all lower than the values found in SSP without fractionating the sample. Figure 5.1 A shows an example of the protein identification statistics for 10 concatenated basic pH fractions analyzed on the Orbitrap HF-X.

Our analysis of the MDC fractions shows that there is an advantage to identifying more protein in samples that detect fewer unique peptides. This is likely due to reducing the number of unique peptides redundantly identifying the same proteins. To study the effects of protein identification redundancies, we define the metric *peptide/protein ratio* that relates the number of unique peptides to the number of unique proteins identified in a sample. We observed that each fraction from this 2D LC-MS/MS run averages a peptide/protein ratio of 2.4. In contrast, with the same LC-MS method on SSP with the unfractionated sample, we observed a 5.7 peptide/protein ratio. Hence, we concluded that decreasing the peptide/protein ratio can increase protein identification.

To understand how SSP's redundancy affects protein identification and how it translates to newer instrumentation, we obtained the MS metrics from a newer Orbitrap Exploris 480 SSP run to perform *in-silico* simulations. Based on a common SSP output, we can expect ~60,000 MS/MS in 90 minutes, with a 70% PSM rate and 78% of the PSMs map to a unique peptide resulting in ~33,000 unique peptides identifying ~4100 proteins. This results in a peptide/protein ratio of 7.7 for the newer instrument. In theory, within a 90-minute acquisition, a 2 peptide/protein ratio can result in ~16,000 proteins identified, assuming the MS is not restricted by sensitivity. At a peptide/protein ratio of 3, the same run can yield ~11,000 proteins, as shown in Figure 5.1 B. Based on these potential values, we seek a method which will allow reproducible reduction of total number of peptides, but still representing the entire population of proteins expressed in the proteome. We use a chromatographic pre-fractionation approach to achieve this and call this method Proteome Selective Isolation Chromatography (P-SLICY).

In our view, to design a pre-fractionation method that can decrease the protein identification redundancy, the requirements are:

(1) The pre-fractionation chromatographic mode should have high orthogonality with the second dimension acidic pH RP.
(2) The method should be fast at isolating peptides that reduce redundant protein identifications.
(3) The procedure should avoid desalting after pre-fractionation to save time and minimize non-specific peptide loss.

We employed Hydrophilic Interaction Liquid Chromatography (HILIC) using the XBridge Amide column based on the requirements. This chromatographic mode has higher orthogonality than basic pH RP,[25] and no desalting is required because the separation uses MS compatible ammonium formate. For a proof-of-concept, the simplest way to isolate a portion of the peptides is to split them based on their retention on the HILIC column. This chromatographic split can be

done in a single step before SSP. Hydrophobic peptides not retained on HILIC constitute the flow-through (FT) fraction. The remaining hydrophilic peptides adsorbed on the column belong to the captured (CAP) fraction. The number of peptides in the FT/CAP fractions can be varied by adjusting the water concentration in the HILIC eluant. We define this elution water concentration as the *slicing point* on the column. Since the basic pH RP concatenated fractions have an average peptide/protein ratio of 2.4, we tried to optimize the slicing point in HILIC to target a peptide/protein ratio between 2 and 3.

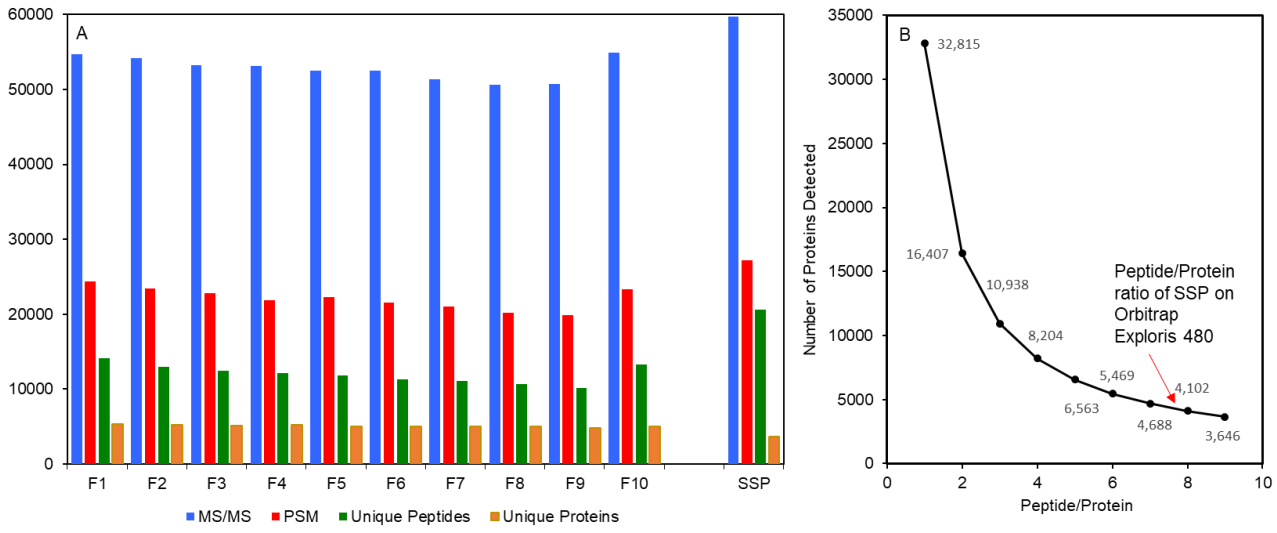


Figure 5.1: (A) MS and protein identification statistics for individual concatenated fractions from basic pH RP with acidic pH RP 2D LC-MS/MS and SSP experiments all performed on Orbitrap HF-X and (B) the simulated protein detection limits for 90 minute LC-MS/MS acquisition for different peptide/protein ratios on the newer Orbitrap Exploris 480.

5.3.2 Selecting slicing points via in-silico methods.

We used *in-silico* approaches to evaluate the slicing points that can yield our target peptide/protein ratio. The human proteome obtained from UniProt[37] was digested *in-silico* using trypsin cleavage rules with no missed cleavage. We then take the list of digested peptides to predict their retention time on the XBridge Amide HILIC column at pH 4.5 using SSRCalc.[38,39] Since the slicing points are expressed as the water concentration in the eluant, we need to convert the peptide retention time to the corresponding water concentration during elution. This conversion is done using the gradient slope of the HILIC 2D LC-MS/MS separation.

Peptides with a retention value below the slicing point belong to the FT fraction, and the remaining peptides belong to the CAP fraction. We perform this FT and CAP triage using the elution water concentration between 5-50% water to calculate the redundancy for the fractions at each simulated slicing point. The peptide-level redundancy is calculated by dividing the number of peptides in the fraction by the number of proteins identified with those peptide sequences. We use the CAP fraction at 5% water as our reference point for redundancy as this fraction has the highest peptide/protein ratio and contains all the peptides of the proteome in the CAP fraction. Therefore, this fraction is equivalent to an SSP run. The *in-silico* values are adjusted accordingly to have a maximum peptide/protein ratio of 7.7, which is the same ratio as SSP on the Orbitrap Exploris 480. Through this evaluation, slicing points of 22-24% water for FT and 30-33% for CAP gave the desired range of peptide/protein ratios between 2 and 3. To evaluate the accuracy of our *in-silico* prediction, we experimentally isolated both FT and CAP fractions eluting at 23, 25, 27, 30, and 33% water for validation.

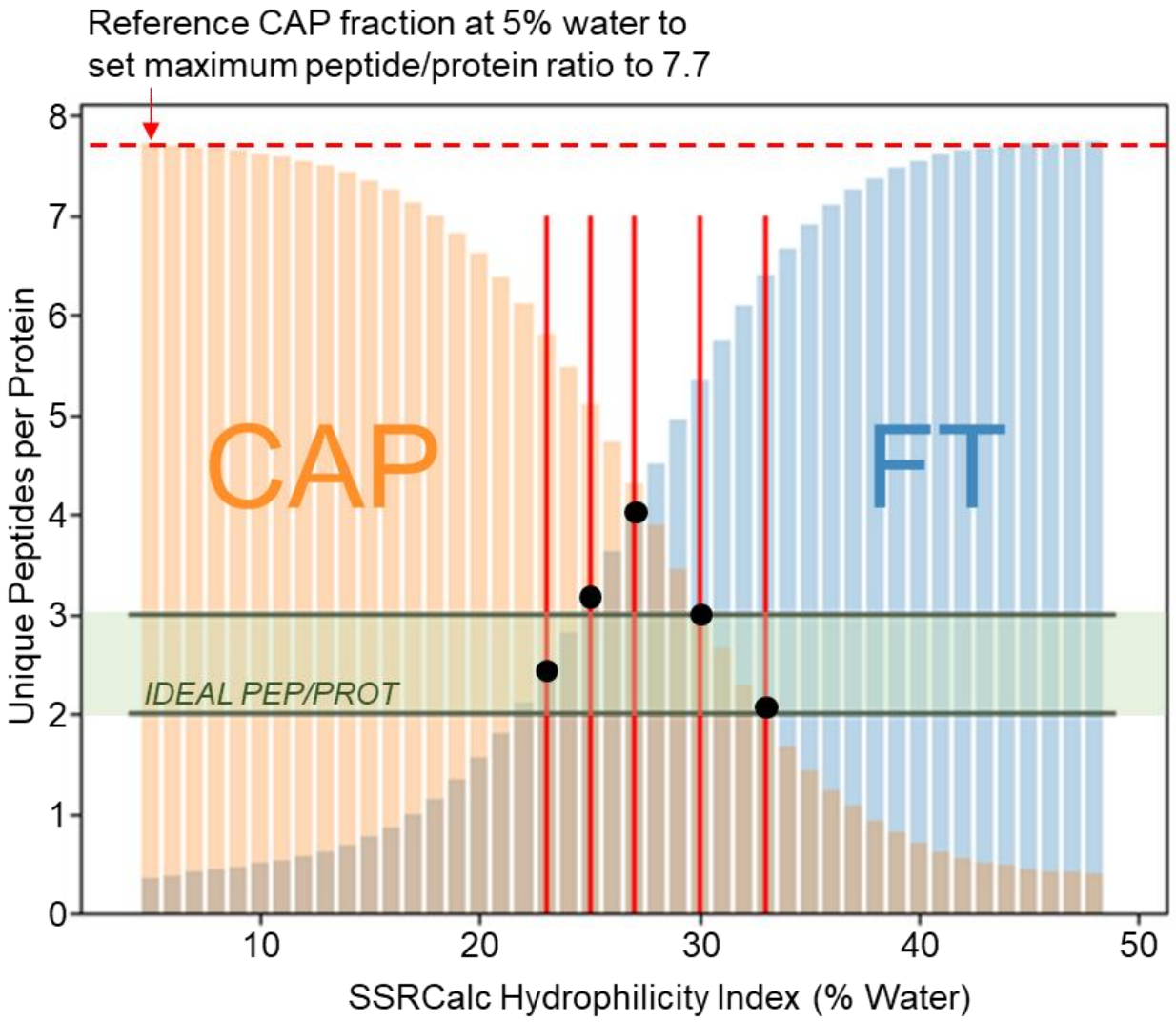


Figure 5.2: In-silico distribution of unique peptide/protein for FT (blue) and CAP (orange) fractions using each water percentage as a potential slicing point. The ideal peptide/protein region is highlighted in green between the peptide/protein ratio of 2 and 3. The selected slicing points for experiments are shown as the vertical lines at 23, 25, 27, 30, and 33% water.

5.3.3 Evaluation of slicing points using 2D fractionation techniques.

The peptide/protein ratios were evaluated at various slicing points by pooling fractions obtained from a standard 2D fractionation procedure using the XBridge Amide column. This was the most approachable option, as we can cleanly "slice" on a specific water concentration corresponding to a particular fraction number. The entire 2D LC-MS/MS experiment resulted in 24 fractions that yielded the identification of 8818 proteins and 157620 unique peptides. The correspondent first dimension chromatogram is shown in Figure 5.3 A.

Upon pooling the fractions to generate FT and CAP P-SLICY fractions (according to the scheme in Figure 5.3 B), the peptide/protein ratios were consistent with the predicted values from the *in-silico* evaluation. Table 5.1 lists the FT fractions at 23 and 25% water with the highest protein identifications and the lowest redundancy based on the peptide/protein ratio. Generally, the FT fraction had higher protein identifications than the CAP fraction of the same complexity, and this phenomenon was independent of the peptide/protein ratio. This is an interesting observation where a physicochemical stratification of peptides yields a fraction more efficacious for protein identification despite having a similar number of unique peptides identified. It is most evident in the CAP fraction at 30% water compared with the FT fraction at 27% water. With only a difference of 200 peptide identifications, this translates to identifying 800 more unique proteins in the FT fraction. Even though FT fractions are more hydrophobic than CAP fractions, the former generally have shorter peptides. It will be interesting to decipher the rationale for a higher number of protein identifications in FT fractions in follow up studies.

The *in-silico* peptide/protein ratios were comparable to the experimentally observed values. This confirms that the *in-silico* analysis is translatable to real experimental findings and is useful to streamline the design process for future P-SLICY studies with different chromatographic modes.

It is important to note that the trend of reducing redundancy increases protein identification for both FT and CAP fractions. Both redundancy reduction and peptide properties are critical. In the ideal scenarios found in FT fractions at 23 and 25% water, we can observe an increase of up to 19% in the number of proteins identified when we use P-SLICY rather than standalone SSP.

