Evaluating the Chromatographic Behaviour of Peptides Carrying Post-Translational Modifications for Proteomics Applications

by

Quinn Kelly Neale

A Thesis submitted to the Faculty of Graduate Studies of the
University of Manitoba
In partial fulfillment of the requirement of the degree of

Master of Science

Department of Chemistry
University of Manitoba
Winnipeg

Copyright © 2024 by Quinn K. Neale

Abstract

Among all types of biomolecules, proteins make up a large portion of the core function of cellular processes. Studying the proteins in a cell is called proteomics, and all the proteins which make up this system, the proteome. Proteomics relies on robust and sophisticated liquid chromatography mass spectrometry (LC-MS) instrumentation with highly developed protocols to achieve confident results in peptide and protein identifications. For such applications, it is useful for the chromatographic and mass spectrometric properties of peptides to be well characterized and consistent across all forms which a peptide can be found. These forms are either native, where peptides are found as simple strings of amino acids or be altered with what are called post-translational modifications (PTMs), where a functional group has been added to the protein endogenously (in the cell) or by chemical treatment. These modifications can have drastic ramifications on the analytical properties of peptides and can even affect the protein identification output by artificially altering the observed abundance of a protein within a sample if the diminished intensities on LC-MS are not taken into consideration. Moreover, endogenous modifications often represent a feature of a protein which is being used to communicate within a signalling pathway or alter the function of that protein in the cell, thereby making PTMs an important choice for biological investigations.

Unfortunately, despite the importance in studying PTMs, their chromatographic characteristics are rarely reported. In this thesis, the chemical characteristics of various PTMs, general and specific, were uncovered and reported, permitting explanations for standard and 'anomalous' retention behaviours for peptides carrying PTMs. The general features across all PTMs studied in our lab over the last 10 years were summarized, while more specific behaviours based on peptide primary structure and the position of the modified amino acid were elucidated for peptides carrying deamidations and HexNAc glycosylations. Moreover, one mode which largely showed promise in investigating PTMs, electrostatic repulsion-hydrophilic interaction chromatography (ERLIC), was studied for its retention properties. This led to the development of a retention time prediction algorithm which explained the sequence-specific behaviour of peptides and general features of PTMs in ERLIC.

Acknowledgments

I would like to first thank my supervisors, Dr. Oleg Krokhin and Dr. Helene Perreault, for guiding me through this process, giving me every opportunity to grow as a researcher, and for providing invaluable knowledge to stimulate my learning experience and improve my research outcomes. These lessons will likely carry through for the rest of my life. I would also like to thank my lab mate Darien Yeung and Vic Spicer for their seemingly infinite help throughout these years, for without them, my training and progress would have been drastically different.

As a primary source of sanity in my life outside the lab, I would like to thank my dog Finn for greeting me at the door each day that I came home. The constant barking did wonders at pulling my attention away from work and back to the reality of being a living and breathing human being. I'd also like to thank my friends, family, and girlfriend for their unending support in pursuing my passions. The fruits of my labour are as much my credit as all those who have supported me along the way. Thank you.

Table of Contents

Abstract ii
Acknowledgments iii
Table of Contents iv
List of Tables vi
List of Figures vii
1 Chapter 1: General Introduction 11
1.1 Proteomics: Peptides, Proteins, and Post-Translational Modifications 2
1.2 Mass Spectrometry of Peptides and PTMs 7
1.3 Chromatography of Peptides and PTMs 14
1.4 Research Objectives 18
1.5 References 19
2 Chapter 2: Compendium of chromatographic behavior of post-translationally and chemically modified peptides in bottom-up proteomic experiments 26
2.1 Article Headline and Authors: 27
2.2 Contribution of Authors 28
2.3 Abstract 29
2.4 Introduction 30
2.5 Materials and Methods 35
2.6 Results and Discussion 38
2.7 Conclusion 48
2.8 Supporting Information 50
2.9 Acknowledgements 51
2.10 Data Availability 52
2.11 Supplementary Information 53
2.12 References 64
3 Chapter 3: Chromatographic Properties of Deamidated Peptides with Asn-Gly Sequences in Proteomic Bottom-Up Experiments 70
3.1 Article Headline and Authors: 71
3.2 Contribution of Authors 72
3.3 Abstract 73
3.4 Introduction 74
3.5 Materials and Methods 79
3.6 Results and Discussion 82

3.7 Conclusion 92
3.8 Acknowledgements 94
3.9 Supplementary Information 95
3.10 References 101
4 Chapter 4: Chromatographic Properties of O-HexNAcylated Peptides in Reversed-Phase and Hydrophilic Interaction Liquid Chromatography 106
4.1 Article Headline and Authors: 107
4.2 Contribution of Authors 108
4.3 Abstract 109
4.4 Introduction 110
4.5 Materials and Methods 114
4.6 Results and Discussion 117
4.7 Conclusion 127
4.8 Acknowledgements 129
4.9 Supplementary Information 130
4.10 References 135
5 Chapter 5: Peptide Retention Time Prediction for Electrostatic Repulsion-Hydrophilic Interaction Chromatography 142
5.1 Article Headline and Authors: 143
5.2 Contribution of Authors 144
5.3 Abstract 145
5.4 Introduction 146
5.5 Materials and Methods 149
5.6 Results and Discussion 152
5.7 Conclusion 164
5.8 Acknowledgements 166
5.9 Data Availability 167
5.10 Supplementary Information 168
5.11 References 172
6 Chapter 6: General Conclusions 178

List of Tables

Table 1.1 Amino acid monoisotopic masses and residual monoisotopic masses 9

Table S2.1 Chromatographic properties of peptides carrying post-translational and chemical modifications in RPLC for proteomic applications. 60

Table 3.1 The effect of exposure time on deamidation rates found for yeast digest (37°C, pH 8.4). 82

Table S3.1 Summary of retention times for synthetic peptides in C18 RP with 0.1% FA 98

Table S3.2 Chromatographic conditions for the first dimension separations used in this study. . 99

Table 4.1 Pair identification information for all runs comprising reversed-phase retention characteristics for serine and threonine O-HexNAcylation. Three normalized collision energies (NCEs) were evaluated for both GlcNAc and GalNAc sets. The average shifts and corresponding standard deviations of ΔHI values are reported with the slopes, intercepts, R^2, and root mean square error (RMSE) values of modified vs nonmodified regressions.* 120

Table S4.1 Summary of O-HexNAx identifications and pairs across various experiments*..... 130

Table S5.1 All amino acid retention coefficients from the SSRCalc ERLIC model. 171

List of Figures

Figure 1.1 Structures, names, 3- and 1- letter codes all 20 amino acids. Structures shown at pH 7.4..... 3

Figure 1.2 A - Notation for peptide fragment ion series initially proposed by Roepstorff and Fohlman.[12] B - examples of discrete fragment ion structures from a singly charged peptide. 5

Figure 1.3 General diagram of an LC-MS junction using ESI technology from reference 26.[26] .. 8

Figure 1.4 Instrument schematic of the Orbitrap Exploris 480.[32] 11

Figure 1.5 Instrument schematic of the timsTOF Pro. A – peptides elute and are ionized by ESI as they are separated on the LC column. B – The ion mobility cell traps and accumulates ions in the 1st analyzer, transfers them to the 2nd analyzer to be ramped while the next set of ions are accumulated in the 1st analyzer. C – The ions from the 2nd analyzer of the TIMS tunnel are measured by the TOF analyzer, curating a list of parent ions for MS/MS. E – The quadrupole rapidly switches between parent masses for MS/MS, in sync with the TIMS elution. D – TIMS elution of parent ions occurs. E – CID fragment ions are detected by the TOF mass analyzer.[35] 12

Figure 1.6 General schematic of a gradient HPLC system.[37] 14

Figure 2.1 Acquisition of retention data for modified/non-modified peptide pairs and retention shift plots from our datasets. The various methods for obtaining retention data in this work include A) monitoring spontaneous degradation and *in-vivo* modifications; B) chemical or enzymatic induction of modifications; C) enzymatic removal of modification; D) co-translational insertion of amino acid analogs (metabolic incorporation); E) synthetic peptides. Representative examples of HI_{mod} vs. $HI_{non\text{-}mod}$ plots are given for F) methionine oxidation, G) N-terminal labeling with iTRAQ(4), H) spontaneous loss of water at the N-terminal glutamic acid. The latter plot shows both 230 true (blue) and 614 in-source (red) modifications. 32

Figure 2.2 Average retention shifts (ΔHI, % ACN) for selected modifications. Supplementary Table S2.1 contains the entire collection of retention data used in this work. 39

Figure 2.3 Sequence dependent features which determine the variation of retention shifts upon modification. A – the dependence of average retention shifts for N-terminal acetylation from the nature of N-terminal residue. Lys is not shown as it carries two acetylation sites: N-terminus and side chain. Amphipathic peptide helicity leads to significant changes in peptide behavior (compared to average values) for B) Met oxidation C) Ser phosphorylation, D) N-terminal acetylation, E), deamidation (iso-Asp product). Hydrophobic (red) and hydrophilic (black)

residues being placed in axial projection of helical wheel, occupy distinctive regions: hydrophobic and hydrophilic faces of amphipathic helix, leading to preferential interaction of the former with C18 phase and higher retention. 44
Figure 2.4 The interplay between retention shifts and slope values for the selected HI_{mod} vs. $HI_{non-mod}$ plots. A – selected modifications highlighting high mannose (1), complex neutral (2), complex with one sialic acid (3) and complex with two sialic acid residues (4) Asn-linked glycan structures, B – correlations for amine-selective labeling groups: each represented by average values for peptides without, with one and with two lysines. 46
Figure S2.1 Non-canonical variants of proline studied in this work. 54
Figure S2.2 Assigning Hydrophobicity Index values to PROCAL peptide retention standard.[43] SSRCalc retention prediction for the PROCAL peptides detected in a mixture of non-modified Arg containing model peptides[14] showed 0.983 R^2-value correlation. HI values have been calculated using projection of experimental retention time values on SSRCalc HI axis: HI = (RT-8.6623) / 2.2453. Peptides eluting outside of linear gradient range (in red) have been excluded. 55
Figure S2.3 Assigning Hydrophobicity Index values to hyper-citrullinated *Mus musculus* peptide library.[20] Retention dataset from this publication contained 3245 modified-non-modified peptide pairs. 310 of them overlap with our in-house retention database showing 0.9938 R^2-value correlation. HI values have been calculated using projection of experimental retention time values on SSRCalc DB HI axis: HI = (RT+2.1989) / 3.8108. 56
Figure S2.4 Identifications of modified peptides and correspondent retention shifts in the 2D LC – MS/MS analysis of zebra fish brain tissue digest. Examples of biased unsupervised assignments of retention shifts highlighted in blue. 57
Figure S2.5 Inter-laboratory reproducibility of retention shifts' measurements for methionine oxidation and arginine citrullination. A-C and D-F show the results of three independent experiments for Met oxidation and Arg citrullination, respectively. Red dots were from in-source oxidation, which needed to be removed for accurate evaluation of average ΔHI. 58
Figure S2.6 The distribution of ΔHI values for citrullination events detected with and without considering detection of non-modified analog.[17] Correspondent ΔHI values have been calculated as following: 59

Figure 3.1 Deamidation of Asn-Gly containing peptides and experimental procedures for retention data collection. A – chemical transformation of NG sequences into DG and iDG products; B – experimental workflows for 1D and 2D LC-MS/MS analyses..... 75
Figure 3.2 Retention orders for non-modified peptides and deamidation products following 0 and 24 hours incubation: A – random coil VMEVNGEGLVK. B – amphipathic helical DVNLTEIVENVNGEFQAK featuring multiple hydrophobic residues on one side of the helical wheel. C – NGITVIQVDEPALR with N-terminal modification site. Axial helical projections are shown with hydrophobic residues highlighted in red. Extracted ion chromatograms are showing traces for non-modified peptide m/z (blue) and deamidation products (m+0.984)/z (pink). 84
Figure 3.3 Distribution of retention shifts (ΔHI, %ACN) in RP HPLC (0.1% formic acid) for corresponding deamidation pairs upon modification. A – retention shifts for the conversion of NG to DG. B – retention shifts for the conversion of NG to iDG. C – Retention shifts for pairs generated by unsupervised assignment. 87
Figure 3.4 RPLC separations for synthetic peptides and their deamidation products. A – amphipathic helical peptide SIVEEINGVIGEGER (FA). B – random coil peptide SSNGLSGGLR. C – NGLSSSGGLLR with N-terminal modification - all with 0.1% FA as ion pairing modifier. D – NGLSSSGGLLR using 0.1% TFA. E – NGLSSSGGLLR (0.1% HFBA)..... 88
Figure 3.5 Summary of retentions shifts for deamidated products across 20 various peptide separation modes. Red error bars indicate the standard deviation for each respective average shift. In this case, the average shifts and standard deviations are not calculated from repeated measurements of the same peptide but rather represent the chemical diversity of all observed modified and nonmodified peptide pairs in each separation mode..... 90
Figure 4.1 A – Structures of serine and threonine residues with O-GlcNAcylation and O-GalNAcylation modifications. B – Fragment ions observed in CID- and HCD-based fragmentation modes.[15] 111
Figure 4.2 Plotted HI mapped retention times (RTHI) of modified (HexNAcylated) versus nonmodified peptide sequence pairs for serine and threonine residues in reversed phase applications. Numbers of pairs and average shifts are reported for the peptide entries within each plot. A – Serine O-GlcNAc. B – Threonine-O-GlcNAc. C – Serine-O-GalNAc. D – Threonine-O-GalNAc. 119

Figure 4.3 Plotted fractions of modified (HexNAcylated) versus nonmodified peptide sequence pairs for serine and threonine residues in HILIC applications. Numbers of pairs and average shifts are reported for the peptide entries within each plot. A – Serine O-GlcNAc. B – Threonine-O-GlcNAc. 119
Figure 4.4 Slopes vs Intercepts for various glycosylations from previous studies.[36,37] High Mannose, Fucosylated, 1 Sialic Acids, and 2 Sialic Acid groups are clustered with the same colours and shapes to facilitated visualization..... 125
Figure 4.5 Retention shifts of various glycosylations from our datasets and previous publications.[36,37] Average shifts are reported in HI values with the error bars representing one standard deviation in each direction. Each N-Glycan group is represented by a reduced name: h = hexose, n = hexosamine, f = fucose, and a = sialic acid. 126
Figure S4.1 Plot of modified versus nonmodified O-HexNAc peptides from a 2D high pH – low pH RP experiment. 131
Figure 5.1 Offset chromatograms from experiments used for the optimization of gradient conditions and fraction collection. Our standard peptides (P1-P6) were used to determine a fraction collection window (blue). Confirmation of the elution window was performed with a BSA digest (orange). Jurkat digest was fractionated at 1-minute intervals between 1 to 45 minutes. 153
Figure 5.2 Retention prediction for ERLIC using a ~170,000 unique peptides retention dataset. Plots represent the observed fraction number of each individual peptide against the models trained and used for prediction. A – ERLIC SSRCalc our standard model with position-dependent retention coefficients with a standard deviation of 2.65 fractions; B – ERLIC GRU model using machine learning approach with a standard deviation of 2.17 fractions..... 154
Figure 5.3 Retention coefficients (RCs) from the ERLIC SSRCalc Model for N-terminal (A), internal (B) and C-terminal (C) positions of amino acids. Terminal RCs are represented by the distance from the termini e.g. N1 = N-terminal position, N2 = second position in from the N-terminus etc. Since tryptic peptides only contain K and R on the C-terminus, C1 coefficients are omitted due to lack of representation for all other amino acids..... 155
Figure 5.4 Plots of normalized internal retention coefficients for ERLIC against other modes. A – ERLIC vs XBridge Amide (HILIC) at pH 2.7. B – ERLIC vs SAX pH 8. C – ERLIC vs ZIC-HILIC at pH 2.7. D – ERLIC vs ZIC-HILIC at pH 2.7. Orange represents the function y = x (Mode 1 = Mode 2) which aids in illustrating the difference in amino acid dependence for peptide retention

in each mode: the closer the values to the line, the more similar two modes are for any amino acid. 159
Figure 5.5 Orthogonality plots for various separation modes based on 10,000 randomly selected peptides against reversed-phase (RP) with formic acid (FA) as an ion-pairing modifier. Orthogonality is scored following the same procedure outlined in our previous publication [43]. A – ERLIC vs 0.1% FA C18 RP. B – Summary of Orthogonality scores and r^2 values. C – XBridge Amide HILIC vs 0.1% FA C18 RP. D – SAX vs 0.1% FA C18 RP. E – high pH reversed phase vs 0.1% FA C18 RP. F – SCX vs 0.1% FA C18 RP. 160
Figure 5.6 Effects of various PTMs on peptide retention in ERLIC. Average fraction shifts were calculated by generating pairs of modified and nonmodified sequence counterparts carrying only one modification. Error bars represent the standard deviation of all fraction shifts for each respective modification. 162
Figure S5.1 Experimental Workflow sample preparation, peptide identification, and RPLC filtering. 168
Figure S5.2 Trends for retention coefficients based on the position of amino acids. Amino acids are represented by their one letter codes. A – aromatic amino acids. B – Hydrophobic amino acids. C – Hydrophilic amino acids. D – Neutral amino acids. E – Basic amino acids. F – Acidic amino acids. 169
Figure S5.3 Orthogonality plot of peptides identification outputs from ERLIC fractionation. RTHI is values were mapped from retention times using our 6-peptide standard. Corresponding fractions were used to visualize orthogonality. 170

1 Chapter 1: General Introduction

1.1 Proteomics: Peptides, Proteins, and Post-Translational Modifications

Liquid chromatography (LC) and mass spectrometry (MS) have been developed into leading tools for deciphering the complex and dynamic matrices of proteins in biological systems. The study of the proteins in these cases is called proteomics, while the entire complement of proteins in a biological system is called the proteome. MS-based proteomics experiments are often conducted in one of two ways: the bottom-up approach which analyzes peptides from digestions of purified protein samples, or top-down where intact proteins are measured, both relying on tandem MS (MS/MS) fragmentation technologies.[1]

Proteins provide function to the cells of all organisms, from single-cell microbes up to the complex cells which make up organs and tissues in animals. They are made of 20 canonical amino acids carrying unique structures and chemical properties. Each amino acid, structure, and 3- and 1-letter codes are shown in **Figure 1.1** for future reference. Proteins can also be altered by what are called post-translational modifications (PTMs) or chemical modifications (CM), i.e. molecular scaffolds which are covalently bonded to their amino acids. PTMs and CMs can have a large range of molecular diversity, often classified into subsets of groups based on their substructures or chemistry. Glycosylation which adds carbohydrate scaffolds to Ser, Thr, and Asn residues, or phosphorylation on Ser, Thr, and Tyr residues are common examples of PTMs. In each case, these modifications can have drastic effects on the specific modified amino acid within a protein. A good example of chemical modification is the alkylation of cysteine residues[2], an important step in digestion procedures[3] with considerable effects on chromatographic elution features.[4]

The major method for identifying PTMs in MS-based proteomics is performed by detecting peptides with modified masses. The mass shift added by a modification, ΔM (Da), corresponds to the monoisotopic mass of a modified peptide minus the monoisotopic mass of the non-modified molecule. A variety of PTMs and corresponding mass shifts are summarized in Chapter 2, **Table S2.1**.

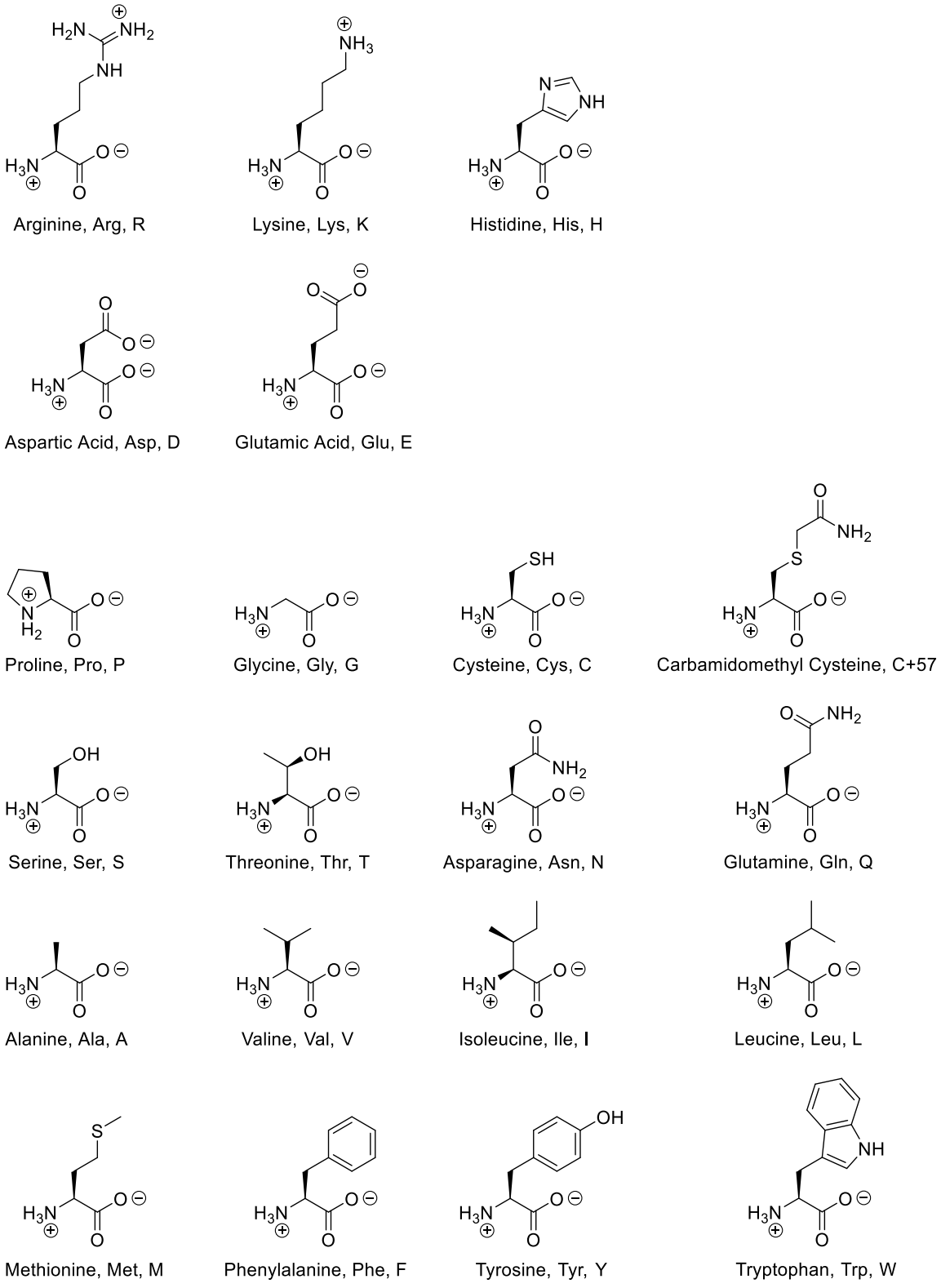


Figure 1.1 Structures, names, 3- and 1- letter codes all 20 amino acids. Structures shown at pH 7.4.

1.1.1 The Biological Context of Proteomics

Ultimately, the goal in proteomics is to interrogate the complex protein profile in a way that answers biological questions. Proteomes are controlled by upstream regulation at the genetic level. This process is referred to as the central dogma in molecular biology where information at the genetic level (deoxyribonucleic acid, DNA) is transcribed (ribonucleic acid, RNA) and then translated into proteins. As the variation of protein levels and their PTMs can interfere with homeostasis and as regulatory dysfunctions are commonly associated with diseases, it is important to characterize protein dynamics in biology. In the case of human studies, this could be between a healthy individual and someone in a diseased state where a set of dysregulated proteins could be the major difference in their proteomic profiles. Unfortunately, comprehensive identification of the entire human proteome is challenging as it is estimated to be as large as 20,000 proteins if we consider that each gene produces one protein. If we include splice variants (instances where some of the genetic information is removed before translation), the number of proteoforms (a variant of a protein encoded in DNA) is estimated to reach up to 70,000 (reviewed[5]).

1.1.2 General Bottom-Up Proteomics Sample Preparation and Data Analysis

Proteomics experiments nowadays are generally performed with mass-spectrometry-based bottom-up approaches: proteins are extracted from samples, cysteine residues are reduced and alkylated, then specific proteases are used to generate peptides which are subsequently purified by chromatography. In special cases, enrichment procedures can be used to generate samples with dominant chemical features, such as N-terminal peptides, C-terminal peptides, peptides with PTMs, and peptides with CMs. Once a peptide sample is generated for its intended purpose, high pressure liquid chromatography (HPLC) is used to separate the complex mixture of peptides by their chemical properties and detection is accomplished with MS. HPLC of peptides commonly employs reversed-phase (RP) chromatography but depending on the sample composition and choice in separation setup, hydrophilic interaction liquid chromatography (HILIC[6]), ion exchange chromatography[7,8], capillary electrophoresis[9], and mixed modes[10] can be employed as well.

During MS, primarily two modes are available to analyze peptide samples: data-dependent acquisition (DDA) and data-independent acquisition (DIA). In DDA mode, peptides are detected as they elute from the column in a survey scan (also referred to as MS1 or Full MS) which detects all eluting analytes (parent ions) within the full mass to charge (m/z) range, then an automated

algorithm is used to select the highest intensity analytes for fragmentation (referred to as MS2 or MS/MS). In DIA mode, a similar survey scan is performed, however fragmentation is performed on m/z windows until the full m/z range is covered. Fragmentation can employ a variety of techniques such as collisional induced dissociation (CID), higher-energy collisional dissociation (HCD), electron-transfer dissociation (ETD), electron-capture dissociation (ECD), and ultraviolet photodissociation (UVPD) to name a few. In each case, the information on a peptide's fragmentation products (daughter ions) is stored to be used for peptide identification. Depending on the choice of fragmentation technique employed[11], peptides will generate one or more fragment ion series (**Figure 1.2**).[12] A more detailed description of MS and chromatography in proteomics can be found in sections 1.2 and 1.3, respectively. Identification of peptides following LC-MS analysis uses search algorithms that perform peptide mass fingerprinting, relying on the predictable fragmentation of peptides and a known proteome.

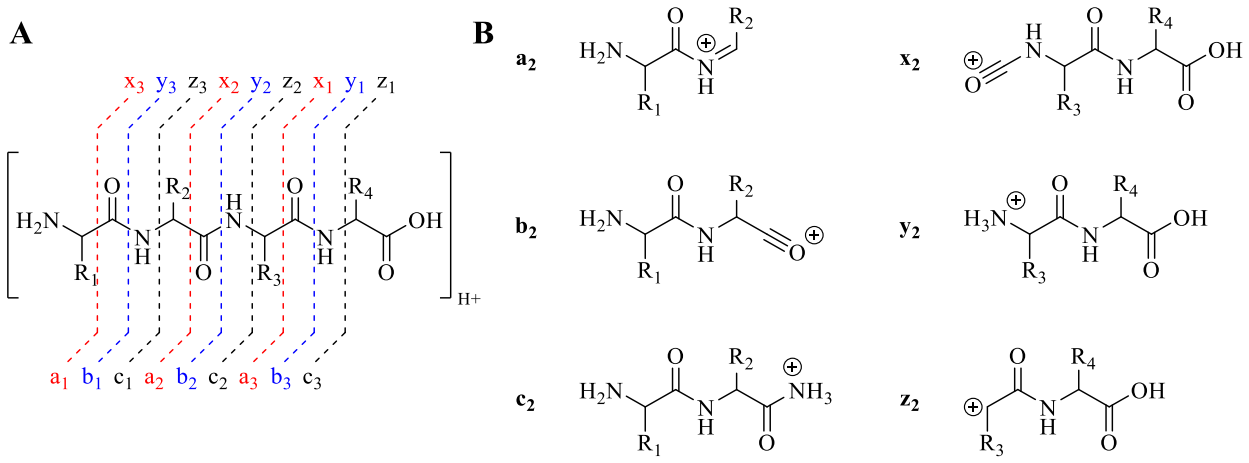


Figure 1.2 A - Notation for peptide fragment ion series initially proposed by Roepstorff and Fohlman.[12] B - examples of discrete fragment ion structures from a singly charged peptide.

1.1.3 Search Algorithms and Data Analysis

There are a variety of search algorithms that can be employed for bottom-up proteomics, including X!Tandem[13], MSFragger[14], Proteome Discoverer[15], MaxQuant[16], and many more. Although each software uses its own computational pipeline, there are underlying similarities between them all. The remainder of this paragraph outlines the major features used by identification algorithms. Briefly, proteomics relies on proteases that will cleave the protein's primary sequence after specific amino acids or sequence motifs.[17–19] The search algorithms use the

known proteome and chosen protease for digestion to generate a set of theoretical peptides for a sample, thereby permitting the generation of predicted fragmentation spectra for each theoretical peptide. In DDA experiments, mass fingerprinting is accomplished by comparing the predicted spectra to experimental spectra. Each pair of predicted and experimental spectra is subjected to a similarity test followed by statistical comparison against other possible matches to determine an expectation value or confidence level for each match. In DIA experiments, the mass fingerprinting is performed similarly, though deconvolution of multiple parent peptides from the same fragmentation window requires sophisticated acquisition procedures and software.[20]

In both cases, predicted retention times of peptides can be folded in as another metric to assess the confidence of peptide identifications. When the predicted and experimental spectra fall above a specific level of confidence for a peptide, they are considered a peptide spectral match (PSM), or in other words, an identification. Subsequent filtering of PSMs is used to reduce the chance of false-positive identifications and to trim the data to the desired level of confidence. Ultimately, an identified peptide can then be mapped back to the known parent protein where quantitative information can be derived from the protein's identified peptides.

The relative or absolute quantitative information of proteins following identification workflows can then be compared amongst technical and biological replicates to assess the statistical relevance of peptide and protein identifications in analytical and biological contexts.[21] Commonly, the fold change of a protein is computed between the biological replicates of a control and a sample which has been perturbed. The identified proteins and their fold changes are scanned for high fold changing, statistically relevant proteins which could serve as potential biomarkers for further investigation. This includes native proteins and those carrying post-translational modifications. Hence, proteomics permits the identification of new lead targets in biological systems.

1.2 –Mass Spectrometry of Peptides and PTMs

Mass spectrometers have become pivotal pieces of instrumentation in analytical chemistry due to their ability to simultaneously detect and identify multiple analytes in samples. A mass spectrometer detects analytes by their mass to charge ratios (m/z) using a mass analyzer. This process can be achieved with a variety of instrument geometries chosen based on the sample complexity, the type of analytes present, and the desired level of mass accuracy.[22] Moreover, mass spectrometers can apply numerous fragmentation techniques on preselected ions for tandem mass spectrometry (MS/MS) to accomplish identification. Under these conditions, each analyte will produce a unique fragmentation pattern in response to instrument stimuli. For peptides, these fragmentation techniques can be limited but provide useful data on peptide sequences and PTMs.[11]

Mass spectrometers can also be coupled to other instruments, in so-called hyphenated strategies, such as for LC-MS/MS. These hyphenated strategies add an extra dimension for analysis, ultimately providing a greater depth and breadth of identification potential. For example, in the case of MS-based proteomics, the application of LC is often required to reduce the sample complexity prior to MS analysis, thereby generating a higher number of identifications and the ability to scrutinize each peptide's retention characteristics. However, careful considerations must be met in hyphenated strategies, as the junction between instruments needs to satisfy the requirement of efficient ionization prior to MS analysis.[23]

1.2.1 LC-ESI-MS/MS

In hyphenated LC-MS strategies, one of the most popular techniques to produce gas-phase ions is electrospray ionization (ESI) as shown in **Figure 1.3**. ESI is a soft ionization technique, permitting the ionization of analytes without the fragmentation of labile and noncovalent bonds. These types of ionization sources are also often coupled to nanoflow LC instrumentation as together they are highly efficient.[24] What is not shown in **Figure 1.3** is the charged state of the capillary which creates a difference of potential between the capillary and the electrode. Simultaneously, the eluent system from the LC carries analytes and flows through the capillary. Together, the eluent and difference of potential result in the formation of a Taylor Cone, which ultimately leads to the generation of charged droplets (reviewed[25]).

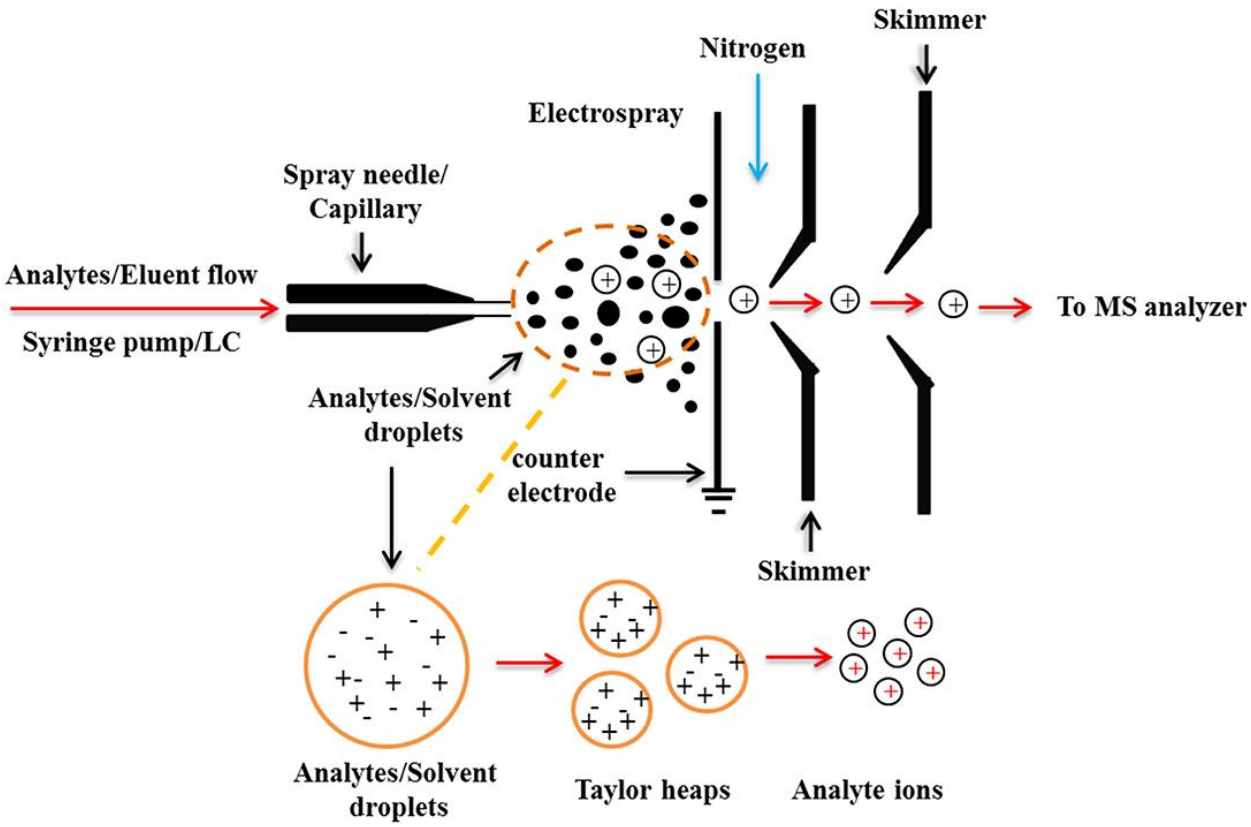


Figure 1.3 General diagram of an LC-MS junction using ESI technology from reference 26.[26]

Following the generation of charged droplets it is believed that a similar process occurs again. As the solvent molecules evaporate, the charge density increases, creating a Taylor Cone on the surface of the droplet, thereby producing smaller charged droplets. This process, often referred to as coulombic explosion, can repeat itself, producing smaller droplets until an analyte can be desolvated, acquiring charge in the process. This results in the formation of singly to multiply charged ions (i.e. $[M+H]^{1+}$, $[M+2H]^{2+}$, etc.). These ions then pass through skimmers to be focused into the mass spectrometer and measured.[27]

1.2.2 Deciphering Peptide MS/MS Spectra

Mass spectrometric analyses of peptides can be performed using a variety of instrument geometries and fragmentation techniques for identification. In such cases, each fragmentation process generates predictable fragment ion series, while the specific choice in technique will determine which of these products will be used for identification (**Figure 1.2**).[11] The a-type, b-type, and c-type series comprise the N-terminal fragment ions, while the x-type, y-type, and z-type

series comprise the C-terminal fragment ions. In each series, the subscript denotes the number of amino acids present for that fragment ion. For practical use, the concept of residual mass is highly useful in sequencing peptides from their fragmentation spectra. This relies on the understanding that each increment of a fragment series (i.e. y_1, y_2, y_3, etc.) differs by the addition of an amino acid. So, for each series in a specific peptide sequence, the mass difference between, $a_2 \rightarrow a_3$, $b_2 \rightarrow b_3$, $c_2 \rightarrow c_3$, etc., is the same. For each amino acid, this mass is referred to as its residual mass, equating to the molecular weight (MW) of an amino acid minus water. The monoisotopic masses of amino acids and their monoisotopic residual masses are shown in **Table 1.1**.

Table 1.1 Amino acid monoisotopic masses and residual monoisotopic masses

Amino Acid	Chemical formula	Monoisotopic Mass (Da)	Residual Chemical formula	Monoisotopic Residual Mass (Da)
Alanine	$C_3H_7NO_2$	89.04768	C_3H_5NO	71.03711
Arginine	$C_6H_{14}N_4O_2$	174.11169	$C_6H_{12}N_4O$	156.10111
Asparagine	$C_4H_8N_2O_3$	132.05351	$C_4H_6N_2O_2$	114.04293
Aspartic Acid	$C_4H_7NO_4$	133.03752	$C_4H_5NO_3$	115.02694
Cysteine	$C_3H_7NO_2S$	121.01975	C_3H_5NOS	103.00919
Glutamic Acid	$C_5H_9NO_4$	147.05318	$C_5H_7NO_3$	129.04259
Glutamine	$C_5H_{10}N_2O_3$	146.06916	$C_5H_8N_2O_2$	128.05858
Glycine	$C_2H_5NO_2$	75.03204	C_2H_3NO	57.02146
Histidine	$C_6H_9N_3O_2$	155.06949	$C_6H_7N_3O$	137.05891
Isoleucine	$C_6H_{13}NO_2$	131.09464	$C_6H_{11}NO$	113.08406
Leucine	$C_6H_{13}NO_2$	131.09464	$C_6H_{11}NO$	113.08406
Lysine	$C_6H_{14}N_2O_2$	146.10554	$C_6H_{12}N_2O$	128.09496
Methionine	$C_5H_{11}NO_2S$	149.05106	C_5H_9NOS	131.04049
Phenylalanine	$C_9H_{11}NO_2$	165.07899	C_9H_9NO	147.06841
Proline	$C_5H_9NO_2$	115.06334	C_5H_7NO	97.05276
Serine	$C_3H_7NO_3$	105.0426	$C_3H_5NO_2$	87.03203
Threonine	$C_4H_9NO_3$	119.05826	$C_4H_7NO_2$	101.04768
Tryptophan	$C_{11}H_{12}N_2O_2$	204.08989	$C_{11}H_{10}N_2O$	186.07931
Tyrosine	$C_9H_{11}NO_3$	181.07391	$C_9H_9NO_2$	163.06333
Valine	$C_5H_{11}NO_2$	117.07899	C_5H_9NO	99.06841

The Thermo Fisher Orbitrap Exploris 480 and the Bruker timsTOF Pro 2 were the instruments used for the experiments discussed in my thesis. These are both examples of instruments utilizing higher-energy collisional dissociation (HCD) and collision induced dissociation (CID), respectively. The peptide fragment ions produced in these cases are predominantly b-type and y-type ions, produced by the fragmentation of amide bonds connecting amino acids. The mobile proton model[28] can be used as a framework to describe this peptide fragmentation mechanism, while a more detailed set of competitive mechanisms have been summarized to satisfy sequence-specific behaviours.[29] For example, further fragmentation can produce a-type ions from the loss of CO on b-ions, and losses of other functional groups can occur for specific amino acids (reviewed[29]). Nonetheless, to identify a peptide from HCD and CID experiments, the expected b-type, y-type, and a-type series can be calculated following equations (1), (2), and (3) below.