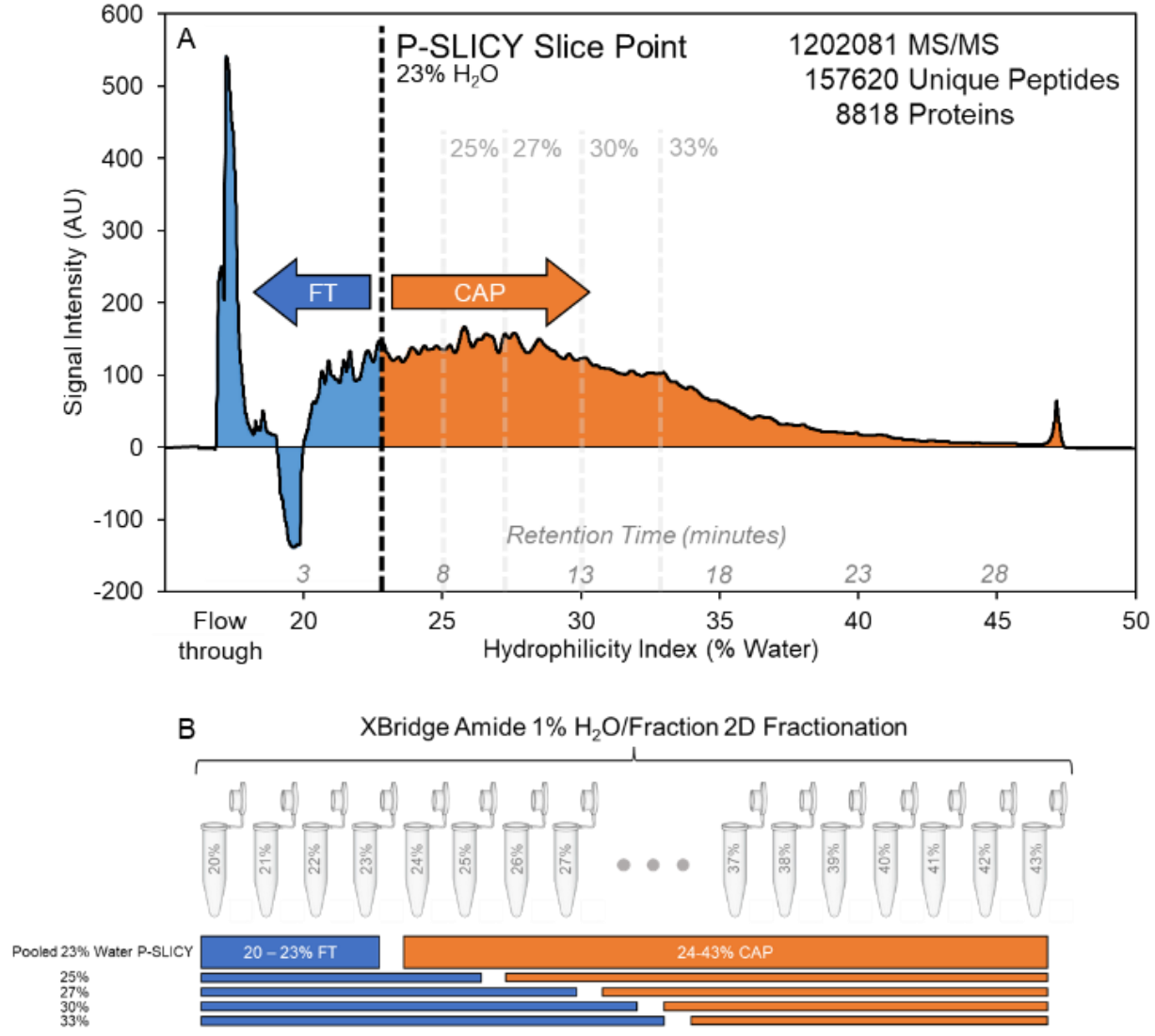


Figure 5.3: (A) 2D fractionation of peptides on XBridge Amide with regions of peptides collected for FT highlighted in blue and CAP in orange at a slice point of 23% water. Other slicing points are indicated in gray. (B) Pooling scheme for proof-of-concept P-SLICY experiments.

Table 5.1: Identification metrics of the P-SLICY fraction at varying slicing points from 2D LC-MS/MS of XBridge Amide in HILIC mode.

	Unique Peptides	**Proteins**	**Experimental Peptide/Protein**	**Predicted Peptide/Protein**
23% FT	12658	4981	2.54	2.06
25% FT	18728	5158	3.63	2.74
27% FT	21414	5027	4.26	3.52
30% FT	25971	4609	5.63	4.74
33% FT	28471	4483	6.35	5.89
23% CAP	26668	4049	6.59	6.14
25% CAP	26317	4116	6.39	5.49
27% CAP	25077	4227	5.93	4.77
30% CAP	21298	4293	4.96	3.56
33% CAP	16008	4258	3.76	2.42
SSP 1D	33373	4330	7.71	N/A

5.3.4 Protein identification performance of P-SLICY.

Having established the feasibility of redundancy reduction to boost protein identification, we designed a step-gradient procedure to consistently isolate peptides at desired slicing points. This should streamline the entire procedure. Injecting the sample at a certain water/ACN ratio will result in the elution of certain peptides in the FT fraction as expected. However, some peptides will be weakly retained on the column at this specific water concentration and elute as wide peaks under isocratic conditions. To avoid isocratic elution of these weakly retained CAP peptides, the water concentration in the eluant is decreased to strengthen the retention of CAP peptides on the column. Subsequently, the CAP peptides are then eluted at a higher water concentration.

To determine the reproducibility of step-gradient P-SLICY for enhancing protein identification, we analyzed the peptides isolated in the five slicing points in triplicate and obtained the values shown in Figure 5.4 A. The results from the step-gradient approach are comparable to the data from the fraction pooling analysis. Once again, the FT fractions at both 23 and 25% water had the highest protein identification enhancement when using step-gradients. This averaged a 13% increase in protein identification. The overlap of proteins observed between P-SLICY and SSP is good, covering ~90% of proteins detected in standalone SSP. The addition of P-SLICY allowed the identification of an additional 1636 proteins, demonstrated in Figure 5.4B, while missing 482 protein identifications.

From Figure 5.4C, the overall trend of reducing the peptide/protein ratio increases the number of proteins identified until a peptide/protein ratio at ~4. At peptide/protein ratios below 4, the trend either plateaus for FT fractions or descends for CAP fractions. The isolated fractions may largely contain peptides below the lower limit of identification and the narrow dynamic range of the instrument operating with DDA can still prevent successful protein identifications. P-SLICY is a pre-fractionation approach independent of MS. Therefore, we expect future improvements to the MS dynamic range with DDA to enhance the observable proteome with P-SLICY.

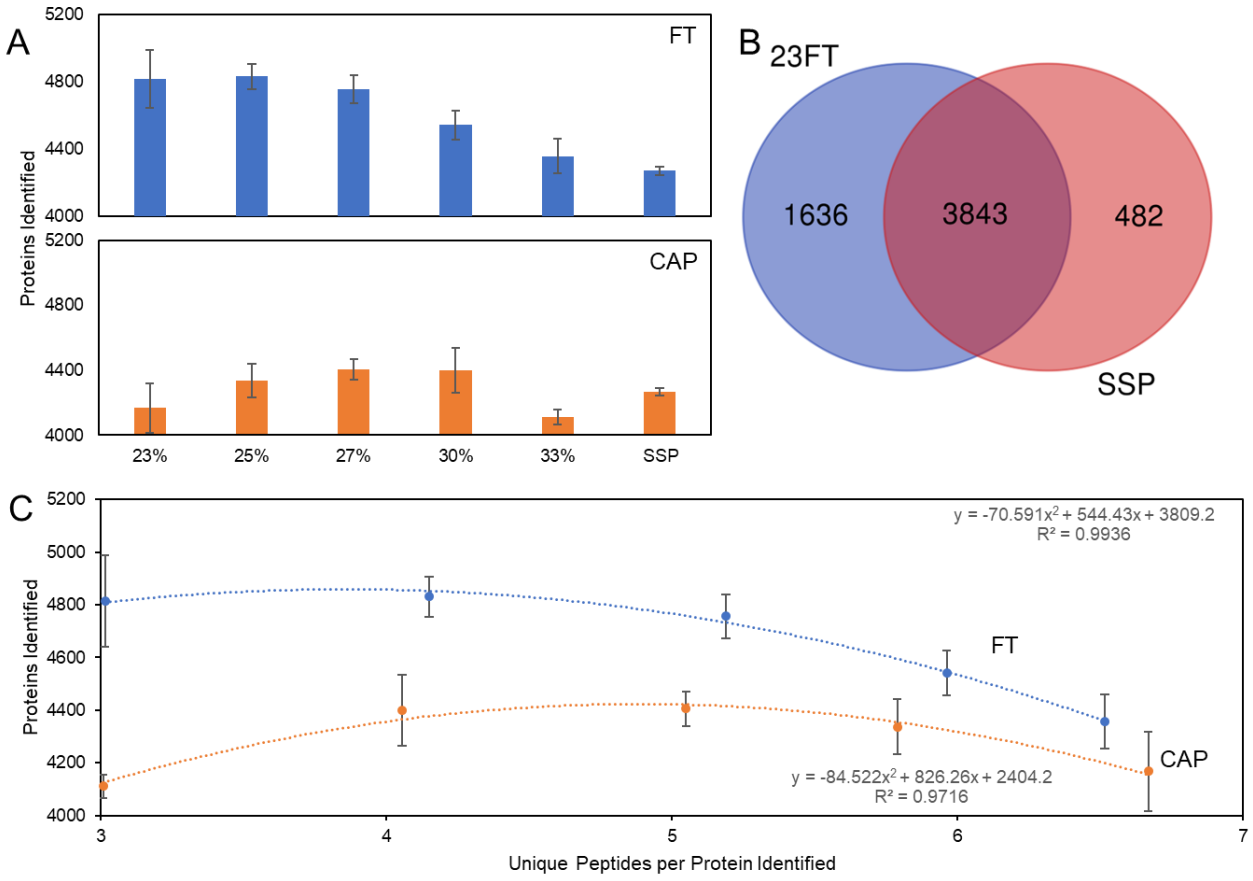


Figure 5.4: Identification performance of step-gradient P-SLICY over standalone SSP. (A) Number of proteins identified per replicate in FT (top) and CAP fractions (bottom). (B) Overlap of proteins identified between 23% water FT and standalone SSP. (C) Difference between protein identification and peptide/protein relation in FT and CAP fractions.

5.3.5 Quantitative stability of P-SLICY using HILIC Tips.

Addressing both feasibility and reproducibility with fraction pooling and step-gradient based P-SLICY, the next objective is throughput. In the current applications of proteomics, high sample numbers make high-throughput methods mandatory. We further enhanced our LC-based P-SLICY approach and adapted the technique to support tip-based separations through centrifugation. Based on concepts from StageTips[40] and MOLEX,[41] we constructed a 'HILIC tip' using bulk sorbent

from an unpacked XBridge Amide column. We packed the sorbent into EvoSep Pure C18 tips to improve the reproducibility of flow rates during centrifugation. The flow rates for the tip-based separation were optimized to match the analytical flow rate of the LC method. The tip-based separation was pursued using one condition. We determined the FT fraction at 23% water was the best condition to follow because it can achieve one of the highest numbers of protein identifications with the lowest peptide/protein ratio.

With the intent for HILIC tips to be applied for quantitation, we successfully achieved a median coefficient of variation (CV) of 11.4% for protein intensities across three replicates. To provide further context, the replicate injection of Jurkat from the same autosampler vial for standalone SSP achieved a median CV of 11.6% shown in Figure 5.5 A. This indicates that HILIC tips do not contribute significantly to inconsistencies during sample preparation and share the same variations as replicate injections. The protein identification statistics in tip-based P-SLICY are comparable to LC-based P-SLICY, averaging 5066 proteins with a peptide/protein ratio of 3.5, as shown in Figure 5.5 B. Furthermore, when we filter for proteins observed in at least 2 out of 3 experiments, the number of proteins identified in tip-based P-SLICY reduces to 4606. When the 2 peptides per protein quantification rule is included in the filtering, tip-based P-SLICY can yield 3839 protein identifications. In comparison, the standalone SSP, averages 4250 proteins identified with a peptide/protein ratio of 7.9, slightly higher than previously observed. The two quantitation filtering steps reduced the protein identification to 3502 proteins. Incorporating tip-based P-SLICY can obtain a 10% improvement over standalone SSP for quantifiable proteins. However, when we include confident protein identifications with only 1 peptide per protein, tip-based P-SLICY can offer a 19% enhancement in the number of proteins identified.

There are 768 proteins identified with only one peptide in tip-based P-SLICY. These proteins span a protein probability range of 0.69-1.00 based on the reports from ProteinProphet. All the peptides identifying these proteins have a probability score above 0.96, resulting in a 1% false discovery rate. Surprisingly, the median CV is 8.2% for this subset of proteins, meaning the extracted intensities are more consistent than typical quantifiable proteins. This can provide an additional ~650 proteins with CVs below 20% for quantitation. We find that for the potential quantitation of proteins with one peptide identification. the CV for a protein is more informative than the protein probability score. For example, the peptide ILTTVLK has a low protein probability score of 0.68, likely due to the peptide's size and limited protein coverage. However, upon performing peptide alignment using the peptide search function in UniProt on this sequence, only one SwissProt annotated protein entry was found, and it matches the protein accession from the MS identification.

Since ILTTVLK has little ambiguity about the protein it identifies, we should be able to use it for quantitation. The CV for this peptide is 4%, which is highly reproducible and better than most protein identifications filtered with the standard quantitation rules. Furthermore, this *m/z* and retention time associated with this peptide are observable in the extracted ion chromatogram from the standalone SSP runs. However, due to the narrow peak width and many co-eluting ions in the standalone SSP, this ion was never selected for MS/MS with the DDA algorithm. Therefore, P-SLICY is able to extract these lower abundance peptides for MS/MS and ultimately, quantitation. Using proteins with one peptide identification is crucial because they can identify proteins with lower abundance that are infrequently selected for MS/MS in standard DDA approaches, as shown in Figure 5.5 C. According to the Generalized Proteomics data Meta-analysis Database (GPMDB),[42] the protein associated with ILTTVLK is rarely observed and only has 598 LC-

MS/MS observations. Compared to a protein commonly observed with DDA in human cell lines with the annotation P78527, there are over ~119,000 LC-MS/MS observations. P-SLICY can enrich under-represented peptides that would not be selected by DDA by lowering the peptide-level redundancy.

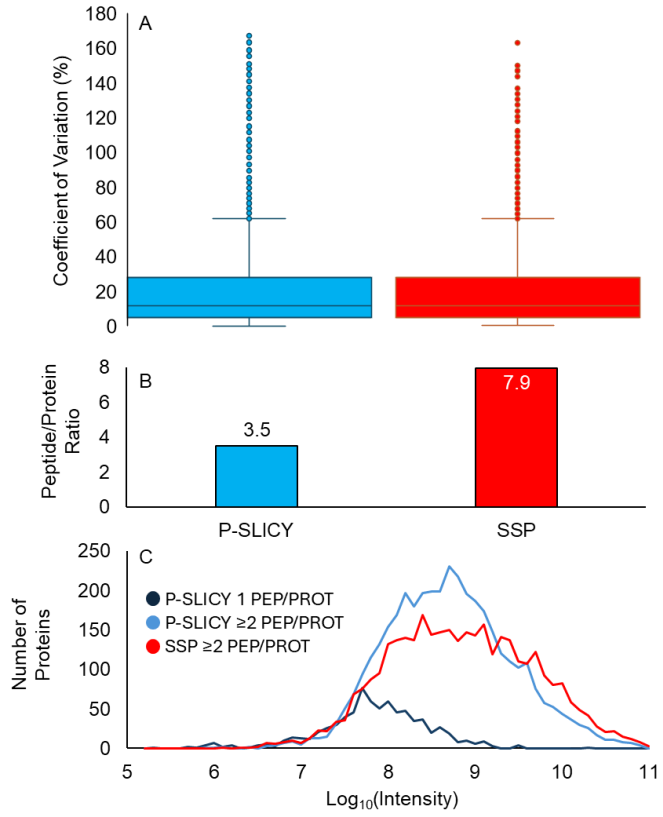


Figure 5.5: Reproducibility of peptide and protein quantification with tip-based P-SLICY. (A) Protein quantity variation for proteins with at least 2 peptide identifications between tip-based P-SLICY (blue) and standalone SSP replicates (red). (B) Average peptide/protein ratios for tip-based P-SLICY and SSP replicates. (C) The distribution of protein intensity values in tip-based P-SLICY for proteins with only 1 peptide identification (dark blue), with at least 2 peptide identifications (blue), and in SSP for proteins with at least 2 peptide identifications (red).

5.4 Conclusions

This manuscript introduces our method P-SLICY, a one-step chromatography-based pre-fractionation approach to enhance the number of proteins identified in an SSP experiment. P-SLICY achieves this improvement by reducing the number of peptides associated with redundant protein identifications, therefore lowering the peptide/protein ratio. We selected the XBridge Amide HILIC column at pH 4.5 to use as the chromatographic mode in P-SLICY because it had the highest orthogonality with the second dimension acidic pH RP. The peptides were split into two fractions in HILIC containing the flowthrough and captured peptide, respectively. Using step-gradients to isolate the FT and CAP fractions for P-SLICY, we found that the FT fractions with 23 and 25% water in the eluant identified the most proteins and had the lowest peptide/protein ratio. Both P-SLICY and standalone SSP have good protein overlap with ~90% of proteins identified in standalone SSP also found in P-SLICY. Incorporating P-SLICY provided an additional 19% of protein identified while reducing the sample's peptide/protein ratio from ~8 to ~2.

We improved the throughput of P-SLICY by adapting the method into a pipette tip format. Packing the HILIC sorbent into the tips allowed for consistent isolation of peptides in the FT fraction with 23% water in the eluant. P-SLICY enhances the number of quantifiable proteins by 9% when stringent rules are applied to filter for proteins with at least 2 peptide identifications. We demonstrated that P-SLICY has the potential for reproducible quantitation of proteins with only 1 peptide identification, which can be useful to investigate under-represented and low abundance proteins in differential expression. We will continue to explore this topic with alternative chromatographic approaches to see their efficacy and other experimental parameters that can increase the number of protein identifications for SSP. In this study, the FT fraction eluted with 23% water resulted in a 3.3-fold reduction in redundancy and this can potentially be achieved

isolating a portion of peptides from alternative chromatographic modes such as basic pH RP and strong anion exchange. It would be interesting to compare the resulting peptide and protein identification with different chromatographic approaches isolating a portion of the peptides in the proteome but may yield different peptide sets that may identify more unique proteins.

5.5 References

(1) Shuken, S.R. An Introduction to Mass Spectrometry-Based Proteomics. *J. Proteome Res.* **2023**, 22, 7, 2151-2171.
(2) Martinez-Val, A.; Guzman, U.H.; Olsen, J.V. Obtaining Complete Human Proteomes. *Annu. Rev. Genomics Hum. Genet.* **2022**, 23, 99-121.
(3) Schwanhausser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M. Global quantification of mammalian gene expression control. *Nature*, **2011**, 473, 337-342.
(4) Zubarev, R. The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics*, **2013**, 13, 5, 723-726.
(5) Whiteaker, J.R.; Lundeen, R.A.; Zhao, L.; Schoenherr, R.M.; Burian, A.; Huang, D.; Voytovich, U.; Wang, T.; Kennedy, J.J.; Ivey, R.G.; Lin, C.; Murillo, O.D.; Lorentzen, T.D.; Thiagarajan, M.; Colantonio, S.; Caceres, T.W.; Roberts, R.R.; Knotts, J.G.; Reading, J.J.; Kaczmarczyk, J.A.; Richardson, C.W.; Garcia-Buntley, S.S.; Bocik, W.; Hewitt, S.M.; Murray, K.E.; Do, N.; Brophy, M.; Wilz, S.W.; Yu, H.; Ajjarapu, S.; Boja, E.; Hiltke, T.; Rodriguez, H.; Paulovich, A.G. Targeted Mass Spectrometry Enables Multiplexed Quantification of Immunomodulatory Proteins in Clinical Biospecimens. *Front. Immunol.* **2021**, 12, 765898.
(6) Shi, T.; Song, E.; Nie, S.; Rodland, K.D.; Liu, T.; Qian, W.J.; Smith, R.D. Advances in targeted proteomics and applications to biomedical research. *Proteomics*. **2016**, 16, 15, 2160-2182.
(7) Kennedy, J.J.; Whiteaker, J.R.; Ivey, R.G., Burian, A.; Chowdhury, S.; Tsai, C.F.; Liu, T.; Lin, C.W.; Murillo, O.D.; Lundeen, R.A.; Jones, L.A.; Gafken, P.R.; Longton, G.; Rodland, K.D.; Skates, S.J.; Landua, J.; Wang, P; Lewis, M.T.; Paulovich, A.G. Internal Standard Triggered-Parallel Reaction Monitoring Mass Spectrometry Enables Multiplexed Quantification of Candidate Biomarkers in Plasma. *Anal. Chem.* **2022**, 94, 27, 9540-9547.
(8) Moruz, L.; Pichler, P.; Stranzi, T.; Mechtler, K.; Kall, L. Optimized Nonlinear Gradients for Reversed-Phase Liquid Chromatography in Shotgun Proteomics. *Anal. Chem.* **2013**, 85, 16, 7777-7785.
(9) Trudgian, D.C.; Fischer, R.; Guo, X.; Kessler, B.M.; Mirzei, H. GOAT – A simple LC-MS/MS gradient optimization tool. *Proteomics* **2014**, 14, 12, 1467-1471.
(10) Bache, N.; Geyer, P.E.; Bekker-Jensen, D.B.; Hoerning, O.; Falkenby, L.; Treit, P.V.; Doll, S.; Paron, I.; Muller, J.B.; Meier, F.; Olsen, J.V.; Vorm, O.; Mann, M. A Novel LC System Embeds

Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics. *Mol. Cell Proteomics.* **2018**, 17, 11, 2284-2296.
(11) Iwasaki, M.; Miwa, S.; Ikegami, T.; Tomita, M.; Tanaka, N.; Ishihama, Y. One-Dimensional Capillary Liquid Chromatographic Separation Coupled with Tandem Mass Spectrometry Unveils the *Escherichia coli* Proteome on a Microarray Scale. *Anal. Chem.* **2010**, 82, 7, 2616-2620.
(12) Thakur, S.S.; Geiger, T.; Chatterjee, B.; Bandilla, P.; Frohlick, F.; Cox, J.; Mann, M. Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol. Cell. Proteomics.* **2011,** 10, 8, M110.003699.
(13) Kocher, T.; Pichler, P.; Swart, R.; Mechtler, K. Analysis of protein mixtures from whole-cell extracts by single-run nanoLC-MS/MS using ultralong gradients. *Nat. Protoc.* **2012**, 7, 5, 882-890.
(14) Zhou, F.; Lu, Y.; Ficarro, S.B.; Webber, J.T.; Marto, J.A. Nanoflow low pressure high peak capacity single dimension LC-MS/MS platform for high-throughput, in-depth analysis of mammalian proteomes. *Anal. Chem.* **2012**, 5, 84, 11, 5133-5139.
(15) Hebert, A.S.; Prasad, S.; Belford, M.W., Bailey, D.J.; McAlister, G.C.; Abbatiello, S.E., Huguet, R.; Wouters, E.R.; Dunyach, J, Brademan, D.R., Westphall, M.S.; Coon, J.J. Comprehensive Single-Shot Proteomics with FAIMS on a Hybrid Orbitrap Mass Spectrometer. *Anal. Chem.* **2018**, 90, 15, 9529-9537.
(16) Pirmoradian, M.; Budamgunta, H.; Chingin, K.; Zhang, B.; Astorga-Wells, J.; Zubarev, R.A. Rapid and deep human proteome analysis by single-dimension shot proteomics. *Mol. Cell. Proteomics.* **2013**, 12, 11, 3330-3338.
(17) Trujillo, E.A.; Hebert, A.S.; Brademan, D.R.; Coon, J.J. Maximizing Tandem Mass Spectrometry Acquisition Rates for Shotgun Proteomics. *Anal. Chem.* **2019**, 91, 20, 12625-12629.
(18) Scheltema, R.A.; Hauschild, J.; Lange, O.; Hornburg, D.; Denisov, E.; Damoc, E.; Kuehn, A.; Makarov, A.; Mann, M. The Q Exactive HF, a Benchtop Mass Spectrometer with a Pre-filter, High performance Quadrupole and an Ultra-high-field Orbitrap Analyzer. *Mol. Cell. Proteomics.* **2014**, 13, 12, 3698-3708.
(19) Kelstrup, C.D.; Bekker-Jensen, D.B.; Arrey, T.N.; Hogrebe, A.; Harder, A.; Olsen, J.V. Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *J. Proteome Res.* **2018**, 17, 1, 727-738.
(20) Kawashima, Y.; Nagai,H.; Konno, R.; Ishikawa, M.; Nakajima, D.; Saito, H.; Nakamura, R.; Furuyashiki, T.; Ohara, O. Single-Shot 10K Proteome Approach: Over 10,000 Protein Identifications by Data-Independent Acquisition-Based Single-Shot Proteomics with Ion Mobility Spectrometry. *J. Prot. Res.* **2022**, 21, 6, 1418-1427.
(21) Demichev, V.; Szyrwiel, L.; Yu, F.; Teo, G.C.; Rosenberger, G.; Niewienda, A.; Ludwig, D.; Decker, J.; Kaspar-Schoenfeld, S.; Lilley, K.S.; Mulleder, M.; Nesvizhskii, A.I.; Ralser, M. dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. *Nat. Commun.* 2022, 3944.
(22) Guzman, U.H.; Martinez-Val, A.; Ye, Z.; Damoc, E.; Arrey, T.N.; Pashkova, A.; Renuse, S.; Denisov, E.; Petzoldt, J.; Peterson, A.C.; Harking, F.; Ostergaard, O.; Rydbirk, R.; Aznar, S.; Stewart, H.; Xuan, Y.; Hermanson, D.; Horning, S.; Hock, C.; Makarov, A.; Zabrouskov, V.; Olsen,

J.V. Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. *Nat Biotechnol*. **2024**, *https://doi.org/10.1038/s41587-023-02099-7*
(23) Wolters, D.A.; Washburn, M.P.; Yates, J.P. An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics. *Anal. Chem.* **2001**, 73, 23, 5683-5690.
(24) Gilar, M.; Olivova, P.; Daly, A.E.; Gebler, J.C. Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. *J. Sep. Sci.*, **2005**, 28, 14, 1694-1703.
(25) Yeung, D.; Mizero, B.; Gussakovsky, D.; Klaassen, N.; Lao, Y.; Spicer, V.; Krokhin, O.V. Separation Orthogonality in Liquid Chromatography–Mass Spectrometry for Proteomic Applications: Comparison of 16 Different Two-Dimensional Combinations. *Anal. Chem.* **2020**, 92, 5, 3904-3912
(26) Betancourt, L.H.; De Bock, P.; Staes, A.; Timmerman, E.; Perez-Riverol, Y.; Sanchez, A.; Besada, V.; Javier Gonzalez, L. Vandekerckhove, J.; Gevaert, K.; SCX charge state selective separation of tryptic peptides combined with 2D-RP-HPLC allows for detailed proteome mapping. *J. Proteom.* **2013**, 8, 91, 164-171.
(27) Spicer, V.; Ezzati, P.; Neustaeter, H.; Beavis, R.C.; Wilkins, J.A.; Krokhin, O.V. 3D HPLC-MS with Reversed-Phase Separation Functionality in All Three Dimensions for Large-Scale Bottom-Up Proteomics and Peptide Retention Data Collection. *Anal. Chem.* **2016**, 88, 5, 2847-2855
(28) Loroch, S.; Schommartz, T.; Brune, W.; Zahedi, R.P.; Sickmann, A. Multidimensional electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) for quantitative analysis of the proteome and phosphoproteome in clinical and biomedical research. *Biochim. Biophys. Acta.* **2015**, 1854, 5, 460-468.
(29) Branca, R.M.M.; Orre, L.M.; Johansson, H.J.; Granholm, V.; Haus, M.; Perez-Bercoff, A.; Forshed, J.; Kall, L.; Lehtio, J. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods*. **2013**, 11, 59-62.
(30) Picotti, P.; Bodenmiller, B.; Mueller, L.N.; Domon, B.; Aebersold, R. Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell*. **2009**, 138, 4, 795-806.
(31) Kong, A.T.; Leprevost, F.V.; Avtonomov, D.M.; Mellacheruvu, D.; Nesvizhskii, A.I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods.* **2017**, 14, 5, 513-520.
(32) Yang, K.L.; Yu, F.; Teo, G.C.; Li, K.; Demichev, V.; Ralser, M.; Nesvizhskii, A.I. MSBooster: improving peptide identification rates using deep learning-based features. *Nat. Commun.* **2023**, 14, 4539-4552
(33) Kall, L.; Canterbury, J.D.; Weston, J.; Noble, W.S.; MacCoss, M.J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, 4, 923-925.
(34) Nesvizhskii, A.I.; Keller, A.; Kolker, E.; Aebersold, R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Anal. Chem.* **2003**, 75, 4646-4658.