(1) b_x = [sum of residual masses for constituent amino acids] + H

(2) y_x = [sum of residual masses for constituent amino acids] + H_3O

(3) $a_x = b_x$ - CO

1.2.3 The Thermo Fisher Orbitrap Exploris 480

Orbitraps are a unique family of mass spectrometers, classified together based on their Orbitrap mass analyzers. In particular, the model used for this thesis is an Exploris 480. A schematic of the Exploris 480 is shown in **Figure 1.4** with major components listed for the instrument. For detailed explanations and further references for each working part of the instrument, the information can be found in references 30-32.[30–32]

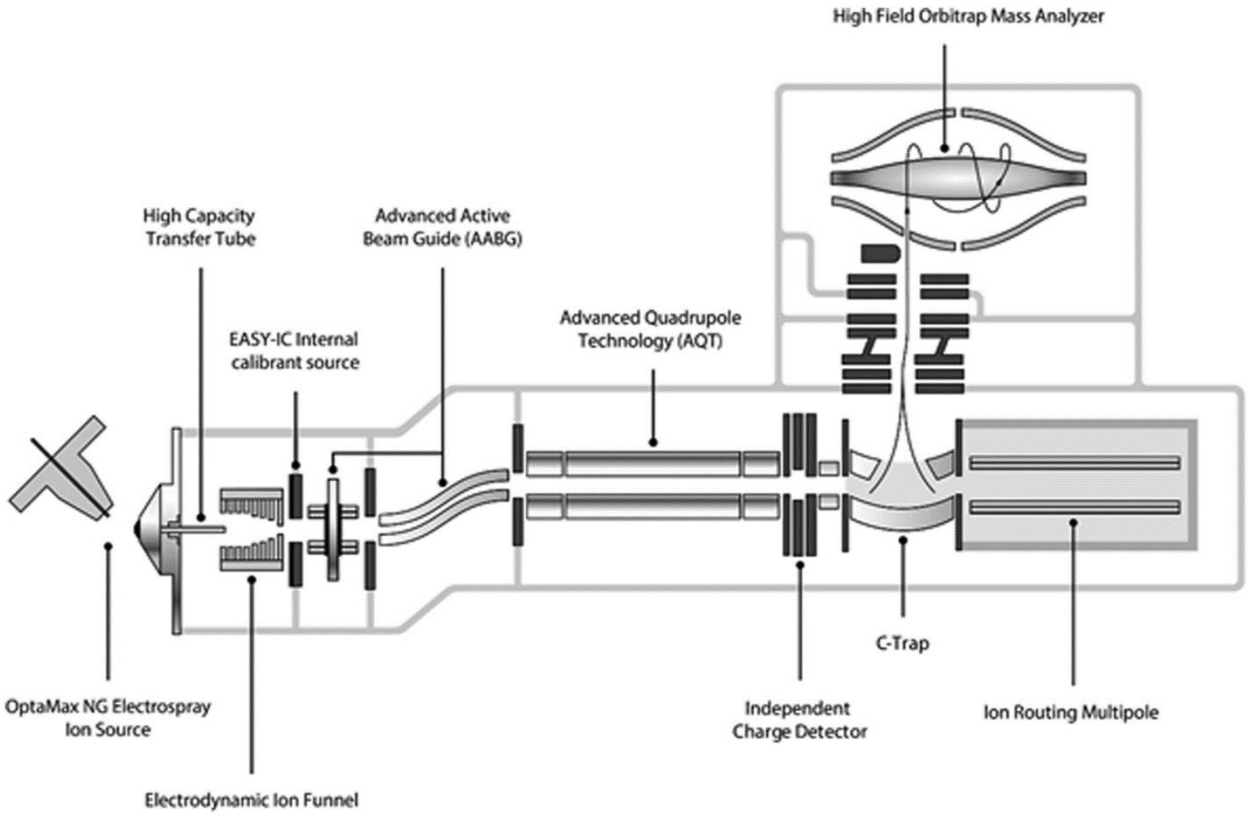


Figure 1.4 Instrument schematic of the Orbitrap Exploris 480.[32]

At the front end of the instrument is the ESI source, followed by the inlet of the instrument where a high capacity tube transfers droplets and ions into the mass spectrometer under high vacuum pressure. The electrodynamic ion funnel, also known as the S-lens, is the first part in the advanced active beam guide (AABG) in **Figure 1.4** and focuses the particle beam. The particles then enter a bent flatapole (second part in the AABG in **Figure 1.4**), where an electrostatic field bends the beam path of ions into the quadrupole while neutral particles and remaining droplets pass through and are deposited in a waste area. During a scan, the ions are guided through the quadrupole to a short octupole integrated with an independent charge detector. Ions are then routed through the C-trap and accumulated in the nitrogen-filled ion routing multipole. Here, ions can be trapped in the case of a survey scan (MS1) or fragmented in the case of MS/MS (MS2). The resulting ion packet is then sent back to the C-trap and through to the Orbitrap mass analyzer for detection. The Orbitrap mass analyzer then measures the frequency of charged particles that are oscillating around the barrel-like structure using split electrodes, generating an image current. The frequency of ions depends on their m/z values, where the fast Fourier transform can be applied to

generate a mass spectrum.[33] This permits the high mass accuracy and high resolution of these Orbitrap instruments.

1.2.4 The Bruker timsTOF Pro 2

The Bruker timsTOF Pro 2 was another instrument used for this thesis. It uses a time of flight (TOF) mass analyzer on the back-end and dual trapped ion-mobility spectrometry (TIMS) on the front-end of the instrument. This combination of techniques when hyphenated with LC is referred to as 4D analysis (chromatography, TIMS, survey scan, and MS/MS).[34,35]

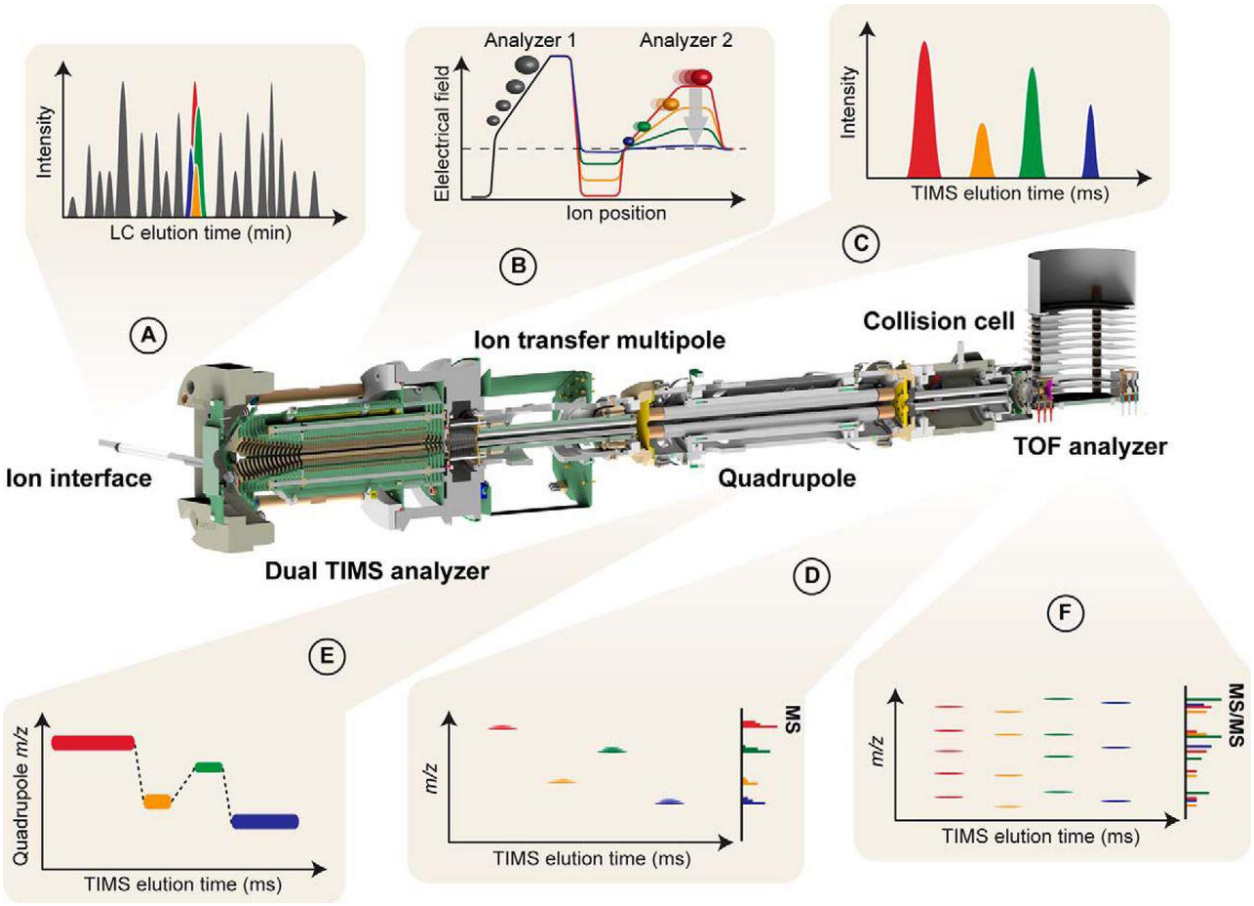


Figure 1.5 Instrument schematic of the timsTOF Pro. A – peptides elute and are ionized by ESI as they are separated on the LC column. B – The ion mobility cell traps and accumulates ions in the 1st analyzer, transfers them to the 2nd analyzer to be ramped while the next set of ions are accumulated in the 1st analyzer. C – The ions from the 2nd analyzer of the TIMS tunnel are measured by the TOF analyzer, curating a list of parent ions for MS/MS. E – The quadrupole rapidly switches between parent masses for MS/MS, in sync with the TIMS elution. D – TIMS elution of parent ions occurs. E – CID fragment ions are detected by the TOF mass analyzer.[35]

A schematic of the timsTOF Pro (a previous version, since a detailed schematic of the current model is not available), is shown in **Figure 1.5**. Analytes eluting from the LC are volatilized and ionized by the ESI source and enter the mass spectrometer through a glass capillary. The ions from this mixture are deflected at 90° into the dual TIMS tunnel and focused by an ion funnel. The ions are collected in the first region of the TIMS tunnel for a specified accumulation time, then transferred to the second TIMS region. During this process, electrostatic fields limit radial movement while gas from the capillary flows through the tunnel, carrying ions axially. This drift is simultaneously counteracted by an electric field in each TIMS region. The ions under these conditions come to rest, are hence "trapped" and separated by their ion mobility (IM) related to their collisional cross section (CCS) values and charge. While ions accumulate in the first TIMS region, the second TIMS region gradually reduces the electrostatic potential, releasing low mobility ions first. Once through the TIMS cell, a survey scan can be accomplished by passing the ions through the ion transfer multipole, the quadrupole, and the collision cell, to the TOF mass analyzer. In the case of MS/MS, the quadrupole can be used to select parent ions for MS/MS in the collision cell. The timsTOF Pro uses an acquisition mode called parallel accumulation-serial fragmentation (PASEF) to synchronize the TIMS and quadrupole for MS/MS.[34,35] This means that for the first IM scan, multiple TOF analyses are performed and summed to generate the full survey scan. Then an algorithm uses the intensities to select precursors for MS/MS, rapidly switching the quadrupole to select for the expected parent ions as they elute from the TIMS cell.

The TOF mass analyzer of the instrument detects m/z values of analytes based on their time of flight.[36] Ions are accelerated into the TOF mass analyzer by an orthogonal electric field. Under these conditions, the acceleration and kinetic energy experienced by each analyte ion will be based on its m/z values and will launch ions into a flight path. As ions can have different positions when they enter the TOF mass analyzer, they can also be subjected to different accelerations in response to the electric field. This presents a problem as it could drastically reduce mass resolution. However, a reflector in the flight path of the TOF mass analyzer corrects for these discrepancies. Once the ions collide with a detector, usually a microchannel plate, the time of each collision is recorded and used to determine the m/z values with their corresponding intensities.

1.3 Chromatography of Peptides and PTMs

High pressure liquid chromatography (HPLC) has played a significant role in proteomics studies due to the need to separate highly complex mixtures of peptides. For the majority of proteomics experiments, RPLC with linear acetonitrile gradients is applied and has remained the gold standard for peptide separation sciences. Shown in **Figure 1.6** is a schematic for a general HPLC used for a gradient separation. There are solvent reservoirs for eluents A and B that need to be degassed and filtered prior to being pumped through solvent lines under high pressure. The pumps are programmed such that the desired composition and flowrate of solvent after mixing satisfies the gradient method. An analytical column is used to separate the analytes in a sample mixture, which are subsequently analysed by a detector.

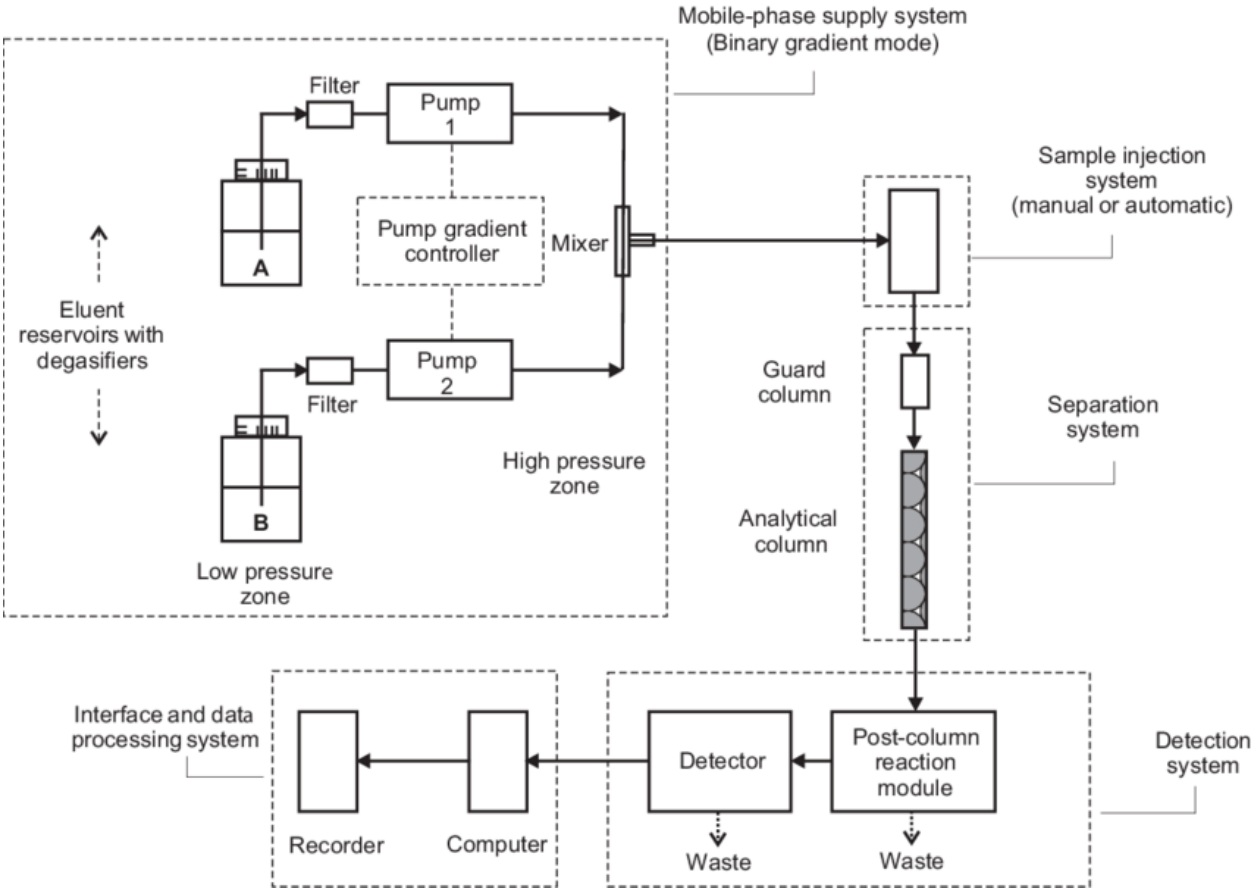


Figure 1.6 General schematic of a gradient HPLC system.[37]

Using modern tandem MS as the detector, proteomics is able to identify over 100,000 peptides from a standard bottom-up proteomics experiment.[38] In the majority of these cases, the

chromatograms of the peptide mixtures do not appear fully resolved, with many coeluting species and overlapping chromatographic peaks. For the most part, this can be circumvented by evaluating peak shapes and widths using extracted ion chromatograms, where the trace of a peak is monitored based on a specified m/z value. Moreover, the coelution of peptides is less of an issue since MS instruments offer fast detection speeds for eluting peptides and their corresponding MS/MS spectra are used for identification. Nonetheless, due to the complex matrix of peptides, chromatographic practices require consistent and accurate quality control metrics to monitor retention drift, column quality, verify peptide retention times, and monitor peak shape.

1.3.1 General Chromatographic Behaviour of Peptides

Early works in the Krokhin group in determining peptide retention properties showed that under isocratic conditions, retention factors of peptides change drastically over the course of a small window of acetonitrile (ACN) composition in aqueous ACN eluents.[39] Applying this to gradient methods, it explains that a peptide's retention time falls within a small window of the ACN gradient, based on the change in these retention factors. This was also the first step in the generation of a 6-peptide mixture (P1-P6) used by the group as internal retention standard in all reversed-phase runs. These peptides were assigned indexed retention values, named hydrophobicity index values (HI, %ACN), relating to the point in the ACN gradient where they start eluting from the column. As a result, by plotting the dependence of retention time against the HI values for each P1 to P6, a linear correlation is observed. Using the slope and intercept of the regression, the HI values can be calculated for the observed peptides in an experiment. By converting retention time to HI values using the indexed retention times of P1 to P6, it accounts for retention drift since all peptides are subjected to under the same instrument conditions, permitting run-to-run comparison of peptide retention times.

In reversed-phase, peptides elute in order of increasing hydrophobicity. In gradient methods, this often employs an increase in ACN composition of the eluent. On the peptide level, hydrophobicity is dependent on a peptide's composition, position of amino acids, and on the ion-pairing modifier used in eluent buffers. These sequence specific properties of peptides can be modelled[40–42] and extended to all forms of chromatography, owing to their heteropolymeric characteristics. These practices have been broadly applied by the Krokhin group. Reversed-phase is not the only mode used for peptide separations, there is also HILIC[6], which separates peptides

in order of increasing hydrophilicity with sequence-specific properties.[43–45] Ion exchange modes such as strong cation exchange and capillary zone electrophoresis (CZE) similarly fall under this umbrella.[46,47] These models provide a framework to describe the retention of peptides chemically, allowing one to explain the observed retention times of peptides from a mechanistic perspective.

For peptides carrying PTMs, the specific structure of a PTM will alter the retention behaviour of a peptide based on the structure of the modification, i.e. its hydrophobicity/hydrophilicity and any change in charge. These are summarized for their general characteristics in Chapter 2, where the retention times of modified peptides are compared to those of their nonmodified sequence counterparts.[48] These are also summarized in a more detailed manner in Chapters 3 and 4, where the sequence specific properties of deamidation and glycosylation were evaluated. Despite the importance of PTMs in proteomics research and their biological implications, information on their retention behaviours is lacking in the literature. For the field of proteomics, where the number of peptide and protein identifications expands and becomes more important each year, understanding the retention behaviours of peptides and their post-translational modifications can be implemented to increase the confidence in these identifications.[49] For this thesis and in the Krokhin group in general, this would be adding the retention models for PTMs to the Sequence-Specific Retention Calculator (SSRCalc, https://proteome.ad.umanitoba.ca/ssrcalc) family of peptide retention prediction algorithms.

1.3.2 Peptide Enrichment Protocols for Targeted Proteomic Measurements

Prior to LC-MS/MS analysis, enrichment protocols can be used to generate samples with desired chemical properties. For example, the enrichment of phosphopeptides can be accomplished using procedures such as immobilized metal ion chromatography[50], generating a sample of peptides predominantly carrying phosphopeptide moieties. Another example is the enrichment of glycopeptides using lectins, proteins which bind carbohydrate scaffolds, known as lectin weak affinity chromatography (LWAC).[51] For glycosylation, applying lectins for broad analysis is challenging as each lectin carries specificity towards particular carbohydrate moieties while glycans (scaffolds of carbohydrates) can have a large chemical diversity in their arrangements and compositions. HILIC and its variants, on the other hand, can be used for broad enrichment of glycopeptides, owing to the fact that these are polar molecules that retain well on HILIC sorbents. These procedures are generally performed using solid phase extraction (SPE) in cartridges or

pipette tips[52], but can also utilize HPLC instruments with gradient methods to generate fractions with a higher abundance of glycopeptides.[53,54]

1.3.3 2-Dimensional and 1-Dimensional Chromatography

Akin to enrichment procedures using HPLC instrumentation, fractionation prior to LC-MS/MS analyses can aid in reducing the complexity of an entire peptide sample, thereby facilitating the identification of lower abundance peptides and therefore proteins. These are called 2-dimensional (2D) chromatography applications due to the application of a first-dimension fractionation, then of a subsequent analysis of the fractions in a second dimension by LC-MS/MS. The first dimension in these cases can use any separation mode, ideally applied based on the sample and research objectives (such as PTMs, etc.). One important condition should be observed when selecting the separation technique for the first dimension: it should be sufficiently different in separation selectivity (orthogonal) to the second dimension LC. Ultimately, these procedures aid in generating a higher number of identifications compared to 1-dimensional (1D) setups but require more instrument time as multiple fractions per sample are analyzed.[55–57]

Commonly, the final step in 2D and 1D setups in proteomics is to apply nanoflow reversed-phase chromatography. These procedures often use formic acid (FA) as an ion-pairing modifier due to a high amenability to MS applications. Peptides from samples or fractions elute in order of their hydrophobicities, then are sequenced by MS/MS and identified using peptide identification search algorithms.

1.4 Research Objectives

This thesis was separated into chapters representing different projects including published works and works in preparation for publication. All projects in this thesis relate to the chromatographic analyses of peptides, with a focus on post-translational modifications. Despite the importance of studying PTMs in proteomics, the understanding of their chromatographic properties is often nuanced by the intrigue of biological implications. Thus, the information on their chromatographic properties is rarely reported in detail. A detailed characterization of properties for PTMs not only permits the explanation of retention times for individual entries, but also permits the filtering of potential false-identifications for an entire data-output. Therefore, this thesis represents steps to characterize PTMs chromatographic properties which can be implemented to generate higher quality data outputs.

Chapter 2 focuses on the general effect of PTMs on peptide retention by comparing modified and nonmodified sequence counterparts for 88 modifications that the Krokhin group has studied in recent years. Chapters 3 and 4 build on strategies explained in Chapter 2, representing focused studies on 2 modifications, deamidation and glycosylation by O-linked N-acetylhexosamine (O-HexNAc). These include the expansion of characterization to sequence-specific properties for each modification in reversed phase and their behaviours in various first-dimensions of 2D chromatography setups.

Chapter 5 focuses on characterizing the retention properties of peptides in a HILIC variant, electrostatic repulsion-hydrophilic interaction chromatography (ERLIC), based on recent instances of this separation mode being applied to PTM analyses.[52–54] Prior to modelling PTM retention in ERLIC, the retention properties of nonmodified peptide properties needed to be comprehensively characterized. In this case, the behaviour of selected PTMs were later summarized using the same strategies as presented in Chapter 2.

1.5 References

1. Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. Top Down Proteomics: Facts and Perspectives. *Biochem Biophys Res Commun* **2014**, *445* (4), 683–693. https://doi.org/10.1016/j.bbrc.2014.02.041.
2. Sechi, S.; Chait, B. T. Modification of Cysteine Residues by Alkylation. A Tool in Peptide Mapping and Protein Identification. *Anal Chem* **1998**, *70* (24), 5150–5158. https://doi.org/10.1021/ac9806005.
3. Suttapitugsakul, S.; Xiao, H.; Smeekens, J.; Wu, R. Evaluation and Optimization of Reduction and Alkylation Methods to Maximize Peptide Identification with MS-Based Proteomics. *Mol Biosyst* **2017**, *13* (12), 2574–2582. https://doi.org/10.1039/C7MB00393E.
4. Jiang, X.; Shamshurin, D.; Spicer, V.; Krokhin, O. V. The Effect of Various S-Alkylating Agents on the Chromatographic Behavior of Cysteine-Containing Peptides in Reversed-Phase Chromatography. *Journal of Chromatography B* **2013**, *915–916*, 57–63. https://doi.org/10.1016/j.jchromb.2012.12.010.
5. Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; Boja, E. S.; Costello, C. E.; Cravatt, B. F.; Fenselau, C.; Garcia, B. A.; Ge, Y.; Gunawardena, J.; Hendrickson, R. C.; Hergenrother, P. J.; Huber, C. G.; Ivanov, A. R.; Jensen, O. N.; Jewett, M. C.; Kelleher, N. L.; Kiessling, L. L.; Krogan, N. J.; Larsen, M. R.; Loo, J. A.; Ogorzalek Loo, R. R.; Lundberg, E.; MacCoss, M. J.; Mallick, P.; Mootha, V. K.; Mrksich, M.; Muir, T. W.; Patrie, S. M.; Pesavento, J. J.; Pitteri, S. J.; Rodriguez, H.; Saghatelian, A.; Sandoval, W.; Schlüter, H.; Sechi, S.; Slavoff, S. A.; Smith, L. M.; Snyder, M. P.; Thomas, P. M.; Uhlén, M.; Van Eyk, J. E.; Vidal, M.; Walt, D. R.; White, F. M.; Williams, E. R.; Wohlschlager, T.; Wysocki, V. H.; Yates, N. A.; Young, N. L.; Zhang, B. How Many Human Proteoforms Are There? *Nat Chem Biol* **2018**, *14* (3), 206–214. https://doi.org/10.1038/nchembio.2576.
6. 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.
7. Davis, M. T.; Beierle, J.; Bures, E. T.; McGinley, M. D.; Mort, J.; Robinson, J. H.; Spahr, C. S.; Yu, W.; Luethy, R.; Patterson, S. D. Automated LC–LC–MS–MS Platform Using Binary Ion-Exchange and Gradient Reversed-Phase Chromatography for Improved Proteomic

Analyses. *J Chromatogr B Biomed Sci Appl* **2001**, *752* (2), 281–291. https://doi.org/10.1016/S0378-4347(00)00547-8.

8. Dai, J.; Wang, L.-S.; Wu, Y.-B.; Sheng, Q.-H.; Wu, J.-R.; Shieh, C.-H.; Zeng, R. Fully Automatic Separation and Identification of Phosphopeptides by Continuous PH-Gradient Anion Exchange Online Coupled with Reversed-Phase Liquid Chromatography Mass Spectrometry. *J Proteome Res* **2009**, *8* (1), 133–141. https://doi.org/10.1021/pr800381w.
9. Fonslow, B. R.; Yates III, J. R. Capillary Electrophoresis Applied to Proteomic Analysis. *J Sep Sci* **2009**, *32* (8), 1175–1188. https://doi.org/10.1002/jssc.200800592.
10. Alpert, A. J. Electrostatic Repulsion Hydrophilic Interaction Chromatography for Isocratic Separation of Charged Solutes and Selective Isolation of Phosphopeptides. *Anal Chem* **2008**, *80* (1), 62–76. https://doi.org/10.1021/ac070997p.
11. 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.
12. Roepstorff, P.; Fohlman, J. Proposal for a Common Nomenclature for Sequence Ions in Mass Spectra of Peptides. *Biomed Mass Spectrom* **1984**, *11* (11), 601. https://doi.org/10.1002/bms.1200111109.
13. 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.
14. 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.
15. Orsburn, B. C. Proteome Discoverer—A Community Enhanced Data Processing Suite for Protein Informatics. *Proteomes* **2021**, *9* (1), 15. https://doi.org/10.3390/proteomes9010015.
16. Tyanova, S.; Temu, T.; Cox, J. The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics. *Nat Protoc* **2016**, *11* (12), 2301–2319. https://doi.org/10.1038/nprot.2016.136.
17. Swaney, D. L.; Wenger, C. D.; Coon, J. J. Value of Using Multiple Proteases for Large-Scale Mass Spectrometry-Based Proteomics. *J Proteome Res* **2010**, *9* (3), 1323–1329. https://doi.org/10.1021/pr900863u.

18. 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.
19. 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.
20. Lou, R.; Shui, W. Acquisition and Analysis of DIA-Based Proteomic Data: A Comprehensive Survey in 2023. *Molecular & Cellular Proteomics* **2024**, *23* (2), 100712. https://doi.org/10.1016/j.mcpro.2024.100712.
21. Rotello, R. J.; Veenstra, T. D. Mass Spectrometry Techniques: Principles and Practices for Quantitative Proteomics. *Curr Protein Pept Sci* **2021**, *22* (2), 121–133. https://doi.org/10.2174/1389203721666200921153513.
22. Haag, A. M. Mass Analyzers and Mass Spectrometers; 2016; pp 157–169. https://doi.org/10.1007/978-3-319-41448-5_7.
23. Medhe, S. Ionization Techniques in Mass Spectrometry: A Review. *Mass Spectrom Purif Tech* **2018**, *04* (01). https://doi.org/10.4172/2469-9861.1000126.
24. Wilm, M.; Mann, M. Analytical Properties of the Nanoelectrospray Ion Source. *Anal Chem* **1996**, *68* (1), 1–8. https://doi.org/10.1021/ac9509519.
25. Wilm, M. Principles of Electrospray Ionization. *Molecular & Cellular Proteomics* **2011**, *10* (7), M111.009407. https://doi.org/10.1074/mcp.M111.009407.
26. Chen, G.; Fan, M.; Liu, Y.; Sun, B.; Liu, M.; Wu, J.; Li, N.; Guo, M. Advances in MS Based Strategies for Probing Ligand-Target Interactions: Focus on Soft Ionization Mass Spectrometric Techniques. *Front Chem* **2019**, *7*. https://doi.org/10.3389/fchem.2019.00703.
27. Banerjee, S.; Mazumdar, S. Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. *Int J Anal Chem* **2012**, *2012*, 1–40. https://doi.org/10.1155/2012/282574.
28. Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. Mobile and Localized Protons: A Framework for Understanding Peptide Dissociation. *Journal of Mass Spectrometry* **2000**, *35* (12), 1399–1406. https://doi.org/10.1002/1096-9888(200012)35:12<1399::AID-JMS86>3.0.CO;2-R.

29. Paizs, B.; Suhai, S. Fragmentation Pathways of Protonated Peptides. *Mass Spectrom Rev* **2005**, *24* (4), 508–548. https://doi.org/10.1002/mas.20024.
30. 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. *Molecular & Cellular Proteomics* **2020**, *19* (4), 716–729. https://doi.org/10.1074/mcp.TIR119.001906.
31. Hauschild, J.-P.; Peterson, A. C.; Couzijn, E.; Denisov, E.; Chernyshev, D.; Hock, C.; Stewart, H.; Hartmer, R.; Grinfeld, D.; Thoeing, C.; Reitemeier, B.; Kreutzmann, A.; Lange, O.; Balschun, W.; Venckus, A.; Kanngiesser, S.; Kholomeev, A.; Quiring, G.; Czemper, F.; Wouters, E.; Belford, M.; Mueller, M.; Grote, J.; Arrey, T. N.; Kraegenbring, J.; Strupat, K.; Mohring, S.; Crone, C.; Harder, A.; Robitaille, A.; Gajadhar, A.; Motamed, K.; Liu, Y.; Ferrer, D. L.; Damoc, N. E.; Criscuolo, A.; Srzentic, K.; George, E.; Yang, C.; Ntai, I.; Souza, A.; Wieghaus, A.; Makarov, A. A Novel Family of Quadrupole-Orbitrap Mass Spectrometers for a Broad Range of Analytical Applications. *Preprints (Basel)* **2020**. https://doi.org/10.20944/preprints202006.0111.v1.
32. Denisov, E.; Damoc, E.; Makarov, A. Exploring Frontiers of Orbitrap Performance for Long Transients. *Int J Mass Spectrom* **2021**, *466*, 116607. https://doi.org/10.1016/j.ijms.2021.116607.
33. 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.
34. 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.
35. 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. *Molecular & Cellular Proteomics* **2018**, *17* (12), 2534–2545. https://doi.org/10.1074/mcp.TIR118.000900.

36. Chernushevich, I. V.; Loboda, A. V.; Thomson, B. A. An Introduction to Quadrupole–Time-of-flight Mass Spectrometry. *Journal of Mass Spectrometry* **2001**, *36* (8), 849–865. https://doi.org/10.1002/jms.207.
37. Verma, S. P.; Santoyo, E. High-Performance Liquid and Ion Chromatography: Separation and Quantification Analytical Techniques for Rare Earth Elements. *Geostand Geoanal Res* **2007**, *31* (3), 161–184. https://doi.org/10.1111/j.1751-908X.2007.00842.x.
38. Guzman, U. H.; Martinez Del Val, A.; Ye, Z.; Damoc, E.; Arrey, T. N.; Pashkova, A.; Denisov, E.; Petzoldt, J.; Peterson, A. C.; Harking, F.; Østergaard, O.; Stewart, H.; Xuan, Y.; Hermanson, D.; Hock, C.; Makarov, A.; Zabrouskov, V.; Olsen, J. V. Narrow-Window DIA: Ultra-Fast Quantitative Analysis of Comprehensive Proteomes with High Sequencing Depth. *bioRxiv* **2023**. https://doi.org/10.1101/2023.06.02.543374.
39. 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. https://doi.org/10.1021/ac9016693.
40. 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.
41. Gussakovsky, D.; Anderson, G.; Spicer, V.; Krokhin, O. V. Peptide Separation Selectivity in Proteomics LC-MS Experiments: Comparison of Formic and Mixed Formic/Heptafluorobutyric Acids Ion-Pairing Modifiers. *J Sep Sci* **2020**, *43* (20), 3830–3839. https://doi.org/10.1002/jssc.202000578.
42. 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.
43. 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.
44. 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. https://doi.org/10.1016/j.chroma.2017.12.046.

45. Yeung, D.; Klaassen, N.; Mizero, B.; Spicer, V.; Krokhin, O. V. Peptide Retention Time Prediction in Hydrophilic Interaction Liquid Chromatography: Zwitter-Ionic Sulfoalkylbetaine and Phosphorylcholine Stationary Phases. *J Chromatogr A* **2020**, *1619*. https://doi.org/10.1016/j.chroma.2020.460909.
46. 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.
47. 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. https://doi.org/10.1021/acs.analchem.6b04544.
48. Neale, Q.; Prefontaine, A.; Battellino, T.; Mizero, B.; Yeung, D.; Spicer, V.; Budisa, N.; Perreault, H.; Zahedi, R. P.; Krokhin, O. V. Compendium of Chromatographic Behavior of Post-Translationally and Chemically Modified Peptides in Bottom-Up Proteomic Experiments. *Anal Chem* **2023**, *95* (39), 14634–14642. https://doi.org/10.1021/acs.analchem.3c02412.
49. 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.
50. Fíla, J.; Honys, D. Enrichment Techniques Employed in Phosphoproteomics. *Amino Acids* **2012**, *43* (3), 1025–1047. https://doi.org/10.1007/s00726-011-1111-z.
51. Goumenou, A.; Delaunay, N.; Pichon, V. Recent Advances in Lectin-Based Affinity Sorbents for Protein Glycosylation Studies. *Front Mol Biosci* **2021**, *8*. https://doi.org/10.3389/fmolb.2021.746822.
52. Zacharias, L. G.; Hartmann, A. K.; Song, E.; Zhao, J.; Zhu, R.; Mirzaei, P.; Mechref, Y. HILIC and ERLIC Enrichment of Glycopeptides Derived from Breast and Brain Cancer Cells. *J Proteome Res* **2016**, *15* (10), 3624–3634. https://doi.org/10.1021/acs.jproteome.6b00429.

53. Cui, Y.; Yang, K.; Tabang, D. N.; Huang, J.; Tang, W.; Li, L. Finding the Sweet Spot in ERLIC Mobile Phase for Simultaneous Enrichment of N-Glyco and Phosphopeptides. *J Am Soc Mass Spectrom* **2019**, *30* (12), 2491–2501. https://doi.org/10.1007/s13361-019-02230-6.
54. Cui, Y.; Tabang, D. N.; Zhang, Z.; Ma, M.; Alpert, A. J.; Li, L. Counterion Optimization Dramatically Improves Selectivity for Phosphopeptides and Glycopeptides in Electrostatic Repulsion-Hydrophilic Interaction Chromatography. *Anal Chem* **2021**, *93* (22), 7908–7916. https://doi.org/10.1021/acs.analchem.1c00615.
55. 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.
56. 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* (18), 7036–7042. https://doi.org/10.1021/ac800984n.
57. Dwivedi, R. C.; Spicer, V.; Krokhin, O. V. Reversed-Phase HPLC of Peptides – A Valuable Sample Preparation Tool in Bottom-Up Proteomics: Separation Selectivity in Single and Multi-Dimensional Separation Modes. In *Sample Preparation in Biological Mass Spectrometry*; Springer Netherlands, 2011; pp 1031–1050. https://doi.org/10.1007/978-94-007-0828-0_48.

2 Chapter 2: Compendium of chromatographic behavior of post-translationally and chemically modified peptides in bottom-up proteomic experiments

2.1 Article Headline and Authors:

Compendium of chromatographic behavior of post-translationally and chemically modified peptides in bottom-up proteomic experiments

Quinn Neale,[1] Alexandre Prefontaine,[1] Taylor Battellino,[1] Benilde Mizero,[1] Darien Yeung,[2] Victor Spicer,[3] Nediljko Budisa,[1] Helene Perreault,[1] Rene P. Zahedi,[2,3,4,5] Oleg V. Krokhin*[2,3,4]

[1]Department of Chemistry, University of Manitoba, 360 Parker Building, Winnipeg, Manitoba R3T 2N2, Canada
[2]Department of Biochemistry and Medical Genetics, University of Manitoba, 336 Basic Medical Sciences Building, 745 Bannatyne Avenue, Winnipeg, Manitoba R3E 0J9, Canada
[3]Manitoba Centre for Proteomics and Systems Biology, University of Manitoba, 799 JBRC, 715 McDermot Avenue, Winnipeg, Manitoba R3E 3P4, Canada
[4]Department of Internal Medicine, University of Manitoba, 799 JBRC, 715 McDermot Avenue, Winnipeg, Manitoba R3E 3P4, Canada
[5]CancerCare Manitoba Research Institute, 675 McDermot Avenue, Winnipeg, Manitoba R3E 0V9, Canada

Published:
Analytical Chemistry
September 22nd, 2023
Volume 95, Issue 39
DOI: https://doi-org.uml.idm.oclc.org/10.1021/acs.analchem.3c02412

2.2 Contribution of Authors

For this chapter, QN was involved in data collection and analysis of retention features of GlcNAc and GalNAc modification of Ser/Thr; summarizing data for all modifications under investigation; finding general trends and correlations between retention shifts and slopes for individual HI_{mod} vs. $HI_{non\text{-}mod}$ dependencies; manuscript writing and preparation of the figures. Other authors: AP, TB, BM, and DY were all involved in data acquisition and analysis for respective modifications. VS performed computational work. NB, HP, RZ and OK were involved in writing. NB, HP and OK supervised their respective students involved in the project.

2.3 Abstract

We have systematically evaluated the chromatographic behavior of post-translationally / chemically modified peptides using data spanning over 70 of the most relevant modifications. These retention properties were measured for standard bottom-up proteomic settings (fully porous C18 separation media, 0.1% formic acid as ion pairing modifier) using collections of modified/non-modified peptide pairs. These pairs were generated by spontaneous degradation, chemical or enzymatic treatment, analysis of synthetic peptides or the co-translational incorporation of non-canonical proline analogues. In addition, these measurements were validated using external data acquired for synthetic peptides and enzymatically induced citrullination. Working in units of hydrophobicity index (HI, % ACN) and evaluating the average retention shifts (ΔHI) represents the simplest approach to describe the effect of modifications from a didactic point of view. Plotting HI values for modified (y-axis) vs. unmodified (x-axis) counterparts generates unique slope and intercept values for each modification defined by the chemistry of the modifying moiety: its hydrophobicity, size, pKa of ionizable groups and position of the altered residue. These composition-dependent correlations can be used for coarse incorporation of PTMs into models for prediction peptide retention. More accurate predictions would require the development of specific sequence-dependent algorithms to predict ΔHI values.

2.4 Introduction

Post-translational modifications (PTM) and chemical peptide modifications (CM), henceforth abbreviated in their entirety as PTCM, are modifications to amino acids that involve the covalent attachment, removal and re-arrangement of amino acids' functional moieties. While protein PTM is usually a specifically targeted process and plays an important role in the regulation of biological systems,[1,2] CM are usually introduced intentionally (as 'fixed' modifications) or unintentionally ('variable' modifications) during the handling and processing of biological samples. CMs are often employed to facilitate protein digestion (reduction of disulfides and alkylation of free cysteines),[3] to improve detection sensitivity[4] or to enable quantitative proteomic analysis,[5,6] to name a few applications. The list of typical spontaneous modifications (due to sample handling) includes oxidation (Met, Trp), deamidation (Asn, Gln), N-terminal loss of ammonia, among others, some of which also occur *in-vivo*.

As a result of their effect on protein and peptide mass, structure, and other physicochemical properties, PTCM have a direct impact on proteomic studies and typically complicate and impair protein and peptide identification. Thus, many PTCMs remain undetected while consuming an essential portion of the available MS/MS acquisition capacity, diminishing the overall output of LC-MS/MS-based studies. There are hundreds of PTCMs that can occur in bottom-up proteomic analyses, including unwanted modifications caused by side reactions.[7-9] For example, peptide formylation has been identified as one of the typical unwanted modifications found in mass-tolerant searches of LC-MS/MS data.[9-11]

The prevalence of PTCMs and their impact on protein and peptide identification, including the potential of false-positive peptide identifications arising from PTCMs that are not accounted for in database searches, merit a deeper study. Mass shifts, changes in MS response,[12] charge state distribution,[13] susceptibility to neutral loss or unique fragmentation[14] are the most important mass-spectrometry properties being evaluated. Apart from these, changes of chromatographic properties can also be leveraged and integrated into data analysis workflows to improve confidence and exclude false positive PSMs. In the proteomics field, the differentiating power of reversed-phase liquid chromatography (RPLC) which is used in virtually all MS-based proteomics studies, is still underutilized. Therefore, it would be very useful to understand how specific PTCMs induce retention shifts in RPLC to accurately predict retention times[15] of PTCM peptides.

While the notion of retention shifts upon modification has long been common knowledge in the field of proteomics, accurate evaluation and introduction into retention prediction models is a more recent trend.[15-18] ***Generation of retention data*** and ***accurate measurement of retention values*** constitute two major prerequisites of these studies. Different approaches have been used to collect retention data of PTCM-containing peptides, including analysis of *in-vivo* modifications and products of spontaneous degradation,[19] peptides modified by targeted chemical or enzymatic treatments,[12, 20-23] as well as synthetic peptide analogues[14, 17]. Co-translational (metabolic) incorporation of modified residues[24] represents a potent method but is still underutilized in the context of studying PTM-carrying peptides. All these approaches differ in terms of throughput and cost. Identification of the most abundant *in-vivo* or spontaneous modifications[19] does not involve additional experimental cost, as each LC-MS run can be reanalyzed to retrieve these data. Identification of rare or infrequent modifications requires specific enrichment protocols, which also helps to reduce the required instrument time.[12] Methods based on synthetic analogs[14, 17] are comparatively more expensive, as they require both peptide synthesis and dedicated instrument time.

The concept of determining retention shifts induced by PTMCs is based on the premise that both modified and nonmodified counterparts have been detected. Zybailov et al.[19] monitored *in-vivo* and spontaneous modifications in a large scale proteomic analysis of *Arabidopsis thaliana*. Retention time shifts were reported for over 30 different modifications. Phosphorylated peptides can be found in shotgun bottom-up proteomics runs along with their non-modified parent molecules, but their low detection sensitivity and the sub stoichiometric character of this modification call for the application of enrichment protocols.[25] Marcantonio et al.[12] used phosphopeptide enrichment followed by dephosphorylation by alkaline phosphatase to generate retention data for 750 peptide pairs. Similarly, we have used deglycosylation with PNGaseF to obtain retention values of N-glycosylated and their deglycosylated (Asp containing) analogs.[23] In contrast to these examples of enzymatic removal of a modification, citrullination of Arg residues with arginine deiminases[20,21] has been used to study the chromatographic behavior of citrullinated peptides. Examples of chemically induced modifications in our dedicated RPLC studies can be found for acetylation,[22] Met oxidation,[16] a number of alkylation modes of Cys,[26] Lys carbamylation,[21] labeling with isotopically coded tags for relative quantitation iTRAQ4/iTRAQ8/TMT/TMTpro.[27,28] A synthetic peptide approach has been applied by Kuster

and coworkers for several modifications.[14,29] A seminal study by Marx et al.[29] described the acquisition of LC and MS fragmentation data for 100,000 designed phosphorylated – non-phosphorylated peptide pairs, with ~71,000 non-modified and 58,000 phosphorylated peptides detected and used to characterize retention shifts. The same laboratory has reported the fragmentation properties and chromatographic behavior of 21 different PTMs on Lys, Arg, Tyr and Pro, comprising a total of ~200 pairs of synthetic peptides for each modification.[14] The different methods for generating retention data used in this work are shown schematically in **Figure 2.1**.

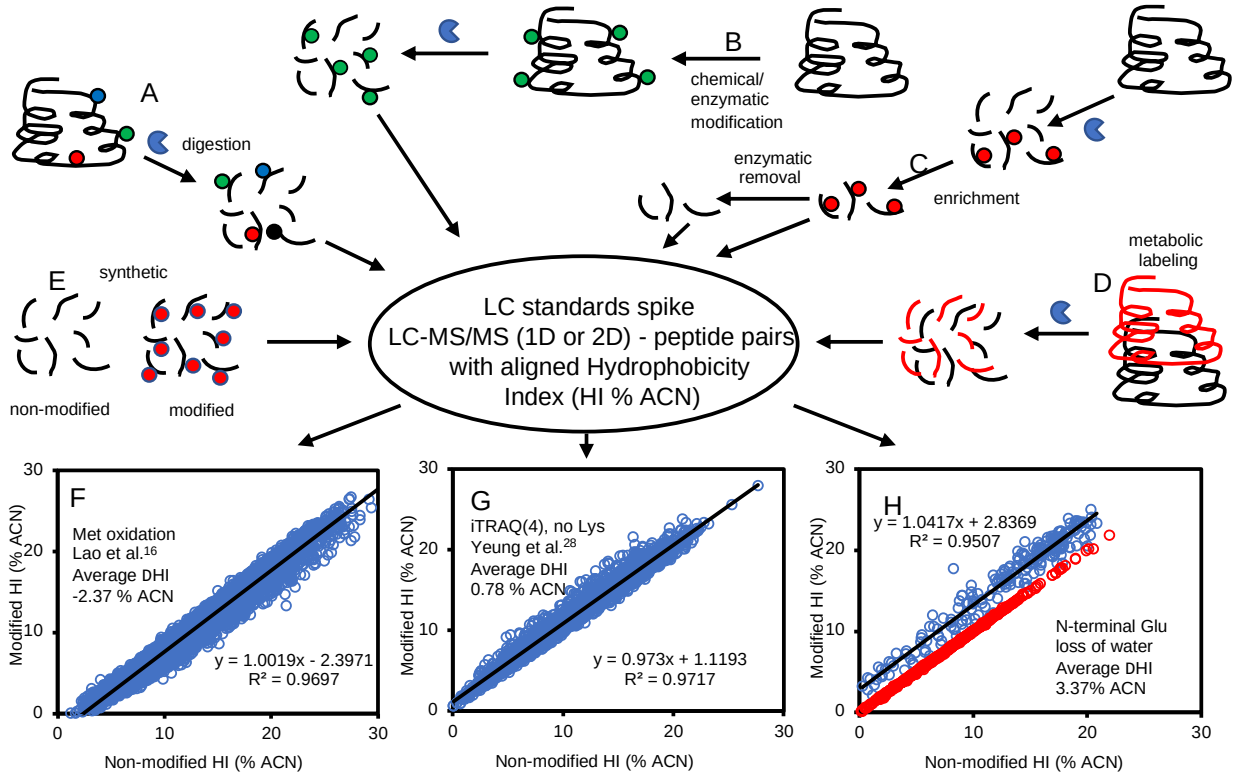


Figure 2.1 Acquisition of retention data for modified/non-modified peptide pairs and retention shift plots from our datasets. The various methods for obtaining retention data in this work include A) monitoring spontaneous degradation and *in-vivo* modifications; B) chemical or enzymatic induction of modifications; C) enzymatic removal of modification; D) co-translational insertion of amino acid analogs (metabolic incorporation); E) synthetic peptides. Representative examples of HI_{mod} vs. $HI_{non\text{-}mod}$ plots are given for F) methionine oxidation, G) N-terminal labeling with iTRAQ(4), H) spontaneous loss of water at the N-terminal glutamic acid. The latter plot shows both 230 true (blue) and 614 in-source (red) modifications.

Accurate measurement and representation of retention shifts is another issue that is often being overlooked in the design of proteomic experiments. The use of ***linear water-acetonitrile***

gradients and ***matching chromatographic data with retention standards*** is a must when it comes to RPLC focused studies. The use of non-linear (segmented) gradients is advantageous for optimal delivery of RPLC-separated peptide mixtures into the mass spectrometer,[30] but significantly complicates the measurement of chromatographic properties. The vast majority of RPLC PTCM studies listed above used multiple LC-MS runs (multi-fraction or multi-sample experimental designs), sacrificing retention time assignment precision due to the poor reproducibility on nano-LC separations. For example, the pioneering reports on the retention of phosphopeptides[12, 29] and modifications in *Arabidopsis thaliana*[19] did not use retention time alignment or, in case of the latter study, linear gradients. Application of retention standards for proteomic studies was first proposed by us and others.[31,32] Currently a wide variety of standards are available[33,34] for routine applications and can be easily adapted for interchangeable use. Variation of other LC parameters such as type of sorbent, column temperature, gradient slope, flow rate, and column size should also be considered.[35]

Time units are the most natural for any HPLC measurement,[19] however when gradient elution is applied in proteomics, the unitless representation in acetonitrile percentage (HI, Hydrophobicity Index)[32] has advantages. Retention time shifts are measured for a particular instrumental setting where the exact slope of the acetonitrile gradient cannot be guaranteed, especially for nanoflow separations. Spiking each sample with a special standard with indexed retention values allows an actual gradient to be measured experimentally,[32] information that might otherwise be lost. The assignment of retention values and the effect of modification in HI also directly informs the expected retention time shifts when planning proteomic experiments. For example, the average retention shift for a single Met oxidation is -2.37 % ACN,[16] so at a gradient of 0.33% per minute (typical for 2 hr LC-MS acquisitions), the corresponding average retention time shift should be approximately -7.2 minutes. An alternative peptide hydrophobicity scale uses arbitrary iRT units.[33] HI and iRT retention scales have been aligned and can be used interchangeably,[36] but for this work we consider the former more appropriate because it is directly related to experimental settings.

Our lab has utilized accurate retention database collection to optimize our Sequence-Specific Retention Calculator prediction model (SSRCalc, SSRCalcDB)[37] and to study PTMs. Since 2008, we have used the same retention standards,[32] linear water/acetonitrile gradients, and identical separation media (Luna C18(2) from Phenomenex) for all our RPLC-focused studies, generating

a unique collection of high-quality data under virtually the same LC conditions. Across three generations of mass-spectrometers, consistent chromatography has been a key feature allowing easy system-to-system transition and reliable chromatographic measurements. Here, we leverage our unique and rich data collection to evaluate retention properties of PTCM peptides (published[16, 21-23, 26-28] and under preparation). We supplemented our in-house data with the collection of high-quality retention data for 22 PTMs extracted from Zolg et al.[14, 38] and for citrullination reported by Fert-Bober et al.[20] In particular, the external studies used different RPLC settings and units to measure retention properties. This was not a drawback but was used as an opportunity to assess the cross-compatibility of retention studies performed under different chromatographic conditions and hydrophobicity scales – a key issue in the continuous expansion of the chromatographic PTCM compendium.

2.5 Materials and Methods

Experimental settings for retention data collection used in this work can be found in Zolg. et al.[14] and Fert-Bober et al.,[20] and our previous publications on various Cys alkylation chemistries,[26] Met oxidation,[16] Arg citrullination and Lys carbamylation,[21] N-glycosylation,[23] acetylation,[22] iTRAQ/TMT labeling.[27,28] This experimental section describes generic procedure for digestion and 1D or 2D LC-MS/MS analysis of zebrafish brain tissue for detection of the most common PTMs, *E. coli* modified to study co-translational metabolic incorporation of non-canonical proline derivatives,[39,40] and synthetic peptide mixtures to evaluate the effect of O-GlcNAc and -GalNAc glycosylation.

2.5.1 Materials, Chromatographic Columns

HPLC-grade water and acetonitrile were used for preparation of the eluents. All chemicals were obtained from Sigma Chemicals (St. Louis, MO) unless otherwise noted. Sequencing grade modified porcine trypsin (Promega, Madison, WI) was used for digestion. The designed standard peptides P1-P6[32] were synthesized by BioSynthesis Inc (Lewisville, TX). SpikeMix PTM-Kit 55 – OGlcNAc-OGalNAc-Unmodified, PTM-Kit 56 – OGalNAc and PTM-Kit 57 – OGlcNAc have been purchased from JPT Peptide Technologies (Berlin, Germany). In-house packed C18-XTerra (Waters, Milford, MA) 5 µm, 100 Å, and 100 µm x 30 cm Luna C18(2) (Phenomenex, Torrance, CA) 3 µm, 100 Å were used for the first and second dimension RPLC separations, respectively.

2.5.2 Co-translational Incorporation of Modified Proline Residues into Escherichia coli

Proline auxotrophic *Escherichia coli* MG1655 (genotype: ΔproBA::frt, ΔproC::frt, ΔargECBH::frt, ΔlysA::frt) was cultured in chemically defined media (see Supporting Information), with a growth-limiting amount of proline (25µM) and an excess of the proline analogs (1mM) in the growth medium. The analogs used were (2S,4R)-4-fluoroproline, (2S, 4S)-4fluoroproline, (2S)-1,3-thiazolidine-2-carboxylic acid, (4R)-1,3-thiazolidine-4-carboxylic acid, (2S,4R)-4-hydroxyproline, (DL)-pipercolic acid, 3,4-dehydro-L-proline, (1S,3S,5S)-2-azabicyclo[3.1.0]hexane-3-carboxylic acid, and (1R,3S,5R)-2-azabicyclo[3.1.0]hexane-3-carboxylic acid (Sigma-Aldrich, **Supplementary Figure S2.1**). Cells were cultured at 30°C for 24 hours, lysed by sonication, and digested as described below.

2.5.3 Preparation of Digests for HPLC-MS Analysis

The zebrafish brain tissue and *E.coli* digests were prepared by denaturation with sodium dodecyl sulfate (SDS), reduction with dithiothreitol (DTT) and alkylation with iodoacetamide (IAA), the proteins were digested with trypsin, using the SP3 method, as described elsewhere.[41] After digestion, samples were C18 purified, analyzed to determine peptide content (Nano Drop 2000, Thermo Fisher Scientific) and lyophilized.

2.5.4 First Dimension Separation Conditions

A 100 µg aliquot of the *Danio rerio* digest was used for the first-dimension separations, under high pH RPLC conditions.[42] A 1% ACN per minute linear gradient was applied at a 150 uL/min flow rate. Both eluents A (water) and B (90% ACN) contained 10 mM ammonium formate at pH 10. One-minute fractions were collected within a peptide elution window of 8-28 minutes, concatenated pairwise, and lyophilized.

2.5.5 Second Dimension LC-MS/MS

An Orbitrap Exploris 480 coupled to NanoEasy1200 nano-LC system (both Thermo Fisher Scientific) was used in a standard data dependent acquisition mode (**Supplementary experimental procedures**). The following chromatographic settings were used: Luna C18(2) 3 µm, 100 µm x 30 cm column, linear water/acetonitrile gradients of ~0.3% per minute with 0.1% formic acid as ion pairing modifier. Lyophilized fractions or *E. coli* digests were re-suspended in 0.1% formic acid in water (buffer A) spiked with P1-P6 peptides[32] to provide injection of ~0.5 µg of the sample with ~200 fmol of the standard peptides for retention time alignment purposes. Synthetic peptide mixtures have been diluted with buffer A and spiked with retention standard to make ~200 fmol of each analyte per injection.

2.5.6 Peptide Identification and Assignment of Modified/Non-Modified Peptide Pairs

Raw LC-MS/MS spectrum files were converted to .mgf format using the tool bundled into Thermo Scientific Proteome Discoverer™. The X!Tandem search engine (Vengeance 2015.12.15.2) was used for peptide/protein identification with a 10 and 20 ppm mass accuracy for parent/daughter ions. The search parameters were as follows: fixed modification of Cys (carbamidomethylation, 57.021@C), variable modifications were deamidation of Asn and Gln

(0.984@N/Q), oxidation of Met and Trp (15.995 and 31.990 @ M/W), phosphorylation of Ser/Thr/Tyr (+79.966 @ S/T/Y) protein N-terminal acetylation (+42.011), N-terminal cyclization for Gln/Cys (-17.027) and Glu (-18.021). An additional search was performed to detect unwanted peptide formylation at Ser and Thr and peptide N-termini (+27.995).[9] Tryptic enzyme specificity was used with up to two missed cleavages allowed. In peptides with multiple potential modification sites, explicit evidence of localization (corresponding b or y ions) was required; those failing this requirement were excluded.

Retention times in RPLC (formic acid) were converted into hydrophobicity index (HI, % acetonitrile) units using linear regression against tabulated retention values of standard peptides spiked in each sample/fraction. A rigorous filtering procedure was used based on confident identification scores (log(e) ≤ -1 or log(e) ≤ -3 for metabolic co-translational incorporation of Pro analogs), accuracy of the parent mass, and retention prediction errors to ensure high quality retention data. X! Tandem search results at log(e) ≤ -1 usually resulted in 0.3-0.5% false positive rate, which was confirmed by a corresponding population of SSRCalc prediction outliers (~5 per 1000 of identified peptides). These retention outliers were subsequently removed.

Retention time alignment for external data was done as follows. The "evidence.txt" results files from Zolg et al.[14,38] contained ID information for both synthetic target peptides and the PROCAL retention standards[43] spiked into each sample. HI values have been assigned for all PROCAL calibrating peptides using SSRCalc retention time prediction correlation plot as shown in **Figure S2.2**. These tabulated values were used to convert retention times for all identified peptides in individual runs and their corresponding retention shifts into % ACN units. Fert-Bober et al.[20] reported aligned retention times for 3245 citrullinated/non-modified peptide pairs without information on iRT standard used for LC calibration. 310 non-modified peptides from this collection overlapped with data from previous runs conducted our lab. Using their HI index values in SSRCalc DB allowed direct RT-HI conversion for all sequences in this dataset as shown in **Figure S2.3.**

2.6 Results and Discussion

2.6.1 Detection of PTCMs in Standard Bottom-Up Proteomic Runs.

The total yield of PTCM peptide identifications is automatically constrained by the applied search parameters. Our analysis of the 2D LC-MS/MS data of zebrafish samples, indicates that ~27% of PSMs can be assigned to modified peptides (**Figure S2.4**). Following a recent report by Lenco et al.,[9] we found that ~3% of PSMs in this dataset represent unwanted peptide formylation. For many PTM-carrying peptides the corresponding non-modified form was identified as well, giving the basis for the unsupervised assignment of retention shifts. A comparison of these shifts against the values from dedicated studies is shown in **Figure S2.4**. Clearly, some of these match well, while the others such as the N-terminal loss of water at Glu (**Figure 2.1h**) deviate significantly, meriting deeper investigation into possible sources of these errors.

2.6.2 Retention Shifts Assignment: Reproducibility, Possible Sources of Errors and Requirements.

Supplementary Table S2.1 shows the average retention shifts, peptide size and hydrophobicity for retention datasets analyzed in this study. The average effect of a modification expressed in HI units (retention shift, ΔHI), is the most informative way to represent the expected overall impact of PTCM on RPLC behavior of peptides (**Figure 2.2**). However, there are, several issues that can affect the accuracy of these evaluations - these should be carefully considered as discussed in this section.

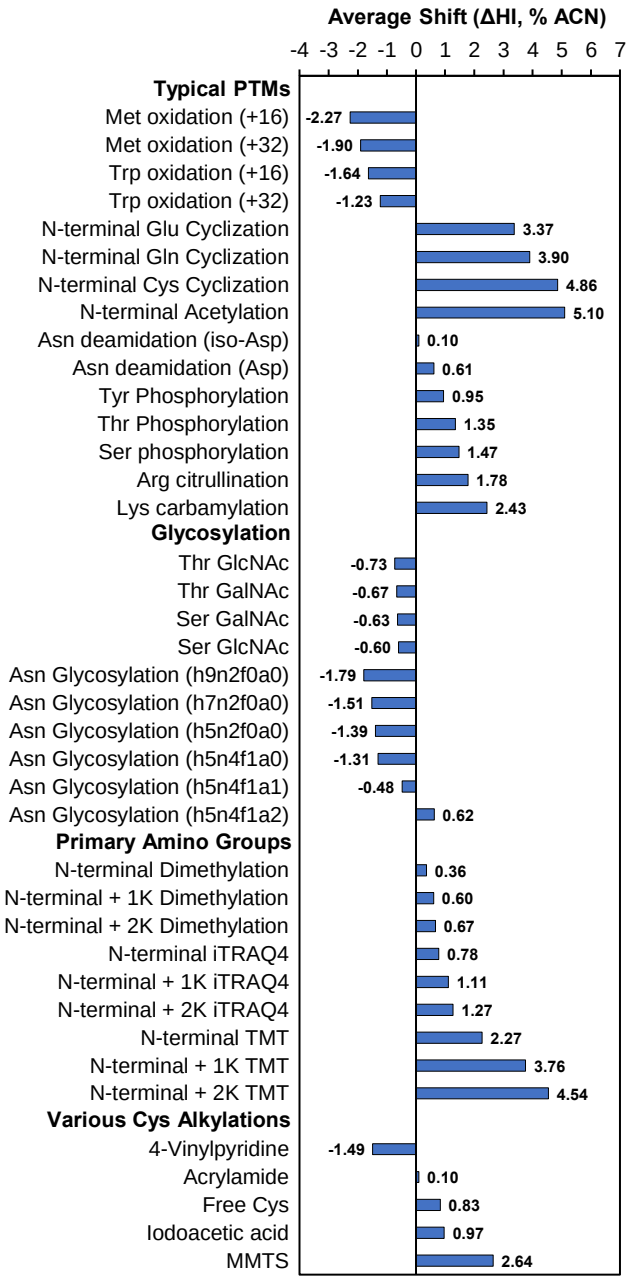


Figure 2.2 Average retention shifts (ΔHI, % ACN) for selected modifications. Supplementary Table S2.1 contains the entire collection of retention data used in this work.

Supplementary Figure S2.5 shows the distribution of retention shifts obtained in three independent experiments each for methionine oxidation and arginine citrullination. Overall distributions and average values for methionine oxidation were consistent for three datasets, except for an erroneous spikes readout near ΔHI~0 for ProteomeTools data[38] and zebrafish analysis (**Figure S2.5 b,c**). This is a consequence of in-source methionine oxidation.[44] These datapoints

must be removed for accurate analysis of the retention shifts. In addition to methionine oxidation, we found other in-source modifications characterized by co-elution of modified and non-modified peptide pairs, such as N-terminal neutral loss of water at Glu and Trp (+15.995 Da) oxidation. Notably, the former modification was affected the most (**Figure 2.1h**): ~73% instances of detected pairs were assigned to in-source generated products. Dealing with such outliers often requires manual interpretation while knowing expected retention time shifts *a-priori*, which is not always the case. Considering retention shifts in both separation dimensions for 2D runs is very helpful as the instances of identical retention of modified/non-modified pair on both LC columns are rare and likely indicate in-source origin of the modified analyte in question. Deamidation represents one of the most complicated instances: two products of this reaction have identical minor mass shift of 0.984 Da and may have retention times identical to the unmodified molecule. Deamidation data used in this work have been manually verified to ensure a complete separation of parent molecules and two degradation products (aspartic and iso-aspartic acid analogs).

Figure S2.5 d-f shows results of three independent measurements for arginine citrullination. Our[21] and Zolg's et al.[14] datasets being smaller in size still showed similar behavior (**Figure S2.5 d,e**). The collection of Fert-Bober et al.[20] data is characterized by larger average shift of 2.13 HI units (**Figure S2.5 f)**. The most likely reason for this discrepancy is the authors' pre-set cut off for retention change determined by the analysis of synthetic peptides.

Additional requirements important for the accurate evaluation of retention properties include: i) definitive evidence of PTMs' localization for peptides with multiple potential modification sites; ii) representative peptides' size and hydrophobicity compared to typical bottom-up proteomic runs as shown in **Table S2.1**; iii) equal (abundance driven) representation of the residues in the proximity to modification site; iv) compatibility of chromatographic conditions for various experimental settings. The former is usually addressed by assigning the false localization rate using statistical approaches.[45-46] In our study, the vast majority of modifications were site specific or targeted rare amino acids, whereas in other cases such as phosphorylation, a strict requirement of having definitive product ion confirming localization was applied. Alternatively, a synthetic peptide approach was a prerequisite for labile modifications such as GlcNAc/GalNAc. To evaluate the influence of chromatographic conditions, we considered identical modifications targeted in Zolg et al.[14,38] and our studies. This provides an insight into how chromatographic settings can affect conclusions, specifically comparing results from Luna

C18(2) separations at room temperature with those of Reprosil C18AQ at 50 ºC with addition of 5% DMSO to the eluent. The three examples shown in **Table S2.1** - citrullination of Arg and oxidation of Met and Pro showed nearly identical behaviors. At the same time, average retention shifts for Tyr phosphorylation were found to be -0.33 and 0.95 % ACN for Reprosil C18AQ and Luna C18(2), respectively. Independent measurements using Reprosil C18AQ and formic acid/water/ACN at room temperature (manuscript in preparation) confirmed the same trend, suggesting that the differences in packing material chemistry may affect behavior of selected modifications.

Knowing retention of non-modified peptides is important, however in some cases predicted hydrophobicity values may be used instead. **Figure S2.6** shows retention shifts for the entire collection of 800 singly citrullinated peptides detected following enzymatic treatment where retention of non-modified peptides was predicted *in-silico*.[21] The distribution of the shifts was found wider due to inherent prediction errors but showing very similar values of average shifts: 1.77 vs. 1.53 % ACN. This approach requires a prediction model, which calculates hydrophobicity in units identical to measured values. SSRCalc RP (formic acid) model and SSRCalc DB both report values in HI (% ACN), thus allowing such transformation in the absence of the data for non-modified peptides.

2.6.3 HI (Modified) vs. HI (Non-Modified) Plots to Summarize RPLC Behavior of Modified Peptides.

Plotting retention values of modified analogs vs. those of non-modified allows a precise evaluation of how retention shift values change with peptide hydrophobicity. Zolg et al.[14] used this approach for 21 targeted modifications. Each plot was characterized by a unique slope and intercept and deviations of datapoints from the trend line. Intercepts expressed in iRT units were used to evaluate the contribution of individual modifications to peptide retention.[14] We applied the same format except for expressing retention values in HI units, changing axis assignments (HI_{mod} on Y-axis, $HI_{non\text{-}mod}$ X-axis) and monitoring both intercepts and average retention shifts. **Figure 2.1f-h** show such plots for Met oxidation, N-terminal labeling with iTRAQ-4plex and N-terminal loss of water at Glu. This way of representing the data (summarized in **Table S2.1** for all modifications) has several advantages. First, using HI units allows for easy interpretation of the data as they tie directly to retention times through experimental gradient slope settings. Second,

average shifts inform on expected retention variations of the entire peptide population, while the intercepts show expected retention shifts for the most hydrophilic molecules. Intercepts and average shift values are identical when the slope equals 1, which is rarely the case.

Several CMs target primary amino groups (N-termini and Lys) and are typically performed at the peptide level. Because of this, retention data for those are shown separately for N-terminal modifications (no Lys), peptides with one Lys (mostly C-terminal) and two (C-terminal and internal) lysines. Although these retention studies are relevant for respective workflows which derivatize at the peptide level, when it comes to *in-vivo* modification of internal lysines, a synthetic peptide approach or protein labeling prior digestion have evident advantage. When using acetylation at the peptide level, the average contribution of internal Lys can be estimated as a difference between 2 Lys and 1 Lys: 11.46 - 9.22 = 2.24, compared to 2.91 % ACN in the results of Zolg et al.[14].

2.6.4 HI vs. HI Plots to Summarize RPLC Behavior of Peptides Under Different Chromatographic Conditions.

The slopes and intercepts in **Table S2.1** can be used for rough incorporation of PTCMs into prediction models: the predicted HI value for a non-modified molecule is adjusted by ΔHI, which depends from peptide hydrophobicity. Interestingly, a similar concept of retention shift can be applied to the variations of chromatographic conditions, when their effect on peptide retention have been similarly measured in HI units. **Table S2.1** shows two of such examples: change of column temperature from 25 to 55 ºC (manuscript in preparation) and switching ion-pairing modifiers from 0.1% formic to 0.5% acetic acid.[47] The latter leads to an average -0.62% ACN retention decrease, which also depends on the number of positively charged residues and can be compartmentalized further, similar to the number of modified Lys residues for acetylation/iTRAQ/TMT.

2.6.5 Deviations from HI (Modified) vs. HI (Non-Modified) Plots: Sequence-Specific Character of Retention Shifts upon Modification.

Since peptide retention in RPLC is driven by a multitude of composition and sequence-dependent factors one should not expect a constant retention shift upon a specific modification. The amplitude of the variations depends on the position and nature of the modified residue. N-

terminal ion-pairing shielding of the residues (**Figure 2.3a**) and stabilization of amphipathic helical structures upon interaction with the C18 surface (**Figure 2.3b-e**) represent two major mechanisms that drive sequence-dependent character of interactions.[37] For example, methionine oxidation and lysine carbamylation are characterized by similar amplitudes of average retention shifts (**Table S2.1**). However, R^2 is larger, and the root mean square error (RSME) value is smaller for the latter modification. This difference is driven by possible involvement of hydrophobic Met in amphipathic helical structures, which tend to be retained very strongly.[16] Met oxidation in hydrophobic face of amphipathic helix may result retention drop as low as -9 % ACN, compared to -2.37 % average value. Subsequently, accurate retention prediction for different modifications may require datasets of various sizes to properly address possible effect of peptide secondary structure. Another notable example is a drastic difference between iso-aspartic and aspartic acids for deamidation. Due to their acidic character and reduced ion-pairing association, both analogues tend to be retained stronger compared to asparagine containing peptides. This is not the case for iso-Asp inside amphipathic helixes (manuscript in preparation). Placing an additional carbon atom into the peptide's backbone leads to extremely low helical propensity of iso-Asp residue and lowers peptide retention (**Figure 2.3e**).

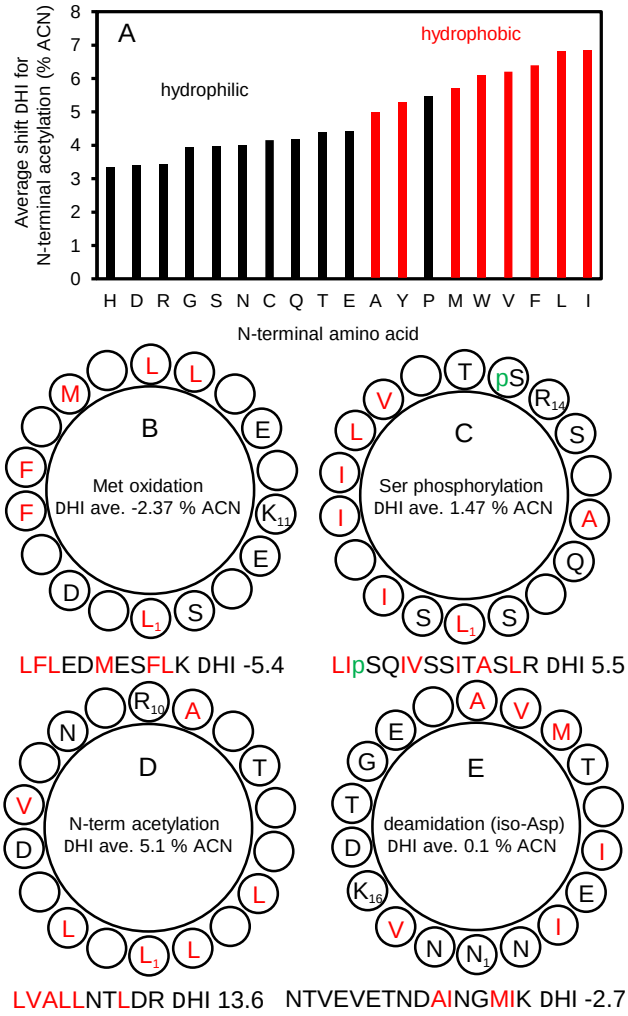


Figure 2.3 Sequence dependent features which determine the variation of retention shifts upon modification. A – the dependence of average retention shifts for N-terminal acetylation from the nature of N-terminal residue. Lys is not shown as it carries two acetylation sites: N-terminus and side chain. Amphipathic peptide helicity leads to significant changes in peptide behavior (compared to average values) for B) Met oxidation C) Ser phosphorylation, D) N-terminal acetylation, E), deamidation (iso-Asp product). Hydrophobic (red) and hydrophilic (black) residues being placed in axial projection of helical wheel, occupy distinctive regions: hydrophobic and hydrophilic faces of amphipathic helix, leading to preferential interaction of the former with C18 phase and higher retention.

2.6.6 Factors Determining Slope and Intercept.

It would be advantageous to be able to predict an expected retention shift based solely on molecular structure. Zolg et al.[14] have shown this may work for modifications of similar chemistry such as acylations of lysine. These modifications all have slopes below 1 and positive retention shifts that could be predicted based on elemental composition. However, the effect of peptide

modifications across the breadth of molecular structures is harder to elucidate. **Figure 2.4** shows the correlation between slopes and average retention shifts for selected modifications in this study. They can be arranged into groups based on the distribution between quadrants where the left/right halves delineate hydrophilic/hydrophobic modifications and top/bottom portions of the graph correspond to modifications with slopes above or below 1.

Generally, the extent of a modification's retention shift is reduced with increasing peptide length and hydrophobicity.[16, 27, 28] This must yield slope values below 1 for hydrophobic modifications such as dimethylation, TMT, and iTRAQ (bottom right quadrant, **Figure 2.4b**), and above 1 for hydrophilic ones such as oxidations of methionine, tryptophan, and proline (top left quadrant, **Figure 2.4a**). Notably, all these modifications do not involve alteration of the net charge at acidic pH of the eluent. Altering the charge state leads to deviations from this trend. The addition of a negatively charged group or the removal of a positive charge increases slope values for hydrophobic modifications, as can be seen for phosphorylation, N-terminal acetylation and cyclization (top right quadrant). In a complementary manner, positive charge inducing hydrophilic modification decreases the slope value such as for cysteine alkylation with vinyl-pyridine vs. standard iodoacetamide (**Figure 2.4a**). Thus, we hypothesize that the resulting slope is a superposition of hydrophobic and charge effects. This hypothesis does not hold for hydrophobic modifications with charge elimination on internal residues: citrullination, carbamylation and acylations of Lys (**Figure 2.4a**, bottom right quadrant). It is possible that the position of modification site (internal vs. terminal) is the decisive factor here. The distribution of data points for acetylation (**Figure 2.4b**) indirectly confirms this. There is a "Two Lys" data point corresponding to a single acetylation of internal Lys which also deviates towards the bottom right quadrant.

Extremely large modifications such as N-linked glycans represent a unique case and show clustering based on their polysaccharide composition. Sequential addition of one and two 5-acetylneuraminic acid groups shifts their clusters towards the upper right quadrant – like phosphorylation. N-glycosylation is also characterized by lower slopes, although it is hydrophilic (bottom left quadrant). This could be a consequence of the large size of N-glycans– similar to the extremely low slope values for bulky TMT/iTRAQ labels. Overall, all these findings suggest that a very complex interplay of hydrophobicity, charge, size and position of modified residue determines the retention properties.

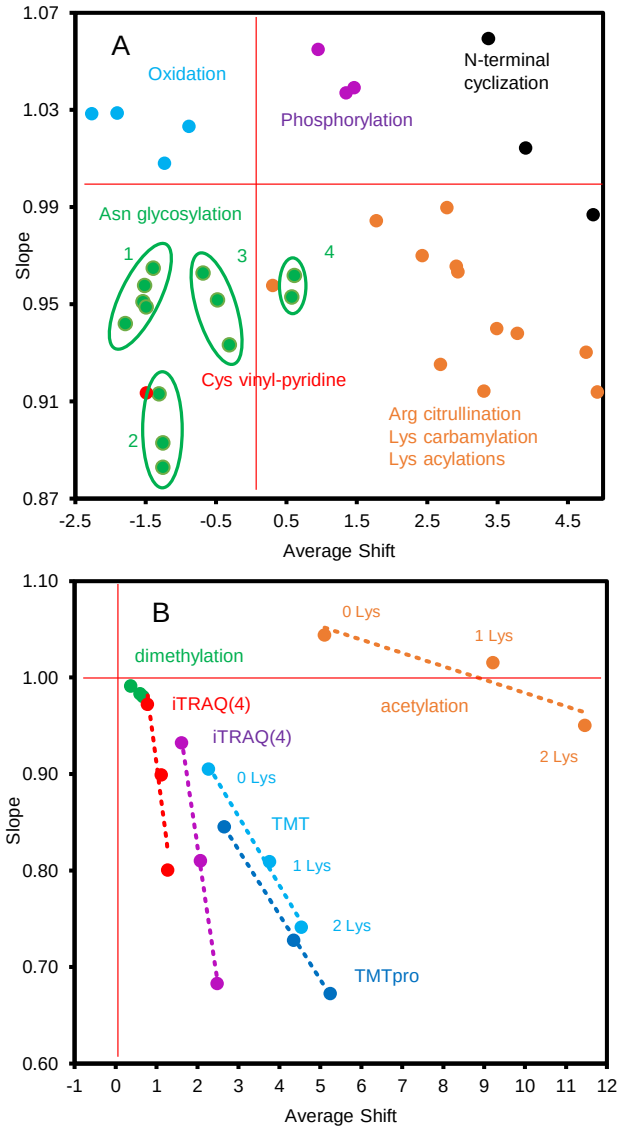


Figure 2.4 The interplay between retention shifts and slope values for the selected HI_{mod} vs. $HI_{non-mod}$ plots. A – selected modifications highlighting high mannose (1), complex neutral (2), complex with one sialic acid (3) and complex with two sialic acid residues (4) Asn-linked glycan structures, B – correlations for amine-selective labeling groups: each represented by average values for peptides without, with one and with two lysines.