(35) Yu, F.; Haynes, S.E.; Nesvizhskii, A.I. IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. *Mol. Cell Proteomics.* **2021**, 20, 100077
(36) Dwivedi, R.C.; Spicer, V.; Harder, M.; Antonovici, M.; Ens, W.; Standing, K.G.; Wilkins, J.A.; Krokhin, O.V. Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics. *Anal. Chem.* **2008**, 15, 80, 18, 7036-7042.
(37) The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, 51, 523-531.
(38) Krokhin, O.V. Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-A pore size C18 sorbents. *Anal. Chem.* **2006**, 78, 22, 7785-7795.
(39) Krokhin, O.V.; Ezzati, P.; Spicer, V. Peptide Retention Time Prediction in Hydrophilic Interaction Liquid Chromatography: Data Collection Methods and Features of Additive and Sequence-Specific Models. *Anal. Chem.* **2017**, 89, 10, 5526-5533
(40) Rappsilber, J.; Mann, M.; Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2007**, 2, 1896-1906
(41) Nishida, H.; Ishihama, Y. One-Step Isolation of Protein C-terminal Peptides from V8 Protease-Digested Proteins by Metal Oxide-Based Ligand-Exchange Chromatography. *Anal. Chem.* **2022,** 94, 2, 944-951
(42) GPM, GPMDB, https://gpmdb.thegpm.org/

Chapter 6 Chromatographic behaviour of peptides modified with amine-reacting tags for relative protein quantitation in proteomic applications

This section contains text and figures from a paper published in *Journal of Chromatography A* in 2022, 1679, 463391. **Darien Yeung**, Geoffrey Anderson, Victor Spicer, and Oleg V. Krokhin.

DY performed data analysis, writing, editing of the manuscript. GA performed the experiments, VS performed data analysis, editing of the manuscript. OVK conceived and supervised the project, performed data analysis, writing, and editing of the manuscript.

Abstract

Reversed-phase (RP) HPLC separation of peptides labeled with amine-reacting tags for relative protein quantitation (iTRAQ4, iTRAQ8 - isobaric tag for relative and absolute quantitation, TMT - tandem mass tag) has been investigated using large-scale proteomics derived retention datasets. These tags have similar chemistry but use linkers of different length and hydrophobicity, moving the positively charged functional groups further from peptide backbone. Peptide hydrophobicity (RP HPLC retention), on average, increases in the following order: non-labeled < iTRAQ4 < iTRAQ8 < TMT under both low pH (0.1% formic acid) and pH 10 eluent conditions. At the same time, the interplay between hydrophobicity and length of the labeling group drives the deviations from this order. Thus, longer and less hydrophobic iTRAQ8 moiety results in greater retention increase for peptides carrying amphipathic helical structures at the N-terminus. Development of a peptide retention prediction models for these modifications was achieved by predicting correspondent retention shifts □HI (hydrophobicity index, % acetonitrile) between unmodified and labelled peptide pairs.

6.1 Introduction

Rapid advances in proteomics have dramatically expanded our capacity for protein/peptide analysis in recent years. Following initial method development geared towards protein identification, researchers are focusing more and more on detailed quantitative proteomics approaches.[1] Due to the inherent complexity of the samples and relatively low reproducibility of mass spectrometry (MS) detection, the use of isotopically coded compounds has emerged as the leading quantitation approach. Applications of isotopically coded tags such as iTRAQ and TMT are based on the same principle and received the most attention.[2,3] The procedure relies on the complete labeling of amino groups on the entire population of proteolytic peptides. These labels consist of two distinct regions (Figure 6.1 A) with unique distributions of isotopically labeled atoms, giving identical monoisotopic masses and chromatographic properties across the set of tags – i.e. they are indistinguishable under LC and MS1 measurements. Upon fragmentation by tandem MS, they produce fragments carrying qualitative (peptide backbone fragmentation) and quantitative features (reporter ion fragments). The number of the samples that can be compared (quantified) in a single assay varies depending on label structure and number of incorporated stable isotopes. Thus, 4-plex iTRAQ consists of four different labels with 114, 115, 116 and 117 Da reporter ion fragments. Currently there are 4 and 8-plex versions of iTRAQ,[4] while TMT has 6 and 11 plex versions (having identical structure and referred to as TMT in this work), 16 and 18 plex versions of TMTpro labels.[5-7] These four chemically distinct labels represent most labeling chemistries used in proteomics today.

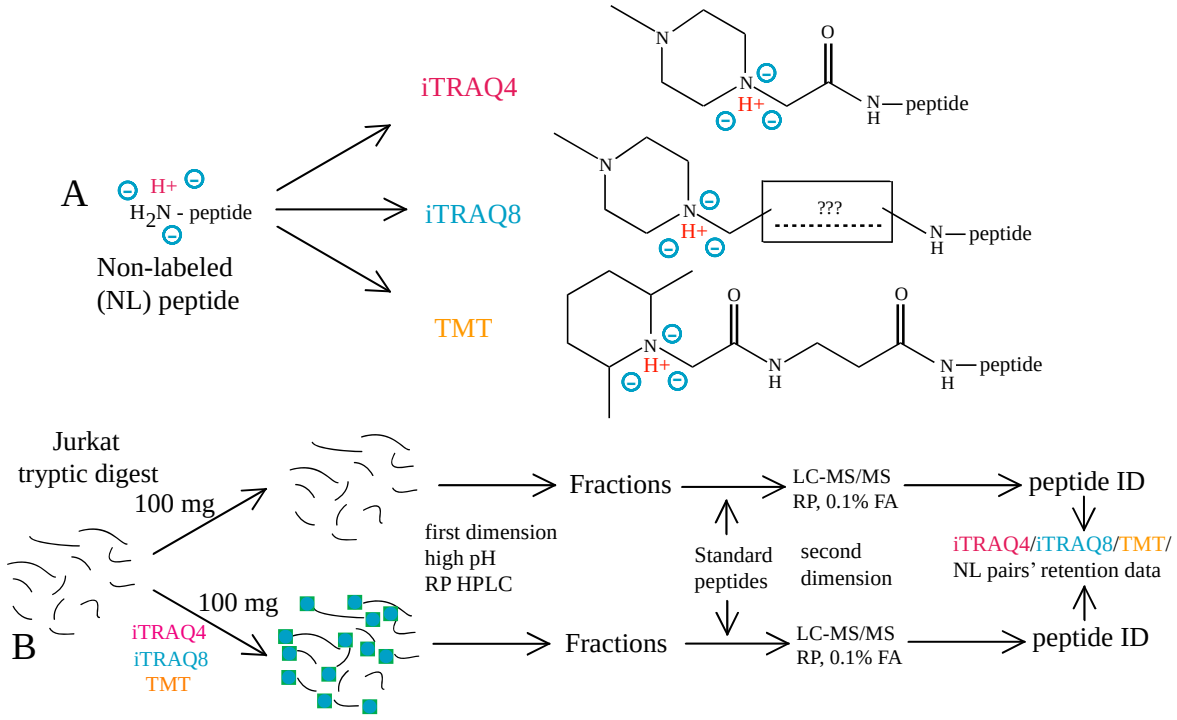


Figure 6.1: iTRAQ/TMT chemistry and experimental procedures for retention data collection. (A) chemical composition of iTRAQ/TMT labels (note that chemical structure of iTRAQ8 was not disclosed by its manufacturer); (B) experimental flow for 2D LC-MS/MS acquisitions. Optimization of collisional energy settings was done by RPLC-MS/MS (1D) analysis of ~1µg of the digests (Table 6.1).

Complete labeling targeting amino groups (N-terminal amines and side chains of Lysine) generates new populations of peptides, differing in their properties compared to original non-labeled digests. This requires additional efforts to evaluate changes in peptide mass-spectrometry and HPLC behaviour to adjust parameters of LC-MS acquisitions, if needed. So far, detailed studies of chromatographic properties of iTRAQ labeled peptides, in particular, are lacking. Tsai et al.[8] noted that TMT labeling improves detectability of hydrophilic phosphopeptides. Pfammatter et al.[9] found that TMT labeling helps improving detectability of MHC I (major histocompatibility

complex class 1) peptides due to increased chromatographic retention. Pichler and co-workers[10] have reported the following order of increased retention: iTRAQ 4-plex < iTRAQ 8-plex < TMT 6-plex and found that observed retention shifts are larger for more hydrophilic peptides. The first detailed study of sequence-dependent retention features of labeled peptides compared retention of non-labeled vs. TMT and TMTpro labeled pairs of peptides using more than 100,000 instances.[11] The authors found that retention shifts upon TMT labeling, while being positive on average, spanned a very wide range, including being negative for some species. These differences were explained by relocation of positive amino group charges upon labeling and the corresponding alteration of the ion-pairing environment. The authors also evaluated the chromatographic properties of TMT-labeled peptides in several alternative peptide separation modes: high pH RP, HILIC (hydrophilic interaction liquid chromatography) and SCX (strong-cation exchange).

Evaluation of chromatographic properties of new class of peptides can be achieved using peptide retention time prediction studies, which gain popularity in the past 20 years due to fast developments in proteomics.[12] Since peptide identifications in proteomic applications are based on statistical analysis of respective tandem MS spectra, it has been plagued by a significant rate of false positive identifications. Retention time prediction has been suggested as a secondary confirmation tool to improve confidence of peptide sequence assignments.[13] Following original developments of additive models in early 1980's[14,15] proteomic researchers used ever-growing high-quality retention dataset for advanced retention modeling. Artificial neural network based (PNNL NET, Pacific Northwest National Laboratory Normalized Elution Time),[16] our semi-empirical SSRCalc (Sequence-Specific Retention Calculator),[17] support-vector regression (SVR) based predictor ELUDE,[18] BioLCCC (Liquid Chromatography of Biomacromolecules at the Critical Conditions) based on physico-chemical modeling[19] are among other models reviewed and

compared by Tarasova et al.[20] This comparison performed in 2016 showed the best models reaching 0.96 R^2-value, which corresponds to prediction error $\Delta t_{95\%}$ ± 2.2% acetonitrile. In other words – retention of 95% of peptides can be predicted with accuracy of ± 2.2% acetonitrile for linear RP HPLC gradients. Further modifications of SSRCalc brought this accuracy to 0.98 R^2 ($\Delta t_{95\%}$ ± 1.5%).[11] The recent improvements in prediction accuracy have been driven by novel machine-learning retention prediction approaches.[21-23] Offering 0.99+ R^2 (($\Delta t_{95\%}$ ± 1%) prediction accuracy, these methods require extremely large datasets (300-400 thousand datapoints) for model training and provide no information on underlying features of retention mechanism, thus limiting their value for peptide separation science.

There is a need to fill the void of missing information on retention properties for iTRAQ labels. In this work we aimed to complete this task by studying the effects of iTRAQ 4 and 8-plex labeling on peptide behaviour in RPLC for proteomic applications. A standard TMT(6/11) was also included in this study for comparison. High pH RP – low pH RP is arguably the most popular 2D LC combination for high throughput proteomics[24] and was used for retention data collection.

6.2 Materials and Methods

6.2.1 Materials, protein digestion

All chemicals were analytical chemistry or LC-MS grade and sourced from Sigma Chemicals (St. Louis, MO) or Thermo Fisher Scientific (Waltham, MA) unless noted otherwise. LC-MS grade acetonitrile and water (Thermo Fisher) were used for preparation of the eluents. Sequencing grade modified trypsin (Promega, Madison, WI) was used for the digestion. Tryptic digests of human Jurkat cells were prepared using SP3 protocol as described elsewhere.[25] Cys residues were reduced

with dithiothreitol and alkylated with iodoacetamide (both from Sigma). The resulting mixture of tryptic peptides were purified using reversed-phase SPE prior to the labeling. iTRAQ 4 (115 version), iTRAQ 8 (121 version, Sciex, Vaughan, ON) and TMT (TMT0, Thermo Fisher) modification of 100 μg of the digested samples according to the manufacturers' protocols. Four different samples were analysed by identical 1D and 2D LC-MS/MS methods and referred in this work as NL (non-labeled), iTRAQ4, iTRAQ8, and TMT modified mixtures (Figure 6.1 B).

6.2.2. LC-MS analysis workflow

A home-packed 1 × 100 mm 5μm Xterra C18 (Waters, Milford, MA) was utilized for the first dimension separations in 2D experiments using Agilent 1100 series HPLC system (Agilent Technologies, Wilmington, DE) with manual 100 μL loop injector and UV detection at 214 nm. High pH RPLC separation employed 10 mM pH 10 ammonium formate as eluent additive with 0.75% ACN per min linear gradient at 150 μL/min.[24] Various number of 1 min fractions were collected based on UV elution profile starting from 10 min (gradient delay time) and concatenated pairwise: 40 (20 concatenated) NL, 42 (21) iTRAQ4, 42 (21) iTRAQ8, 52 (26) TMT. Fractions were lyophilised and dissolved in buffer A for RPLC-MS dimension.

Second dimension LC-MS settings featured sample loading through 100 μm × 5 mm Pep Map 100 trap-column (Thermo Fisher) and separation on 100 μm × 200 mm column packed with 3μm Luna C18 (2) (Phenomenex, Torrance, CA). 2D LC Ultra System (Eksigent, Dublin, CA) was used with both eluents A (water) and B (acetonitrile) containing 0.1% formic acid (Optima LC-MS, Thermo Fisher) as ion-pairing modifier. A linear water/acetonitrile gradient of 0.5 % per min was applied (0.5% to 40% in 79 min), followed by a column wash at 80% B for 5 minutes and a 6 min pre-equilibration step at 0.5% phase B. The samples containing ~1 μg of peptides were spiked

with 6 standard peptides[26] for the retention data alignment and conversion of retention times into HI (Hydrophobicity Index, % acetonitrile) units.

A Triple TOF5600 mass spectrometer (Sciex, Concord, ON) was used in a standard data-dependent acquisition mode with two different collision energy settings: normal and high recommended for the analysis of NL and iTRAQ labeled peptides, respectively.

6.2.3. Data analysis and peptide identifications

The X!Tandem search engine[27] was used for peptide identifications. Mass errors for parents and fragment ions were set to ± 20 ppm and 50 ppm, respectively. Search parameters included static Cys modification (+57.021 Da) along with iTRAQ4/iTRAQ8/TMT offset (+144.1/304.2/224.15 Da), and standard collection of potential (variable) post-translational modifications (PTMs). Additional searches omitting labeled settings have been done for all modified digests to determine completeness of modifications. All identifications with log (e) ≤ -1 were additionally filtered using SSRCalc retention prediction. Creation of the retention datasets used for the analysis of the chromatographic behaviour involved removing variable PTMs and reducing the peptides to single most-intense instances, giving 41833 NL, 17376 iTRAQ4, 29141 iTRAQ8 and 54627 TMT peptides.

6.3 Results and discussion

6.3.1. Peptide/protein identification in 1D and 2D LC-MS experiments.

The identification (ID) output in bottom-up proteomic runs is driven by multiple factors. RPLC, MS performance, consistency in sample preparation, and the completeness of labeling are among

the most important ones. All RPLC separations in this work have been performed using identical conditions where separation efficiency and reproducibility of water/acetonitrile gradients have been controlled using spikes of standard peptides. Given the consistent performance of the Triple TOF 5600 mass spectrometer, the only MS parameter affecting identification output was collision energy settings. High collision energy setting is recommended by MS vendor for the analysis of iTRAQ labeled peptides, leading to increased intensity of reporter ions and better quantitation. At the same time, this reduces the abundance of the heavier peptide backbone fragment ions, which may lead to poorer identification output. To test this hypothesis, we performed 1D LC-MS/MS analyses using both normal and high collision energy settings. Table 6.1 shows that higher collision energy results in a consistent ~10% reduction in the number of identified peptides, so we opted to use normal collision energy settings for the 2D LC-MS/MS acquisitions to maximise identification rate.

Table 6.1 shows that identification outputs increased in the order: iTRAQ4 < iTRAQ8 < NL < TMT. This wide variability in the number of peptides maybe deceiving, as it is affected by multiple factors: number of fractions, completeness of the labeling, and the isotopic purity of the label. The latter benefits NL and TMT analyses, since we used a TMT0 reagent without stable isotopes designed for testing sample preparation protocols. The lowest ID output for iTRAQ 4 was observed mostly due to insufficient labeling: we saw ~16% of the peptides being unlabeled, compared to ~3% for iTRAQ8 and TMT. Increased dynamic range during 2D LC-MS/MS resulted in even larger proportion of low abundance unlabeled peptides.

The most informative evidence of the effect of labeling on ID output can be seen by comparing X!Tandem scores of identical peptides assigned in the same charge states (Figure 6.2 A). Analysing ~4,900 species detected across all four 2D LC-MS/MS runs we conclude that confidence of

identification increases in the order: NL < iTRAQ4 < iTRAQ8 < TMT. The interplay of these factors, plus increased peptide retentivity following modifications determined variations in observed ID performance. It should be noted, however, that Triple TOF 5600 mass spectrometer settings do not allow optimization of collision energy for individual labeling chemistry. Therefore, performing similar measurement using another mass spectrometer may result in somewhat different results. At the same time, simple fact that collision energy settings optimized for iTRAQ favor identification of TMT-labeled peptides confirm a general trend shown in Figure 6.2 A.

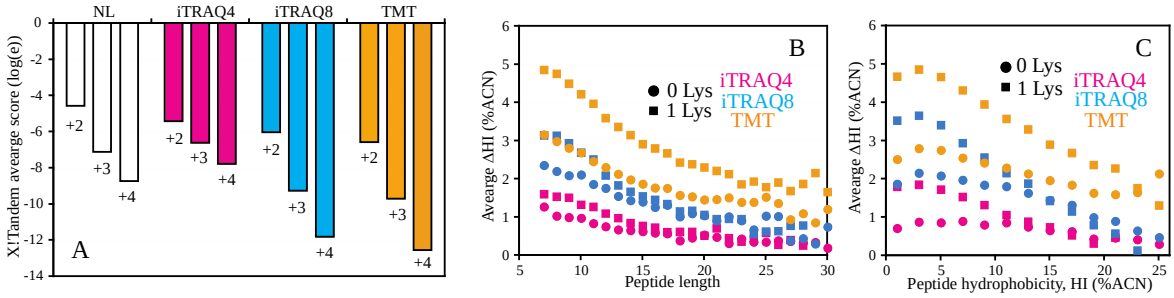


Figure 6.2: The effect of labeling on MS/MS identification and chromatographic behavior of peptides. (A) variation of average X!Tandem identification scores depending on the labeling status (lover log (e) values correspond to more confident identifications); (B, C) the effect of peptide length and hydrophobicity on RP HPLC retention shift following modification, respectively.

Table 6.1: Identification output for 1D and 2D LC-MS/MS acquisitions

Run	# of MS/MS	# of peptide IDs	# of unique peptide IDs	# of Protein ID	Unlabeled peptides %
1D LC-MS/MS					
NL, normal CE	17559	12007	8278	1870	

NL, high CE	17657	10747	7434	1683	
iTRAQ4, normal CE	15862	4598	3774	1019	16.4
iTRAQ4, high CE	15949	4261	3462	917	16.2
iTRAQ8, normal CE	18109	5047	3247	832	3.4
iTRAQ8, high CE	18165	4567	2815	719	3.2
TMT, normal CE	19309	10020	6963	1429	2.7
TMT, high CE	19211	10163	6935	1403	2.4
2D LC-MS/MS					
NL (20 fractions)	146178	94762	43841	6189	
iTRAQ4 (21 fractions)	188644	29389	19493	4158	48
iTRAQ8 (21 fractions)	215323	74501	30629	4586	7.4
TMT (26 fractions)	261693	127706	56562	6526	5

6.3.2. Overall effect of iTRAQ/TMT labeling on chromatographic behaviour of tryptic peptides.

Analysis of the chromatographic properties using high-throughput proteomic methods is affected by the size of retention dataset. The four 2D LC-MS/MS acquisitions performed in this work has resulted in different number of identifications, therefore datasets of various size can be used for assigning retention shifts. Table 6.2 compares all labels using non-labeled / labeled peptide pairs in all three chemistries. Average retention increases in the following order: NL < iTRAQ4 (0.91) < iTRAQ8 (1.79) < TMT (2.69% acetonitrile) considering 4,901 identical peptide pairs. To confirm the consistency of our measurements, we also provide retention shifts for 24,218 and 143,993 NL-TMT pairs found in this study and our previous work.[11] It is interesting to note that this earlier

study was done for a different stationary phase (PepMap, RSLC C18, 2μm, 100Å, Thermo Fisher) and yields similar results.

Similar trends were found for high pH RPLC separation: iTRAQ4 (1.5), iTRAQ8 (2.0%), TMT (9.5 %) acetonitrile average retention shifts.

Table 6.2: RPLC retention shifts (ΔHI, %ACN) observed following iTRAQ/TMT labeling.

# of peptides	iTRAQ4*	iTRAQ8*	TMT*	TMT**	TMT***	TMTpro***
0 Lys (3199)	0.79	1.61	2.19	2.23 (10095)	2.27 (56580)	2.66 (46385)
1 Lys (1672)	1.25	2.11	3.6	3.54 (12454)	3.77 (73990)	4.37 (59684)
2 Lys (40)	1.42	2.69	4.48	4.28 (1579)	4.57 (11943)	5.30 (5613)
All (4911)	0.91	1.79	2.69	3.05 (24218)	3.26 (143993)	3.71 (111867)

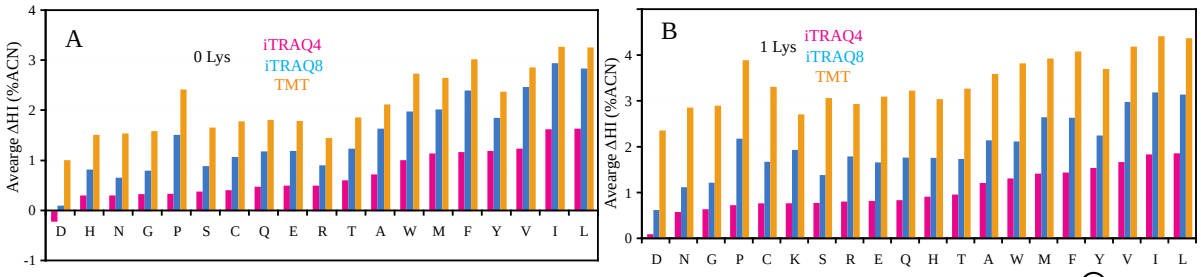
* - average retention shifts for 4911 identical peptide pairs detected for all three labels; ** - average retention shifts for 24,218 peptide pairs detected in NL and TMT runs; *** - TMT and TMTpro retention shifts from [11].

6.3.3. Composition- and sequence-dependent properties of labeled peptides affecting RPLC retention.

Similar to other peptide modifications, we monitored several composition and sequence-dependent features affecting peptide retention. Retention shifts (ΔHIs) decrease with peptide length and hydrophobicity (Figure 6.2 B, C). Considering that there is a correlation between peptide mass and hydrophobicity, we note that the character of these dependencies is slightly different. ΔHI values reduced uniformly by peptide length. ΔHI values peak for all labels at hydrophobicity ~3 HI units, except for single (N-terminal labeling) of iTRAQ4, which characterized by nearly constant average retention shifts within ~3-12 % acetonitrile range. Retention shifts decrease faster for peptides carrying two modification sites: N-terminus and lysine. For iTRAQ8 labeled peptides, those with

one Lys exhibit higher ΔHI compared to no Lys species below HI 15, and lower retention shifts are observed for more hydrophobic analytes.

Major sequence-dependent features explaining variation of retention shifts are driven by relocation of positively charged group following modifications (Figure 6.1A). This leads to alteration in ion-pairing environment and increased hydrophobic interactions of N-terminal amino acids, thus significant variability in retention shifts. For example, the ΔHIs in an identical set of 3199 peptides pairs containing no Lys vary wildly: iTRAQ4 (-1.24 – 5.5), iTRAQ8 (-1.23 – 7.47), TMT (-0.18 – 7.34 % acetonitrile). Figure 6.3 A, B shows the dependence of average retention shifts from the nature of N-terminal residue. Largest retention shifts are observed for compact hydrophobic residues Leu and Ile; smallest - for acidic Asp – in complete agreement with previously reported data for TMT and TMTpro.[11] Three Arg terminated peptides of similar length (8-10 residues) and hydrophobicity (9-10 HI units) were selected to demonstrate variations in retention shifts upon labeling (Figure 6.3 C). Due to relocation of hydrophilic counterions, the largest increase in retention is observed for IAQLICER featuring one turn amphipathic helical structure at N-terminus (Figure 6.3 D).



C

Sequence	HI (NL)	ΔHI iTRAQ4	ΔHI iTRAQ8	ΔHI TMT
DMMSDLTR	10.28	-1.05	-0.73	0.35
ICNQVLVCER	8.72	1.73	2.71	3.33
IAQLICER	8.59	3.66	6.65	5.98

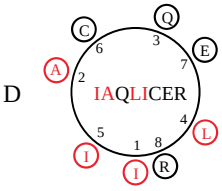


Figure 6.3: Sequence-dependent effects of iTRAQ/TMT labeling on RP HPLC peptide retention. (A, B) Average retention shifts depending on N-terminal residue for peptides with no Lys and one Lys residue, respectively; (C) Representative peptide sequences of similar hydrophobicity exhibiting various retention shifts upon labeling; (D) Axial helical projection of IAQLICER exhibiting anomalously large retention shifts.

Figure 6.4 A-C shows correlations of retention values between NL peptides and peptides modified with all three labels. While the largest average retention increase is typical for TMT, the largest variation (lowest R^2-value compared to NL peptides) is observed for iTRAQ8. To explain this, we inspected subset with largest ΔHI. While the iTRAQ4 < iTRAQ8 < TMT ordering holds across a general peptide population, these molecules usually elute in the order of iTRAQ4 < TMT < iTRAQ8. All of them feature hydrophobic residues at the N-terminal and at sequence positions 4 or 5; this leads to preferential alpha-helical arrangement upon interaction with C18 phase (Figure 6.3 D). In NL versions of these peptides stabilization of amphipathic helical structures is suppressed due to presence of hydrophilic formate counterions. Following charge relocation retention increases in accordance with the length of the linker. Another confirmation of this difference can be found by comparing the average shifts differences (Figure 6.3 A) between Leu

and Gly: iTRAQ4 (1.31), iTRAQ8 (2.04), TMT (1.67). Clearly, the iTRAQ8 label's linker is longer than that of TMT.

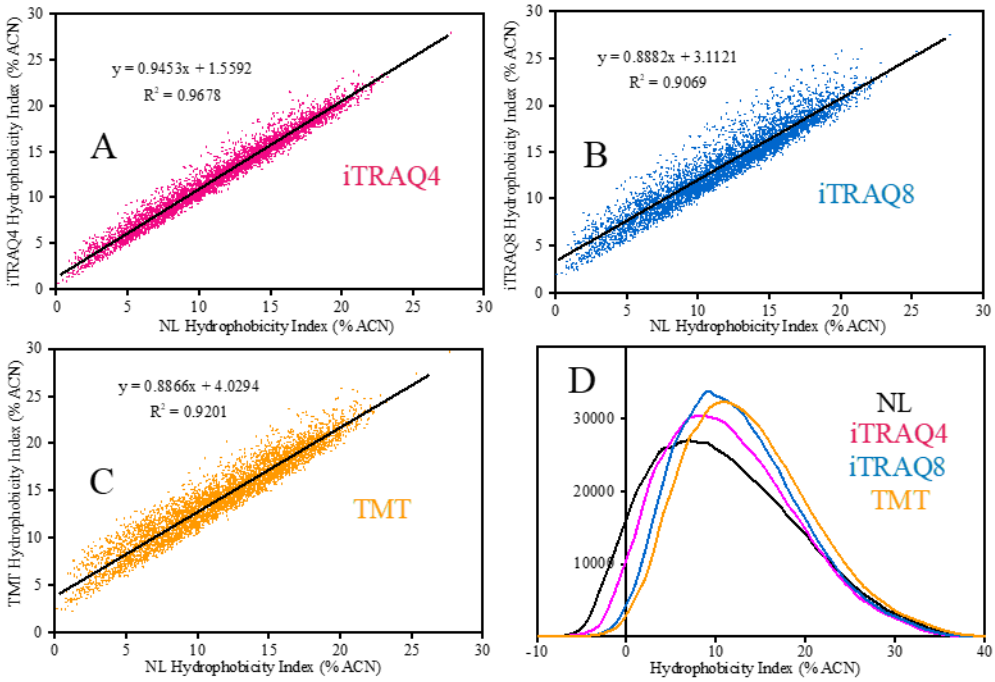


Figure 6.4: Variation of RPLC retention following iTRAQ/TMT labeling. (A-C) Hydrophobicity values correlations for various tags, compared to NL peptides; (D) Predicted distribution HI (% acetonitrile) values of no missed cleavage tryptic peptides 6 residues or longer from human proteome.

Three versions of the SSRCalc model have been developed to accommodate changes in retention upon labeling. This was achieved by predicting ΔHI retention shifts. Resulting hydrophobicity values were calculated as: $SSRCalcHI_{iTRAQ/TMT} = SSRCalcHI_{NL} + \Delta HI_{iTRAQ/TMT}$. Similar to the core SSRCalc for formic acid as ion pairing modifier, they yield ~0.98 R^2-value correlations ($\Delta t_{95\%}$ ± 1.5% acetonitrile), with the iTRAQ8 giving the lowest correction performance due to the larger contribution of alpha-helical stabilization.