2.6.7 Metabolic Co-Translational Incorporation of Proline Analogues into E. coli

Co-translational incorporation has been confirmed for more than 200 non-canonical amino acid analogs in organisms ranging from procaryotes to animal models.[48] Our method of metabolic co-translational incorporation (also known as Selective Pressure Incorporation, SPI[49]) requires a proline (Pro) auxotrophic *E. coli* microbial strain.[50] Fermentation in a defined synthetic (i.e., minimal) medium allows incorporation of Pro analogs that are structurally like proline, as they

must be taken up by native cell membrane transporters and processed in the protein translation machinery. In particular, the SPI method exploits the substrate promiscuity of aminoacyl-tRNA synthetases, crucial enzymes in canonical amino acid recognition, activation, and tRNA loading.[51] For this reason, misincorporations are best suited for amino acid analogs with small modifications on the side chains such as hydroxylation, alkylation, ring expansions or constrictions, and single-atom substitutions by halogenations (which usually introduce local polarity inversions). All these features are present in our nine proline analogs (**Figure S2.1**), which were used in the Pro-auxotrophic *E. coli* MG1566 strain to replace Pro residues with related analogs. Our goal was to assess such co-translational modifications by using PTCM peptides retention data acquisition. The observed shifts were comparatively modest (**Table S2.1**) with the largest ΔHI amplitude observed for 4*R*-hydroxyproline, -0.88 % ACN. While the number of peptide pairs identified by 1D LC-MS/MS varied for these compounds, ranging from 50 to 3736, the identification count remained constant for biological replicates for the same compound (not shown here).

2.7 Conclusion

Dramatic advances in the performance of high-throughput proteomics have provided virtually unlimited retention data sets for modeling peptide retention in RPLC. Further developments aim at continuous improvement of prediction accuracy and expanding it to the entire peptide collection found in proteomics runs. These includes a wide variety of peptides carrying PTCMs. Here, we summarize recent collective efforts (more than 70 modifications) toward initial retention data collection as a predecessor to widespread RPLC modeling of modified peptides and reviewed potential pitfalls affecting the accuracy of these evaluations. While standard bottom-up LC-MS acquisitions provide access to the most abundant modifications, less abundant or rare ones require specialized sample preparation protocols in which various combinations of enrichment, chemical/enzymatic reactions, or peptide synthesis can be applied to obtain the required retention information. The co-translational incorporation of non-canonical proline derivatives performed in this work represents another cost-effective alternative that allows the acquisition of retention data for hundreds to thousands of modified/non-modified peptide pairs in a standard 1D LC-MS acquisition. The identification result can be easily scaled up by deeper 2D LC fractionation if needed. The incorporation of non-canonical amino acids into protein and proteome is a rapidly growing field of research that will eventually produce synthetic cells.[52] The hallmark of these synthetic cells (and thus artificial life) will be massively modified proteomes whose evaluation and content quantification will require a standard analytical approach. This study, as well as ongoing work in our laboratories, aims in this direction.

We have found that it is critical to perform consistent chromatographic measurements, which should preferably use linear water/acetonitrile gradients and calibration of RPLC separation space using standard peptide mixtures. The latter are critical for an accurate evaluation of retention properties. At the same time, the role of retention standards will decrease due to the steady increase in the size of high-quality retention datasets and accuracy of prediction models. With this in mind, the output of our SSRCalc model and SSRCalc DB indexes are both expressed in Hydrophobicity Index HI units, allowing the alignment of any data to the % ACN scale either by predicting for confidently identified peptides or by using database indexes for overlapping peptides.

We propose to use the average retention shift, slope, and intercept of HI_{mod} vs. $HI_{non\text{-}mod}$ plots for evaluating the effect of individual modification on RPLC behavior. RMSE or R^2 statistics can be used for evaluating the extent of sequence-dependent variations of the shifts. Reporting the

average shift in acetonitrile percentage units is the most informative way to assess expected retention features of modified peptides. In addition, a similar method can be used for variations of chromatographic conditions to monitor (predict) their effect on peptide retention. In the future we envision incorporating different PTCMs and chromatographic conditions into the modeling of peptide retention by developing relatively simple modules to predict ΔHI shift for a particular modification of peptide sequence or variation in chromatographic LC settings. These modules can be knowledge-based models or various applications of machine-learning methods,[18] which gain popularity in recent years. The modified retention values $HI_{mod} = HI_{non\text{-}mod} + \Delta HI$ are based on the application of a highly accurate core prediction model ($HI_{non\text{-}mod}$) adjusted by the predicted shift value.

We believe that a community-wide collection of retention data is the only viable option to dramatically expand the coverage of PTCM by predictive models. The pooling of external data with our collection will be following strict quality controls, as demonstrated for selected datasets in this work. The additions of the various PTCMs to our models will be available on the SSRCalc web portal (https://proteome.ad.umanitoba.ca/ssrcalc) to facilitate retention prediction tasks for modified peptides in the future.

2.8 Supporting Information

Experimental procedures for culturing *E. coli* used for incorporation of non-canonical proline variants (structures provided) and mass spectrometry instrumental settings; protocols for chromatographic alignment of external data and results for inter-laboratory comparisons; statistics for identification of common PTMs in 2D-LC MS/MS and summary of chromatographic properties of all studied modifications provided in .doc format.

2.9 Acknowledgements

This work was supported by the grants from the Natural Sciences and Engineering Research Council of Canada (RGPIN-05669-2020 A.P., RGPIN-2016-05963 and RGPIN-2022-05015, O.K.). N.B. is grateful for financial support by the Canada Research Chairs Program grant agreement No. 950-23197.

2.10 Data Availability

Mass spectrometry data in MGF and .raw formats, are available at the MassIVE repository (massive.ucsd.edu) under accession MSV00086214.

2.11 Supplementary Information

Supplementary Experimental procedures.

Media for *E.coli* culture:

The synthetic medium used was New Minimal Medium (NMM) containing: 22.5mM KH2PO4, 50 mM K2HPO4, 7.5 mM (NH4)2SO4, 8.5 mM NaCl, 1 mM MgSO4, 10ng/L trace elements (Cu2+, Zn2+, Mn2+, and MoO42-), 20 mM glucose, 10 mg/L biotin and thiamine. 50 mg/L of arginine and Lysine and 1mM of the analog being incorporated.

Mass spectrometry settings:

An Orbitrap Exploris 480 was used in a standard data-dependent acquisition mode. The MS1 resolution was set to 60,000 at m/z 200 with the RF Lens set to 40% and fixed at a scan range of m/z 380-1500. The normalized automatic gain control (AGC) target was set to 300% with a maximum ion injection time of 50 milliseconds collecting profile spectra. The intensity threshold for DDA was set to 10,000 isolating only precursors of charge +2 to +6 with an exclusion window of 20 seconds and the DDA cycle time was fixed to 3 seconds. For MS2 acquisitions, the resolution was set to 30,000 at m/z 200 and the isolation window to a width of m/z 1.6 with a normalized collision energy of 30%. MS2 AGC target was at 100% with the maximum ion injection time set to automatic.

HPLC settings for the second dimension RPLC-MS:

Luna C18(2) 3 µm, 100 µm x 30 cm column, have been used at 500 nl/min flow rate with linear water/acetonitrile gradients of ~0.3% per minute (0.5-35% ACN) with 0.1% formic acid as ion pairing modifier in both eluents A (100% water) and B (20:80 water: acetonitrile). 3 µL direct injection have been used.

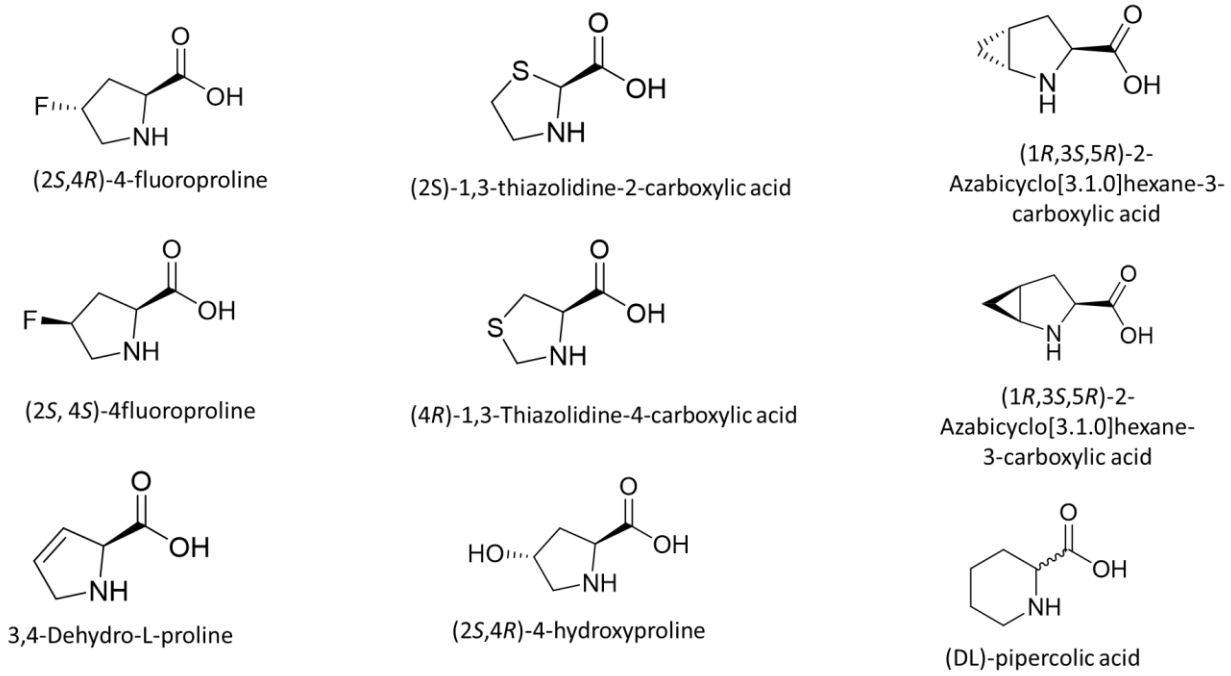
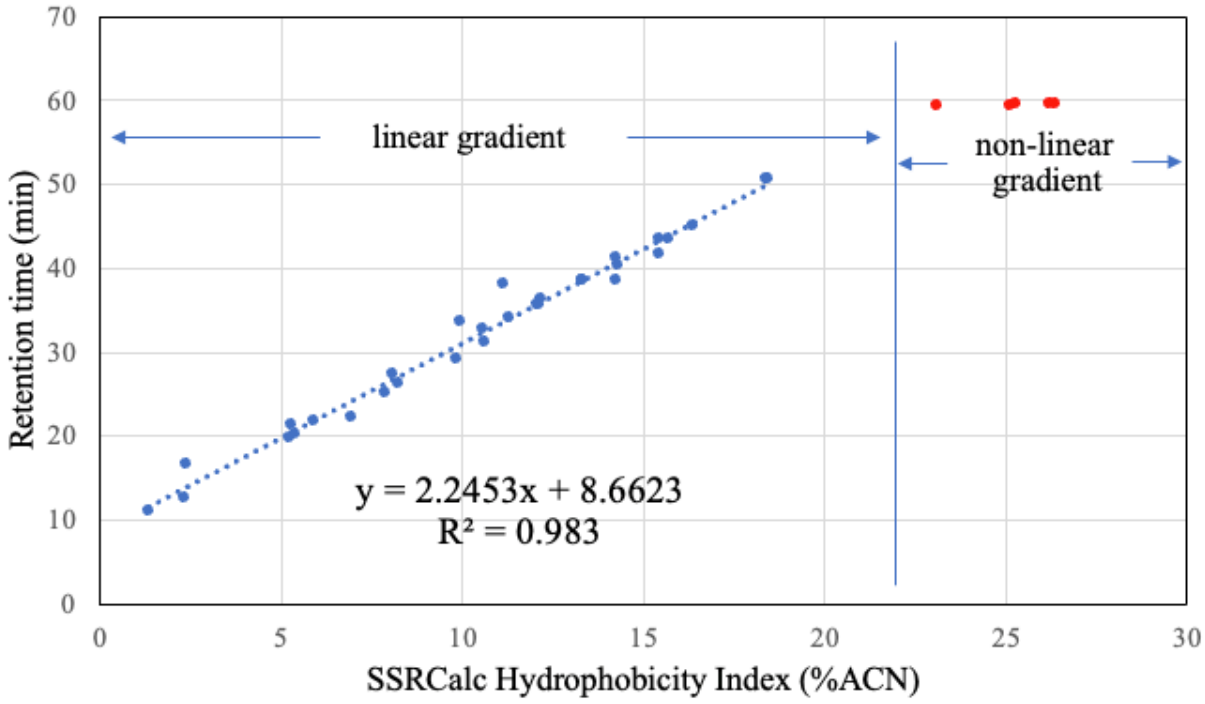


Figure S2.1 Non-canonical variants of proline studied in this work.



Peptide	SSRCalc HI (%ACN)	Retention time (min)	HI (%ACN)	Peptide	SSRCalc HI (%ACN)	Retention time (min)	HI (%ACN)
HEHISSDYAGK	1.32	11.15	1.11	ALFSSITDSEK	12.15	36.43	12.37
TFAHTESHISK	2.33	12.64	1.77	HFALFSTDVTK	11.13	38.14	13.13
ISLGEHEGGGK	2.40	16.77	3.61	VSGFSDISIYK	13.30	38.59	13.33
FGTGTYAGGEK	5.22	19.94	5.02	TFGTETFDTFK	14.25	38.73	13.39
LSSGYDGTSYK	5.37	20.24	5.16	TFTGTTDSFFK	13.33	38.73	13.39
TASGVGGFSTK	5.28	21.40	5.67	ASDLLSGYYIK	14.30	40.35	14.11
VGASTGYSGLK	5.93	21.82	5.86	FLFTGYDTSVK	14.25	41.36	14.56
SYASDFGSSAK	6.91	22.32	6.08	GFVIDDGLITK	15.42	41.76	14.74
LYSYYSSTESK	7.88	25.11	7.33	GIFGAFTDDYK	15.69	43.44	15.49
LYTGAGYDEVK	8.24	26.27	7.84	VYAETLSGFIK	15.44	43.45	15.49
TLIAYDDSSTK	8.07	27.35	8.32	GASDFLSFAVK	16.37	45.22	16.28
FLASSEGGFTK	9.83	29.20	9.15	VSSIFFDTFDK	18.36	50.60	18.68
FVGTEYDGLAK	10.62	31.14	10.01	FFLTGTSIFVK	18.46	50.64	18.69
YALDSYSLSSK	10.54	32.70	10.71	GDFTFFIDTFK	23.12	59.43	n/a
GFLDYESTGAK	9.95	33.76	11.18	SLFFIIDGFVK	25.09	59.53	n/a
HLTGLTFDTYK	11.31	34.11	11.33	IDVYILALLLK	25.27	59.56	n/a
HDTVFGSYLYK	12.06	35.61	12.00	LFISALVDFFK	26.34	59.56	n/a
YFGYTSDTFGK	12.10	35.61	12.00	SILAFLYLYFK	26.19	59.63	n/a

Figure S2.2 Assigning Hydrophobicity Index values to PROCAL peptide retention standard.[43] SSRCalc retention prediction for the PROCAL peptides detected in a mixture of non-modified Arg containing model peptides[14] showed 0.983 R^2-value correlation. HI values have been calculated using projection of experimental retention time values on SSRCalc HI axis: HI = (RT-8.6623) / 2.2453. Peptides eluting outside of linear gradient range (in red) have been excluded.

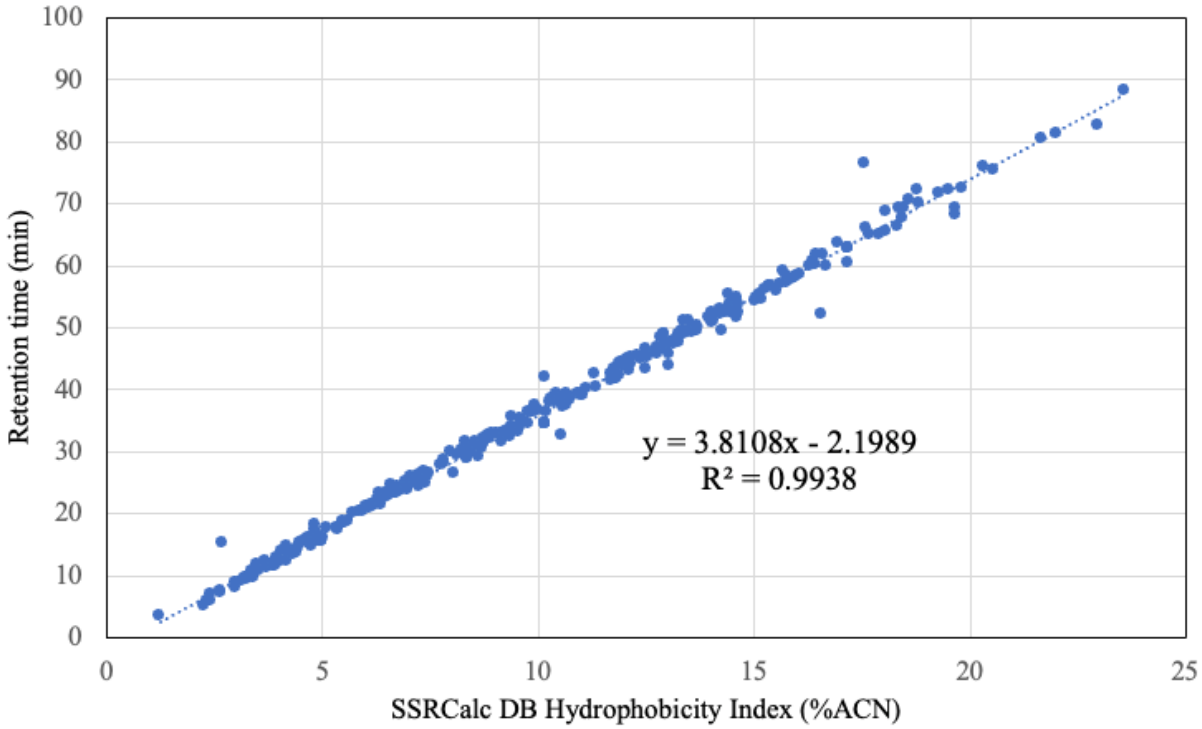
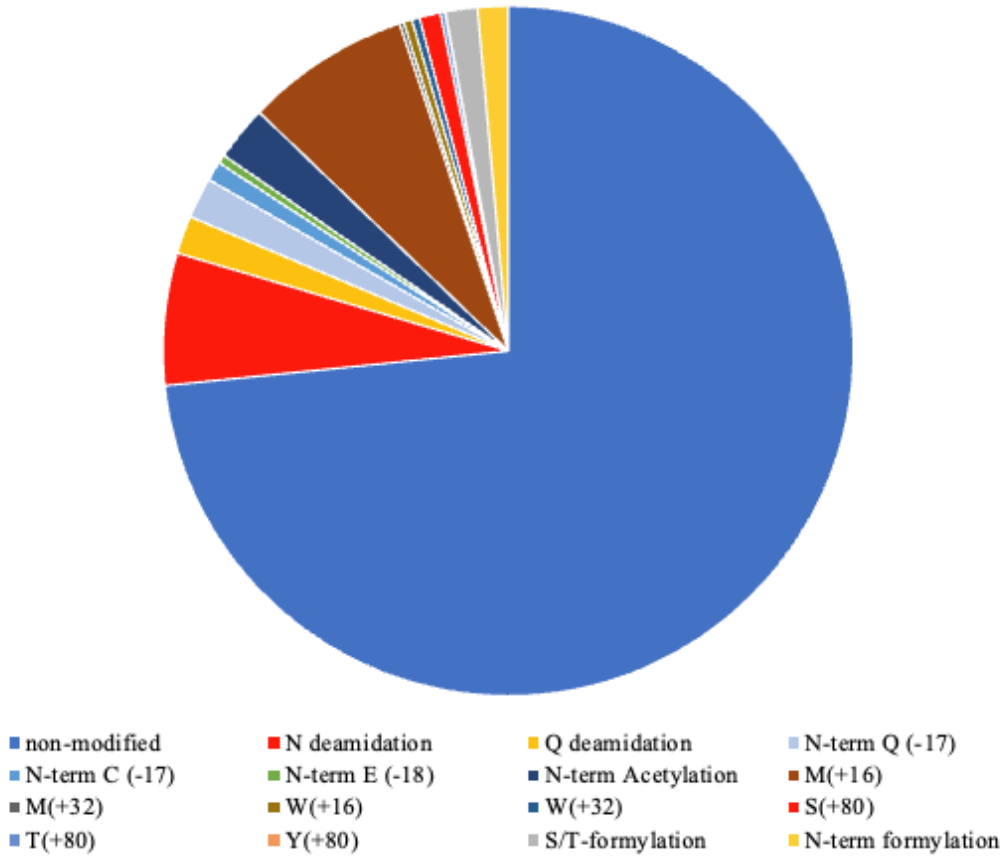


Figure S2.3 Assigning Hydrophobicity Index values to hyper-citrullinated *Mus musculus* peptide library.[20] Retention dataset from this publication contained 3245 modified-non-modified peptide pairs. 310 of them overlap with our in-house retention database showing 0.9938 R^2-value correlation. HI values have been calculated using projection of experimental retention time values on SSRCalc DB HI axis: HI = (RT+2.1989) / 3.8108.



Modification	# of PSMs	# of pairs	ΔHI ave unsupervised	ΔHI ave verified	Comment
non-modified	243730				
N deamidation	20529	7366	0.68	0.1/0.61	
Q deamidation	5907	3135	0.42	n/a	
N-term Q (-17)	6355	3428	3.88	3.9	
N-term C (-17)	3030	1428	4.84	4.86	
N-term E (-18)	1344	857	0.93	3.37	in-source loss of water
N-term Acetylation	8549	176	4.28	5.1	not representative population
M(+16)	25650	9671	-2.27	-2.37	in-source oxidation
M(+32)	696	72	-1.91	-1.9	
W(+16)	1307	337	-1.47	-1.64	in-source oxidation
W(+32)	1212	599	-1.24	-1.23	
S(+80)	3266	2601	1.4	1.47	
T(+80)	701	492	1.25	1.35	
Y(+80)	111	80	0.88	0.95	
S/T-formylation	4948	n/a	n/a	n/a	
N-term formylation	4734	n/a	n/a	n/a	

Figure S2.4 Identifications of modified peptides and correspondent retention shifts in the 2D LC – MS/MS analysis of zebra fish brain tissue digest. Examples of biased unsupervised assignments of retention shifts highlighted in blue.

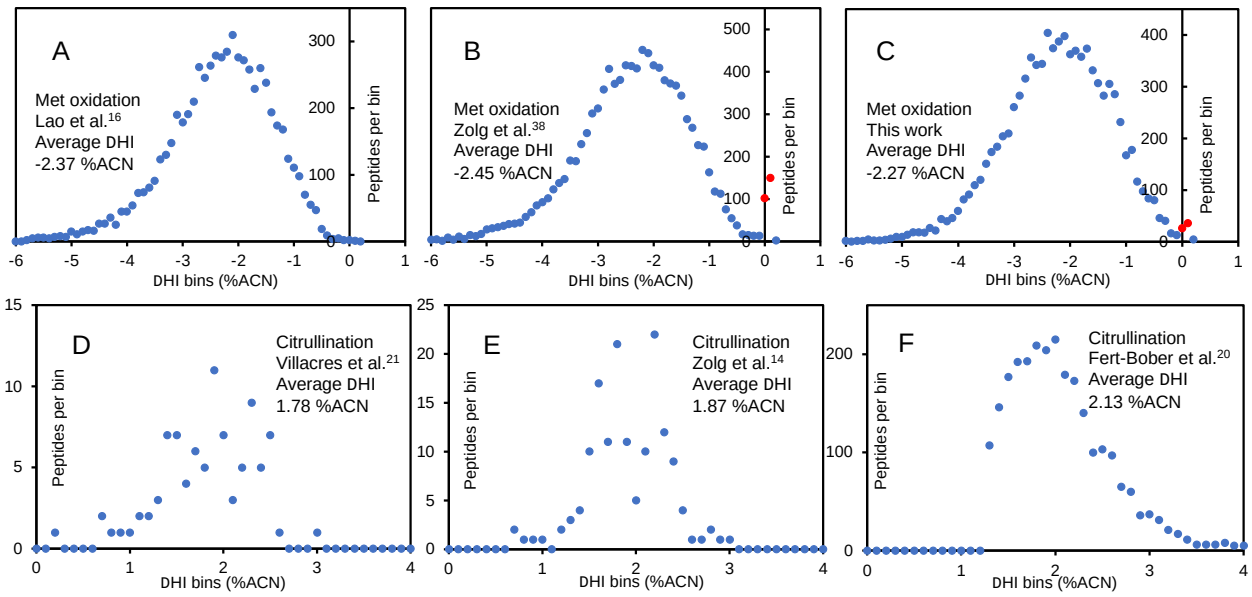


Figure S2.5 Inter-laboratory reproducibility of retention shifts' measurements for methionine oxidation and arginine citrullination. A-C and D-F show the results of three independent experiments for Met oxidation and Arg citrullination, respectively. Red dots were from in-source oxidation, which needed to be removed for accurate evaluation of average ΔHI.

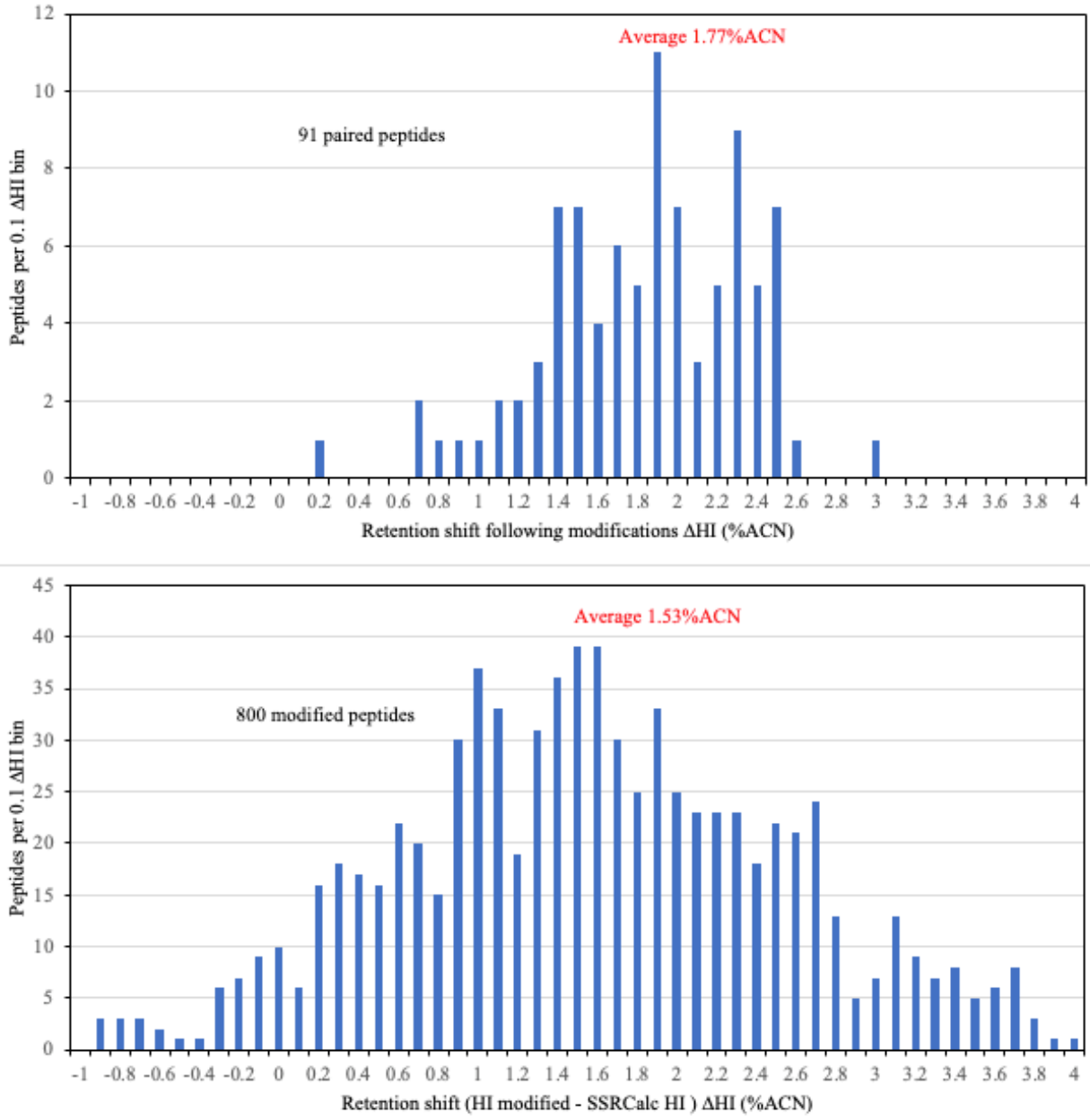


Figure S2.6 The distribution of ΔHI values for citrullination events detected with and without considering detection of non-modified analog.[17] Correspondent ΔHI values have been calculated as following:

- peptide pairs (91) $\Delta HI = HI_{mod} - HI_{non\text{-}mod}$;
- citrullinated peptides unpaired (800) $\Delta HI = HI_{mod} - SSRCalc\ HI$.

Table S2.1 Chromatographic properties of peptides carrying post-translational and chemical modifications in RPLC for proteomic applications.

Modification	DM (Da)	# of pairs	Average Hyd—phobicity, HI (%ACN)	Average length	Average shift, DHI (%ACN)	Intercept, HI (%ACN)	Slope	R^2	RMSE
Non-modified peptides in 2D zebra fish runs		126951	10.62	13.99					
Typical PTMs found in bottom-up proteomics run (zebra fish)									
Met oxidation (+16)	+15.995	9282	11.42	14.53	-2.27	-2.5912	1.0285	0.9716	0.9330
Met oxidation (+32)	+31.990	66	15.25	20.35	-1.90	-2.3442	1.0289	0.9830	0.7698
Trp oxidation (+16)	+15.995	302	13.21	13.79	-1.64	-2.58	1.071	0.9707	0.9236
Trp oxidation (+32)	+31.990	594	14.97	14.73	-1.23	-1.3493	1.0080	0.9875	0.5180
N-terminal Gln Cyclization	-17.027	3381	10.71	14.04	3.90	3.7481	1.0144	0.9545	1.1539
N-terminal Glu Cyclization	-18.011	217	13.09	14.02	3.37	2.5937	1.0594	0.9699	0.9971
N-terminal Cys Cyclization	-17.027	1401	10.21	13.70	4.86	4.9942	0.9870	0.9255	1.3991
Additional modifications targeted in dedicated studies									
Asn deamidation (Asp)	+0.984	348	14.72	17.40	0.61	0.8227	0.9855	0.9969	0.3447
Asn deamidation (iso-Asp)	+0.984	348	14.72	17.40	0.10	1.0007	0.9394	0.9777	0.8883
Ser phosphorylation	+79.966	7166	8.77	15.86	1.47	1.1212	1.0392	0.9851	0.6346
Thr Phosphorylation	+79.966	1018	9.91	17.46	1.35	0.9807	1.0372	0.9851	0.6215
Tyr Phosphorylation*	+79.966	126	11.02	16.48	0.95	0.3481	1.0548	0.9759	0.8208
Arg citrullination*	+0.984	91	8.28	13.91	1.78	1.9079	0.9844	0.9873	0.4962
Lys carbamylation	+43.006	1199	10.46	15.21	2.43	2.7453	0.9699	0.9838	0.6115
Met oxidation (+16)*	+15.995	6407	13.47	14.4	-2.37	-2.4183	1.0037	0.9705	0.9160
Met oxidation (+32)	+31.990	262	14.09	13.62	-1.91	-1.8909	0.9989	0.9832	0.6673
Glycosylation									
Ser GlcNAc	+203.079	55	9.66	13.62	-0.60	-0.2386	0.9623	0.9933	0.4316
Thr GlcNAc	+203.079	33	7.99	13.73	-0.73	-0.135	0.9251	0.9793	0.6512
Ser GalNAc	+203.079	56	9.57	13.71	-0.63	-0.3108	0.9668	0.987	0.5334

Thr GalNAc	+203.079	35	7.64	13.60	-0.67	-0.0265	0.9164	0.9789	0.6622
Asn Glycosylation (h4n5f1a0)	+1809.666	26	8.14	15.35	-1.25	-0.3008	0.8829	0.9889	0.3423
Asn Glycosylation (h5n2f0a0)	+1216.423	47	10.28	16.81	-1.39	-1.0291	0.9648	0.9912	0.4261
Asn Glycosylation (h5n4f0a1)	+1913.677	23	12.14	15.61	-0.68	-0.2317	0.9629	0.9911	0.4639
Asn Glycosylation (h5n4f0a2)	+2204.772	30	11.74	15.10	0.56	1.1317	0.9529	0.9909	0.4211
Asn Glycosylation (h5n4f1a0)	+1768.640	44	7.96	15.41	-1.31	-0.6147	0.9131	0.9874	0.4079
Asn Glycosylation (h5n4f1a1)	+2059.735	40	8.59	14.65	-0.48	-0.0633	0.9518	0.9966	0.2617
Asn Glycosylation (h5n4f1a2)	+2350.830	27	9.86	15.81	0.62	0.9926	0.9618	0.9950	0.2994
Asn Glycosylation (h5n5f1a0)	+1971.719	36	8.14	14.81	-1.25	-0.4000	0.8928	0.9892	0.3666
Asn Glycosylation (h5n5f1a1)	+2262.814	22	8.83	16.14	-0.31	0.2790	0.9333	0.9959	0.2690
Asn Glycosylation (h6n2f0a0)	+1378.476	28	11.60	17.89	-1.49	-0.8962	0.9489	0.9960	0.3281
Asn Glycosylation (h7n2f0a0)	+1540.529	30	9.57	15.37	-1.51	-1.1049	0.9577	0.9959	0.2742
Asn Glycosylation (h8n2f0a0)	+1702.581	33	9.51	15.91	-1.53	-1.0677	0.9509	0.9950	0.2966
Asn Glycosylation (h9n2f0a0)	+1864.634	14	12.34	16.36	-1.79	-1.0707	0.9420	0.9971	0.2536
Primary amino groups' modifications (N-terminus plus Lys)									
N-terminal Acetylation	+42.011	10632	12.75	13.39	5.10	4.5404	1.0442	0.8728	1.9869
N-terminal + 1K Acetylation	+84.021	11791	11.87	13.16	9.22	9.0318	1.0156	0.7616	2.6036
N-terminal + 2K Acetylation	+126.032	2390	9.63	14.45	11.46	11.9375	0.9505	0.6840	2.6889
N-terminal Dimethylation	+28.031	26310	13.12	15.01	0.36	0.4807	0.9912	0.9968	0.3181
N-terminal + 1K Dimethylation	+56.063	32839	12.82	14.79	0.60	0.8096	0.9834	0.9961	0.3488
N-terminal + 2K Dimethylation	+84.094	5286	10.96	16.82	0.67	0.8824	0.9804	0.9971	0.3070
N-terminal iTRAQ4	+144.106	3183	12.03	12.96	0.78	1.1071	0.9725	0.9745	0.7280
N-terminal + 1K iTRAQ4	+288.212	1662	11.55	12.26	1.11	2.2784	0.8993	0.9693	0.7364
N-terminal + 2K iTRAQ4	+432.318	38	8.98	13.34	1.27	3.0644	0.8004	0.9819	0.4712
N-terminal iTRAQ8	+304.205	3199	12.03	12.95	1.61	2.4229	0.9326	0.9166	1.3011
N-terminal + 1K iTRAQ8	+608.411	1659	11.52	12.27	2.07	4.2559	0.8103	0.9143	1.1424
N-terminal + 2K iTRAQ8	+912.616	38	8.98	13.34	2.48	5.3318	0.6828	0.9506	0.6754
N-terminal TMT	+224.152	56080	10.89	13.29	2.27	3.2996	0.9053	0.9622	1.0623
N-terminal + 1K TMT	+448.305	73311	10.74	13.40	3.76	5.8098	0.8092	0.9441	1.1592

N-terminal + 2K TMT	+672.457	11640	8.31	15.02	4.54	6.6864	0.7414	0.9373	0.9805
N-terminal TMTpro	+304.207	45979	10.25	12.50	2.65	4.2364	0.8455	0.9420	1.1858
N-terminal + 1K TMTpro	+608.414	59194	10.42	12.51	4.35	7.1899	0.7277	0.9176	1.2373
N-terminal + 2K TMTpro	+912.621	5475	8.21	14.28	5.24	7.9318	0.6724	0.9152	1.0484
Analogs of proline: Pro-Xxx substitution									
4*R*-fluoroproline	+17.991	3736	14.05	17.41	-0.08	-0.35389	1.0185	0.9969	0.2657
4*S*-fluororpoline	+17.991	3612	14.04	17.24	0.07	-0.0880	1.0103	0.9939	0.3158
4R-hydroxyproline*	+15.995	122	14.99	19.04	-0.88	-1.2560	1.0232	0.9918	0.3508
3-thiaproline	+17.956	125	16.18	15.32	0.22	0.1525	0.9945	0.9945	0.3058
4-thiaproline	+17.956	366	17.65	15.69	0.63	0.8144	0.9957	0.9957	0.2858
Trans[1.5]cyclopropane	+12.000	2168	13.56	13.80	0.48	0.4428	0.9965	0.9964	0.2741
Cis[1.5]cyclopropane	+12.000	502	14.65	14.47	0.50	0.6268	0.9959	0.9959	0.2825
Pipercolic acid	+14.016	50	17.12	15.68	0.69	1.4274	0.9897	0.9897	0.3859
Dehydroproline	- 2.016	1199	13.68	16.28	-0.23	-0.3653	0.9970	0.9970	0.2643
Various Cys alkylation chemistries: carbamidomethyl Cys – Xxx substitution									
Free Cys	-57.021	60	11.34	14.93	0.83	0.7112	1.0107	0.9960	0.3424
Iodoacetic acid	+0.984	68	11.07	14.66	0.97	0.9789	0.9994	0.9970	0.2782
Acrylamide	+14.016	64	11.34	14.80	0.10	0.3058	0.9820	0.9997	0.0791
4-Vinylpyridine	+48.036	66	11.46	14.95	-1.49	-0.4940	0.9135	0.9887	0.4980
MMTS	-11.034	63	11.22	14.52	2.64	3.0518	0.9636	0.9713	0.8481
External datasets[14, 20, 38]									
Arg citrullination*	+0.984	2575	12.19	16.95	2.13	1.8359	1.0238	0.9899	1.0484
Met oxidation (+16)*	+15.995	10201	13.29	14.18	-2.45	-2.3235	0.9906	0.9631	0.9500
Lys Acetylation	+42.011	172	7.76	13.87	2.91	3.1809	0.9656	0.9829	0.5445
Lys Biotinylation	+226.078	177	8.00	13.82	4.85	6.3550	0.8122	0.9585	0.7464
Lys Butyrylation	+70.042	208	8.17	13.66	4.92	5.6227	0.9139	0.9557	0.8797
Lys Crotonylation	+68.026	207	8.20	13.64	4.76	5.3307	0.9303	0.9619	0.8392
Lys Dimethylation	+28.031	198	8.12	13.43	0.20	0.3627	0.9802	0.9982	0.1836
Lys Formylation	+27.995	204	8.15	13.70	2.78	2.8638	0.9898	0.9839	0.5644

Lys Glutarylation	+114.032	200	8.35	13.68	3.31	4.0223	0.9143	0.9730	0.6782
Lys GlyGlycylation	+114.043	224	8.14	13.46	0.31	0.6722	0.9577	0.9940	0.2473
Lys Hydroxyisobutylation	+86.037	199	8.15	13.75	3.49	3.9777	0.9400	0.9760	0.6718
Lys Malonylation	+86.000	193	8.05	13.81	2.94	3.2332	0.9634	0.9822	0.5858
Lys Methylation	+14.016	214	8.33	13.56	0.19	0.1617	1.0039	0.9978	0.2099
Lys Propionylation	+56.026	208	8.28	13.73	3.78	4.2963	0.9380	0.9717	0.7185
Lys Succinylation	+100.016	204	8.13	13.73	2.69	3.2983	0.9253	0.9771	0.6291
Lys Trimetylation	+42.047	186	8.29	13.39	0.14	0.0699	1.0091	0.9976	0.2205
R									
Arg Citrullination	+0.984	153	7.65	14.24	1.87	1.7239	1.0193	0.9928	0.4089
Arg Dimethylation_asymmetric	+28.031	169	7.96	13.95	0.32	0.4394	0.9854	0.9980	0.2066
Arg Dimethylation_symmetric	+28.031	160	7.99	14.03	0.39	0.5534	0.9803	0.9978	0.2154
Arg Methylation	+14.016	173	7.93	13.96	0.17	0.2919	0.9846	0.9992	0.1319
P									
Pro Hydroxyproline*	+15.995	94	8.10	14.70	-0.94	-1.0713	1.0157	0.9909	0.4453
Y									
Tyr Nitrotyrosine	+44.985	156	10.31	11.78	2.31	2.9483	0.9377	0.9794	0.7622
Tyr Phosphorylation*	+79.966	154	10.24	11.76	-0.33	-0.7198	1.0382	0.9806	0.7900
Various chromatographic conditions									
25 ºC to 55 ºC column temperature	NA	59312	10.07	13.47	-1.69	-1.3882	0.9695	0.9884	0.5020
0.1% formic to 0.5% acetic acid	NA	41117	12.04	13.82	-0.70	-0.8860	1.0159	0.9942	0.4036

* - datasets used for interlaboratory comparisons

2.12 References

1. Doll, S.; Burlingame, A. L. Mass Spectrometry-Based Detection and Assignment of Protein Posttranslational Modifications. *ACS Chem Biol* **2015**, *10* (1), 63–71. https://doi.org/10.1021/cb500904b.
2. Aebersold, R.; Mann, M. Mass-Spectrometric Exploration of Proteome Structure and Function. *Nature* **2016**, *537* (7620), 347–355. https://doi.org/10.1038/nature19949.
3. Sechi, S.; Chait, B. T. Modification of Cysteine Residues by Alkylation. A Tool in Peptide Mapping and Protein Identification. *Anal Chem* **1998**, *70* (24), 5150–5158. https://doi.org/10.1021/ac9806005.
4. Leitner, A.; Lindner, W. Chemistry Meets Proteomics: The Use of Chemical Tagging Reactions for MS-based Proteomics. *Proteomics* **2006**, *6* (20), 5418–5434. https://doi.org/10.1002/pmic.200600255.
5. 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.
6. 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. *Molecular & Cellular Proteomics* **2004**, *3* (12), 1154–1169. https://doi.org/10.1074/mcp.M400129-MCP200.
7. Boja, E. S.; Fales, H. M. Overalkylation of a Protein Digest with Iodoacetamide. *Anal Chem* **2001**, *73* (15), 3576–3582. https://doi.org/10.1021/ac0103423.
8. Kollipara, L.; Zahedi, R. P. Protein Carbamylation: In Vivo Modification or in Vitro Artefact? *Proteomics* **2013**, *13* (6), 941–944. https://doi.org/10.1002/pmic.201200452.
9. Lenčo, J.; Khalikova, M. A.; Švec, F. Dissolving Peptides in 0.1% Formic Acid Brings Risk of Artificial Formylation. *J Proteome Res* **2020**, *19* (3), 993–999. https://doi.org/10.1021/acs.jproteome.9b00823.
10. Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P. A Mass-Tolerant Database Search Identifies a Large Proportion of Unassigned Spectra in

Shotgun Proteomics as Modified Peptides. *Nat Biotechnol* **2015**, *33* (7), 743–749. https://doi.org/10.1038/nbt.3267.

11. 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.
12. Marcantonio, M.; Trost, M.; Courcelles, M.; Desjardins, M.; Thibault, P. Combined Enzymatic and Data Mining Approaches for Comprehensive Phosphoproteome Analyses. *Molecular & Cellular Proteomics* **2008**, *7* (4), 645–660. https://doi.org/10.1074/mcp.M700383-MCP200.
13. Pichler, P.; Köcher, 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* (15), 6549–6558. https://doi.org/10.1021/ac100890k.
14. Zolg, D. P.; Wilhelm, M.; Schmidt, T.; Médard, G.; Zerweck, J.; Knaute, T.; Wenschuh, H.; Reimer, U.; Schnatbaum, K.; Kuster, B. ProteomeTools: Systematic Characterization of 21 Post-Translational Protein Modifications by Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) Using Synthetic Peptides. *Molecular & Cellular Proteomics* **2018**, *17* (9), 1850–1863. https://doi.org/10.1074/mcp.TIR118.000783.
15. Moruz, L.; Staes, A.; Foster, J. M.; Hatzou, M.; Timmerman, E.; Martens, L.; Käll, L. Chromatographic Retention Time Prediction for Posttranslationally Modified Peptides. *Proteomics* **2012**, *12* (8), 1151–1159. https://doi.org/10.1002/pmic.201100386.
16. Lao, Y. W.; Gungormusler-Yilmaz, M.; Shuvo, S.; Verbeke, T.; Spicer, V.; Krokhin, O. V. Chromatographic Behavior of Peptides Containing Oxidized Methionine Residues in Proteomic LC–MS Experiments: Complex Tale of a Simple Modification. *J Proteomics* **2015**, *125*, 131–139. https://doi.org/10.1016/j.jprot.2015.05.018.
17. Kim, J.; Petritis, K.; Shen, Y.; Camp, D. G.; Moore, R. J.; Smith, R. D. Phosphopeptide Elution Times in Reversed-Phase Liquid Chromatography. *J Chromatogr A* **2007**, *1172* (1), 9–18. https://doi.org/10.1016/j.chroma.2007.09.032.
18. 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.

19. Zybailov, B.; Sun, Q.; van Wijk, K. J. Workflow for Large Scale Detection and Validation of Peptide Modifications by RPLC-LTQ-Orbitrap: Application to the *Arabidopsis Thaliana* Leaf Proteome and an Online Modified Peptide Library. *Anal Chem* **2009**, *81* (19), 8015–8024. https://doi.org/10.1021/ac9011792.
20. Fert-Bober, J.; Venkatraman, V.; Hunter, C. L.; Liu, R.; Crowgey, E. L.; Pandey, R.; Holewinski, R. J.; Stotland, A.; Berman, B. P.; Van Eyk, J. E. Mapping Citrullinated Sites in Multiple Organs of Mice Using Hypercitrullinated Library. *J Proteome Res* **2019**, *18* (5), 2270–2278. https://doi.org/10.1021/acs.jproteome.9b00118.
21. Villacrés, 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. https://doi.org/10.1021/acs.jproteome.0c00775.
22. Mizero, B.; Yeung, D.; Spicer, V.; Krokhin, O. V. Peptide Retention Time Prediction for Peptides with Post-Translational Modifications: N-Terminal (α-Amine) and Lysine (ε-Amine) Acetylation. *J Chromatogr A* **2021**, *1657*, 462584. https://doi.org/10.1016/j.chroma.2021.462584.
23. Ang, E.; Neustaeter, H.; Spicer, V.; Perreault, H.; Krokhin, O. Retention Time Prediction for Glycopeptides in Reversed-Phase Chromatography for Glycoproteomic Applications. *Anal Chem* **2019**, *91* (21), 13360–13366. https://doi.org/10.1021/acs.analchem.9b02584.
24. Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Molecular & Cellular Proteomics* **2002**, *1* (5), 376–386. https://doi.org/10.1074/mcp.M200025-MCP200.
25. Wakabayashi, M.; Kyono, Y.; Sugiyama, N.; Ishihama, Y. Extended Coverage of Singly and Multiply Phosphorylated Peptides from a Single Titanium Dioxide Microcolumn. *Anal Chem* **2015**, *87* (20), 10213–10221. https://doi.org/10.1021/acs.analchem.5b01216.
26. Jiang, X.; Shamshurin, D.; Spicer, V.; Krokhin, O. V. The Effect of Various S-Alkylating Agents on the Chromatographic Behavior of Cysteine-Containing Peptides in Reversed-Phase Chromatography. *Journal of Chromatography B* **2013**, *915–916*, 57–63. https://doi.org/10.1016/j.jchromb.2012.12.010.
27. Mizero, B.; Villacrés, 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* (5), 1218–1228. https://doi.org/10.1021/acs.jproteome.1c00833.

28. Yeung, D.; Anderson, G.; Spicer, V.; Krokhin, O. V. Chromatographic Behaviour of Peptides Modified with Amine-Reacting Tags for Relative Protein Quantitation in Proteomic Applications. *J Chromatogr A* **2022**, *1679*, 463391. https://doi.org/10.1016/j.chroma.2022.463391.
29. Marx, H.; Lemeer, S.; Schliep, J. E.; Matheron, L.; Mohammed, S.; Cox, J.; Mann, M.; Heck, A. J. R.; Kuster, B. A Large Synthetic Peptide and Phosphopeptide Reference Library for Mass Spectrometry–Based Proteomics. *Nat Biotechnol* **2013**, *31* (6), 557–564. https://doi.org/10.1038/nbt.2585.
30. 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.
31. Eyers, C. E.; Simpson, D. M.; Wong, S. C. C.; Beynon, R. J.; Gaskell, S. J. QCAL—a Novel Standard for Assessing Instrument Conditions for Proteome Analysis. *J Am Soc Mass Spectrom* **2008**, *19* (9), 1275–1280. https://doi.org/10.1016/j.jasms.2008.05.019.
32. 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. https://doi.org/10.1021/ac9016693.
33. Escher, C.; Reiter, L.; MacLean, B.; Ossola, R.; Herzog, F.; Chilton, J.; MacCoss, M. J.; Rinner, O. Using i <scp>RT</Scp> , a Normalized Retention Time for More Targeted Measurement of Peptides. *Proteomics* **2012**, *12* (8), 1111–1121. https://doi.org/10.1002/pmic.201100463.
34. Thermo Fisher Scientific Inc. *Pierce™ Peptide Retention Time Calibration Mixture* (Catalog No. 88320 and 88321). 2011.
35. 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.

36. Grigoryan, M.; Shamshurin, D.; Spicer, V.; Krokhin, O. V. Unifying Expression Scale for Peptide Hydrophobicity in Proteomic Reversed Phase High-Pressure Liquid Chromatography Experiments. *Anal Chem* **2013**, *85* (22), 10878–10886. https://doi.org/10.1021/ac402310t.
37. 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.
38. Zolg, D. P.; Wilhelm, M.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Delanghe, B.; Bailey, D. J.; Gessulat, S.; Ehrlich, H.-C.; Weininger, M.; Yu, P.; Schlegl, J.; Kramer, K.; Schmidt, T.; Kusebauch, U.; Deutsch, E. W.; Aebersold, R.; Moritz, R. L.; Wenschuh, H.; Moehring, T.; Aiche, S.; Huhmer, A.; Reimer, U.; Kuster, B. Building ProteomeTools Based on a Complete Synthetic Human Proteome. *Nat Methods* **2017**, *14* (3), 259–262. https://doi.org/10.1038/nmeth.4153.
39. Nickling, J. H.; Baumann, T.; Schmitt, F.-J.; Bartholomae, M.; Kuipers, O. P.; Friedrich, T.; Budisa, N. Antimicrobial Peptides Produced by Selective Pressure Incorporation of Non-Canonical Amino Acids. *Journal of Visualized Experiments* **2018**, No. 135. https://doi.org/10.3791/57551.
40. Baumann, T.; Nickling, J. H.; Bartholomae, M.; Buivydas, A.; Kuipers, O. P.; Budisa, N. Prospects of In Vivo Incorporation of Non-Canonical Amino Acids for the Chemical Diversification of Antimicrobial Peptides. *Front Microbiol* **2017**, *8*. https://doi.org/10.3389/fmicb.2017.00124.
41. 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* (10). https://doi.org/10.15252/msb.20145625.
42. 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* (18), 7036–7042. https://doi.org/10.1021/ac800984n.
43. Zolg, D. P.; Wilhelm, M.; Yu, P.; Knaute, T.; Zerweck, J.; Wenschuh, H.; Reimer, U.; Schnatbaum, K.; Kuster, B. PROCAL: A Set of 40 Peptide Standards for Retention Time Indexing, Column Performance Monitoring, and Collision Energy Calibration. *Proteomics* **2017**, *17* (21). https://doi.org/10.1002/pmic.201700263.

44. Chen, M.; Cook, K. D. Oxidation Artifacts in the Electrospray Mass Spectrometry of Aβ Peptide. *Anal Chem* **2007**, *79* (5), 2031–2036. https://doi.org/10.1021/ac061743r.
45. Beausoleil, S. A.; Villén, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. A Probability-Based Approach for High-Throughput Protein Phosphorylation Analysis and Site Localization. *Nat Biotechnol* **2006**, *24* (10), 1285–1292. https://doi.org/10.1038/nbt1240.
46. Taus, T.; Köcher, T.; Pichler, P.; Paschke, C.; Schmidt, A.; Henrich, C.; Mechtler, K. Universal and Confident Phosphorylation Site Localization Using PhosphoRS. *J Proteome Res* **2011**, *10* (12), 5354–5362. https://doi.org/10.1021/pr200611n.
47. Battellino, T.; Ogata, K.; Spicer, V.; Ishihama, Y.; Krokhin, O. Acetic Acid Ion Pairing Additive for Reversed-Phase HPLC Improves Detection Sensitivity in Bottom-up Proteomics Compared to Formic Acid. *J Proteome Res* **2023**, *22* (1), 272–278. https://doi.org/10.1021/acs.jproteome.2c00388.
48. Dumas, A.; Lercher, L.; Spicer, C. D.; Davis, B. G. Designing Logical Codon Reassignment – Expanding the Chemistry in Biology. *Chem Sci* **2015**, *6* (1), 50–69. https://doi.org/10.1039/C4SC01534G.
49. Minks, C.; Alefelder, S.; Moroder, L.; Huber, R.; Budisa, N. Towards New Protein Engineering: In Vivo Building and Folding of Protein Shuttles for Drug Delivery and Targeting by the Selective Pressure Incorporation (SPI) Method. *Tetrahedron* **2000**, *56* (48), 9431–9442. https://doi.org/10.1016/S0040-4020(00)00827-9.
50. Budisa, N.; Minks, C.; Medrano, F. J.; Lutz, J.; Huber, R.; Moroder, L. Residue-Specific Bioincorporation of Non-Natural, Biologically Active Amino Acids into Proteins as Possible Drug Carriers: Structure and Stability of the *per* -Thiaproline Mutant of Annexin V. *Proceedings of the National Academy of Sciences* **1998**, *95* (2), 455–459. https://doi.org/10.1073/pnas.95.2.455.
51. Budisa, N.; Minks, C.; Alefelder, S.; Wenger, W.; Dong, F.; Moroder, L.; Huber, R. Toward the Experimental Codon Reassignment *in Vivo* : Protein Building with an Expanded Amino Acid Repertoire. *The FASEB Journal* **1999**, *13* (1), 41–51. https://doi.org/10.1096/fasebj.13.1.41.
52. Budisa, N. Xenobiology, New-to-Nature Synthetic Cells and Genetic Firewall. *Curr Org Chem* **2014**, *18* (8), 936–943. https://doi.org/10.2174/1385272818081406161543301.

3 Chapter 3: Chromatographic Properties of Deamidated Peptides with Asn-Gly Sequences in Proteomic Bottom-Up Experiments

3.1 Article Headline and Authors:

Chromatographic Properties of Deamidated Peptides with Asn-Gly Sequences in Proteomic Bottom-Up Experiments

Quinn Neale[1,2], Haley Neustaeter[1], Vic Spicer[2], Oleg V. Krokhin[2,3]*

[1]Department of Chemistry, University of Manitoba, 360 Parker Building, 144 Dysart Road, Winnipeg, R3T 2N2, Canada
[2]Manitoba Centre for Proteomics and Systems Biology, 799 JBRC, 715 McDermot Avenue, Winnipeg, R3E 3P4, Canada
[3]Department of Internal Medicine, University of Manitoba, 799 JBRC, 715 McDermot Avenue, Winnipeg, R3E 3P4, Canada

Published:
The Journal of Chromatography A
November 9th, 2024
Volume 1738
DOI: https://doi.org/10.1016/j.chroma.2024.465513

3.2 Contribution of Authors

For this chapter, QN was involved in the data collection and analysis to characterize the retention features of deamidated peptides; analysing synthetic peptides to confirm previously observed trends; analysing the distribution of retention shifts associated with deamidation from supervised and unsupervised workflows; evaluating the retention behaviours of isoaspartic acid containing peptides in various 2D chromatographic setups; manuscript writing and preparation of figures and tables for publication. HN was involved in data collection and analysis for sample digests to establish deamidation trends in reversed-phase. VS was responsible for computational work. HP and OK were involved in writing and supervised all students involved.

3.3 Abstract

Studies surrounding deamidation have relied on the chromatographic and mass spectrometric differentiation of Asn containing peptides and their isomeric Asp and iso-Asp products. The development of mass spectrometry analytical techniques and characterization of isomer specific fragmentation patterns has permitted the investigation of some deamidation species but has struggled to remain effective when applied and on complex samples or in high throughput scenarios. On the other hand, chromatographic separations can provide additional information to facilitate detection of deamidation. In this work the retention characteristics of deamidation products have been reported in reversed-phase separations using formic acid as an ion-pairing modifier. We found three major elution patterns depending on primary and secondary structure of Asn-Gly containing tryptic peptides. Random coil, helical conformations, and N-terminal positioning of Asn usually result in Asn < iso-Asp < Asp, iso-Asp < Asn < Asp, and Asn < Asp < iso-Asp elution order, respectively. These trends, found from the analyses of proteomic samples, were subsequently confirmed via analytical scale UV-HPLC. Additionally, we determined the retention shifts following deamidation for twenty various separation settings used as a first-dimension fractionation for high-throughput proteomic 2D LC-MS/MS analyses.

3.4 Introduction

Bottom-up proteomics relies on using two analytical techniques to separate (HPLC) and analyse (mass spectrometry, MS) complex peptide mixtures. While the general assumption of having twenty natural amino acids dictates a certain population of peptides to be analysed, having post-translational modifications (PTMs) or protein splice variants dramatically increases the size of analytes' pool. There is an urgent need to study both, MS and HPLC properties of modified peptides. This allows to improve the overall performance of LC-MS technique, increase the coverage of proteomic studies, and develop a much needed analytical protocols for studying biologically relevant modifications. Reversed-phase HPLC (RPLC) serves as a method of choice for most proteomic researchers. Not surprisingly, studying RPLC behaviour of modified peptides has been a hot topic in recent years.[1]

Deamidation is an irreversible PTM affecting the properties of peptides and proteins. It can occur through enzymatic or spontaneous degradation of Asparagine (Asn) or glutamine (Gln) residues in-vivo and in-vitro and results in a mass shift of +0.984 Da. It has been well established that deamidation reactions under acidic pH undergo direct hydrolysis while at physiological or basic conditions, typical for proteomic sample preparation, undergo a two-step process (**Figure 3.1A**).[2,3] Unsurprisingly, proteomic practitioners deal with consequences of deamidation on a daily basis. During deamidation, the side chain amino groups of Asn and Gln are lost via internal cyclization forming aminosuccinimide and glutarimide intermediates, respectively[4–7] Then subsequent hydrolysis leads to the formation of two isomeric products. For Asparagine, this produces the native α-Aspartic acid (Asp) and alternative β-Aspartic acid (iso-Asp).[8] Similarly, glutamine produces the native α-glutamic (Glu) and alternative γ-glutamic acid (iso-Glu) products.[9]

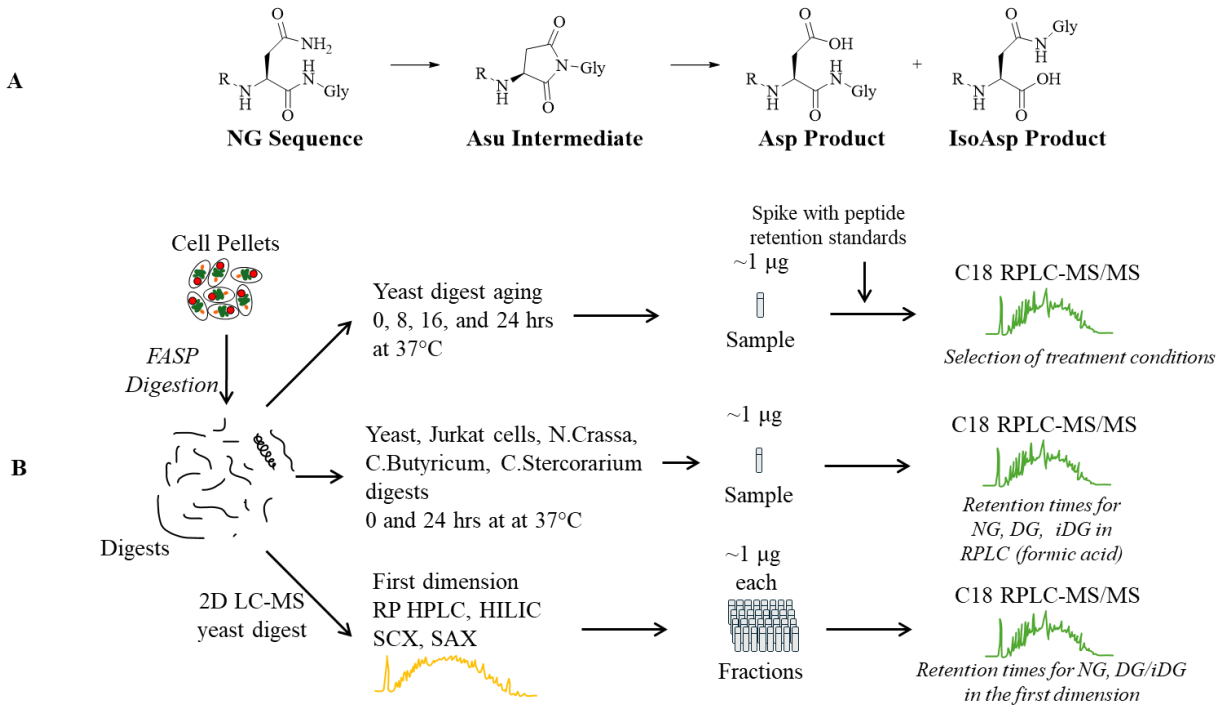


Figure 3.1 Deamidation of Asn-Gly containing peptides and experimental procedures for retention data collection. A – chemical transformation of NG sequences into DG and iDG products; B – experimental workflows for 1D and 2D LC-MS/MS analyses.

When studying deamidation, it is important to understand what factors are affecting spontaneous reaction rates. Protein deamidation rates were shown in early studies to rely primarily on structural features, which are subsequently influenced by environmental conditions. It was found that Asparagine containing sequences had faster rates (half-life of ~1-500 days) than glutamine (half-life of ~100-5000 days), while the 3-dimensional structure and the identity of the neighbouring amino acids on the carboxyl side was ultimately responsible for modulation of these rates.[8,10] Proteins with conformations lacking constraints, as a result of denaturation in non-folded regions, appeared to have increased deamidation rates compared to rigid or folded structures.[11] For asparagine containing sequences, the Asn-Gly motif produced the fastest deamidation rate with a half-life of less than 6 days in sterically unhindered positions, followed by Asn-His and Asn-Ser motifs.[12]

The environmental conditions which were determined to further influence deamidation are pH, buffer type, ionic strength, and temperature. Increased ionic strength, pH and temperature were all found to augment deamidation rates.[13–15] Using phosphate and ammonia buffers resulted in higher deamidation rates while also following an increase in rates with respect to pH, suggesting

that deamidation also depends on the ability of buffers to act as base catalysts.[14] This presents a potent hazard for proteomics experiments where deamidation rates are known to accelerate during workup. For example, the Asn-Gly sequence was found to undergo 80% conversion under standard overnight digestion conditions: 37°C at pH 8.4.[16] In fact, various digestions buffers can alter deamidation rates, where it was found that reactions occur fastest in descending order of triethyl ammonium bicarbonate, ammonium bicarbonate, Tris-HCL, and ammonium acetate.[17] Thus, proper control of environmental conditions and knowledge surrounding the susceptibility to deamidation is crucial when designing experiments which study this modification.

Nonetheless, deamidation cannot be broadly eliminated and still presents a challenge since the formation of isomeric products can result in redundant identification of peptides, reducing the potential breadth of protein identification in complex samples. Identification of endogenous deamidation is another issue, hindered by the acceleration of spontaneous artificial deamidation during sample preparation. Correct annotation is important since biological processes could be mediated by protein deamidation, acting as molecular clocks.[10,18,19] In addition, deamidation of proteinaceous samples, such as in monoclonal antibodies, serves as an important quality control metric in industry to provide higher quality products.[20,21] As such, methods have been developed to characterize proteins and their deamidation products, although often only for Asparagine deamidation due to the slow reaction rates of glutamine containing species.

Various mass-spectrometry techniques have been used to distinguish between deamidation products (Supplementary materials). Chromatography, often complementary to mass spectrometry, has also been applied to analyze deamidation through differentiation of Asp and iso-Asp by their retention characteristics. One powerful separation method, electrostatic repulsion hydrophilic interaction chromatography (ERLIC), was developed by combining the benefits of both ion exchange chromatography (IEC) and hydrophilic interaction chromatography (HILIC), allowing the separation of tryptic peptides based on hydrophilicity and pI.[22,23] It was first applied to separate phosphorylated peptides from complex mixtures but was later used to study deamidation because of the negative charge acquired by Asp and iso-Asp products.[24] In this study, tryptic digests of bovine serum albumin, chicken ovalbumin and rat liver tissue were investigated for deamidation using a 2-dimensional scheme. A first dimension reversed-phase liquid chromatography (RPLC) fractionation was performed under the assumption that Asp and iso-Asp coelute in a standard RPLC separation, then a secondary ERLIC dimension was used to separate

Asp and iso-Asp within the same fraction. They determined that separation in ERLIC afforded the retention order of Asn < Asp < iso-Asp, a similar result found in other HILIC methods later.[25] The notion that Asp and iso-Asp peptides always coelute in RPLC, although sometimes observed, is not always correct. For example, another study which used RPLC to investigate deamidation of trastuzumab in human plasma resolved the peaks of the peptide IYPTNGYTR and deamidation products. The separation was performed using ammonium acetate as ion-pairing modifier and showed an elution order of Asn < iso-Asp < Asp.[26] In a later study by a different group using TFA, this same peptide exhibited iso-Asp < Asn < Asp elution order.[27] Similarly, Krokhin et al.[16] reported Asn < iso-Asp < Asp and iso-Asp < Asn < Asp retention orders for SLNGEWR peptides when using formic (FA) and trifluoroacetic acid (TFA) as ion pairing modifiers, respectively. The same study found the retention order iso-Asp < Asn < Asp for 38 instances and coelution of iso-Asp/Asn pairs for 22 instances for tryptic Asn-Gly containing peptides when using TFA based eluents. These examples suggest that the retention properties of deamidated peptides in RPLC could be dependent from many factors, including eluent composition and peptide sequence.

Two-dimensional (2D) chromatography applications are another set of popular methods due to the increase in identifications and deeper coverage of low abundance PTM-containing peptides. Our lab has routinely used 2D formats to investigate peptide retention properties and developed retention models for several separation settings used in the first dimension.[28–34] These data can serve as a valuable information source on modified peptides' retention under additional (to the second dimension RPLC) chromatographic conditions / separation modes.

Whether through use of diagnostic fragments or a 3:1 ratio between isomeric products, a robust LC-MS method for validation does not exist without the analysis of synthetic peptides to differentiate isomeric products by retention times. We proposed using retention time prediction models as a tool to increase the confidence in identifications of PTM-containing peptides. We have previously characterized the sequence-specific properties of other PTMs and chemical modifications (CM), thereafter generating retention models which are being incorporated into the Sequence-Specific Retention Calculator (SSRCalc, https://proteome.ad.umanitoba.ca/ssrcalc).[1] In this study, we aim to characterize the sequence-specific and structural features that drive retention characteristics for Asn-Gly, Asp-Gly, and isoAsp-Gly (referred to hereafter as NG, DG, and iDG, respectively) containing tryptic peptides in RPLC with formic acid as ion pairing modified. Their

behavior in alternative separation modes have been evaluated in lesser details focusing on non-modified and the most abundant deamidation product (likely iso-Asp).

3.5 Materials and Methods

3.5.1 Materials, Protein Digestion, Chromatographic Columns

All chemicals were analytical chemistry or LC-MS grade and sourced from Sigma Chemicals (St. Louis, MO, USA) or Thermo Fisher Scientific (Waltham, MA, USA) unless noted otherwise. LC-MS grade water and acetonitrile were used for preparation of the eluents. Sequencing-grade modified trypsin (Promega, Madison, WI, USA) was used for the digestion. Tryptic digests of *Saccharomyces cerevisiae* (yeast), human Jurkat cells, *Neurospora crassa*, *Clostriduim butyricum*, *Clostridium stercorarium* were prepared using FASP digestion procedure as described elsewhere.[35] The resulting mixtures of tryptic peptides were purified using reversed-phase HPLC and lyophilized. Designed standard peptides P1-P6[36] and synthetic analogs containing NG and DG sequences shown in Supplementary Table S1 were synthesized by BioSynthesis Inc. (Lewisville, TX, USA). LC-UV measurements have been performed using 3x100 XTerra MS C18 3.5 µm column (Waters, Milford, MA). Nanoflow LC-MS/MS measurements employed trap column set up: 100µm x 200mm analytical column packed with 3 µm Luna C18(2) (Phenomenex, Torrance, CA) and 300 µm x 5 mm PepMap 100 trap-column (Thermo Fisher).

3.5.2 Spontaneous Deamidation Induced by the Storage at 37 °C pH 8.4

Tryptic digest of yeast was stored at 37°C pH 8.4 conditions for 0 (no treatment), 8, 16, and 24 hours and analysed by LC-MS/MS to verify the effect of deamidation on protein / peptide identification output. Subsequently, the rest of the samples have been aged under the same conditions for 24 hours (**Figure 3.1B**). Both non-treated and 24 hr samples were analysed in an attempt to extract retention values for NG, DG, and iDG variants. Synthetic peptides containing NG have been aged for ~12 hours to generate the situation where all three components of the mixture will be detectable by LC-UV measurements to confirm retention time assignments for all three components of the reaction mixture.

3.5.3 Chromatographic Conditions and Apparatus.

LC-UV measurement for synthetic peptides and first dimension separation for 2D LC-MS/MS. Agilent 1100 series HPLC system with UV detector (214 nm) with a 20 or 100 µL injection loop and 150-300 µL/min flow rate was used for the separations. The separations of synthetic peptides

have been done using 2% per minute water – acetonitrile (ACN) gradient on 3x100 mm 3.5 µm XTerra MS C18 column (Waters, Milford, MA) at 300 µL/min flow rate. Three different ion-pairing systems have been used (0.1% v/v): formic (FA), trifluoroacetic (TFA), heptafluorobutyric (HFBA). Each peptide sequence was represented by two synthetic analogs containing NG or DG. They were analysed individually and compared with aged NG peptides (12 hours at 37 ^{o}C pH 8.4). This allowed assignment of retention times for all 3 components of the reaction mixture: NG, DG and iDG. The identity of the latter was additionally confirmed based on ~3:1 iDG:DG ratio.

First dimension retention data have been taken from a series of our previous publications on modeling various peptide separation modes and studying peptide separation orthogonality in 2D LC proteomic applications.[28–34] The separation conditions: columns, gradients, flow rates are given in supplementary **Table S3.2**. All separations have been performed using 100-200 µg yeast digests. Fractions from different separation modes and model tryptic digest from this study were dissolved in buffer A for LC-MS/MS (0.1% formic acid in water) to provide ~1 ug injections and spiked with P1-P6 standard[36] for retention time alignment purposes.

3.5.4 LC-MS/MS Acquisitions.

LC-MS/MS runs for 5 tryptic digests (non-treated and 24 hours exposure) in this work and the second-dimension separations in 2D LC-MS/MSs have been done using 2D LC Ultra system (Eksigent, Dublin, CA). 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 ~0.29% acetonitrile per minute gradient. Both eluents A (water) and B (acetonitrile) contained 0.1% formic acid. The gradient program consisted of following steps: a linear increase from 0.4 to 31% buffer B (acetonitrile) in 107 minutes, 5 minutes at 80% B and then 8 minutes at 0.4% B for column equilibration (120 min total analysis time). All MS/MS acquisitions were done using a standard data-dependent acquisition protocol using a TripleTOF 5600 mass spectrometer (Sciex, Concord, ON) detailed in Supplementary Experimental conditions.

3.5.5 Data Analysis and Retention Time Assignment.

The X!Tandem search engine 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 peptides (Log (e) < -3) were considered.

Retention times in the first dimension were assigned as the fraction number in which this peptide was found. When the 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 is sufficient for evaluation of retention properties on the large scale. Retention times of modified and non-modified peptides for LC-MS/MS (C18, formic acid) were assigned by manual inspection of extracted ion chromatograms for the model digests or through automatic (non-supervised) analysis for high pH RP – low pH RP two-dimensional run. Since no distinction has been made between DG and iDG products in the latter case, we can assume that most of these assignments correspond to more intense iDG product. Experimental retention times for RPLC formic acid have been converted into hydrophobicity index HI (acetonitrile %) units using tabulated HI values for P1-P6 peptides.[36]

3.6 Results and Discussion

3.6.1 Retention Data Collection Using Extended Exposure to Spontaneous Deamidation.

Our approach to studying the effect of PTMs on peptides' chromatographic properties is based on measuring retention values for the modified and non-modified peptide pairs. The way to collect this information depends on the chemical properties and abundance of the modification. Monitoring in-vivo or spontaneous modifications during sample preparation, enzymatic / chemical modification of peptides, co-translational incorporation of modified residues, analysis of synthetic peptides – all can be employed to generate correspondent pairs of peptides.[1]

Deamidation is very abundant in proteomic samples. Previous reports on deamidation showed an 80% conversion of NG sequences during a standard overnight digestion, reflecting the importance of considering deamidation during proteomics analyses.[16] **Table 3.1** shows the summary of all identified deamidated (N and Q) and oxidized (M and W) peptides from yeast digest at 0-, 8-, 16-, and 24-hour intervals of exposure to favourable reaction conditions (10mM ammonium bicarbonate, pH=8.4, 37°C). Analysis immediately following sample preparation (no exposure) resulted in a benchmark of 716 identified deamidations (~7.7% of total identification output), encompassing *in-vivo* and *in-vitro* conversion. For comparison, a study on deamidation using mouse brain tissue by Nepomuceno et al. showed a similar result where 7.2% of 44,392 peptides were deamidated (N and Q).[37] At 8- and 16-hour intervals of exposure, 1030 peptides (~11.6% of output) and 1158 peptides (~12.9% of output) were identified with deamidation, respectively. After 24 hours of exposure, 1523 deamidations were identified. Based on this data we used 24 hours, for the "artificial aging" of remainder of samples.

Table 3.1 The effect of exposure time on deamidation rates found for yeast digest (37°C, pH 8.4).

Sample	# of deamidated peptides (N and Q)	# of oxidized peptides (M and W)	# of proteins	# of unique peptides	# of peptides	# of MS/MS
Yeast 0 hr	716	843	1728	9311	22021	41433
Yeast 8 hr	1030	867	1612	8916	21613	41107
Yeast 16 hr	1158	892	1646	8982	21819	41444
Yeast 24 hr	1523	834	1647	8977	22903	41976

Tryptic digests of yeast, Jurkat human cells, *N. crassa*, *C. butyricum*, and *C. stercorarium* were analysed in identical fashion for 0 and 24 hours exposure by LC-MS/MS. Following procedure was used to ensure a high quality of manual assignment and improve confidence of peptide identification:

i. Only deamidated peptides carrying NG sequences with confidence score log(e) < -3 were considered. This eliminated most false identifications and incorrect assignments made based on incorrect selection of monoisotopic peaks for fragmentation.
ii. Extracted ion chromatograms for non-modified and modified fragments (M+0.984 Da) have been inspected manually in both 0 and 24 hrs exposure samples.
iii. The decision to include respective peptides in the final collection was made only if correspondent peak intensities decreased/increased following sample aging.
iv. Modified peptides identities were assigned based on the ratio of signal intensities ~3:1 iDG:DG.

Figure 3.2 shows the selected examples of extracted ion chromatogram profiles. Samples without additional exposure (0 hours) showed a smaller intensity of deamidated products, thus adding confidence to our evaluations. These three examples correspond to 3 cases of elution order observed and discussed in the following sections.

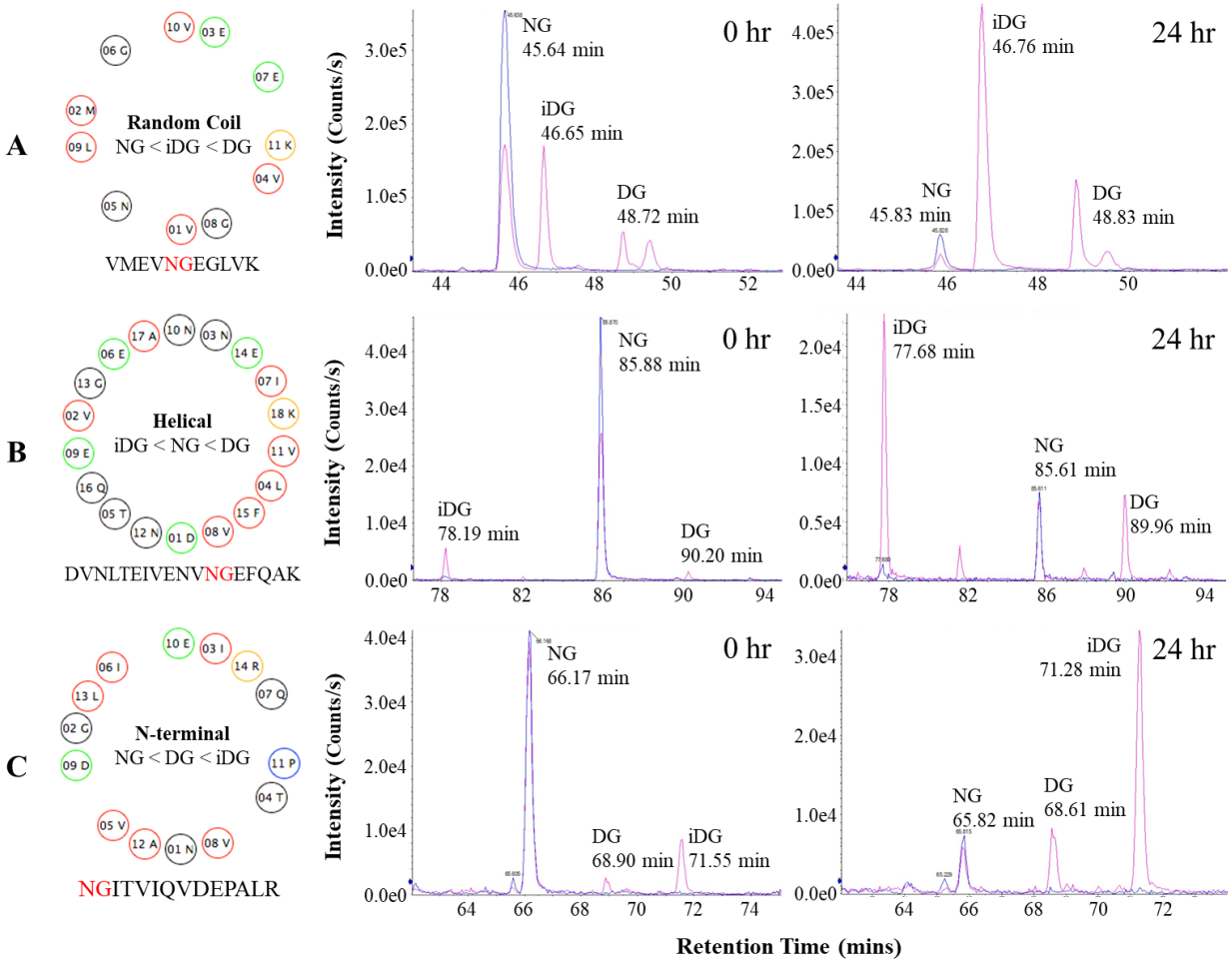


Figure 3.2 Retention orders for non-modified peptides and deamidation products following 0 and 24 hours incubation: A – random coil VMEVNGEGLVK. B – amphipathic helical DVNLTEIVENVNGEFQAK featuring multiple hydrophobic residues on one side of the helical wheel. C – NGITVIQVDEPALR with N-terminal modification site. Axial helical projections are shown with hydrophobic residues highlighted in red. Extracted ion chromatograms are showing traces for non-modified peptide m/z (blue) and deamidation products (m+0.984)/z (pink).