6.4 Conclusions

Relative quantitation bottom-up proteomics is now in regular use in research facilities worldwide, meaning that very often complex peptide mixtures are no longer analyzed in their native form, but rather as modified versions where all amino groups are labeled. These modifications must be accounted for by adjustments to the chromatographic settings of acquisition methods. This study was focused mostly on iTRAQ chemistry, providing a complete coverage of the most popular isobaric tags for relative quantitation in proteomics. Retention of peptides in RPLC with formic acid as ion pairing modifier on average increases in the following order: NL < iTRAQ4 < iTRAQ8 < TMT < TMTpro. Variations in the hydrophobicity and length of the labels drive variations in separation selectivity. Thus, we find that while being less hydrophobic compared to TMT, the longer iTRAQ8 tags are characterized by relatively large retention shifts for hydrophobic N-terminal residues and for peptides carrying amphipathic helical sequences.

Figure 6.4 D shows the distribution of tryptic peptides from the human proteome across the entire RPLC gradient separation scale, based on our predictive SSRCalc models, confirming that all labels increase the detectability (hydrophobicity) of species, some of which are not retained in non-labeled format. Approximately 8.5% of non-labeled tryptic peptides (6 amino acids or longer) have HI < 0 and will not be retained under the normal RPLC conditions used in proteomics. This value reduces continuously across the labels: iTRAQ4 (4.2), iTRAQ8 (1.37), TMT (0.92). Contrary to the usual suggestion to increase the upper range of the gradient separation to ensure the elution of more hydrophobic iTRAQ/TMT modified peptides, we suggest maintaining the usual ~0-35% acetonitrile elution range, as the longer hydrophobic peptides show insignificant retention increase upon labeling. At the same time, we observe a less uniform distribution of analytes across the chromatogram (compressed peptide distribution) following the modification

and strongly recommend using non-linear gradients to make the delivery of peptides to the mass analyser more uniform.

6.5 References

(1) Ong, S.E.; Mann, M. Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.*, **2005**, 1, 252-262.
(2) Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A.K.; Hamon, C. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem*., **2003**, 75, 1895-1904.
(3) Ross, P.L.; Huang, Y.N.; Marchese, J.N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D.J. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics*, **2004**, 3, 1154-1169.
(4) Pottiez, G.; Wiederin, J.; Fox, H.S.; Ciborowski, P. Comparison of 4-plex to 8-plex iTRAQ quantitative measurements of proteins in human plasma samples. *J. Proteome Res.*, **2012**, 11, 3774-3781.
(5) Thompson, A.; Wolmer, N.; Koncarevic, S.; Selzer, S.; Bohm, G.; Legner, H.; Schmid, P.; Kienle, S.; Penning, P.; Hohle, C.; Berfelde, A.; Martinez-Pinna, R.; Farztdinov, V.; Jung, S.; Kuhn, K.; Pike, I. TMTpro: Design, Synthesis, and Initial Evaluation of a Proline-Based Isobaric 16-Plex Tandem Mass Tag Reagent Set. *Anal. Chem.*, **2019**, 91, 15941-15950.
(6) Li, J.; Van Vranken, J.G.; Pontano Vaites, L.; Schweppe, D.K.; Huttlin, E.L.; Etienne, C.; Nandhikonda, P.; Viner, R.; Robitaille, A.M.; Thompson, A.H.; Kuhn, K.; Pike, I.; Bomgarden, R.D.; Rogers, J.C.; Gygi, S.P.; Paulo, J.A. TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat. Methods*, **2020**, 17, 399-404.
(7) Li, J.; Cai, Z.; Bomgarden, R.D.; Pike, I.; Kuhn, K.; Rogers, J.C.; Roberts, T.M.; Gygi, S.P.; Paulo, J.A. TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing. *J. Proteome Res.*, **2021**, 20, 2964-2972.
(8) Tsai, C.F.; Smith, J.S.; Krajewski, K.; Zhao, R.; Moghieb, A.M.; Nicora, C.D.; Xiong, X.; Moore, R.J.; Liu, T.; Smith, R.D.; Jacobs, J.M.; Rajagopal, S.; Shi, T. Tandem Mass Tag Labeling Facilitates Reversed-Phase Liquid Chromatography-Mass Spectrometry Analysis of Hydrophilic Phosphopeptides. *Anal. Chem.*, **2019**, 91, 11606-11613.
(9) Pfammatter, S. ; Bonneil, E.; Lanoix, J.; Vincent, K.; Hardy, M.P.; Courcelles, M.; Perreault, C.; Thibault, P. Extending the Comprehensiveness of Immunopeptidome Analyses Using Isobaric Peptide Labeling. *Anal. Chem.*, **2020**, 92, 9194-9204.
(10) Pichler, P.; Kocher, T.; Holzmann, J.; Mazanek, M.; Taus, T.; Ammerer, G.; Mechtler, K. Peptide labeling with isobaric tags yields higher identification rates using iTRAQ 4-plex compared to TMT 6-plex and iTRAQ 8-plex on LTQ Orbitrap. *Anal. Chem.*, **2010**, 82, 6549-6558.
(11) Mizero, B.; Villacres, C.; Spicer, V.; Viner, R.; Saba, J.; Patel, B.; Snovida, S.; Jensen, P.; Huhmer, A.; Krokhin, O.V. Retention Time Prediction for TMT-Labeled Peptides in Proteomic LC-MS Experiments. *J. Proteome Res.*, **2022**, 21, 1218-1228.

(12) Moruz, L.; Kall, L. Peptide retention time prediction. *Mass. Spectrom. Rev.*, **2017**, 36, 615-623.
(13) Strittmatter, E.F.; Kangas, L.J.; Petritis, K.; Mottaz, H.M.; Anderson, G.A.; Shen, Y.; Jacobs, J.M.; Camp, D.G.; Smith, R.D. Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. *J. Proteome Res*., **2004**, 3, 760-769.
(14) Meek, J.L. Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. *Proc. Natl. Acad. Sci. USA*, **1980**, 77, 1632-1636.
(15) Guo, D.; Mant, C.T.; Taneja, A.K.; Parker, J.M.R.; Hodges, R.S. Prediction of peptide retention times in reversed-phase high-performance liquid chromatography. I. Determination of retention coefficients of amino acid residues of model synthetic peptides. *J. Chromatogr*., **1986**, 359, 499-517.
(16) Petritis, K.; Kangas, L.J.; Yan, B.; Monroe, M.E.; Strittmatter, E.F.; Qian, W.J.; Adkins, J.N.; Moore, R.J.; Xu, Y.; Lipton, M.S.; Camp, D.G.; Smith, R.D. Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Anal. Chem.*, **2006**, 78, 5026-5039.
(17) Krokhin, O.V. Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-A pore size C18 sorbents, Anal. Chem., **2006**, 78, 7785-7795.
(18) Moruz, L.; Tomazela, D.; Kall, L. Training, selection, and robust calibration of retention time models for targeted proteomics. *J. Proteome Res.*, **2010**, 9, 5209-5216.
(19) Gorshkov, A.V.; Tarasova, I.A.; Evreinov, V.V.; Savitski, M.M.; Nielsen, M.L.; Zubarev, R.A.; Gorshkov, M.V. Liquid chromatography at critical conditions: comprehensive approach to sequence-dependent retention time prediction. *Anal. Chem.*, **2006**, 78, 7770-7777.
(20) Tarasova, I.A.; Masselon, C.D.; Gorshkov, A.V.; Gorshkov, M.V. Predictive chromatography of peptides and proteins as a complementary tool for proteomics. *Analyst*, **2016**, 141, 4816-4832.
(21) Ma, C.; Ren, Y.; Yang, J.; Ren, Z.; Yang, H.; Liu, S. Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. *Anal. Chem.*, **2018**, 90, 10881-10888.
(22) Gessulat, S.; Schmidt, T.; Zolg, D.P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; Reimer, U.; Ehrlich, H.C.; Aiche, S.; Kuster, B.; Wilhelm, M. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods*, **2019**, 16, 509-518.
(23) Muller, J.B.; Geyer, P.E.; Colaco, A.R.; Treit, P.V.; Strauss, M.T.; Oroshi, M. ; Doll, S.; Virreira Winter, S.; Bader, J.M.; Kohler, N.; Theis, F.; Santos, A.; Mann, M. The proteome landscape of the kingdoms of life. *Nature*, **2020**, 582, 592-596.
(24) Dwivedi, R.C.; Spicer, V.; Harder, M.; Antonovici, M.; Ens, W.; Standing, K.G.; Wilkins, J.A.; Krokhin, O.V. Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics. *Anal. Chem.*, **2008**, 80, 7036-7042.
(25) Hughes, C.S.; Foehr, S.; Garfield, D.A.; Furlong, E.E.; Steinmetz, L.M.; Krijgsveld, J. Ultrasensitive proteome analysis using paramagnetic bead technology. *Mol. Syst. Biol.*, **2014**, 10, 757.
(26) Krokhin, O.V.; Spicer, V. Peptide retention standards and hydrophobicity indexes in reversed-phase high-performance liquid chromatography of peptides. *Anal. Chem.*, **2009**, 81, 9522-9530.

(27) Craig, R.; Beavis, R.C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **2004**, 20, 1466-1467.

Chapter 7 Concluding Remarks and Future Direction

7.1 Conclusion of thesis

In this thesis, I have described the development of Proteome-Selective Isolation Chromatography to selectively enrich a subset of peptides from the proteome. In contrast to prior peptide-centric approaches, we are selecting for a group of peptides within a chromatographic retention region that can identify a wide diversity of proteins with minimal redundancy rather than using amino acid specific enrichment. This was our strategy to improve protein identification and accelerate high-throughput proteomics. To determine a suitable chromatographic approach for selective peptide enrichment, in Chapter 2, we compared the peptide separation orthogonality between sixteen different chromatographic modes paired with RP LC-MS/MS. We identified XBridge Amide HILIC and ZIC-HILIC at pH 4.5 as suitable chromatographic modes due to their high separation orthogonality values and difference in separation mechanism with RP. We further characterized the peptide retention behaviour in these chromatographic systems using the sequence specific retention time calculator (SSRCalc) model.

We previously modelled the peptide retention properties on XBridge Amide HILIC[1] and here we implemented the same procedure for zwitterionic phases ZIC-HILIC and ZIC-cHILIC. In Chapter 3 and 4, we explored the peptide retention behaviour on ZIC-HILIC at pH 4.5 and also pH 2.7 using formic acid modifier in the mobile phase, since Boersema et al.[2] had suggested that pH $\leq$ 3 exhibited higher separation orthogonality with this sorbent. We found that changing the pH will dramatically change the retention properties where the overall apparent hydrophilicity of the columns decreases. Furthermore, peptides with more basic amino acids exhibited larger retention. Using these retention time prediction models, we were able to perform *in-silico* separation of

peptides derived from human whole cell lysate and determine ideal chromatographic regions to isolate analytes for selective peptide enrichment.

In Chapter 5, we demonstrated the development, optimization, and application of P-SLICY using XBridge Amide HILIC at pH 4.5. We reported an increase in protein identification as a result of the three-fold redundancy reduction on a peptide level that we measured by comparing the number of unique peptide identification as a ratio with number of unique protein identifications. Our P-SLICY technique was further adapted to a pipette tip-based format that improved consistency to allow for a 20% increase in number of quantifiable proteins and enhanced throughput of analysis. Using the P-SLICY XBridge Amide pipette tips, we accelerated the selective peptide enrichment workflow from 45 minutes per sample on an HPLC system to 5 minutes per sample for potentially 96 concurrent samples. With the aim to eventually integrate the isobaric tag-based quantitative workflows to P-SLICY, we modelled the retention shifts of peptides that have been modified with the amine reactive isobaric tags that we described in Chapter 6. The peptides exhibit an average increase in hydrophobicity of +4% ACN after labelling with isobaric tags. This indicates that in future studies involving the enrichment of tagged peptides using HILIC P-SLICY, we will need to change the chromatographic enrichment protocol to favor more hydrophobic peptides.

7.2 Significance and limitations of Proteome Selective Isolation Chromatography

The significance of our research with P-SLICY is our demonstration of overcoming the logarithmic relationship between the number of identified peptides and proteins. Furthermore, we are demonstrating that peptide-centric approaches are applicable to modern proteomic workflows

and specifically, we simplified the processes to a one-step enrichment process to allow 20% more quantifiable protein identifications.

However, there are limitations to enriching a subset of peptides for analysis. We can only observe the peptides that are in this subset for P-SLICY. In a scenario that requires the identification of a peptide that is not enriched in HILIC P-SLICY, then an orthogonal PSLICY method will be required, or the enrichment step may need to be removed from the workflow. Furthermore, P-SLICY may encounter dynamic range issues with certain bio-fluid samples such as plasma. The high dynamic range of plasma from the high abundance of albumin versus the low abundance of secreted chemokines will still affect P-SLICY's performance if the peptides belonging to the high abundance proteins are enriched in the analyzed fractions. If there is a way to avoid enriching the peptides belonging to high abundance proteins, then this problem can be ameliorated. This will require further optimization of P-SLICY in terms identifying unanimous physiochemical properties that are enriched in peptides of high abundance proteins and changing the HILIC column for another type of stationary phase that can favor this enrichment. We can potentially isolate peptides from lower abundance proteins by avoiding this chromatographic region of peptides from high abundance proteins.

There are still other types of chromatographic modes that we should explore for utilization in P-SLICY methods including acidic pH 2.7 HILIC, pH 10 RP, SCX, SAX, and many more. Since HILIC P-SLICY can increase the number of quantifiable proteins by 20%, we can further increase this percentage by incorporating additional P-SLICY enrichments in parallel using different chromatographic mode. This can enhance the coverage of the proteome in fewer fractions than traditional 2D LC-MS/MS. In traditional 2D LC-MS/MS, up to 40-60% overlap in protein identifications can be expected between fractions. We can decrease the percentage of overlapping

protein identifications in P-SLICY enrichments and increase the total number of unique protein identifications by selecting orthogonal peptide isolation procedures for each fraction in this proposed parallel P-SLICY protocol. Resulting from this, we can expect a wider breadth of peptides that will be identified for discovery proteomics.

Furthermore, there is a need to appropriately define the maximal depth that can be achieved for a particular sample to benchmark various 2D separation techniques versus P-SLICY. With one of the goals for P-SLICY being to improve the depth of protein identifications for protein analysis, it is useful to determine the proportion of proteins expressed in the cells that were in fact identified via P-SLICY. This will allow us to establish a metric for proteome coverage, which we can express as a percentage of protein that were identified out of the proteins that were expressed. It is hard to concretely define the denominator because as a community, we do not have a quantifiable approach to determine the expressed proteome in its entirety. Hence, we will need to infer from the genes that were identified as transcripts in RNA sequencing. We can assume in standard proteomic approaches that the canonical proteins will be translated from mRNA transcripts. By counting the number of genes that were transcribed, we can use this number as a surrogate for the number of proteins expressed in the cell of interest if there is a corresponding RNA sequencing data. In the future, it would be beneficial to obtain a reference RNA sequencing dataset for JURKAT E6.1 cell line, which we used for P-SLICY experiments, and extract the number of genes identified to determine the proteome coverage attained by P-SLICY versus 2D separations.