3.6.2 RP HPLC Behaviour of NG, DG and iDG Analog Peptides with Formic Acid as Ion Pairing Modifier.

Retention values for 343 peptide triads have been confidently assigned using the method outlined above. We found that correspondent NG, DG and iDG analogs exhibit three major elution trends. The elution of iDG was found to depend on primary and secondary peptide structure, resulting in all three possible positioning of iDG relative to the NG and DG pair. Whereas the NG < DG elution order was found exclusively. The latter agrees with the fact that acidic Asp and Glu have higher hydrophobicity values compared to their Asn and Gln analogs.[34,38]

The most common trend observed for iDG peptides, which was found in 50.9% of the triads, comprising random coil peptides where the NG sequence is found internally (peptide VMEVNGEGLVK in **Figure 3.2A**). This corresponded to an elution order of NG < iDG < DG. The decrease in retention of iDG compared to DG can be attributed to differences in the peptide backbone structure. The constitutional isomers differed in structure by one methylene unit (-CH_2-), positioned in the backbone for iDG peptides and in the side chain for DG peptides (**Figure 3.1A**). As a result, two assumptions can be made for iDG peptides: reduced flexibility in the orientation of their side chain disfavours conformations which increase hydrophobic interactions with the non-polar stationary phase, and better steric availability of methylene group in DG increases retention on C18 phase.

In 31.6% of cases, a trend for peptides primarily containing NG in the middle of the amphipathic helical stretch featured an elution order of iDG < NG < DG (Peptide DVNLTEIVENVNGEFQAK in **Figure 3.2B**). Retention of amphipathic helical peptides have been found to be driven by their additional stabilization upon interaction of hydrophobic face of the helix with C18 surface, leading to increased retention in RP HPLC.[39] The negative retention shift for iDG peptides compared to NG helical peptides can be explained by the drastic change of the helix structure (angles between atomic bonds) due to addition of extra carbon atom in the peptide backbone. This effect is similar to the behaviour of iso-Asp in proteins[40], where deamidation could induce change of conformation from helical to beta-sheet structures.

The elution order of NG < DG < iDG was observed in 15.8% of cases. This trend is represented predominantly by peptides with NG sequences primarily at or near N-terminal position (Peptide NGITVIQVDEPALR in **Figure 3.2C**). Our study on modeling peptide capillary zone electrophoresis at acidic pH showed a significant decrease in apparent charge of N-terminal amino group for terminal positions of Asp (by ~0.27).[41] The effect was found smaller when modified residue is places into positions #2, 3, etc. from N-terminus. This leads to reduced ion-pairing formation and increased retention due to the loss of hydrophilic formate counterions. In case of iDG, this effect is likely even more pronounced since a carboxyl group of the iso-Asp side chain is located closer to peptide backbone compared to Asp. It should be noted that N-terminal positioning of modified residue results in the largest positive retention time shifts for both Asp and iso-Asp derivatives. Apparently, as the distance of the modified residue from the N-terminus

increases, a corresponding reduction in magnitude of the shift and reversal of DG < iDG elution order occurs, thus resulting in the characteristic behaviour of the first category of elution orders.

3.6.3 The Distribution of Peptide Retention Shifts as Determined Using Manual Annotation and Unsupervised Assignments.

Plotting the distribution of retention shifts for all manually identified pairs (**Figure 3.3A-B**) and pairs generated from unsupervised workflows (from X! Tandem identifications **Figure 3.3C)** reflects a set of general characteristics which should be noted. The overall trend observed for deamidated peptides is that DG peptides increase retention in a much more controlled manner, while iDG peptides shift in a broader range as a result of their structural effects on peptide conformation and ion pairing as discussed previously. Blind analysis of deamidated peptides is expected to favour the identification and pair generation of the more intense product, isoaspartic acid. As can be seen from **Figure 3.3C**, the unsupervised generation of pairs represents an intermediate case: a broader distribution compared to aspartic acid, and narrower compared to iDG. It should be noted that the total number of modified peptide identifications is much larger in the blind run (994 compared to 348) based on 2D LC-MS/MS analysis. Since the vast majority of elution orders correspond to case 1 where DG and iDG elute very closely, the correct assignment of retention values could be compromised.

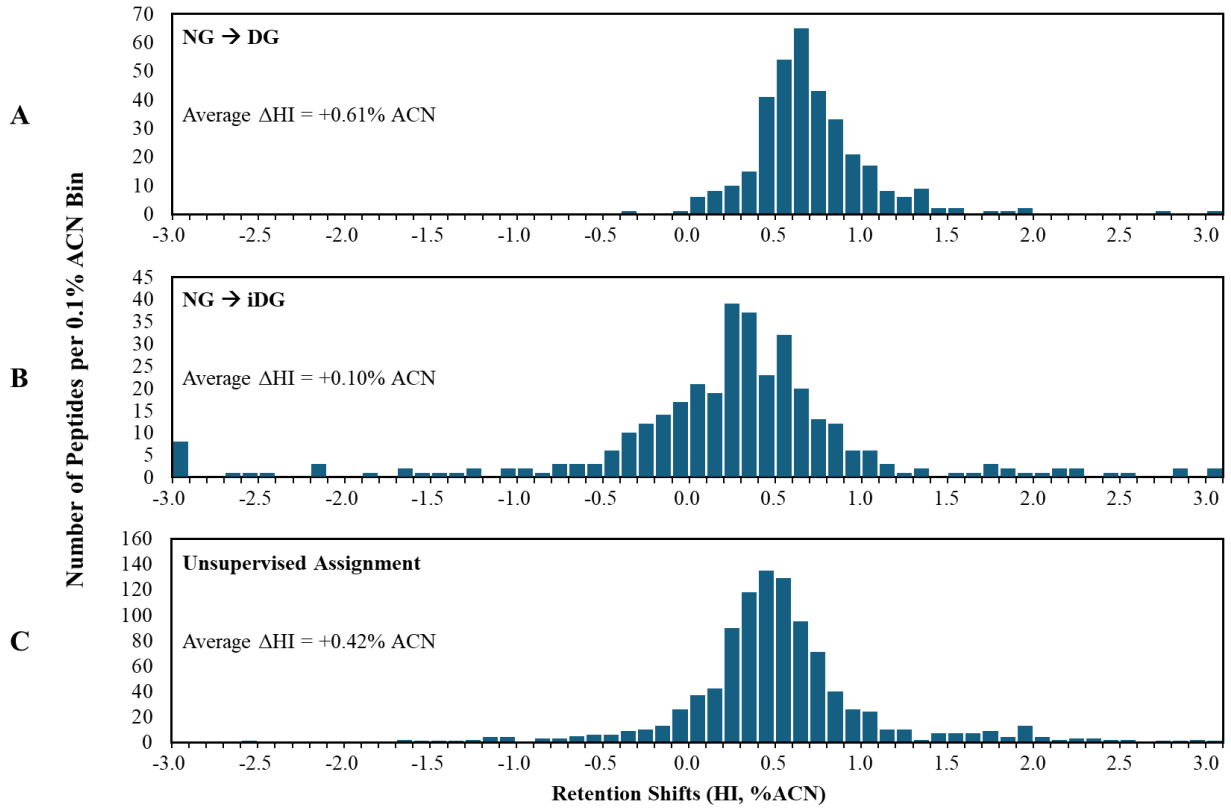


Figure 3.3 Distribution of retention shifts (ΔHI, %ACN) in RP HPLC (0.1% formic acid) for corresponding deamidation pairs upon modification. A – retention shifts for the conversion of NG to DG. B – retention shifts for the conversion of NG to iDG. C – Retention shifts for pairs generated by unsupervised assignment.

3.6.4 Validation of Retention Features Found by LC-MS/MS Analyses Using Synthetic Peptides

A series of synthetic peptides have been designed to confirm three major elution orders described above: random coil, amphipathic helical and sequences with N-terminal positioning of modified residues. It should be noted that three representative cases sometimes could be hard to design properly as they have properties, which gradually transition positioning of respective species. Thus, there is no clear distinction between 100% helical and 100% random coil peptides as the boundaries for helical stabilization have yet to be quantitively defined. The same applies to positions #2, 3, 4… at N-terminus. At which point the loss of ion-pairing environment become less dominating compared to the difference in hydrophobicity between Asp and iso-Asp side chains? In some instances, all three features driving retention properties (helicity, basicity and hydrophobicity of a side chain) could be changed simultaneously. In addition to it, it was shown

that identical modifications incur smaller retention shifts for larger peptides.[1] All together these makes de-novo design of peptides with specific properties very challenging.

Table S1 shows retention times determined for all 3 components (NG, DG, and iDG peptide triads) of the reaction mixtures through LC-UV measurements for several selected peptides. Retention times for NG and DG variants were assigned for the analyses of these individual synthetic molecules. iDG variants have been detected in 12-hour aged samples based on the appearance of new peak and relative intensity of a signal for Asp and iso-Asp peaks. **Figure 3.4** shows the elution profiles for A) amphipathic helix SIVEEINGVIGEGER, B) random coil SSNGLSGGLLR and C) NGLSSSGGLLR with N-terminal deamidation site – all for 12-hour aged peptides. Clearly, examples A and C are characterized by the largest changes in the retention properties. SSNGLSGGLLR exhibit a modest retention increase with overlapping peaks of two deamidated products.

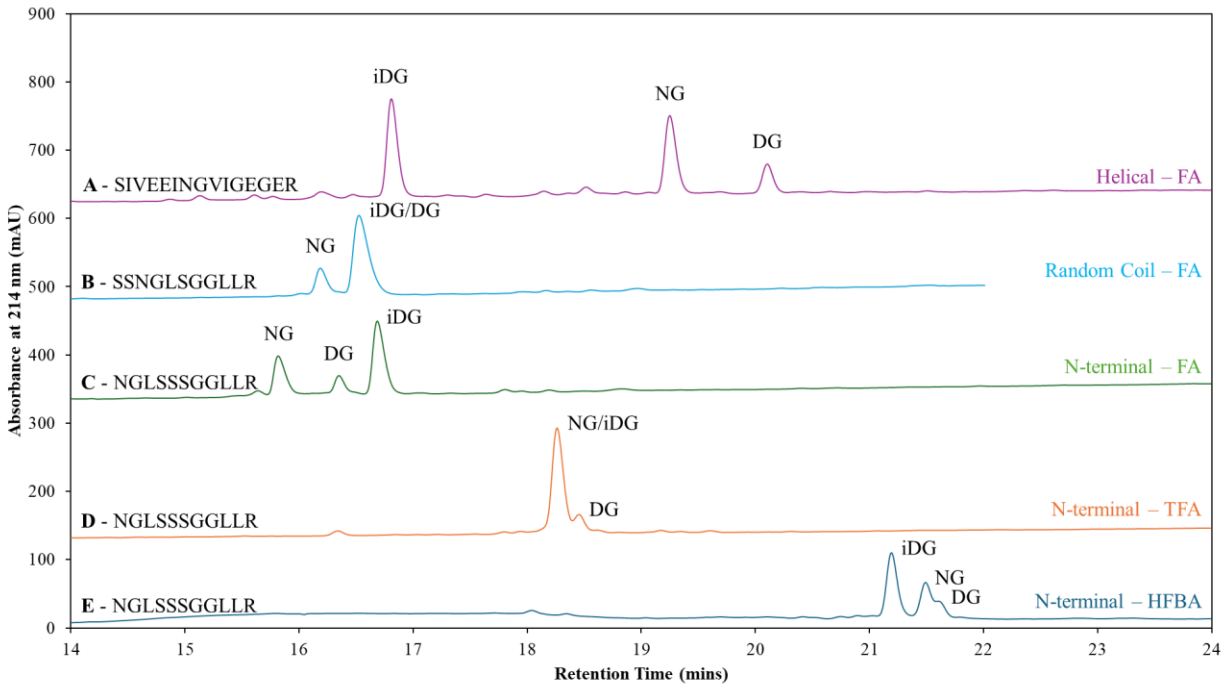


Figure 3.4 RPLC separations for synthetic peptides and their deamidation products. A – amphipathic helical peptide SIVEEINGVIGEGER (FA). B – random coil peptide SSNGLSGGLR. C – NGLSSSGGLLR with N-terminal modification - all with 0.1% FA as ion pairing modifier. D – NGLSSSGGLLR using 0.1% TFA. E – NGLSSSGGLLR (0.1% HFBA).

It was of interest to monitor the effect of ion-pairing modifier's hydrophobicity as shown in **Figure 3.4C-E**. Changing ion-pairing additive increases retention of all peptides in the order FA < TFA < HFBA, as expected. At the same time relative retention of the original NG and deamidation products are changing. Retention of more acidic DG and iDG grows slower in hydrophobic ion-pairing environments, thus confirming our explanation of the retention orders which was based on ion-pairing properties. The loss of more hydrophobic HFBA- counter-ions due to decreased peptide basicity leads to reduced retention of deamidated products relative to NG. This effect is more pronounced for iDG due to a closer positioning of acidic side chain to the peptide backbone. This results in complete reversal of iDG elution: the last in FA to the first eluting component with HFBA.

3.6.5 Chromatographic Behaviour of Deamidated Peptides in Various Separation Modes Used in the First Dimension of 2D Methods.

The use of alternative separation modes in 2-dimensional chromatography where the second dimension is reserved for RP HPLC with formic acid as ion-pairing modifier, results in an increase in peptide identification output.[42] The advantage of using a particular separation mode is emphasized by its orthogonality compared to RP HPLC in the second dimension.[31] Deamidation results in the formation of acidic Asp and iso-Asp, therefore this specific chemical property can be used to improve the separation of deamidated products from the non-modified peptide. This feature can be utilized by eluent pH control, the selection alternative separation mode and ion pairing. Such separations have been demonstrated on several occasions[22–27], providing a comprehensive chromatographic toolset when good peak resolution is required. Our previous studies on peptide separation orthogonality using 2D LC-MS/MS provided us with much more diverse datasets to study the effect of deamidation.[31] At the same time, since retention values in the first dimension are assigned using discrete fraction numbers it complicates distinguishing between identical-mass Asp and iso-Asp analogs. Nevertheless, in this work we attempted large scale "unsupervised" assignment of retention properties in various separation modes used in proteomic 2D LC-MS/MS.

Figure 3.5 summarizes all retention shifts found for NG containing deamidated peptides using alternative separation conditions (to RP HPLC / formic acid). Since no distinction has been made between Asp and iso-Asp, we assume that retention values assigned based on peak intensity likely correspond to the iso-Asp analog. Retention shifts are expressed as the change in eluent

composition upon elution between NG containing peptides and their corresponding iDG modified sequences in water, acetonitrile % units, and mM of salt for HILIC, RPLC, and ion-exchange separations, respectively.

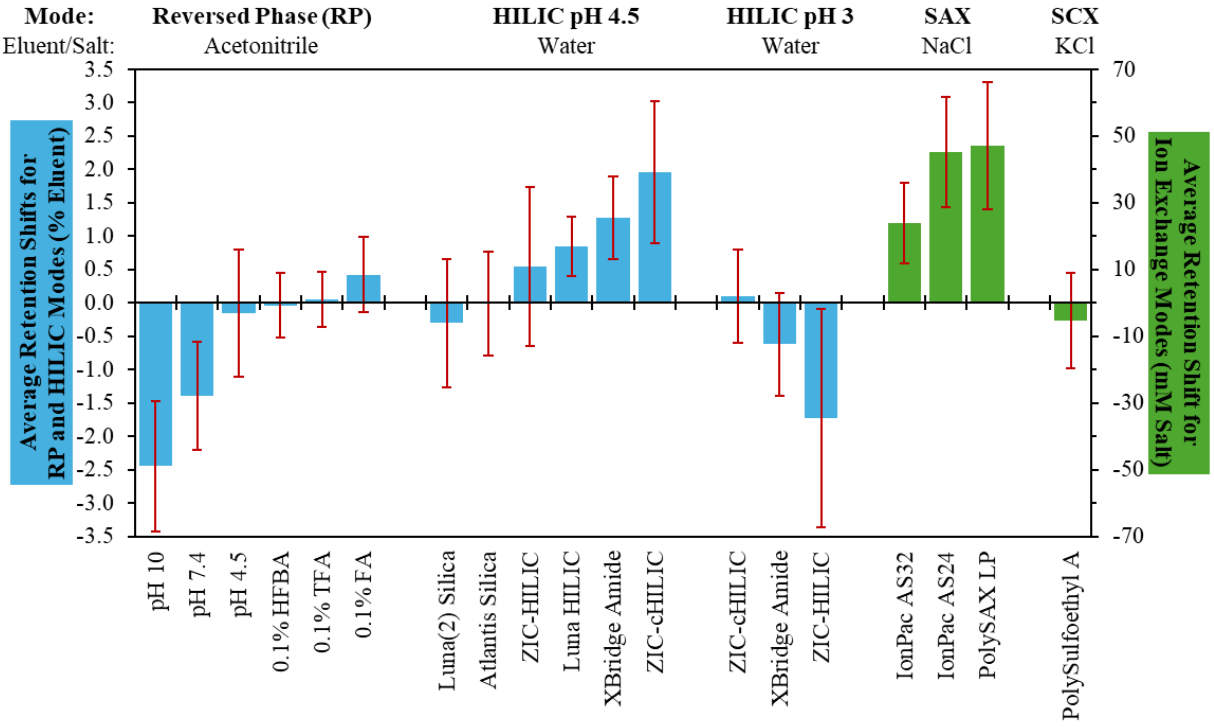


Figure 3.5 Summary of retentions shifts for deamidated products across 20 various peptide separation modes. Red error bars indicate the standard deviation for each respective average shift. In this case, the average shifts and standard deviations are not calculated from repeated measurements of the same peptide but rather represent the chemical diversity of all observed modified and nonmodified peptide pairs in each separation mode.

Across the reversed-phase modes that were evaluated, eluent pH (predominantly) and the ion-pairing properties appear to drive the observed retention shifts. Increasing pH of the eluent leads to dissociation of carboxy group on Asp and iso-Asp side chains creating very hydrophilic functional group. Not surprisingly, we observed a steady decrease in the retention shifts from slightly positive (acidic pH, formic acid) to increasingly negative at pH 10. The effect of the ion-pairing modifier on hydrophobicity at acidic pH was the same as shown in **Figure 3.4C-E**: deamidated species retain higher when employing a more hydrophilic modifier.

HILIC represents another mode which could be utilized for the separation of deamidated peptides due to their increased hydrophilicity upon dissociation of Asp and iso-Asp side chains.[25]

HILIC separations were investigated using pH 4.5 and 3 conditions which provided significantly different ionization statuses and must be considered separately. It is important to note that at pH 4.5, the aspartic acid side chains are expected to be ~80% deprotonated, resulting in large positive retention shifts under HILIC conditions. The differences we observed are driven by relative hydrophilicity of the stationary phases and the charge of their functional groups. The Luna and Atlantis silica have negatively charged silanols resulting in electrostatic repulsion for Asp and iso-Asp and a slight retention decrease. ZIC-cHILIC being zwitterionic, features a distal position of a positively charged group. This leads to additional electrostatic attraction and largest retention increase. Three sorbents on this list ZIC-HILIC, XBridge amide, ZIC-cHILIC possess about the same hydrophilicity.[29] Their relative positioning in **Figure 3.5** was driven by different orientation of zwitter-ionic pairs on ZIC sorbents. Three neutral HILIC phases Luna HILIC < XBridge amide < Plyhydroxyethyl A exhibited retention increase orders in accordance with their relative hydrophilicity[32], as well.

Under acidic HILIC conditions (0.1% formic acid) studied for three sorbents, the retention shifts reduced drastically: there is less Asp and isoAsp side chain dissociation, and they are more hydrophobic compared to Asn. At the same time, ZIC-HILIC, XBridge amide, ZIC-cHILIC showed the same effect when it comes to the orientation of zwitterionic pairs: ZIC-HILIC exhibited the most negative (large) and ZIC-cHILIC the most positive (slight) retention shifts.

On top of the conceptually more complicated modes such as HILIC, ion-exchange ones showed unsurprising results. Anion exchange separations at pH 8 saw a dramatic increase in retention due to the additional negative charge appearing upon deamidation. Strong cation-exchange at acidic pH showed modest negative shifts, as a result of reduced peptide basicity as was described earlier for capillary zone electrophoresis modeling study.[41]

3.7 Conclusion

Asparagine deamidation is one of the most abundant protein/peptide modifications targeted by many researchers for the development of analytical techniques to detect it. There are several chromatographic techniques amenable to separate non-modified peptides and two deamidation products. At the same time, most of these publications targeted selected molecules, thus limiting generalization of common chromatographic trends. In this work we attempted to address this issue on a larger scale by overviewing peptide retention for deamidated species (all from peptides carrying NG sequences) relative to their non-modified counterparts for RP HPLC separation with formic acid as ion-pairing modifier. These chromatographic settings are dominant in the field of proteomics, thus providing wide applicability of our findings.

Aspartic acid is more hydrophobic compared to asparagine. Therefore an Asn < Asp retention order was established, as expected. The behavior of iso-Asp derivatives varied widely driven by various features related to alterations in residue's hydrophobicity, acidity, and propensity to form helical structures. Depending on the combination of these factors, the iso-Asp derivative can elute in all three possible positions relative to the Asn < Asp pair. We conclude that most peptides carrying NG sequences exhibit an NG < iDG < DG elution order. This situation occurs for deamidation sites situated in the middle of peptide, which reside in random coil conformation. This suggests that iso-Asp is more hydrophilic compared to Asp: corresponding to the structural re-arrangement by moving one methylene group from side chain into peptide back bone. This makes the hydrophobic methylene group less accessible for interaction with C18 stationary phase. The iDG < NG < DG elution order was found for amphipathic helical peptides in which the addition of extra carbon into the backbone reduces helix stability and hence RP HPLC retention. The third possible elution order, NG < DG < iDG, was found for peptides carrying modified residues at or near N-termini. The same peptides exhibited the largest retention increase for the transformation of Asn to Asp. We explain this observation by the effect of the acidic Asp side chain on the basicity and ion-pairing ability of N-terminal amino groups. Loosing hydrophilic formate counterions due to decrease of its basicity leads to increased retention. The side chain carboxyl group of iso-Asp is located closer to the peptide backbone, thus increasing this charge effect resulting in increased retention of iso-Asp derivatives.

The effect of ion-pairing was indirectly confirmed by changing the hydrophobicity of ion-pairing agent through using formic, trifluoroacetic and heptafluorobutyric acid. The loss of more

hydrophobic counterions lead to a smaller increase of retention times upon deamidation. This was shown for both measurements using synthetic peptides and analyses of large scale 2D LC proteomics data. The latter employed RPLC (TFA, HFBA) in the first and RPLC FA in the second separation dimension. According to our findings, using FA-based eluents provides a better change to achieve a complete resolution of all 3 peaks constituting the components of the deamidation reaction.

The effect of deamidation was evaluated for twenty different separation modes varying in mechanism, pH of the eluents, chemistry of stationary and mobile phase. This was done, however, without distinguishing between Asp and iso-Asp variants. Our findings follow general ideas of dominant effects of ionization state and separation mechanism. The largest retention shifts between Asn and its Asp and iso-Asp pairs have been found for high pH RPLC (negative shift), pH 8 anion-exchange and pH 4.5 HILIC separation on neutral stationary phases.

Although our experimental procedures focused on NG containing peptides, current knowledge of peptide chromatography allows us to extrapolate our findings for deamidation of Asn and Gln in RP HPLC with formic acid in general. Replacing Gly with other residues will not change observed trends dramatically in case of Asn, except for amphipathic helical peptides. Both Gly and Pro (in particular) are characterized by a lower propensity to form helical structures. Therefore, any other replacement of residues will likely result in a higher helicity and a larger negative retention shifts for the iso-Asp product in helical conformation. The difference between Gln and Glu, and Asn and Asp retention coefficients is virtually the same for formic acid RP HPLC.[34] Because of this, we anticipate similar retention differences between these pairs based solely on hydrophobicity. At the same time, N-terminal positioning of Glu will result in smaller positive peptide retention shifts due to the diminished effect of acidic Glu vs. Asp on the basicity of a peptide.[41] The retention shifts of iso-Glu in the amphipathic helix will likely be even more negative compared to iso-Asp due to insertion of two extra methylene groups in peptide backbone.

3.8 Acknowledgements

This work was supported by grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2016-05963 and RGPIN-2022-05015 O.V.K).

3.9 Supplementary Information

Distinguishing between deamidation products using mass spectrometry

Mass spectrometry (MS)-based bottom-up proteomics is one of the most popular tools to investigate deamidation since it can be applied to complex samples. In these cases, modified and non-modified peptides are differentiated by their mass shifts and corresponding tandem mass spectra (MS/MS). However, differentiating between the two isomeric modification products remains challenging since they produce identical mass shifts and corresponding b and y ions. The simplest approaches assigned iso-Asp and Asp based on a previously reported 3:1 (iso-Asp:Asp) intensity ratio.[1] The development of collision induced dissociation (CID)-based methods has helped uncover isomer-specific characteristics, where iso-Asp containing synthetic peptides generated $b_{n-1}+H_2O$ and y_{l-n+1}-46 products, where l is the length of the peptide and n is the distance of iso-Asp from the N terminus.[2,3] However, the intensities of these fragment products were found to be inconsistent, varying based on sequence, composition, and ionization methods.[3] Another problem with confident identification of deamidation raises from incorrect automatic selection of monoisotopic parent ions for fragmentation in data-dependent MS/MS acquisition. This is true especially for larger peptides in which the second peak in the isotopic envelope has larger intensity than monoisotopic peak. This often results in false identification of deamidation.

Electron transfer dissociation (ETD) and electron capture dissociation (ECD) methods have also been shown to differentiate Asp and iso-Asp products based on c and z ion. In these methods, the diagnostic iso-Asp fragment ions c_n•+58 and z_{l-n}-57 in ECD, and c_n+57 and z_{l-n}•-57 in ETD, have been used for identification, where l is the length of the peptide and n is the distance of iso-Asp from the N terminus.[4,5] Additionally, a neutral loss of 60 Da was found corresponding to the loss of the Aspartic acid side chain, allowing the identification of Aspartic acid products.[6] Unfortunately, applying ECD- and ETD-based strategies to high throughput proteomics analyses is challenging since there is a preference for peptides which carry higher charged states, ideally carrying 3 or more positive charges.[7–9] This is an important consideration given that most tryptic peptides are found in the +2 charged state, meaning that proteome-wide applications of these methods can become insufficient.[10] Supplemental activation was shown to increase the abundance of diagnostic c and z ions in ETD analysis of +2 charged peptides, however this required specialized instrumentation on top of the pre-existing requirement for ECD/ETD capabilities.[11,12] Moreover, the diagnostic c and z ions from a study using ECD were found to be unreliable, where

signals were believed to arise from instrument noise or potential co-isolation during MS/MS.[9] Thus, there is yet to exist an irrefutable MS method of analysis.

Supplementary References

1. Geiger, T.; Clarke, S. Deamidation, Isomerization, and Racemization at Asparaginyl and Aspartyl Residues in Peptides. Succinimide-Linked Reactions That Contribute to Protein Degradation. *Journal of Biological Chemistry* **1987**, *262* (2), 785–794. https://doi.org/10.1016/S0021-9258(19)75855-4.
2. Papayannopoulos, I. A.; Biemann, K. Fast Atom Bombardment and Tandem Mass Spectrometry of Synthetic Peptides and Byproducts. *Pept Res* **1992**, *5* (2), 83–90.
3. González, L. J.; Shimizu, T.; Satomi, Y.; Betancourt, L.; Besada, V.; Padrón, G.; Orlando, R.; Shirasawa, T.; Shimonishi, Y.; Takao, T. Differentiating α- and β-Aspartic Acids by Electrospray Ionization and Low-Energy Tandem Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **2000**, *14* (22), 2092–2102. https://doi.org/10.1002/1097-0231(20001130)14:22<2092::AID-RCM137>3.0.CO;2-V.
4. Cournoyer, J. J.; Pittman, J. L.; Ivleva, V. B.; Fallows, E.; Waskell, L.; Costello, C. E.; O'Connor, P. B. Deamidation: Differentiation of Aspartyl from Isoaspartyl Products in Peptides by Electron Capture Dissociation. *Protein Science* **2005**, *14* (2), 452–463. https://doi.org/10.1110/PS.041062905.
5. O'Connor, P. B.; Cournoyer, J. J.; Pitteri, S. J.; Chrisman, P. A.; McLuckey, S. A. Differentiation of Aspartic and Isoaspartic Acids Using Electron Transfer Dissociation. *J Am Soc Mass Spectrom* **2006**, *17* (1), 15–19. https://doi.org/10.1016/j.jasms.2005.08.019.
6. Haselmann, K. F.; Budnik, B. A.; Kjeldsen, F.; Polfer, N. C.; Zubarev, R. A. Can the (M• — X) Region in Electron Capture Dissociation Provide Reliable Information on Amino Acid Composition of Polypeptides? *European Journal of Mass Spectrometry* **2002**, *8* (6), 461–469. https://doi.org/10.1255/ejms.519.
7. Iavarone, A. T.; Paech, K.; Williams, E. R. Effects of Charge State and Cationizing Agent on the Electron Capture Dissociation of a Peptide. *Anal Chem* **2004**, *76* (8), 2231–2238. https://doi.org/10.1021/ac035431p.
8. Pitteri, S. J.; Chrisman, P. A.; Hogan, J. M.; McLuckey, S. A. Electron Transfer Ion/Ion Reactions in a Three-Dimensional Quadrupole Ion Trap: Reactions of Doubly and Triply Protonated Peptides with $SO_2^{\bullet -}$. *Anal Chem* **2005**, *77* (6), 1831–1839. https://doi.org/10.1021/ac0483872.
9. Yang, H.; Fung, E. Y. M.; Zubarev, A. R.; Zubarev, R. A. Toward Proteome-Scale Identification and Quantification of Isoaspartyl Residues in Biological Samples. *J Proteome Res* **2009**, *8* (10), 4615–4621. https://doi.org/10.1021/PR900428M.
10. Sun, W.; Wu, S.; Wang, X.; Zheng, D.; Gao, Y. A Systematical Analysis of Tryptic Peptide Identification with Reverse Phase Liquid Chromatography and Electrospray Ion Trap Mass Spectrometry. *Genomics Proteomics Bioinformatics* **2004**, *2* (3), 174–183. https://doi.org/10.1016/S1672-0229(04)02023-6.
11. Chan, W. Y. K.; Chan, T. W. D.; O'Connor, P. B. Electron Transfer Dissociation with Supplemental Activation to Differentiate Aspartic and Isoaspartic Residues in Doubly Charged Peptide Cations. *J Am Soc Mass Spectrom* **2010**, *21* (6), 1012–1015. https://doi.org/10.1016/j.jasms.2010.02.002.

12. Swaney, D. L.; McAlister, G. C.; Wirtala, M.; Schwartz, J. C.; Syka, J. E. P.; Coon, J. J. Supplemental Activation Method for High-Efficiency Electron-Transfer Dissociation of Doubly Protonated Peptide Precursors. *Anal Chem* **2007**, *79* (2), 477–485. https://doi.org/10.1021/ac061457f.

Table S3.1 Summary of retention times for synthetic peptides in C18 RP with 0.1% FA

Type	***Peptide Sequence***	***RT NG (mins)***	***RT DG (mins)***	***RT iDG (mins)***
Random Coil	GTIEVDGADLVVXXK	17.33	17.61	17.61
Random Coil	IAIXXFGR	16.15	16.88	16.15
Random Coil	LDHFIASXXTLFNQK	17.02	17.39	17.39
Random Coil	SSXXLSGGLLR	16.21	16.51	16.51
N-Terminal	XXLSSSGGLLR	15.84	16.37	16.68
N-Terminal	XXEPGIIFLDR	19.06	19.68	20.24
Helical	SIVEEIXXVIGEGER	19.2	20.12	16.82
Helical	SIVXXIEEVIGEGER	20.47	21.16	18.78

Table S3.2 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 in water B: 10 mM ammonium formate pH 4.5 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
Acidic HILIC (gradient: % water increase starting from 10:90 water: acetonitrile, 0.3 mL/min)			
XBridge Amide 3.5µm 100Å (3x50)	A: Water 0.1% formic acid B: 90:10 ACN:water with 0.1% formic acid	1%	38
ZIC-HILIC 3.5 µm 100Å (2.1x100)		1%	47
ZIC-cHILIC ZIC 3 µm 100Å (2.1x100)		1%	35
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 as A 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)	20mM ammonium bicarbonate pH 7.4	1%	70 (0.5 min)
Luna C18(2) 5 µm 100Å (1x100)	10mM ammonium formate pH 4.5	1%	70 (0.5 min)
Luna C18(2) 5 µm 100Å (1x100)	0.1% formic acid	1%	70 (0.5 min)
Luna C18(2) 5 µm 100Å (1x100)	0.1% trifluoroacetic acid	1%	70 (0.5 min)
Luna C18(2) 5 µm 100Å (1x100)	0.1% heptafluorobutyric acid	1%	70 (0.5 min)

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); ZIC-HILIC, ZIC-cHILIC (SeQuant Merck, Darmstadt, Germany)

Experimental conditions for MS acquisition using a TripleTOF 5600 mass spectrometer

A TripleTOF5600 mass spectrometer (Sciex, Concord, ON) in standard MS/MS mode was used for data-dependent acquisition - settings used were: 250 ms survey MS spectra (m/z 375-1250) followed by up to 20 MS/MS measurements on the most intense parent ions (400 counts/second threshold for charged states between +2 through +5, m/z 100-1600 mass range for MS/MS, and 100 ms each). Previously targeted parent ions were excluded for 12 seconds from repetitive MS/MS acquisition.

3.10 References

1. Neale, Q.; Prefontaine, A.; Battellino, T.; Mizero, B.; Yeung, D.; Spicer, V.; Budisa, N.; Perreault, H.; Zahedi, R. P.; Krokhin, O. V. Compendium of Chromatographic Behavior of Post-Translationally and Chemically Modified Peptides in Bottom-Up Proteomic Experiments. *Anal Chem* **2023**, *95* (39), 14634–14642. https://doi.org/10.1021/acs.analchem.3c02412.
2. Patel, K.; Borchardt, R. T. Chemical Pathways of Peptide Degradation. II. Kinetics of Deamidation of an Asparaginyl Residue in a Model Hexapeptide. *Pharmaceutical Research: An Official Journal of the American Association of Pharmaceutical Scientists* **1990**, *7* (7), 703–711. https://doi.org/10.1023/A:1015807303766.
3. Patel, K.; Borchardt, R. T. Chemical Pathways of Peptide Degradation. III. Effect of Primary Sequence on the Pathways of Deamidation of Asparaginyl Residues in Hexapeptides. *Pharm Res* **1990**, *07* (8), 787–793. https://doi.org/10.1023/A:1015999012852.
4. Geiger, T.; Clarke, S. Deamidation, Isomerization, and Racemization at Asparaginyl and Aspartyl Residues in Peptides. Succinimide-Linked Reactions That Contribute to Protein Degradation. *Journal of Biological Chemistry* **1987**, *262* (2), 785–794. https://doi.org/10.1016/S0021-9258(19)75855-4.
5. Capasso, S.; Mazzarella, L.; Sica, F.; Zagari, A. Deamidation via Cyclic Imide in Asparaginyl Peptides. *Pept Res* **1989**, *2* (2), 195–200.
6. Capasso, S.; Mazzarella, L.; Sica, F.; Zagari, A. First Evidence of Spontaneous Deamidation of Glutamine Residue via Cyclic Imide to α- and γ-Glutamic Residue under Physiological Conditions. *J Chem Soc Chem Commun* **1991**, No. 23, 1667–1668. https://doi.org/10.1039/C39910001667.
7. Capasso, S.; Mazzarella, L.; Sica, F.; Zagari, A.; Salvadori, S. Kinetics and Mechanism of Succinimide Ring Formation in the Deamidation Process of Asparagine Residues. *Journal of the Chemical Society, Perkin Transactions 2* **1993**, No. 4, 679. https://doi.org/10.1039/p29930000679.
8. Robinson, A. B.; Rudd, C. J. Deamidation of Glutaminyl and Asparaginyl Residues in Peptides and Proteins; 1974; pp 247–295. https://doi.org/10.1016/B978-0-12-152808-9.50013-4.

9. Riggs, D. L.; Silzel, J. W.; Lyon, Y. A.; Kang, A. S.; Julian, R. R. Analysis of Glutamine Deamidation: Products, Pathways, and Kinetics. *Anal Chem* **2019**, *91* (20), 13032–13038. https://doi.org/10.1021/acs.analchem.9b03127.
10. Robinson, N. E.; Robinson, A. B. Molecular Clocks. *Proceedings of the National Academy of Sciences* **2001**, *98* (3), 944–949. https://doi.org/10.1073/PNAS.98.3.944.
11. Wright, H. T. Sequence and Structure Determinants of the Nonenzymatic Deamidation of Asparagine and Glutamine Residues in Proteins. *Protein Engineering, Design and Selection* **1991**, *4* (3), 283–294. https://doi.org/10.1093/PROTEIN/4.3.283.
12. Robinson, N. E. Protein Deamidation. *Proceedings of the National Academy of Sciences* **2002**, *99* (8), 5283–5288. https://doi.org/10.1073/pnas.082102799.
13. McKerrow, J. H.; Robinson, A. B. Deamidation of Asparaginyl Residues as a Hazard in Experimental Protein and Peptide Procedures. *Anal Biochem* **1971**, *42* (2), 565–568. https://doi.org/10.1016/0003-2697(71)90074-1.
14. Tyler-Cross, R.; Schirch, V. Effects of Amino Acid Sequence, Buffers, and Ionic Strength on the Rate and Mechanism of Deamidation of Asparagine Residues in Small Peptides. *Journal of Biological Chemistry* **1991**, *266* (33), 22549–22556. https://doi.org/10.1016/S0021-9258(18)54607-X.
15. Pace, A. L.; Wong, R. L.; Zhang, Y. T.; Kao, Y. H.; Wang, Y. J. Asparagine Deamidation Dependence on Buffer Type, PH, and Temperature. *J Pharm Sci* **2013**, *102* (6), 1712–1723. https://doi.org/10.1002/jps.23529.
16. Krokhin, O. V.; Antonovici, M.; Ens, W.; Wilkins, J. A.; Standing, K. G. Deamidation of -Asn-Gly- Sequences during Sample Preparation for Proteomics: Consequences for MALDI and HPLC-MALDI Analysis. *Anal Chem* **2006**, *78* (18), 6645–6650. https://doi.org/10.1021/ac061017o.
17. Hao, P.; Ren, Y.; Datta, A.; Tam, J. P.; Sze, S. K. Evaluation of the Effect of Trypsin Digestion Buffers on Artificial Deamidation. *J Proteome Res* **2015**, *14* (2), 1308–1314. https://doi.org/10.1021/pr500903b.
18. Robinson, A. B.; McKerrow, J. H.; Cary, P. Controlled Deamidation of Peptides and Proteins: An Experimental Hazard and a Possible Biological Timer. *Proceedings of the National Academy of Sciences* **1970**, *66* (3), 753–757. https://doi.org/10.1073/pnas.66.3.753.