7.3 Future direction for redundancy reduction approaching total proteome quantitation

There is an urgent need to establish procedures for protein quantitation covering the entire proteome. With this thesis work, we have demonstrated that analyzing an enriched subset of

peptides from the proteome can improve the number of unique protein identifications. The next step would be to select a set of peptides in the human proteome to perform targeted whole proteome quantitation. Large-scale targeted proteomics panels have been explored in the past,[3–9] but to the best of our knowledge, no one has yet attempted the whole proteome targeted proteomics. With all the advancements in LC and MS technology, we should have an active mindset towards our data acquisition techniques and start thinking about the analytes, peptides, and proteins we want to analyze for. Currently, the MS acquisition algorithm and MS instrument characteristics largely dictates the proteins we can observe in each LC-MS/MS experiment. This study is directed towards putting chromatographic technology on the same level of importance for the overall output of bottom-up proteomic experiments.

The concept of targeted whole proteome quantitation is feasible with current generation instrumentation. There are mass spectrometers that can analyze over 100 MS2 scans per second, of which, the new Orbitrap Astral MS system can support 200 MS2 scans per second. If we assume the human proteome consists of 20,000 proteins, we will need to quantify 40,000 peptides to abide by the 2 peptides per protein quantitation rule. In a theoretical case of 1 MS2 spectrum per peptide, we would be able to acquire 40,000 MS2 scans in 3.3 minutes. Increasing the number of spectra per peptide to 10 to improve quantitative accuracy will result in 400,000 MS2 spectra acquired in 30 minutes. However, to achieve these values, the selected peptides from the proteome will need to fit within dynamic range of detectable concentration and to span the entire chromatogram evenly. Therefore, retention time prediction algorithms are important tools to help decide which peptides to select at various parts of the chromatogram to improve peptide distribution during LC separation.

Peptide selection will be the most critical part of the targeted assay and is not trivial. In addition to the selection criteria of retention time, these peptides must be readily observed in a variety of proteomics experiments (measured by proteotypicity), uniquely identifies the associated protein, and is free of amino acids motifs that are prone to artificial PTMs (such as - deamidation, N-cyclization, or oxidation). There previously has been literature outlining prediction algorithms that can provide predicted proteotypicity for sequences of peptides.[10,11] However, these algorithms cannot appropriately account for artificial PTM propensity and potential peptide sequences that are truncated endogenously as a protein modification (like the adiponectin protein[12]). Resources like the Global Proteome Machine (GPM) contain peptide data from 600,000+ public experimental datasets that we can select peptides that can be the most representative of the associated protein's abundance. GPM provides the number of observations for each peptide, and we want to select peptides with the most observations for each protein. We can also determine the propensity for the peptide to have artificial PTMs by comparing their respective number of observations for the modified versus the non-modified peptide. Furthermore, we can establish the *protein representation* metric for a peptide based on the pearson correlation R^2 value of the extracted signal intensities of the peptide versus the protein abundance calculated (usually by sum of intensities from its constituent peptides). We want to reduce the possibility of selecting peptides with low correlation because they will have poor representation of their protein quantities due to a variety of experimental or biological factors. Some of such reasons can be poor recovery in sample preparation or biological degradation of the peptide as a result of protein truncation (ie. zymogen activation)

Moving further upstream in the same workflow from MS to LC to sample preparation, choosing the appropriate protease is critical. Traditionally, membrane proteins do not generate

proteotypic peptides in the transmembrane region when using trypsin as the protease. For example, the protein encoded by the gene KMT2D has a 227 amino acid long stretch without Arg/Lys as the cleavage site.[13] In our conceptual targeted assay, we will need to explore alternative proteases that will provide a reasonable selection of proteotypic peptides for traditionally unfavourable proteins like membrane proteins. Recent publications involving specific proteases like Asp-N,[14,15] proalanase,[16] Krakatoa[17] have been attractive options due to their orthogonality of cleavage with respect to trypsin, but also non-specific proteases like subtilisin and thermolysin are also promising candidates[13,18] to access regions of the proteome that have not been previously characterized. Upon selecting this set of peptides, a targeted scouting run can be used to establish the experimental retention time of these peptides that can be validated with the retention time prediction algorithms for further refinement of the peptide set. Subsequently, after validation, a parallel reaction monitoring (PRM) acquisition method can be established as a standardized assay for the relative quantitation of the whole proteome. Absolute quantitation can be attained by obtaining the correspondent set of stable isotopically labelled peptides with known concentration that will be used to construct calibration curves for converting the mass spectrometry intensity values to concentrations.

There are several advantages to establishing this quantitative method. Discovery proteomics, since its conception, have issues with missing values due to the stochasticity of the acquisition methods.[9,19] Missing values require statistical imputation to *fill in the gaps* but proteins that only have observations in few samples out of the batch will not have sufficient information for imputation.[20,21] Data-independent acquisition (DIA) can be used as an alternative to the data-dependent acquisition strategies that were used in the entirety of this thesis. DIA has been shown to reduce the missing values problem relative to DDA and can achieve more protein identification.

However, since this technique was popularized after the start of experiments for this thesis, we chose to continue with DDA for the sake of consistency amongst all experiments we have established. Future work in integrating DIA to P-SLICY will be interesting to compare the improvements to proteome coverage. Another alternative to reducing missing values is targeted mass spectrometry. Target approaches generally guarantee a value for each selected protein in each experimental condition. In the past, also due to the stochasticity of the MS methods, "not detected" cannot infer quantitative information because the MS algorithm may have missed it due to DDA schedule saturation. Using this targeted approach, "not detected" is a useful piece of information indicating that the peptide is below the lower limit of quantitation. Establishing a targeted quantitative panel of proteins allows for standardization to improve the reproducibility of results from proteomics assays. These methods can further improve accuracy and precision by allowing more measurements per peptide over the chromatographic peak. A non-technical advantage to this type of targeted approach is that a proteomics lab can provide a guaranteed list of proteins that can be observed to biological researchers.

On the other hand, these target panels have limited observations only towards the list of peptides selected. Aberrations that arise outside this list of peptides will not be observed. Therefore, protein isoforms, post-translational modifications, single amino acids variants will be lost unless a specific targeted panel has an additional group of peptides that can encompass the protein modifications. There will always be a bias for detection given that we are selecting a subset of peptides from the proteome. However, the peptide list can easily be changed depending on the study requirements.

The concept of targeted whole proteome quantitation is not meant to replace discovery proteomics tools, but to streamline the conventional proteomics analyses that focus on achieving

proteome quantitation. Our idea of parallel P-SLICY will still be an important utility when paired with DDA, especially in characterizing the dark proteome[22–24] consisting of non-canonical proteins,[25] alternative isoforms,[26] and proteins with uncharacterized PTMs.[27] However, in order to take a *deep dive* into the unknown regions of the proteome, we need smarter DDA algorithms and protein identification tools that can allow us to explore these diverse MS2 spectra that have not been previously annotated. Conceptually, smart DDA algorithms need to be able to learn the characteristics of peptidic signals in an MS1 spectra and prioritize uncharacterized peptidic signals rather than peptides that are well annotated. Distinguishing between characterized and uncharacterized peptide signal can be done by matching against a peptide database using MS1 and retention time. Since uncharacterized peptide are unlikely found in annotated databases, protein identification tools that depend on database searching will not be applicable for unknown or non-canonical proteins. Therefore, *de novo* sequencing tools will be important for this task and recent resurgence of such tools with DeepNovo,[28] DeepNovor,[29] Casanovo,[30] and GLEAMS[31] are exciting to chart the uncharted waters of the expressed proteome. However, interrogating each MS2 signal with *de novo* sequencing is time-consuming with modern tools. Thus, I would propose aligning MS2 spectra between different experiments based on precursor MS1 *m/z*, retention time, and MS2 similarity and perform preliminary differential analysis strictly using the spectral information. The MS2 spectra that exhibits differential expression will then be selected for *de novo* sequencing.

7.4 Future direction for ZIC-HILIC separations at pH 2.7

Current sample preparation schemes are laborious and complex, therefore, extensive training and troubleshooting skills are necessary to achieve high-quality results. One of the critical problems is the number of steps that are necessary to complete the preparation of one sample and

each of these steps will contribute to the consistency of the downstream quantitative results obtained from the LC-MS/MS analysis. The single-pot solid-phase-enhanced sample preparation (SP3)[32] technique we use involves at least 3 sample transfer steps that can potentially lose proteins to the beads, tips, or walls of the plastic to affect consistency. This is more apparent when we use alternative methods such as suspension trapping (S-Trap)[33] that demonstrated higher consistency (measured by CV values) by having sample transfer than SP3 after cell lysis and reduction and alkylation. Furthermore, current methodologies are time-consuming. This is not only a strain on the operator, but long processing times also incurs the issue of proteins undergoing artificial post-translational modifications like oxidation and deamidation and affecting the resultant CV values. From our observation in Chapter 4 regarding the peptide retention properties on HILIC at pH 2.7, there can be potential to adapt this chromatographic mode for bottom-up proteomics. The sample preparation will trap cells in single-use HILIC columns and the subsequent sample preparation procedure starting from lysis can be done in column without sample transfer steps. The retention of peptides using HILIC mode at pH 2.7 increases for peptides with higher number of positive charges. This alludes to the possibility of trapping proteins on the stationary phase under acidic conditions to allow in-column digestion for proteomic sample processing and can be compatible with alternative proteases. In one of my recent works outside the scope the thesis, we accelerated the digestion procedure in the standard sample preparation workflow from 18 hours to 1 minute using a class of proteases called non-specific proteases.[18] This is one of the most significant time reduction steps that can aid in reducing the time that proteins are exposed to the strenuous environmental conditions. The resultant protein identification performance is similar to trypsin which is a good indicator for researchers to consider transitioning their workflows to incorporate these non-specific proteases. Simultaneously, we also experimented with accelerating

reduction/alkylation procedures from 90 minutes to 5 minutes which shows promise for protein quantitation. With these two projects, we are shortening the time needed for sample processing. By further incorporating HILIC column-based sample preparation, reducing the sample transfer steps will further improve the precision and accuracy of analysis.

In the future, we can create methodologies that accommodates sample preparation starting from lysis to peptides using a single-use HILIC column at pH 2.7. Single-use pipette-tip columns were previously used for peptide desalting prior to LC-MS/MS analysis,[34] but there have been limited reports demonstrating the entire sample processing workflow in-tip.[35] From the Olsen group, the One-Tip procedure uses an EvoSep Pure C18 tip to perform low-cell count sample processing.[36] The traditional packing material inside these tip-based columns are C18 RP sorbents, which is not compatible with conventional harsh detergents like sodium dodecylsulfate (SDS). Thus, the authors used the neutral surfactant, n-Dodecyl β-D-maltoside (DDM). However, SDS is a better surfactant than DDM for protein solubilization.[37] When using HILIC at pH 2.7, SDS does not retain on the column therefore is directly compatible with cell lysis in-tip. Therefore, we should be able to load cells with a into the tip and lyse the cells with a short incubation period using 4% SDS for lysis. The proteins from the lysed cells would be trapped in the tip and the SDS and inhibitor cocktails would elute from the column along with the lipids and metabolites. The lipids and metabolites can be subsequently recovered with an orthogonal chromatographic separation with a separate tip, if necessary. All the remaining steps in the proteomic workflow is compatible with the HILIC tips and peptides can be isolated after for LC-MS/MS analysis. Although, these procedures can be used for single-cell processing like One-Tip, the main intention is to streamline sample preparation and minimizing sample transfer. The simplification of this procedure should decrease the probability of error and improve precision and accuracy. The aim is to establish an

efficient, simple, and standardized sample preparation method that can allow any technician without formal molecular biology training to perform routinely. This would help facilitate the integration of proteomics to clinical research.

Progressively proteomics-based LC-MS/MS techniques are migrating into clinical practice as a part of molecular diagnostics.[38] Currently, the accepted LC-MS/MS-based laboratory developed tests (LDT) involve target proteins that are well defined biomarkers in literature and have established immunoassays that clinical practices depend on to guide therapeutic approaches. Thyroglobulin is an example of such a target protein where it is currently used to aid in clinical decision making for treatment of patients with thyroid cancer.[39] Autoantibodies that are present in the sample can interfere with the current immunoassays that try to detect thyroglobulin and proteomic LC-MS/MS assays are able to overcome these issues and provide improved CVs for quantitation. Most LC-MS/MS assays developed by clinical labs for LDTs are extensions of currently available techniques from literature to avoid re-inventing the wheel, so to say. Therefore, they often still involve many of the laborious sample processing steps. Demonstrating the applicability of HILIC sample processing at pH 2.7 with the accelerated sample procedures could provide faster turnaround times for clinical diagnostics. Potentially, we hope to be able to attain a sample processing workflow that can be completed in 5 minutes for a person with minimal proteomics training. This can aid in increasing uptake of our methods in clinical practice; however, this would be a challenging task beyond just the science of the techniques but also constructing proteomic output reports that are clinically interpretable and actionable[40] and navigating the regulatory landscape for these clinical tests.[41]

7.5 References

(1) Krokhin, O. V.; Ezzati, P.; Spicer, V. Peptide Retention Time Prediction in Hydrophilic Interaction Liquid Chromatography: Data Collection Methods and Features of Additive and Sequence-Specific Models. *Anal. Chem.* **2017**, *89* (10), 5526–5533. https://doi.org/10.1021/acs.analchem.7b00537.

(2) Boersema, P. J.; Divecha, N.; Heck, A. J. R.; Mohammed, S. Evaluation and Optimization of ZIC-HILIC-RP as an Alternative MudPIT Strategy. *J. Proteome Res.* **2007**, *6* (3), 937–946. https://doi.org/10.1021/pr060589m.