19. Robinson, A. B. Molecular Clocks, Molecular Profiles, and Optimum Diets: Three Approaches to the Problem of Aging. *Mech Ageing Dev* **1979**, *9* (3–4), 225–236. https://doi.org/10.1016/0047-6374(79)90101-5.
20. Gervais, D. Protein Deamidation in Biopharmaceutical Manufacture: Understanding, Control and Impact. *Journal of Chemical Technology and Biotechnology* **2016**, *91* (3), 569–575. https://doi.org/10.1002/JCTB.4850.
21. Gupta, S.; Jiskoot, W.; Schöneich, C.; Rathore, A. S. Oxidation and Deamidation of Monoclonal Antibody Products: Potential Impact on Stability, Biological Activity, and Efficacy. *J Pharm Sci* **2022**, *111* (4), 903–918. https://doi.org/10.1016/J.XPHS.2021.11.024.
22. Alpert, A. J. Electrostatic Repulsion Hydrophilic Interaction Chromatography for Isocratic Separation of Charged Solutes and Selective Isolation of Phosphopeptides. *Anal Chem* **2008**, *80* (1), 62–76. https://doi.org/10.1021/ac070997p.
23. Hao, P.; Guo, T.; Li, X.; Adav, S. S.; Yang, J.; Wei, M.; Sze, S. K. Novel Application of Electrostatic Repulsion-Hydrophilic Interaction Chromatography (ERLIC) in Shotgun Proteomics: Comprehensive Profiling of Rat Kidney Proteome. *J Proteome Res* **2010**, *9* (7), 3520–3526. https://doi.org/10.1021/pr100037h.
24. Hao, P.; Qian, J.; Dutta, B.; Cheow, E. S. H.; Sim, K. H.; Meng, W.; Adav, S. S.; Alpert, A.; Sze, S. K. Enhanced Separation and Characterization of Deamidated Peptides with RP-ERLIC-Based Multidimensional Chromatography Coupled with Tandem Mass Spectrometry. *J Proteome Res* **2012**, *11* (3), 1804–1811. https://doi.org/10.1021/PR201048C.
25. Badgett, M. J.; Boyes, B.; Orlando, R. The Separation and Quantitation of Peptides with and without Oxidation of Methionine and Deamidation of Asparagine Using Hydrophilic Interaction Liquid Chromatography with Mass Spectrometry (HILIC-MS). *J Am Soc Mass Spectrom* **2017**, *28* (5), 818–826. https://doi.org/10.1007/s13361-016-1565-z.
26. Bults, P.; Bischoff, R.; Bakker, H.; Gietema, J. A.; van de Merbel, N. C. LC-MS/MS-Based Monitoring of *In Vivo* Protein Biotransformation: Quantitative Determination of Trastuzumab and Its Deamidation Products in Human Plasma. *Anal Chem* **2016**, *88* (3), 1871–1877. https://doi.org/10.1021/acs.analchem.5b04276.
27. Sandra, K.; Vandenheede, I.; Sandra, P. Modern Chromatographic and Mass Spectrometric Techniques for Protein Biopharmaceutical Characterization. *J Chromatogr A* **2014**, *1335*, 81–103. https://doi.org/10.1016/j.chroma.2013.11.057.

28. 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.
29. Yeung, D.; Klaassen, N.; Mizero, B.; Spicer, V.; Krokhin, O. V. Peptide Retention Time Prediction in Hydrophilic Interaction Liquid Chromatography: Zwitter-Ionic Sulfoalkylbetaine and Phosphorylcholine Stationary Phases. *J Chromatogr A* **2020**, *1619*. https://doi.org/10.1016/j.chroma.2020.460909.
30. Gussakovsky, D.; Anderson, G.; Spicer, V.; Krokhin, O. V. Peptide Separation Selectivity in Proteomics LC-MS Experiments: Comparison of Formic and Mixed Formic/Heptafluorobutyric Acids Ion-Pairing Modifiers. *J Sep Sci* **2020**, *43* (20), 3830–3839. https://doi.org/10.1002/jssc.202000578.
31. 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.
32. 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. https://doi.org/10.1016/j.chroma.2017.12.046.
33. 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.
34. 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* (18), 7036–7042. https://doi.org/10.1021/ac800984n.
35. Wiśniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal Sample Preparation Method for Proteome Analysis. *Nat Methods* **2009**, *6* (5), 359–362. https://doi.org/10.1038/nmeth.1322.

36. 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. https://doi.org/10.1021/ac9016693.
37. Nepomuceno, A. I.; Gibson, R. J.; Randall, S. M.; Muddiman, D. C. Accurate Identification of Deamidated Peptides in Global Proteomics Using a Quadrupole Orbitrap Mass Spectrometer. *J Proteome Res* **2014**, *13* (2), 777–785. https://doi.org/10.1021/pr400848n.
38. 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.
39. Blondelle, S. E.; Ostresh, J. M.; Houghten, R. A.; Pérez-Payá, E. Induced Conformational States of Amphipathic Peptides in Aqueous/Lipid Environments. *Biophys J* **1995**, *68* (1), 351–359. https://doi.org/10.1016/S0006-3495(95)80194-3.
40. Yokoyama, H.; Mizutani, R.; Noguchi, S.; Hayashida, N. Structural and Biochemical Basis of the Formation of Isoaspartate in the Complementarity-Determining Region of Antibody 64M-5 Fab. *Sci Rep* **2019**, *9* (1), 18494. https://doi.org/10.1038/s41598-019-54918-0.
41. 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. https://doi.org/10.1021/acs.analchem.6b04544.
42. 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.

4 Chapter 4: Chromatographic Properties of O-HexNAcylated Peptides in Reversed-Phase and Hydrophilic Interaction Liquid Chromatography

4.1 Article Headline and Authors:

Chromatographic Properties of O-Hexnacylated Peptides in Reversed-Phase and Hydrophilic Interaction Liquid Chromatography

Quinn Neale[1,2], Darien Yeung[2,3], Victor Spicer[2], Helene Perreault[1], Oleg Krokhin[2,4]*

[1]Department of Chemistry, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada
[2]Manitoba Centre for Proteomics and Systems Biology, Health Science Centre, Winnipeg, MB, R3E 3P4, Canada
[3]Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, MB, R3E 0J9, Canada
[4]Department of Internal Medicine, University of Manitoba, Winnipeg, MB, R3E 3P4, Canada

To be submitted:
The Journal of Chromatography A
Status: manuscript in preparation
DOI: N/A

4.2 Contribution of Authors

QN performed data collection and analysis for all samples; generating HI_{mod} vs HI_{nonmod} plots; characterizing trends from retention outliers within data outputs; comparisons to larger glycan species; manuscript writing and preparation of figures for publication. DY aided during sample preparation and the generation of new ideas for enrichment methods and adapted procedures. VS performed the computational work. OK and HP were involved in writing and the supervision of students involved.

4.3 Abstract

In this work, the chromatographic behaviour of peptides with simple O-glycosylations (O-GlcNAc and O-GalNAc) was studied in C18 reversed-phase (RP) and in hydrophilic interaction liquid chromatography (HILIC). Instrumentation and protocol setbacks lead to a lack of certainty in the modified site localization and quantification of these modifications in real sample digests. Synthetic peptides were used as an alternative providing extensive collection of modified/non-modified peptide pairs and the information on site localization. Across both O-GlcNAc and O-GalNAc, subtle differences were observed between serine and threonine residues in RP which can be explained by the shielding of a modified amino acid's side chain by the larger sugar moiety. Shielding was also found to occur with nearest neighbours, resulting in more negative retention shifts for peptides with hydrophobic residues surrounding modified sites and smaller retention shifts for those with surrounding hydrophiles. This translated to average retention shifts expressed in % acetonitrile gradient scale of -0.57% for Ser-O-GlcNAc, -0.90% for Thr-O-GlcNAc, -0.59% for Ser-O-GalNAc, and -0.73% ACN for Thr-O-GalNAc. The Properties of O-GlcNAc modified synthetic peptides were also evaluated in HILIC, resulting in positive average retention shifts of +1.95% H_2O for Ser-O-GlcNAc and +1.89% H_2O for Thr-O-GlcNAc. Comparing to previous reports on N-glycosylation, we also show that single monosaccharide modifications produce very similar to the neutral branched polysaccharide retention shifts in RPLC despite the drastic difference in size.

4.4 Introduction

Post translational modifications (PTMs) have garnered a large interest over the last 30 years due to their involvement in the regulation of biological systems. One of such PTMs, O-linked N-acetylglucosamine (O-GlcNAc), is a single sugar moiety which modifies serine and threonine residues and was discovered fortuitously in the early 1980s.[1] Recent analysis of the cellular distribution of O-GlcNAcylated proteins revealed that targets are mainly cytosolic and nuclear, while a smaller proportion are mitochondrial.[2] Consequently, the O-GlcNAc modification has been found to be involved in a variety of biological functions including cellular regulation and chronic diseases.[3] Another attractive feature is the complex interplay and similarities with protein phosphorylation, such as the dynamic cycling and targeting of the same or neighbouring residues.[4] However, contrary to phosphorylation the O-GlcNAc cycling process is only controlled by two enzymes: O-GlcNAc transferase (OGT)[5] and O-GlcNAc hydrolase (OGA).[6] The former adds the modification to proteins using uridine diphosphate-N-acetylglucosamine as a substrate, while the latter hydrolyzes the modification from proteins.

To study O-GlcNAcylation, detection and identification is commonly performed on the peptide level following digestion and LC-MS/MS analysis. These procedures rely on good chromatographic separation to reduce the complexity of samples such that the eluting modified peptides can be selected for MS/MS analysis. In this case, the mass shift corresponding to the mass of the sugar fragment (+203.0794 Da) is used for peptide mass fingerprinting. However, this process is faced with a set of challenges when applied to O-GlcNAc:

i. The low abundance of O-GlcNAc modified peptides within samples results in ion suppression and low detectability.
ii. The presence of an N-acetyl hexosamine (HexNAc) epimer, O-linked N-acetylgalactosamine (O-GalNAc), leads to decreased confidence in correct identification of the sugar. **(Figure 4.1A)**
iii. O-HexNAc modifications are very labile, leading to complications with modification site assignments and requiring application of dedicated acquisition methods.

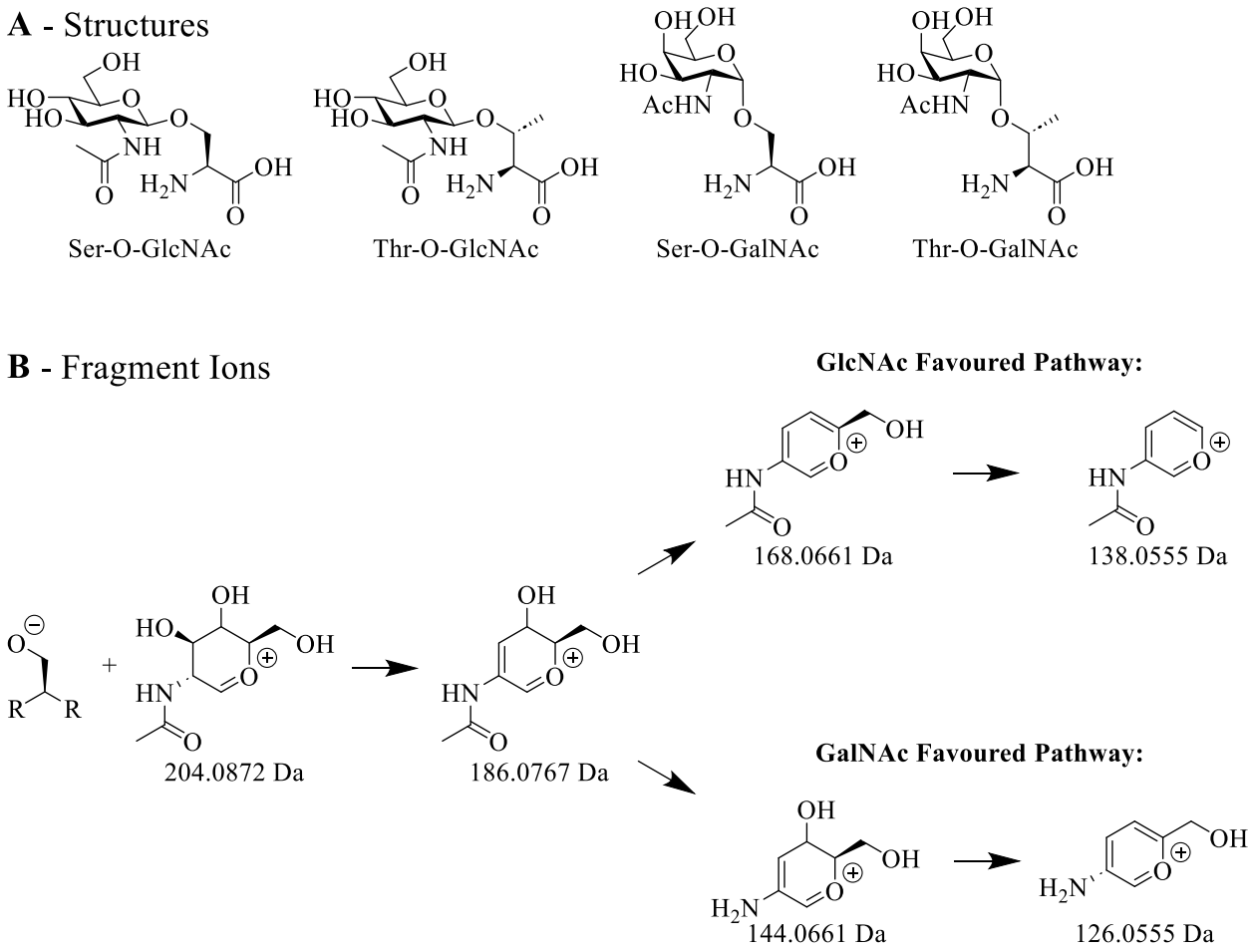


Figure 4.1 A – Structures of serine and threonine residues with O-GlcNAcylation and O-GalNAcylation modifications. B – Fragment ions observed in CID- and HCD-based fragmentation modes.[15]

Detection, differentiation, and site localization of O-GlcNAc and O-GalNAc modified residues are often accomplished using diagnostic fragmentation patterns and supplemented instrumentation. The sugar moiety is often lost during fragmentation in collision induced dissociation (CID) or higher energy collision dissociation (HCD) modes.[7] This results in a mixture of modified and nonmodified species which undergo fragmentation, resulting in a superimposed spectrum of modified, nonmodified, and sugar fragments which can be used for identification.[8,9] While some modified fragments remain on the peptide and permit site localization by mass fingerprinting, significant loss of O-GlcNAc moieties has precluded confident site localization in previous studies[10,11] except for in the case where alternative fragmentation methods are used, such as electron transfer dissociation (ETD).[12] ETD is known to retain modifications during the

fragmentation process[13] and methods were developed accordingly to study O-GlcNAcylation.[14] However, ETD alone cannot differentiate the two epimer sugars, and thus resulting identifications are labelled as O-HexNAc. Differentiation of O-GalNAc from O-GlcNAc can be accomplished, but mainly by comparing the intensities of oxonium fragment ions produced in CID and HCD settings (**Figure 4.1B**).[15] This has led to the generation of EThcD[16] and HCD-pdt-ETD[17] methods which combine the benefits of both HCD and ETD to assign the identity and site of O-HexNAc modified residues. Another consideration has been the presence of N-GlcNAc identifications within analysis results.[14,18–22] These can be omitted by the observation of O-GlcNAc fragments during LC-MS/MS or by consideration of the N-X-S/T motif for N-glycans, where X can be all amino acids except proline.[23]

The low abundance of O-GlcNAc peptides within any sample is a common knowledge in studies surrounding PTMs, thus calling for application of the enrichment protocols. Fahie et al.[24] have summarized methods for detection and analysis of protein O-GlcNAcylation, and many other groups have emphasized alternative enrichment approaches. These include, affinity-based chromatography, immunoprecipitation, metabolic labelling, enzymatic labelling, Beta Elimination followed by Michael Addition of DTT (BEMAD), and other chemical transformation strategies for enrichment. However, these methods have thus far been lackluster, sharing benefits and drawbacks in each of their affinities and specificities (reviewed[25–27]), highlighting that a comprehensive means of enrichment remains to be discovered. Ultimately, the ideal scenario for O-GlcNAc analyses would be a one-pot method where native O-GlcNAc modified peptides can be identified without the need for chemical derivatization, thereby facilitating the application to clinical samples and large-scale proteomics experiments. As such, chromatographic or affinity-based methods would intuitively be the most sensical approach.

Chromatographic attempts at enrichment of native O-GlcNAc structures have largely been approached using lectins, sugar binding proteins, which when bound to material and packed in a column can be used for lectin weak affinity chromatography (LWAC). These applications were initially developed starting from 2006, where enrichment was performed by isocratic elution using wheat-germ agglutinin (WGA) coupled to agarose packed in a 0.1mm x 12m column. This afforded the enrichment and identification of 145 unique O-GlcNAc identifications in the post-synaptic density of murine brains.[28] In 2009, a similar procedure from the same group using a 1mm x 3m column resulted in no visible UV peak for glycosylated peptides. However, analysis of later

fractions permitted the identification of 47 and 87 GlcNAc modified peptides in CID and ETD setups, respectively.[14] In 2012, this procedure was then adapted by the same group to a 2 x 250 mm column, applying three rounds of enrichment and a subsequent high pH fractionation, ultimately leading to the identification of 2434 O-GlcNAc peptides.[18] This dramatic increase was not solely accredited to the LWAC procedure, but also to the use of a more sensitive LC-MS setup, the high-pH separation and lastly the increased enrichment efficiency by coupling WGA to POROS AL resin.[29] These all represent advancements to making lectin enrichment procedures more friendly to proteomics workflows, having been applied to other projects thereafter.[21,30] Another lectin, *Agrocybe aegerita* GlcNAc-specific lectin (AANL), has also been applied to POROS AL and a different packing material Sepharose 4B, showing some similarities as 29 O-HexNAc peptides were identified from mouse liver following enrichment.[31] This method was later improved by generation of an AANL6 mutant which resulted in the improved identification of 79 O-HexNAc peptides.[32] However, more development in this procedure would be required to be comparable to WGA.

Despite these advancements, the information on the retention properties of modified peptides is sparse for conventional proteomics workflows. The retention properties of O-GlcNAc peptides in HILIC have been evaluated where retention time prediction was performed, but this was limited to a subset of synthetic peptides.[33] The reversed-phase (RP) properties have also not been characterized on a large scale despite the numerous applications in the literature. One study, which encompassed numerous modifications from the ProteomeTools project[34] and O-GlcNAc, described the ion-mobility shifts associated with select PTMs while also comparing that to their retention shifts.[35] While an exact numerical value for retention shifts was not provided, it was apparent that the median shift observed across all O-GlcNAc modified and nonmodified peptide pairs was relatively small. Nonetheless, a detailed description of the retention properties for these peptides was not shared.

Our group has previously characterized the RPLC retention shifts of a variety of N-Glycan structures[36] which were later summarized alongside the collection of all PTMs investigated by our group.[37] In this work we targeted the evaluation of chromatographic behaviours of O-GlcNAc and OGalNAc – modified peptides in the most common glycomics chromatographic modes – reversed phase HPLC with formic acid as ion-pairing modifier and HILIC.

4.5 Materials and Methods

4.5.1 Materials and Chromatographic Columns

HILIC buffers were prepared using deionized water, LC-MS grade acetonitrile, and Optima LC/MS ammonium formate (Fisher Scientific, Pittsburgh, PA). RPLC-MS/MS analyses used LC-MS grade water and acetonitrile with 0.1% formic acid (FA). Chemicals were sourced from Millipore Sigma (St. Louis, MO) unless otherwise noted. Sequencing grade modified trypsin (Promega, Madison, WI) was used for digestion. Siliconized 1.5 mL tubes (BioPlas, San Rafael, CA) were used for fraction collection. The custom designed standard peptides P1−P6[38] were synthesized by BioSynthesis Inc. (Lewisville, TX). The nonmodified (SpikeMix™ PTM-Kit 55 – OGlcNAc-OGalNac-unmodified), O-GalNAc (SpikeMix™ PTM-Kit 56 – OGalNAc), and O-GlcNAc (SpikeMix™ PTM-Kit 57 – OGlcNAc) synthetic peptides were sourced from JPT Peptide Technologies GmbH (Berlin, Germany) containing 10 pmol/peptide.

All samples are spiked with the standard peptides P1-P6 before analysis by LC-MS/MS. For 1D RPLC-MS/MS runs, the synthetic peptides were injected without any prior chemical treatment. To limit sample loss for synthetic peptides in 2D chromatographic formats, an in-solution tryptic digest of 10-protein mixture was used as a carrier. 2 mg of the mixture was prepared at equimolar concentrations from bovine serum albumin, chicken serum albumin, human serum albumin, mouse serum albumin, porcine serum albumin, bovine apo-transferrin, human apo-transferrin, human fibrinogen, bovine lactoferrin, and human lactoferrin. In-solution digestion was performed by reduction and alkylation using dithiothreitol (DTT) and iodoacetamide (IAA), respectively, followed by quenching with an additional equivalent of DTT. The pH was adjusted to 8.1 followed by the addition of trypsin for overnight digestion at 37°C. The digestion was quenched after 18 hours using TFA and the resulting peptide mixture was purified RPLC, aliquoted in 50 ug vials, and lyophilized.

4.5.2 Separation Conditions for 1D RPLC-MS/MS Runs

Samples were injected at 300 fmol/peptide for each sample and analyzed using an Easy-nLC 1200 system coupled to an Exploris 480. Samples were separated on an in-house packed Luna C18 (2) column (100 µm x 300 mm, 3 µm, 100Å) using buffer A (0.1% FA in water) and buffer B (0.1% FA in 80:20 acetonitrile/water). The gradient profile followed a linear increase in acetonitrile of 1%B to 45%B over 74 minutes, then a linear 45%B to 90%B over 1 minute, followed by 15

minutes of column washing at 90%B. MS acquisition was in data-dependent acquisition (DDA) mode. Full MS was performed at a resolution of 60,000 with the automatic gain control (AGC) set to 300% and maximum injection time on auto. Precursor ion charge states between 2 and 6 were selected for isolation with a window of 1.6 m/z and fragmented with the normalized collision energy (NCEs) set at 30%. 20 MS/MS spectra were acquired following Full MS at a resolution of 15,000 with AGC set to standard, automated max injection time, and dynamic exclusion set at 14s. 1D RPLC-MS/MS runs for O-GlcNAc and O-GalNAc samples were repeated for NCEs at 25% and 35%.

4.5.3 First Dimension HILIC Separation Conditions for 2D LC-MS/MS Runs

An Agilent 1100 series HPLC with a UV detector (214 nm) and 200 μL injection loop was used for HILIC separations. An XBridge Amide 3 x 50 mm 3.5 μm column (Waters, Milford, MA) was used with a flow rate of 300 μL/min. Buffer A (water) and buffer B (90:10 acetonitrile/water) were prepared 10 mM ammonium formate with pH 4.5. For HILIC fractionation, 1 pmol of each O-GlcNAc and nonmodified synthetic peptides was added to 50 μg of carrier peptides, suspended in 200 μL buffer B. The gradient elution employed a linear increase of 0.7%/min water following 99.9%B to 44.4%B for 71.43 minutes. The eluent system was then ramped to 11.1%B over 1 minute and sustained over 5 minutes for column washing. Sixty 1-minute fractions were collected between 10 and 70 minutes of the gradient profile and lyophilized.

4.5.4 Second Dimension RPLC Separation Conditions for 2D LC-MS/MS Runs

Each sample was analyzed following the same procedure and instrument setup as described in section 2.2 but with changes to injection concentrations and instrument parameters. As most peptides were expected to elute in the first 35 fractions, these fractions were assumed to have more intensity and injected at an estimated 60 fmol/peptide. For fractions 35 to 45 an estimated 120 fmol/peptide was injected. The gradient profile was adjusted for reduced complex sample following a linear increase in acetonitrile of 1%B to 45%B over 24 minutes, then a linear 45%B to 90%B over 1 minute, followed by 15 minutes of column washing at 90%B. Full MS scans were performed at a resolution of 45,000 with the AGC set to 300% and a maximum injection time of 45 ms. Precursor ion charge states between 2 and 5 were selected for MS/MS scans with an isolation window of 1.6 m/z. Fragmentation was performed with the NCE set at 30%. 10 MS/MS

spectra were acquired following Full MS at a resolution of 45,000 with AGC set to 200%, automated max injection time, and dynamic exclusion set at 5s.

4.5.5 Peptide Identification and Data Analysis.

Peptide identification was performed using the X!Tandem search engine[39] mass tolerances set at 10 and 20 ppm for parent and daughter ions, respectively. Carbamidomethyl cysteine (+57.02146 Da) was set as a fixed modification with a standard catalog of spontaneous PTMs and desired modifications set as variable modifications: methionine/tryptophan oxidation (+15.9949 and +31.9898 Da), asparagine/glutamine deamidation (+0.9840 Da), glutamine/cysteine N-terminal cyclization (-17.0265 Da), and serine/threonine O-linked HexNAcylation (+203.0794 Da). The identification output was filtered for peptide entries with logarithmic expectation scores less than -3. Retention times of identified peptides were converted to hydrophobicity index units (HI, %acetonitrile) using the indexed HI values for our standard peptides P1-P6.[38] Nonmodified and glycosylated pairs for a given peptide sequence were generated to determine retention shifts, excluding peptides with multiple modifications.

4.6 Results and Discussion

4.6.1 Identification of O-GlcNAc and O-GalNAc Peptides: Retention Data Acquisition

The first step in addressing the retention properties of GlcNAcylated and GalNAcylated peptides in RPLC was the collection of extensive data on retention of modified/non-modified peptide pairs. Using cell lysate samples was determined to be unsatisfactory for the purposes of amassing a large enough dataset of peptide species. **Table S1** summarizes a variety of experiments which were conducted in an attempt to detect O-GlcNAc-bearing peptides from real peptide digests. These experiments comprise 1-dimensional and 2-dimensional chromatography experiments, alongside enrichment protocols which have been reported previously.[40,41] As a result of applying HCD fragmentation without discriminated between GlcNac and GalNAc, the identified glycopeptides were all labelled O-HexNAc without site localization. Ultimately, the 2-dimensional (2D) setup employing high pH fractionation followed by low pH RPLC-MS analyses produced the highest number of identifications. Unfortunately, these were still insufficient for detailed analysis of the retention features.

Subsequently, the 100 synthetic JPT SpikeMix™ PTM-Kit peptides were chosen as the model sample for the remainder of this project. These collections consist of 100 component mixtures of synthetic peptides in non-modified, GlcNAc modified Ser/Thr, and GalNAc modified Ser/Thr forms: three mixtures in total having identical sequences, but different modification status. Solid phase peptide synthesis can produce low concentrations of side products (deletion of amino acids) that can still be detected and identified by MS approaches. This benefitted our study and increased the number of potential peptide species for the analysis of retention properties. Moreover, since synthetic peptides have known positions of modified residues, this permitted the detection and identification of GlcNAcylated and GalNAcylated peptides with known modified positions using HCD fragmentation.

4.6.2 Chromatographic Behavior of O-GlcNAc and O-GalNAc Modified Peptides in RPLC

Studying peptide separation mechanisms through the development of retention time prediction models is a long-standing research project in our laboratory summarized in the Sequence-Specific Retention Calculator (SSRCalc, https://proteome.ad.umanitoba.ca/ssrcalc) family. It relies on the large-scale retention data collection for each separation mode using proteomics technologies. Datasets of 25,000 peptides and more are usually used – sufficient for

detailed evaluation of retention contributions of all 20 natural amino acids, independently of their abundance. When it comes to PTMs, the modified residue's properties can be evaluated when a comparably sized retention dataset of modified peptides is available. The new modified residue (#21) can then be encoded in the model. This approach is viable only for the most abundant modifications amenable to confident assignment of modification sites: methionine oxidation[42], chemical modifications such as acetylation or labeling with iTRAQ/TMT mass tags.[43–45]

For most endogenous modifications, the yield of modified peptide identifications can be low due to the naturally low abundance of PTMs, leading to insufficiently sized datasets for higher level retention modelling. In this case, our analysis pipeline relies on the generation of modified and nonmodified peptide pairs sharing the same sequence of amino acids, differing only by the modification of interest. This results in a coarse prediction model of expected retention time shift as a function of its nonmodified hydrophobicity (reviewed[37]). This approach is usually visualized by plotting the retention of modified peptides (Y-axis) against their nonmodified counterparts (X-axis), generating plots such as in **Figure S4.1**, **Figure 4.2**, and **Figure 4.3**. Utilizing such dependencies also allows for filtering of extreme chromatographic outliers which can still be present in an automatically generated dataset. The coarse prediction is performed using the slopes and intercepts of the regressions in these cases if the relationship is found to have a sufficiently high R^2 value.

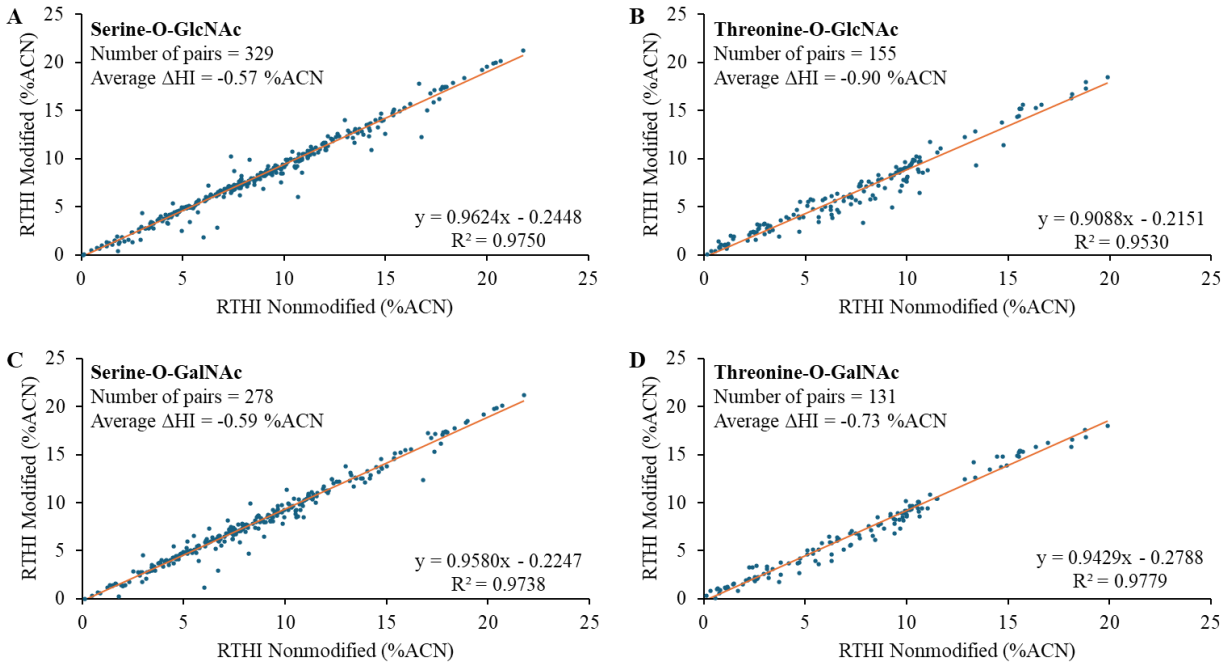


Figure 4.2 Plotted HI mapped retention times (RTHI) of modified (HexNAcylated) versus nonmodified peptide sequence pairs for serine and threonine residues in reversed phase applications. Numbers of pairs and average shifts are reported for the peptide entries within each plot. A – Serine O-GlcNAc. B – Threonine-O-GlcNAc. C – Serine-O-GalNAc. D – Threonine-O-GalNAc.

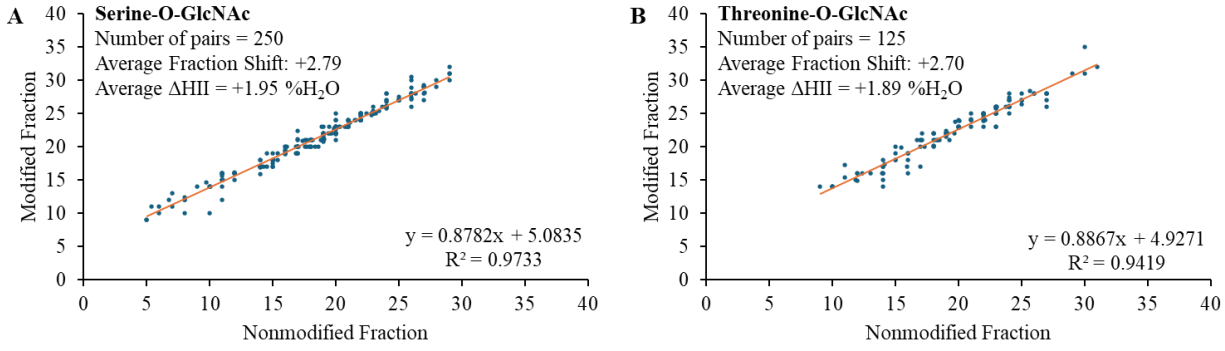


Figure 4.3 Plotted fractions of modified (HexNAcylated) versus nonmodified peptide sequence pairs for serine and threonine residues in HILIC applications. Numbers of pairs and average shifts are reported for the peptide entries within each plot. A – Serine O-GlcNAc. B – Threonine-O-GlcNAc.

The slopes, intercepts, and average retention shifts for each RPLC experiment on JPT peptides in this study are shown in **Table 4.1** where the chromatographic properties of modified and nonmodified pairs were found to be mostly consistent between different LC-MS runs with

different NCE settings. As each of the NCE experiments yielded slightly different sets of pairs for retention analysis, these outputs were combined for each modification then filtered for redundant entries to generate a dataset of unique entries. As the modification sites for the synthetic peptides are known, we were able to discriminate between Ser and Thr modifications for both GlcNAc and GalNAc, generating 4 unique correlation plots shown in **Figure 4.2**. Ultimately, this afforded 329 pairs for Ser-O-GlcNAc, 155 pairs for Thr-O-GlcNAc, 278 pairs for Ser-O-GalNAc, and 131 pairs for Thr-O-GalNAc modifications. Comparing between Ser-O-GlcNAc and Ser-O-GalNAc, the average shift when a serine is modified with either sugar is around -0.58% ACN, reflecting that these modifications show virtually identical behavior in RPLC systems. In the case of threonine, the average retention shifts were larger than serine but were not consistent between Thr-O-GlcNAc and Thr-O-GalNAc, showing approximately -0.90% ACN and -0.73% ACN retention shifts, respectively. These differences in the case of threonine can likely be attributed to the varying number of confidently identified peptide pairs between both sets. Overall, the relatively small retention shift afforded from an O-HexNAc modifications observed on synthetic peptides was consistent with our analyses of real digests with 97 modified-nonmodified pairs, which presented an average retention shift of -0.47% ACN (**Figure S4.1**).

Table 4.1 Pair identification information for all runs comprising reversed-phase retention characteristics for serine and threonine O-HexNAcylation. Three normalized collision energies (NCEs) were evaluated for both GlcNAc and GalNAc sets. The average shifts and corresponding standard deviations of ΔHI values are reported with the slopes, intercepts, R^2, and root mean square error (RMSE) values of modified vs nonmodified regressions.*

NCE	Modification	Pairs	Average Shift (ΔHI, % ACN)	Standard Deviation ΔHI	Slope	Intercept	R^2	RMSE
25	S-O-GlcNAc	187	-0.52	0.8232	0.9646	-0.1951	0.9616	0.8075
	T-O-GlcNAc	100	-0.86	0.8019	0.9270	-0.2888	0.9733	0.7206
30	S-O-GlcNAc	247	-0.54	0.5910	0.9639	-0.2351	0.9822	0.5682
	T-O-GlcNAc	104	-0.87	0.9737	0.9196	-0.2682	0.9515	0.9035
35	S-O-GlcNAc	230	-0.53	0.5381	0.9755	-0.3192	0.9855	0.5257
	T-O-GlcNAc	104	-0.83	0.7928	0.9148	-0.1807	0.9734	0.6874
25	S-O-GalNAc	165	-0.55	0.7691	0.9592	-0.1855	0.9696	0.7455
	T-O-GalNAc	84	-0.71	0.6770	0.9360	-0.1952	0.9844	0.5912

30	S-O-GalNAc	201	-0.55	0.6254	0.9629	-0.2273	0.9814	0.6008
	T-O-GalNAc	98	-0.78	0.7261	0.9398	-0.3005	0.9792	0.6613
35	S-O-GalNAc	196	-0.57	0.7374	0.9537	-0.1841	0.9705	0.7085
	T-O-GalNAc	91	-0.75	0.7227	0.9441	-0.2941	0.9780	0.6685

*An example of modified vs nonmodified regressions can be found in **Figure 4.2**.

Given that each O-HexNAc modification is comprised of a single moiety which is conventionally considered hydrophilic, it was unsurprising to find that the average shifts determined from each experiment were negative and relatively low in reversed-phase measurements. It was also unsurprising to find that the retention properties of O-GlcNAc and O-GalNAc are similar since they share a common structure, differing only as diastereomers (**Figure 4.1A**). In all cases, modified and nonmodified threonine pairs tend to have larger amplitude of negative retention shifts than serine despite their similar structure. We speculate that the difference in retention shifts is a result of shielding on the additional methylene group of the more hydrophobic threonine. When threonine residues are modified, their effective hydrophobicity is reduced to a greater extent due to the large hydrophilic sugar moiety reducing the ability of threonine to interact with the C18 sorbent.

4.6.3 Chromatographic Behavior of O-GlcNAc and O-GalNAc Modified Peptides in HILIC

HILIC fractionation of O-GlcNAc / non-modified synthetic peptides mixtures have been used to elucidate the effect of this modification in an alternative separation mode to RPLC. This part of the study was also used to confirm (disconfirm) the widely spread belief that HILIC separation can be used for selective enrichment of glycosylated peptides[46], including O-HexNAc.[47]

The average retention shifts for serine and threonine O-GlcNAc in HILIC was found to be +2.79 and +2.70 fractions, or in terms of eluent composition, +1.95% and +1.89% water, respectively. The plots representing modified vs. nonmodified fractions can be found in **Figure 4.3**. The broad elution ranges shown in **Figure 4.3** suggest that the selective enrichment procedures, which use pipette tips and cartridges, are likely losing the more hydrophobic peptides which would be represented in our early fractions. Moreover, contrary to what was observed in RP, the difference between serine and threonine retention shifts was found to be insignificant. This finding attests that only hydrophilic groups are the major contributors of peptide retention on

HILIC stationary phases. Thus, the hydroxyl groups on both Ser and Thr are being shielded in the same degree upon modification with GlcNAc.

Our laboratory has previously modelled retention in HILIC which ultimately determined the contribution of each amino acid similarly to our RP model.[48,49] The retention coefficients for serine and threonine were determined to be 4.57 and 3.25 in HILIC, compared to 1.75 and 4.25 in RP, respectively. This represents a larger difference in retention contribution for RP than HILIC, which could explain why the O-GlcNAc modification, while having a relatively large retention shift, does not discriminate greatly between serine and threonine residues in HILIC. A previous study on predicting glycosylated peptide retention in HILIC by Badgett et al. showed that there were differences between serine and threonine where modified threonine had a slightly higher retention[33], as would be expected based on our shielding explanation earlier. In this case, the retention times of single modified and nonmodified peptide pairs in HILIC were assigned using MS. In our case, using fractions based on the second dimension MS/MS identification with 1-minute fraction window likely does not provide sufficient resolution to account for delineations between serine and threonine in HILIC. Despite the lack of difference between serine and threonine residues in our case, sequence-specific properties were generally consistent between HILIC and RP and will be discussed in the next section.

4.6.4 Sequence-Specific Features Driving Variations of the Effect of O-HexNAc Modifications.

While analyzing the outliers from the regressions in **Figure 4.2**, a common trend was that negative outliers generally had hydrophobic resides surrounding the modification site, while positive outliers generally had hydrophilic residues surrounding the modified residue. This observation is also supported by our assumption on shielding presented earlier. Due to the large size of HexNAc moiety, it may play a more dominant role in peptide retention on reversed-phase columns. The retention shielding in this case could influence the entire peptide to some extent. This would explain the slopes below 1 representing larger retention shifts for peptides which are more hydrophobic. Intuitively, this could also explain the outliers as this effect should be more pronounced for the nearest amino acids to the modified site. Therefore, peptides carrying hydrophobic residues near the modified site would exhibit larger retention shifts than peptides carrying hydrophilic residues. These observations are further supported by the results of our HILIC

separation. The largest deviations in the regressions from **Figure 4.3** presented negative outliers with hydrophilic amino acids near the modified residues, while positive outliers had hydrophobic amino acids near the modification site. This reflects a similar mechanism in HILIC that the sugar moiety in glycosylated peptides could be shielding the hydrophobicity of hydrophobic groups, thereby increasing retention, while the opposite is true for hydrophilic amino acids.