(3) Kennedy, J. J.; Whiteaker, J. R.; Ivey, R. G.; Burian, A.; Chowdhury, S.; Tsai, C.-F.; Liu, T.; Lin, C.; Murillo, O. D.; Lundeen, R. A.; Jones, L. A.; Gafken, P. R.; Longton, G.; Rodland, K. D.; Skates, S. J.; Landua, J.; Wang, P.; Lewis, M. T.; Paulovich, A. G. Internal Standard Triggered-Parallel Reaction Monitoring Mass Spectrometry Enables Multiplexed Quantification of Candidate Biomarkers in Plasma. *Anal. Chem.* **2022**, *94* (27), 9540–9547. https://doi.org/10.1021/acs.analchem.1c04382.

(4) Yu, Q.; Xiao, H.; Jedrychowski, M. P.; Schweppe, D. K.; Navarrete-Perea, J.; Knott, J.; Rogers, J.; Chouchani, E. T.; Gygi, S. P. Sample Multiplexing for Targeted Pathway Proteomics in Aging Mice. *Proc. Natl. Acad. Sci.* **2020**, *117* (18), 9723–9732. https://doi.org/10.1073/pnas.1919410117.

(5) Yu, Q.; Liu, X.; Keller, M. P.; Navarrete-Perea, J.; Zhang, T.; Fu, S.; Vaites, L. P.; Shuken, S. R.; Schmid, E.; Keele, G. R.; Li, J.; Huttlin, E. L.; Rashan, E. H.; Simcox, J.; Churchill, G. A.; Schweppe, D. K.; Attie, A. D.; Paulo, J. A.; Gygi, S. P. Sample Multiplexing-Based Targeted Pathway Proteomics with Real-Time Analytics Reveals the Impact of Genetic Variation on Protein Expression. *Nat. Commun.* **2023**, *14* (1), 555. https://doi.org/10.1038/s41467-023-36269-7.

(6) Picotti, P.; Bodenmiller, B.; Mueller, L. N.; Domon, B.; Aebersold, R. Full Dynamic Range Proteome Analysis of S. Cerevisiae by Targeted Proteomics. *Cell* **2009**, *138* (4), 795–806. https://doi.org/10.1016/j.cell.2009.05.051.

(7) Mohammed, Y.; Goodlett, D.; Borchers, C. H. Absolute Quantitative Targeted Proteomics Assays for Plasma Proteins. In *Serum/Plasma Proteomics*; Greening, D. W., Simpson, R. J., Eds.; Methods in Molecular Biology; Springer US: New York, NY, 2023; Vol. 2628, pp 439–473. https://doi.org/10.1007/978-1-0716-2978-9_27.

(8) Gao, Y.; Fillmore, T. L.; Munoz, N.; Bentley, G. J.; Johnson, C. W.; Kim, J.; Meadows, J. A.; Zucker, J. D.; Burnet, M. C.; Lipton, A. K.; Bilbao, A.; Orton, D. J.; Kim, Y.-M.; Moore, R. J.; Robinson, E. W.; Baker, S. E.; Webb-Robertson, B.-J. M.; Guss, A. M.; Gladden, J. M.; Beckham, G. T.; Magnuson, J. K.; Burnum-Johnson, K. E. High-Throughput Large-Scale Targeted Proteomics Assays for Quantifying Pathway Proteins in Pseudomonas Putida KT2440. *Front. Bioeng. Biotechnol.* **2020**, *8*, 603488. https://doi.org/10.3389/fbioe.2020.603488.

(9) Gallien, S.; Kim, S. Y.; Domon, B. Large-Scale Targeted Proteomics Using Internal Standard Triggered-Parallel Reaction Monitoring (IS-PRM)*. *Mol. Cell. Proteomics* **2015**, *14* (6), 1630–1644. https://doi.org/10.1074/mcp.O114.043968.

(10) Chiva, C.; Elhamraoui, Z.; Solé, A.; Serret, M.; Wilhelm, M.; Sabidó, E. Assessment and Prediction of Human Proteotypic Peptide Stability for Proteomics Quantification. *Anal. Chem.* **2023**, *95* (37), 13746–13749. https://doi.org/10.1021/acs.analchem.3c02269.

(11) Mallick, P.; Schirle, M.; Chen, S. S.; Flory, M. R.; Lee, H.; Martin, D.; Ranish, J.; Raught, B.; Schmitt, R.; Werner, T.; Kuster, B.; Aebersold, R. Computational Prediction of Proteotypic

Peptides for Quantitative Proteomics. *Nat. Biotechnol.* **2007**, *25* (1), 125–131. https://doi.org/10.1038/nbt1275.
(12) Zahradka, P.; Taylor, C. G.; Tworek, L.; Perrault, R.; M'Seffar, S.; Murali, M.; Loader, T.; Wigle, J. T. Thrombin-Mediated Formation of Globular Adiponectin Promotes an Increase in Adipose Tissue Mass. *Biomolecules* **2022**, *13* (1), 30. https://doi.org/10.3390/biom13010030.
(13) Gonczarowska-Jorge, H.; Loroch, S.; Dell'Aica, M.; Sickmann, A.; Roos, A.; Zahedi, R. P. Quantifying Missing (Phospho)Proteome Regions with the Broad-Specificity Protease Subtilisin. *Anal. Chem.* **2017**, *89* (24), 13137–13145. https://doi.org/10.1021/acs.analchem.7b02395.
(14) Dau, T.; Bartolomucci, G.; Rappsilber, J. Proteomics Using Protease Alternatives to Trypsin Benefits from Sequential Digestion with Trypsin. *Anal. Chem.* **2020**, *92* (14), 9523–9527. https://doi.org/10.1021/acs.analchem.0c00478.
(15) Giansanti, P.; Tsiatsiani, L.; Low, T. Y.; Heck, A. J. R. Six Alternative Proteases for Mass Spectrometry–Based Proteomics beyond Trypsin. *Nat. Protoc.* **2016**, *11* (5), 993–1006. https://doi.org/10.1038/nprot.2016.057.
(16) Samodova, D.; Hosfield, C. M.; Cramer, C. N.; Giuli, M. V.; Cappellini, E.; Franciosa, G.; Rosenblatt, M. M.; Kelstrup, C. D.; Olsen, J. V. ProAlanase Is an Effective Alternative to Trypsin for Proteomics Applications and Disulfide Bond Mapping. *Mol. Cell. Proteomics* **2020**, *19* (12), 2139–2157. https://doi.org/10.1074/mcp.TIR120.002129.
(17) Yannone, S. M.; Tuteja, V.; Goleva, O.; Leung, D. Y. M.; Stotland, A.; Keoseyan, A. J.; Hendricks, N. G.; Van Eyk, J. E.; Kreimer, S. Blood to Biomarker Quantitation in Under One Hour with Rapid Proteomics Using a Hyperthermoacidic Protease. June 2, 2024. https://doi.org/10.1101/2024.06.01.596979.
(18) Jiang, X.; Yeung, D.; Liu, Y.; Spicer, V.; Afshari, H.; Lao, Y.; Lin, F.; Krokhin, O.; Zahedi, R. P. Accelerating Proteomics Using Broad Specificity Proteases. *J. Proteome Res.* **2024**, *23* (4), 1360–1369. https://doi.org/10.1021/acs.jproteome.3c00852.
(19) Saltzman, A. B.; Chan, D. W.; Holt, M. V.; Wang, J.; Jaehnig, E. J.; Anurag, M.; Singh, P.; Malovannaya, A.; Kim, B.-J.; Ellis, M. J. Kinase Inhibitor Pulldown Assay (KiP) for Clinical Proteomics. *Clin. Proteomics* **2024**, *21* (1), 3. https://doi.org/10.1186/s12014-023-09448-3.
(20) Harris, L.; Fondrie, W. E.; Oh, S.; Noble, W. S. Evaluating Proteomics Imputation Methods with Improved Criteria. *J. Proteome Res.* **2023**, *22* (11), 3427–3438. https://doi.org/10.1021/acs.jproteome.3c00205.
(21) Jin, L.; Bi, Y.; Hu, C.; Qu, J.; Shen, S.; Wang, X.; Tian, Y. A Comparative Study of Evaluating Missing Value Imputation Methods in Label-Free Proteomics. *Sci. Rep.* **2021**, *11* (1), 1760. https://doi.org/10.1038/s41598-021-81279-4.
(22) Posner, Z.; Yannuzzi, I.; Prensner, J. R. Shining a Light on the Dark Proteome: Non-canonical Open Reading Frames and Their Encoded Miniproteins as a New Frontier in Cancer Biology. *Protein Sci.* **2023**, *32* (8), e4708. https://doi.org/10.1002/pro.4708.
(23) Bhowmick, A.; Brookes, D. H.; Yost, S. R.; Dyson, H. J.; Forman-Kay, J. D.; Gunter, D.; Head-Gordon, M.; Hura, G. L.; Pande, V. S.; Wemmer, D. E.; Wright, P. E.; Head-Gordon, T. Finding Our Way in the Dark Proteome. *J. Am. Chem. Soc.* **2016**, *138* (31), 9730–9742. https://doi.org/10.1021/jacs.6b06543.
(24) Cao, X.; Sun, S.; Xing, J. A Massive Proteogenomic Screen Identifies Thousands of Novel Peptides from the Human "Dark" Proteome. *Mol. Cell. Proteomics* **2024**, 100719. https://doi.org/10.1016/j.mcpro.2024.100719.
(25) Kwan, T.; Thompson, S. R. Noncanonical Translation Initiation in Eukaryotes. *Cold Spring Harb. Perspect. Biol.* **2019**, *11* (4), a032672. https://doi.org/10.1101/cshperspect.a032672.

(26) Sinitcyn, P.; Richards, A. L.; Weatheritt, R. J.; Brademan, D. R.; Marx, H.; Shishkova, E.; Meyer, J. G.; Hebert, A. S.; Westphall, M. S.; Blencowe, B. J.; Cox, J.; Coon, J. J. Global Detection of Human Variants and Isoforms by Deep Proteome Sequencing. *Nat. Biotechnol.* **2023**, *41* (12), 1776–1786. https://doi.org/10.1038/s41587-023-01714-x.
(27) Bagwan, N.; El Ali, H. H.; Lundby, A. Proteome-Wide Profiling and Mapping of Post Translational Modifications in Human Hearts. *Sci. Rep.* **2021**, *11* (1), 2184. https://doi.org/10.1038/s41598-021-81986-y.
(28) Tran, N. H.; Zhang, X.; Xin, L.; Shan, B.; Li, M. De Novo Peptide Sequencing by Deep Learning. *Proc. Natl. Acad. Sci.* **2017**, *114* (31), 8247–8252. https://doi.org/10.1073/pnas.1705691114.
(29) Ma, B. Novor: Real-Time Peptide de Novo Sequencing Software. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (11), 1885–1894. https://doi.org/10.1007/s13361-015-1204-0.
(30) Yilmaz, M.; Fondrie, W. E.; Bittremieux, W.; Oh, S.; Noble, W. S. *De Novo* Mass Spectrometry Peptide Sequencing with a Transformer Model. February 9, 2022. https://doi.org/10.1101/2022.02.07.479481.
(31) Bittremieux, W.; May, D. H.; Bilmes, J.; Noble, W. S. A Learned Embedding for Efficient Joint Analysis of Millions of Mass Spectra. *Nat. Methods* **2022**, *19* (6), 675–678. https://doi.org/10.1038/s41592-022-01496-1.
(32) Hughes, C. S.; Moggridge, S.; Müller, T.; Sorensen, P. H.; Morin, G. B.; Krijgsveld, J. Single-Pot, Solid-Phase-Enhanced Sample Preparation for Proteomics Experiments. *Nat. Protoc.* **2019**, *14* (1), 68–85. https://doi.org/10.1038/s41596-018-0082-x.
(33) Zougman, A.; Selby, P. J.; Banks, R. E. Suspension Trapping (STrap) Sample Preparation Method for Bottom-up Proteomics Analysis. *PROTEOMICS* **2014**, *14* (9), 1006–1000. https://doi.org/10.1002/pmic.201300553.
(34) Rappsilber, J.; Mann, M.; Ishihama, Y. Protocol for Micro-Purification, Enrichment, Pre-Fractionation and Storage of Peptides for Proteomics Using StageTips. *Nat. Protoc.* **2007**, *2* (8), 1896–1906. https://doi.org/10.1038/nprot.2007.261.
(35) Su, Y.; Wang, X.; Yang, Y.; Yang, L.; Xu, R.; Tian, R. Zwitterionic Monolith-Based Spintip Column Coupled with Evosep One Liquid Chromatography for High-Throughput Proteomic Analysis. *J. Chromatogr. A* **2022**, *1675*, 463122. https://doi.org/10.1016/j.chroma.2022.463122.
(36) Ye, Z.; Sabatier, P.; Martin-Gonzalez, J.; Eguchi, A.; Lechner, M.; Østergaard, O.; Xie, J.; Guo, Y.; Schultz, L.; Truffer, R.; Bekker-Jensen, D. B.; Bache, N.; Olsen, J. V. One-Tip Enables Comprehensive Proteome Coverage in Minimal Cells and Single Zygotes. *Nat. Commun.* **2024**, *15* (1), 2474. https://doi.org/10.1038/s41467-024-46777-9.
(37) Brown, K. A.; Chen, B.; Guardado-Alvarez, T. M.; Lin, Z.; Hwang, L.; Ayaz-Guner, S.; Jin, S.; Ge, Y. A Photocleavable Surfactant for Top-down Proteomics. *Nat. Methods* **2019**, *16* (5), 417–420. https://doi.org/10.1038/s41592-019-0391-1.
(38) Lin, Y.; Thomas, S. N. Impact of VALID Act Implementation on Mass Spectrometry-Based Clinical Proteomic Laboratory Developed Tests. *J. Mass Spectrom. Adv. Clin. Lab* **2023**, *28*, 30–34. https://doi.org/10.1016/j.jmsacl.2023.02.001.
(39) Netzel, B. C.; Grant, R. P.; Hoofnagle, A. N.; Rockwood, A. L.; Shuford, C. M.; Grebe, S. K. G. First Steps toward Harmonization of LC-MS/MS Thyroglobulin Assays. *Clin. Chem.* **2016**, *62* (1), 297–299. https://doi.org/10.1373/clinchem.2015.245266.
(40) C. Moore, PharmD, Bcps, Bcop, Dpla, Fccp, D.; S. Guinigundo, Msn, Rn, Cnp, Anp-Bc, A. The Role of Biomarkers in Guiding Clinical Decision-Making in Oncology. *J. Adv. Pract. Oncol.* **2023**, *14* (3), 15–37. https://doi.org/10.6004/jadpro.2023.14.3.17.

(41) Gu, Q.; Yu, L.-R. Proteomics Quality and Standard: From a Regulatory Perspective. *J. Proteomics* **2014**, *96*, 353–359. https://doi.org/10.1016/j.jprot.2013.11.024.