The last class of outliers in our sample were helical peptides. In previous studies, we have discussed the effects of amphipathic helicity on retention time predictions. These peptides can be characterized by formation of the structure where one face of the helix barrel carries hydrophilic residues, while the other carries hydrophobic residues.[50] The stability of the helical structures is further increased by a specific N-cap motif prior to the helix.[51] In RP, the hydrophobic face maximizes its interaction with the C18 surface thereby resulting in anomalously large retention compared to predicted values.[51] Our group has previously observed modifications which are assumed to alter helical propensity and thus retention times [Journal of Chromatography A under review]. A previous report has also shown that O-GlcNAc can stabilize a helical structure through interaction with a neighbouring threonine group, but that these interactions are highly depending on peptide primary structure.[52] This led to the assumption that if the O-HexNAc falls in the hydrophobic face of an amphipathic helix, it will dramatically decrease peptide retention. Previously we observed this feature for methionine oxidation[42] where retention drop upon oxidation can be down to -9% ACN, compared to average value of ~-2.4%. Applying similar logic, a modification falling in the hydrophilic face would result in positive retention shifts. This effect could explain one outlier observed in our RP studies for the peptide SPVEES(-O-HexNAc)EDVSNK. In this case, the modification falls in the hydrophobic face of the helix additionally stabilized by SPVE N-cap motif. This resulted in the large negative shift of -3.85 %ACN for O-GlcNAc and -3.72 %ACN for O-GalNAc. Unfortunately, an instance of an amphipathic helical peptide carrying the modification in the hydrophilic face of the peptide was not observed in this study. The effect of peptide amphipathic helicity in HILIC was previously reported to be less profound compared to RPLC.[48] Therefore, this explains why significant outliers in HILIC mode were not observed.

4.6.5 The EFFECT of O-HexNAc Glycosylations vs. N-Glycosylation in RPLC

As we have shown in a previous study, glycosylations in RP appear to have distinct properties compared to other hydrophilic modifications in terms of how they affect chromatographic properties.[37] In particular, the slopes of their modified vs nonmodified regressions are below 1 (**Figure 4.4**) compared to other compact hydrophilic modifications such as oxidation. In addition to it, their average retention shifts are very similar despite the diversity in sugar composition (**Figure 4.5**). The major deviations in these cases occur where the charge of the peptide is altered, such is the case for glycans carrying sialic acids. The previous explanation on shielding of hydrophobic interactions can also be applied to explain the slopes of the larger N-glycans, where a similar effect should occur. However, what is striking is that when comparing the simple O-HexNAc with the larger N-Glycans, their retention shifts are very similar despite the fact that their compositions can differ by up to 10 monosaccharide subunits. This suggests that for neutral glycosylations, elongation and branching does not drastically alter their retention behaviour, and increasing the number of sugar moieties results in a declining effect on the magnitude of retention shifts in RPLC.

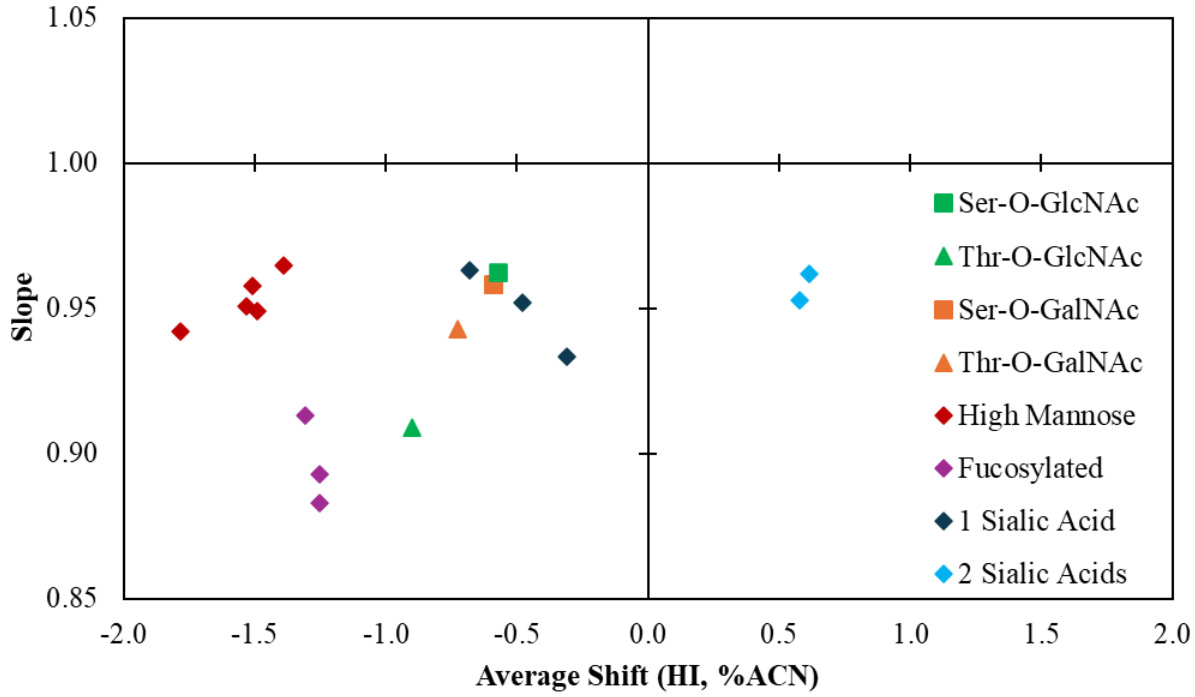


Figure 4.4 Slopes vs Intercepts for various glycosylations from previous studies.[36,37] High Mannose, Fucosylated, 1 Sialic Acids, and 2 Sialic Acid groups are clustered with the same colours and shapes to facilitated visualization.

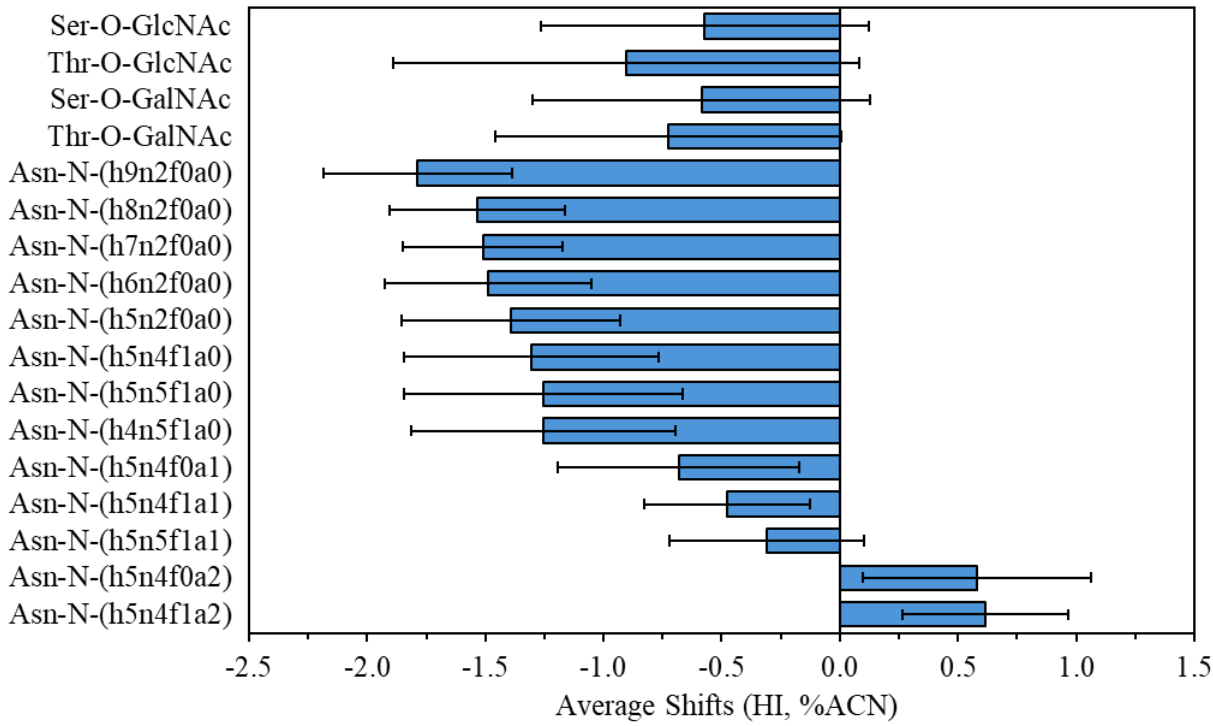


Figure 4.5 Retention shifts of various glycosylations from our datasets and previous publications.[36,37] Average shifts are reported in HI values with the error bars representing one standard deviation in each direction. Each N-Glycan group is represented by a reduced name: h = hexose, n = hexosamine, f = fucose, and a = sialic acid.

4.7 Conclusion

This study represents the first detail investigation of chromatographic properties of O-HexNAc (Ser/Thr) modified peptides using large scale datasets. We have applied two distinct approaches for measuring retention shifts for modified peptides: endogenous peptide enrichment from tryptic digests and using synthetic peptide libraries. The former showed rather disappointing results with the largest number of identified peptide pairs under one hundred using advanced 2D LC-MS/MS. Another disadvantage was found in poor confidence of modification site assignment using HCD fragmentation. Nevertheless, we find both methods in good agreement on measuring retention differences between modified and non-modified peptides in RPLC with formic acid as ion-pairing modifier. O-GlcNAc and O-GalNAc glycosylated peptides in RP were found to have relatively small retention shifts compared to their non-modified analogs despite their large size: ~-0.58 and ~-0.82 for Ser and Thr, respectively. On the other hand, we find that the addition of a single HexNAc moiety result in the hydrophobicity decrease very similar to those of larger N-Glycans. In one representative case, the average magnitude of the retention shifts for the large high mannose structure was only twice as large compared to O-GlcNAc with approximately $1/11^{th}$ of the number of monosaccharide units: Asn-N-h9n2f0a0 with a -1.79 %ACN shift vs. Thr-O-GlcNAc with a -0.90 %ACN in **Figure 4.5**. Moreover, the retention shifts of glycosylations increased with the magnitude of a peptide's innate hydrophobicity, reflecting a unique characteristic for these PTMs on peptide retention properties. Ultimately, we speculated that this is the result of a shielding effect where the sugar moiety blocks the interactions between a peptide and the surface of the sorbent. This was supported by our observations that modified sites with surrounding hydrophobic residues possessed a larger negative retention shifts than peptides with surrounding hydrophilic residues.

Previous reports were able to identify differences between serine and threonine residues in HILIC[33], but these delineations were not observed in our HILIC results. In our case, these could likely be limited by the resolution of our fractionation given that the difference in hydrophobicity between serine and threonine is even smaller in HILIC than RPLC.[48] The sequence-specific effects observed in RPLC were also found in HILIC but bearing opposite consequences for the observed retention features. In this case, we observed that the peptides with more negative retention shifts feature hydrophilic amino acids surrounding the modifies residues, while the positive outliers had hydrophobic amino acids. This supports our assumption of shielding, but to fully characterize the

extent of this effect would require a larger dataset of peptides and synthetic variants for validations. Thus, future studies should focus on generating quality and larger datasets with site localization for advanced retention modelling.

4.8 Acknowledgements

This work was supported by grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2016-05963 and RGPIN-2022-05015 O.V.K).

4.9 Supplementary Information

Table S4.1 Summary of O-HexNAx identifications and pairs across various experiments*

Analysis Format	Treatment	# of O-HexNAc IDs	# pairs
1D RPLC-MS/MS	Untreated	6	1
2D LC-MS/MS high pH-low pH RP concatenated fractions	Untreated[1]	210	97
1D RPLC-MS/MS	OGT Labelling[2]	44	21
1D RPLC-MS/MS	Wang et al. PBA Enrichment[3]	30	NA
1D RPLC-MS/MS	OGT Labelling & Wang et al. PBA Enrichment[2,3]	103	42
1D RPLC-MS/MS	Potel et al. PBA Enrichment[4]	1	0
1D RPLC-MS/MS	Adapted Wang et al. PBA Enrichment[3]	15	1

*Detailed descriptions of each treatment method were provided in the Supplementary Experimental Procedures section.

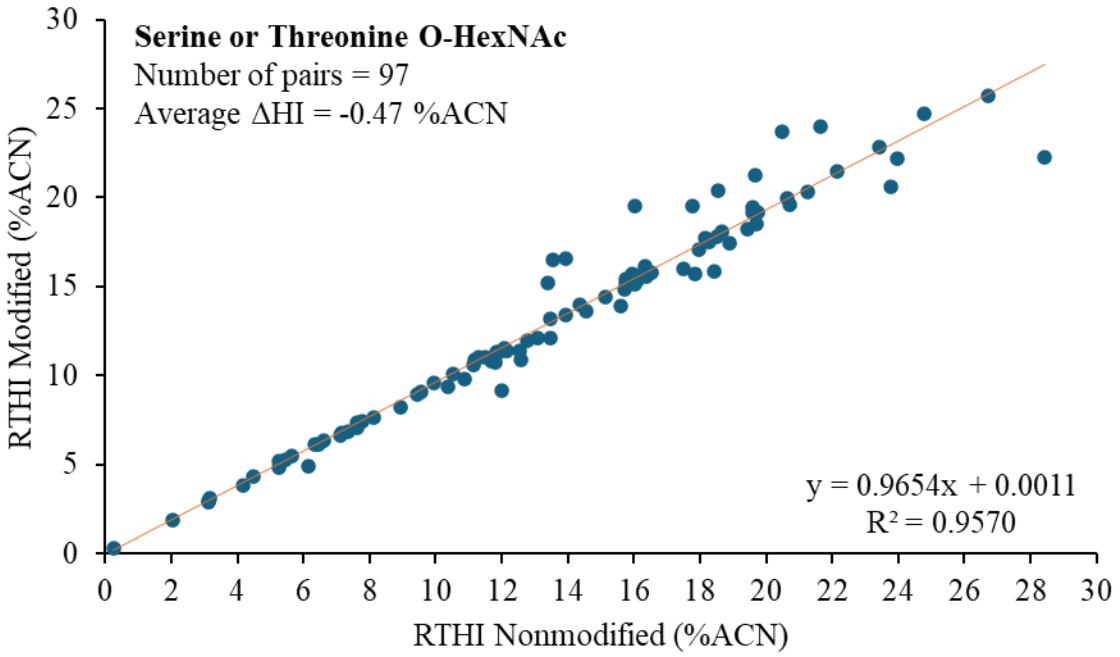


Figure S4.1 Plot of modified versus nonmodified O-HexNAc peptides from a 2D high pH – low pH RP experiment.

Supplementary Experimental Procedures

1D C18 RPFA

All 1D C18 RPFA samples using enrichment procedures were analyzed by LCMS following the 1D RPLC-MS/MS conditions listed in the Materials and Methods section. PNGaseF (Promega) was used when specified to cleave N-glycans from peptide samples following manufacturers specifications.

2D high-pH low-pH reversed phase with sample concatenation

Sample concatenation was performed following a similar procedure as previously reported.[1] Briefly, an Agilent 1100 series HPLC with a UV detector (214 nm), a 100 µL injection loop, and an XTerra 5 µm 100Å (1x100) column was used in the first dimension. Eluent A (water) and eluent B (acetonitrile) were prepared to 20 mM ammonium formate and adjusted to pH 10. A linear acetonitrile gradient was used to separate peptides following an increase of 2%/min eluent B and 1-minute fractions were collected for 20 minutes. The fractions following collection were concatenated (combining fractions 1 + 11, 2 + 12, 3 + 13, etc.) prior to second dimension LCMS analysis employing the same conditions listed under 1D RPLC-MS/MS conditions in the Materials and Methods section.

OGT Labelling

O-GlcNAc Transferase (OGT) labelling adapted from a previous report[2] and was performed in an attempt to artificially increase the number of O-GlcNAcylations. This involved using recombinant human OGT (Biotechne, R&D Systems), Uridine 5′-diphospho-N-acetylglucosamine (UDP-GlcNAc) sodium salt (Millipore Sigma), and peptide digests. 1 µg of OGT was mixed with 10 µg of UDP-GlcNAc and 20 µg of peptides in 25 mM Tris buffer (pH 7.8). The samples were incubated at 37°C for 3 hours, then desalted prior to LCMS analysis.

Wang et al. PBA Enrichment

The procedure adapted from Wang et al.[3] used anhydrous dimethyl sulfoxide (DMSO) (ThermoFisher Scientific), 100 mg 40µm BondElute PBA cartridges (Agilent). Briefly, 1 mg of peptides were treated with PNGase F and suspended in DMSO. This mixture was added to the

PBA cartridge and incubated at 37°C for 2 hours. Nonbound peptides were washed with 100µL of acetonitrile 3 times. Bound peptides were eluted with 0.1% TFA solution after incubation at room temperature for 1 hour. Samples were concentrated in vacuo and prepared for LCMS analysis.

Potel et al. PBA Enrichment

This procedure was adapted from Potel et al.[4] where they used PBA enrichmet in a 96 well plate to investigate glycosylations. This used unpacked 40µm BondElute PBA (Agilent), enrichment buffer (50 mM carbonate buffer, pH 10.5 in 50% acetonitrile), and elution buffer (1% TFA in 50% acetonitrile). Briefly, PNGase F treated peptides were resuspended at 5mg/mL in enrichment buffer. The functionalized PBA beads were washed and sample was added at the 1 : 2.5 w/w ratio (Sample:PBA). Samples were incubated with the beads at room temperature for 1 hour with sufficient agitation to stop sample sedimentation. Samples were washed 7 times with enrichment buffer, centrifuging (15000 x g f 5-mins) between each wash. 50 uL of elution buffer was added and incubated for 15 minutes at room temperature, then collected leaving the beads behind. Elution was repeated 3 times and samples were concentrated in vacuo and prepared for LCMS analysis.

Adapted Wang et al. PBA Enrichment

This procedure adapted the Wang et al. method to similar conditions used in the Potel et al method. This used unpacked 40µm BondElute PBA (Agilent). Briefly, 1 mg of peptides treated with PNGase F weres suspended in DMSO. 2.5 mg of PBA was added for a ratio of 2.5:1 (PBA:Sample) and incubate at 37°C for 2 hours while shaking. Nonbound peptides were washed with acetonitrile 3 times. Sample was centrifuged at 15000 x g for 5 minutes between each wash to pelletize PBA. Peptides were eluted with 150 uL 0.1% TFA solution, incubating for 1 hour at room temperature. Samples were concentrated in vacuo and prepared for LCMS analysis.

Supplementary References

1. 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* (18), 7036–7042. https://doi.org/10.1021/ac800984n.
2. Pathak, S.; Alonso, J.; Schimpl, M.; Rafie, K.; Blair, D. E.; Borodkin, V. S.; Schüttelkopf, A. W.; Albarbarawi, O.; van Aalten, D. M. F. The Active Site of O-GlcNAc Transferase Imposes Constraints on Substrate Sequence. *Nat Struct Mol Biol* **2015**, *22* (9), 744–750. https://doi.org/10.1038/nsmb.3063.
3. Wang, X.; Yuan, Z. F.; Fan, J.; Karch, K. R.; Ball, L. E.; Denu, J. M.; Garcia, B. A. A Novel Quantitative Mass Spectrometry Platform for Determining Protein O-GlcNAcylation Dynamics. *Molecular and Cellular Proteomics* **2016**, *15* (7), 2462–2475. https://doi.org/10.1074/mcp.O115.049627.
4. Potel, C. M.; Burtscher, M. L.; Garrido-Rodriguez, M.; Brauer-Nikonow, A.; Becher, I.; Typas, A.; Zimmermann, M.; Savitski, M. M. Deep Quantitative Glycoproteomics Reveals Gut Microbiome Induced Remodeling of the Brain Glycoproteome. *bioRxiv* **2023**. https://doi.org/10.1101/2023.09.13.557529.

4.10 References

1. Torres, C. R.; Hart, G. W. Topography and Polypeptide Distribution of Terminal N-Acetylglucosamine Residues on the Surfaces of Intact Lymphocytes. Evidence for O-Linked GlcNAc. *J Biol Chem* **1984**, *259* (5), 3308–3317.
2. Wulff-Fuentes, E.; Berendt, R. R.; Massman, L.; Danner, L.; Malard, F.; Vora, J.; Kahsay, R.; Olivier-Van Stichelen, S. The Human O-GlcNAcome Database and Meta-Analysis. *Scientific Data 2021 8:1* **2021**, *8* (1), 1–11. https://doi.org/10.1038/S41597-021-00810-4.
3. Hart, G. W. Nutrient Regulation of Signaling and Transcription. *Journal of Biological Chemistry* **2019**, *294* (7), 2211–2231. https://doi.org/10.1074/jbc.AW119.003226.
4. Hart, G. W.; Slawson, C.; Ramirez-Correa, G.; Lagerlof, O. Cross Talk Between O-GlcNAcylation and Phosphorylation: Roles in Signaling, Transcription, and Chronic Disease. *Annu Rev Biochem* **2011**, *80* (1), 825–858. https://doi.org/10.1146/annurev-biochem-060608-102511.
5. Haltiwanger, R. S.; Holt, G. D.; Hart, G. W. Enzymatic Addition of O-GlcNAc to Nuclear and Cytoplasmic Proteins. Identification of a Uridine Diphospho-N-Acetylglucosamine:Peptide Beta-N-Acetylglucosaminyltransferase. *Journal of Biological Chemistry* **1990**, *265* (5), 2563–2568. https://doi.org/10.1016/S0021-9258(19)39838-2.
6. Dong, D. L.; Hart, G. W. Purification and Characterization of an O-GlcNAc Selective N-Acetyl-Beta-D-Glucosaminidase from Rat Spleen Cytosol. *Journal of Biological Chemistry* **1994**, *269* (30), 19321–19330. https://doi.org/10.1016/S0021-9258(17)32170-1.
7. Ma, J.; Hart, G. W. Analysis of Protein O-GlcNAcylation by Mass Spectrometry. *Curr Protoc Protein Sci* **2017**, *87* (1). https://doi.org/10.1002/cpps.24.
8. Greis, K. D.; Hayes, B. K.; Comer, F. I.; Kirk, M.; Barnes, S.; Lowary, T. L.; Hart, G. W. Selective Detection and Site-Analysis of O-GlcNAc-Modified Glycopeptides by β-Elimination and Tandem Electrospray Mass Spectrometry. *Anal Biochem* **1996**, *234* (1), 38–49. https://doi.org/10.1006/abio.1996.0047.
9. Haynes, P. A.; Aebersold, R. Simultaneous Detection and Identification of O-GlcNAc-Modified Glycoproteins Using Liquid Chromatography−Tandem Mass Spectrometry. *Anal Chem* **2000**, *72* (21), 5402–5410. https://doi.org/10.1021/ac000512w.

10. Chalkley, R. J.; Burlingame, A. L. Identification of GlcNAcylation Sites of Peptides and α-Crystallin Using Q-TOF Mass Spectrometry. *J Am Soc Mass Spectrom* **2001**, *12* (10), 1106–1113. https://doi.org/10.1016/S1044-0305(01)00295-1.
11. Chalkley, R. J.; Burlingame, A. L. Identification of Novel Sites of O-N-Acetylglucosamine Modification of Serum Response Factor Using Quadrupole Time-of-Flight Mass Spectrometry. *Molecular & Cellular Proteomics* **2003**, *2* (3), 182–190. https://doi.org/10.1074/mcp.M300027-MCP200.
12. Myers, S. A.; Daou, S.; Affar, E. B.; Burlingame, A. Electron Transfer Dissociation (ETD): The Mass Spectrometric Breakthrough Essential for *O*-GlcNAc Protein Site Assignments—a Study of the *O*-GlcNAcylated Protein Host Cell Factor C1. *Proteomics* **2013**, *13* (6), 982–991. https://doi.org/10.1002/pmic.201200332.
13. Syka, J. E. P.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and Protein Sequence Analysis by Electron Transfer Dissociation Mass Spectrometry. *Proceedings of the National Academy of Sciences* **2004**, *101* (26), 9528–9533. https://doi.org/10.1073/PNAS.0402700101.
14. Chalkley, R. J.; Thalhammer, A.; Schoepfer, R.; Burlingame, A. L. Identification of Protein O-GlcNAcylation Sites Using Electron Transfer Dissociation Mass Spectrometry on Native Peptides. *Proceedings of the National Academy of Sciences* **2009**, *106* (22), 8894–8899. https://doi.org/10.1073/pnas.0900288106.
15. Halim, A.; Westerlind, U.; Pett, C.; Schorlemer, M.; Rüetschi, U.; Brinkmalm, G.; Sihlbom, C.; Lengqvist, J.; Larson, G.; Nilsson, J. Assignment of Saccharide Identities through Analysis of Oxonium Ion Fragmentation Profiles in LC–MS/MS of Glycopeptides. *J Proteome Res* **2014**, *13* (12), 6024–6032. https://doi.org/10.1021/pr500898r.
16. Yu, Q.; Wang, B.; Chen, Z.; Urabe, G.; Glover, M. S.; Shi, X.; Guo, L.-W.; Kent, K. C.; Li, L. Electron-Transfer/Higher-Energy Collision Dissociation (EThcD)-Enabled Intact Glycopeptide/Glycoproteome Characterization. *J Am Soc Mass Spectrom* **2017**, *28* (9), 1751–1764. https://doi.org/10.1007/s13361-017-1701-4.
17. Zhao, P.; Viner, R.; Teo, C. F.; Boons, G.-J.; Horn, D.; Wells, L. Combining High-Energy C-Trap Dissociation and Electron Transfer Dissociation for Protein O-GlcNAc Modification Site Assignment. *J Proteome Res* **2011**, *10* (9), 4088–4104. https://doi.org/10.1021/pr2002726.

18. Trinidad, J. C.; Barkan, D. T.; Gulledge, B. F.; Thalhammer, A.; Sali, A.; Schoepfer, R.; Burlingame, A. L. Global Identification and Characterization of Both O-GlcNAcylation and Phosphorylation at the Murine Synapse. *Molecular & Cellular Proteomics* **2012**, *11* (8), 215–229. https://doi.org/10.1074/mcp.O112.018366.
19. Trinidad, J. C.; Schoepfer, R.; Burlingame, A. L.; Medzihradszky, K. F. N- and O-Glycosylation in the Murine Synaptosome. *Molecular & Cellular Proteomics* **2013**, *12* (12), 3474–3488. https://doi.org/10.1074/mcp.M113.030007.
20. (20) Kim, Y.-C.; Jahren, N.; Stone, M. D.; Udeshi, N. D.; Markowski, T. W.; Witthuhn, B. A.; Shabanowitz, J.; Hunt, D. F.; Olszewski, N. E. Identification and Origin of *N* -Linked β- <scp>d</Scp> - *N*- Acetylglucosamine Monosaccharide Modifications on Arabidopsis Proteins . *Plant Physiol* **2012**, *161* (1), 455–464. https://doi.org/10.1104/pp.112.208900.
21. Kim, S.; Maynard, J. C.; Sasaki, Y.; Strickland, A.; Sherman, D. L.; Brophy, P. J.; Burlingame, A. L.; Milbrandt, J. Schwann Cell O-GlcNAc Glycosylation Is Required for Myelin Maintenance and Axon Integrity. *The Journal of Neuroscience* **2016**, *36* (37), 9633–9646. https://doi.org/10.1523/JNEUROSCI.1235-16.2016.
22. Malaker, S. A.; Penny, S. A.; Steadman, L. G.; Myers, P. T.; Loke, J. C.; Raghavan, M.; Bai, D. L.; Shabanowitz, J.; Hunt, D. F.; Cobbold, M. Identification of Glycopeptides as Posttranslationally Modified Neoantigens in Leukemia. *Cancer Immunol Res* **2017**, *5* (5), 376–384. https://doi.org/10.1158/2326-6066.CIR-16-0280.
23. Pless, D. D.; Lennarz, W. J. Enzymatic Conversion of Proteins to Glycoproteins. *Proceedings of the National Academy of Sciences* **1977**, *74* (1), 134–138. https://doi.org/10.1073/pnas.74.1.134.
24. Fahie, K.; Narayanan, B.; Zahra, F.; Reeves, R.; Fernandes, S. M.; Hart, G. W.; Zachara, N. E. Detection and Analysis of Proteins Modified by O-Linked *N* -Acetylglucosamine. *Curr Protoc* **2021**, *1* (5). https://doi.org/10.1002/cpz1.129.
25. Maynard, J. C.; Chalkley, R. J. Methods for Enrichment and Assignment of N-Acetylglucosamine Modification Sites. *Molecular & Cellular Proteomics* **2021**, *20*, 100031. https://doi.org/10.1074/mcp.R120.002206.
26. Saha, A.; Bello, D.; Fernández-Tejada, A. Advances in Chemical Probing of Protein: O -GlcNAc Glycosylation: Structural Role and Molecular Mechanisms. *Chemical Society*

Reviews. Royal Society of Chemistry September 21, 2021, pp 10451–10485. https://doi.org/10.1039/d0cs01275k.

27. Dupas, T.; Betus, C.; Blangy-Letheule, A.; Pelé, T.; Persello, A.; Denis, M.; Lauzier, B. An Overview of Tools to Decipher O-GlcNAcylation from Historical Approaches to New Insights. *Int J Biochem Cell Biol* **2022**, *151*, 106289. https://doi.org/10.1016/j.biocel.2022.106289.
28. Vosseller, K.; Trinidad, J. C.; Chalkley, R. J.; Specht, C. G.; Thalhammer, A.; Lynn, A. J.; Snedecor, J. O.; Guan, S.; Medzihradszky, K. F.; Maltby, D. A.; Schoepfer, R.; Burlingame, A. L. O-Linked N-Acetylglucosamine Proteomics of Postsynaptic Density Preparations Using Lectin Weak Affinity Chromatography and Mass Spectrometry. *Molecular & Cellular Proteomics* **2006**, *5* (5), 923–934. https://doi.org/10.1074/mcp.T500040-MCP200.
29. Afeyan, N. B.; Gordon, N. F.; Mazsaroff, I.; Varady, L.; Fulton, S. P.; Yang, Y. B.; Regnier, F. E. Flow-through Particles for the High-Performance Liquid Chromatographic Separation of Biomolecules: Perfusion Chromatography. *J Chromatogr A* **1990**, *519* (1), 1–29. https://doi.org/10.1016/0021-9673(90)85132-F.
30. Maynard, J. C.; Fujihira, H.; Dolgonos, G. E.; Suzuki, T.; Burlingame, A. L. Cytosolic N-GlcNAc Proteins Are Formed by the Action of Endo-β-N-Acetylglucosaminidase. *Biochem Biophys Res Commun* **2020**, *530* (4), 719–724. https://doi.org/10.1016/j.bbrc.2020.06.127.
31. Liu, W.; Han, G.; Yin, Y.; Jiang, S.; Yu, G.; Yang, Q.; Yu, W.; Ye, X.; Su, Y.; Yang, Y.; Hart, G. W.; Sun, H. AANL (Agrocybe Aegerita Lectin 2) Is a New Facile Tool to Probe for O-GlcNAcylation. *Glycobiology* **2018**, *28* (6), 363–373. https://doi.org/10.1093/glycob/cwy029.
32. Su, Y.; Ye, X.; Xu, B.; Li, Y.; Yang, Q.; Yu, W.; Song, J.; Guo, C.; Wang, X.; Hart, G. W.; Sun Hui. CBS Homogenization Mutation Strategy Narrows the Glycan Binding Profile of a GlcNAc-Specific Lectin, AANL. *Glycobiology* **2020**, *30* (3), 159–173. https://doi.org/10.1093/glycob/cwz089.
33. Badgett, M. J.; Boyes, B.; Orlando, R. Predicting the Retention Behavior of Specific O-Linked Glycopeptides. *J Biomol Tech* **2017**, *28* (3), 122–126. https://doi.org/10.7171/jbt.17-2803-003.
34. Paul Zolg, D.; Wilhelm, M.; Schmidt, T.; Médard, G.; Zerweck, J.; Knaute, T.; Wenschuh, H.; Reimer, U.; Schnatbaum, K.; Kuster, B. Proteometools: Systematic Characterization of 21

Post-Translational Protein Modifications by Liquid Chromatography Tandem Mass Spectrometry (Lc-Ms/Ms) Using Synthetic Peptides. *Molecular and Cellular Proteomics* **2018**, *17* (9), 1850–1863. https://doi.org/10.1074/mcp.TIR118.000783.

35. Will, A.; Oliinyk, D.; Bleiholder, C.; Meier, F. Peptide Collision Cross Sections of 22 Post-Translational Modifications. *Anal Bioanal Chem* **2023**, *415* (27), 6633–6645. https://doi.org/10.1007/s00216-023-04957-4.
36. Ang, E.; Neustaeter, H.; Spicer, V.; Perreault, H.; Krokhin, O. V. Retention Time Prediction for Glycopeptides in Reversed Phase Chromatography for Glycoproteomic Applications. *Anal Chem* **2019**. https://doi.org/10.1021/acs.analchem.9b02584.
37. Neale, Q.; Prefontaine, A.; Battellino, T.; Mizero, B.; Yeung, D.; Spicer, V.; Budisa, N.; Perreault, H.; Zahedi, R. P.; Krokhin, O. V. Compendium of Chromatographic Behavior of Post-Translationally and Chemically Modified Peptides in Bottom-Up Proteomic Experiments. *Anal Chem* **2023**, *95* (39), 14634–14642. https://doi.org/10.1021/acs.analchem.3c02412.
38. 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. https://doi.org/10.1021/ac9016693.
39. 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.
40. Wang, X.; Yuan, Z. F.; Fan, J.; Karch, K. R.; Ball, L. E.; Denu, J. M.; Garcia, B. A. A Novel Quantitative Mass Spectrometry Platform for Determining Protein O-GlcNAcylation Dynamics. *Molecular and Cellular Proteomics* **2016**, *15* (7), 2462–2475. https://doi.org/10.1074/mcp.O115.049627.
41. Potel, C. M.; Burtscher, M. L.; Garrido-Rodriguez, M.; Brauer-Nikonow, A.; Becher, I.; Typas, A.; Zimmermann, M.; Savitski, M. M. Deep Quantitative Glycoproteomics Reveals Gut Microbiome Induced Remodeling of the Brain Glycoproteome. *bioRxiv* **2023**. https://doi.org/10.1101/2023.09.13.557529.
42. Lao, Y. W.; Gungormusler-Yilmaz, M.; Shuvo, S.; Verbeke, T.; Spicer, V.; Krokhin, O. V. Chromatographic Behavior of Peptides Containing Oxidized Methionine Residues in

Proteomic LC-MS Experiments: Complex Tale of a Simple Modification. *J Proteomics* **2015**, *125*, 131–139. https://doi.org/10.1016/j.jprot.2015.05.018.

43. Mizero, B.; Yeung, D.; Spicer, V.; Krokhin, O. V. Peptide Retention Time Prediction for Peptides with Post-Translational Modifications: N-Terminal (α-Amine) and Lysine (ε-Amine) Acetylation. *J Chromatogr A* **2021**, *1657*. https://doi.org/10.1016/j.chroma.2021.462584.
44. Yeung, D.; Anderson, G.; Spicer, V.; Krokhin, O. V. Chromatographic Behaviour of Peptides Modified with Amine-Reacting Tags for Relative Protein Quantitation in Proteomic Applications. *J Chromatogr A* **2022**, *1679*. https://doi.org/10.1016/j.chroma.2022.463391.
45. Mizero, B.; Villacrés, 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* (5), 1218–1228. https://doi.org/10.1021/acs.jproteome.1c00833.
46. Selman, M. H. J.; Hemayatkar, M.; Deelder, A. M.; Wuhrer, M. Cotton HILIC SPE Microtips for Microscale Purification and Enrichment of Glycans and Glycopeptides. *Anal Chem* **2011**, *83* (7), 2492–2499. https://doi.org/10.1021/ac1027116.
47. Shen, B.; Zhang, W.; Shi, Z.; Tian, F.; Deng, Y.; Sun, C.; Wang, G.; Qin, W.; Qian, X. A Novel Strategy for Global Mapping of O-GlcNAc Proteins and Peptides Using Selective Enzymatic Deglycosylation, HILIC Enrichment and Mass Spectrometry Identification. *Talanta* **2017**, *169*, 195–202. https://doi.org/10.1016/j.talanta.2017.03.049.
48. 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.
49. 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. https://doi.org/10.1016/j.chroma.2017.12.046.
50. 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.

51. Blondelle, S. E.; Ostresh, J. M.; Houghten, R. A.; Pérez-Payá, E. Induced Conformational States of Amphipathic Peptides in Aqueous/Lipid Environments. *Biophys J* **1995**, *68* (1), 351–359. https://doi.org/10.1016/S0006-3495(95)80194-3.
52. Hashimoto, Y.; Toma, K.; Nishikido, J.; Yamamoto, K.; Haneda, K.; Inazu, T.; Valentine, K. G.; Opella, S. J. Effects of Glycosylation on the Structure and Dynamics of Eel Calcitonin in Micelles and Lipid Bilayers Determined by Nuclear Magnetic Resonance Spectroscopy. *Biochemistry* **1999**, *38* (26), 8377–8384. https://doi.org/10.1021/bi983018j.

5 Chapter 5: Peptide Retention Time Prediction for Electrostatic Repulsion-Hydrophilic Interaction Chromatography

5.1 Article Headline and Authors:

Peptide Retention Time Prediction for Electrostatic Repulsion-Hydrophilic Interaction Chromatography

Quinn Neale[1,2], Darien Yeung[2,3], Victor Spicer[2], Helene Perreault[1], Oleg Krokhin[2,4]*

[1]Department of Chemistry, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada

[2]Manitoba Centre for Proteomics and Systems Biology, Health Science Centre, Winnipeg, MB, R3E 3P4, Canada

[3]Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, MB, R3E 0J9, Canada

[4]Department of Internal Medicine, University of Manitoba, Winnipeg, MB, R3E 3P4, Canada

Published:

The Journal of Chromatography A

November 8th, 2024

Volume 1736

DOI: https://doi.org/10.1016/j.chroma.2024.465414

5.2 Contribution of Authors

The project idea was formed and developed by QN, DY, and OK. Data collection and analysis were performed by QN; sample preparation; evaluation of retention modelling; identification of prediction outliers; validation of ERLIC's assumed retention mechanism based on orientation-specific behaviour from retention coefficients; manuscript writing and preparation of figures for publication. Retention model training was performed by VS. HP and OK were involved in writing and supervision.

5.3 Abstract

Electrostatic Repulsion-Hydrophilic Interaction Chromatography (ERLIC) is one of the legacy separation tools developed by Dr. Andrew Alpert and has been used for developing unique separation methods of hydrophilic compounds, including peptides. In the past it has been studied using designed peptide libraries to elucidate major features of its separation mechanism, while comprehensive peptide retention modeling for ERLIC is still lacking. In this work we employed a proteomics-derived ~170,000 peptide retention datasets to evaluate major ERLIC retention features using the framework of our Sequence-Specific Retention Calculator model. The separation conditions were adjusted to obtain a wider proteome coverage, particularly for non-modified peptides, resulting in a superior separation orthogonality for a 2D LC combination with reversed-phase C18 LC-MS in the second dimension. The SSRCalc ERLIC model presents a consistent theme with the existing ERLIC retention mechanism, reflecting a dependence on peptide orientation and the position of charged and hydrophilic residues across the peptide backbone. R^2 values of 0.935 and 0.955 for correlations were demonstrated for the standard interpretable SSRCalc model and machine learning algorithm, respectively. The effects of various PTMs on peptide retention were evaluated in this study, covering spontaneous (oxidation, deamidation) and enzymatic (N-terminal acetylation, phosphorylation, glycosylation) modifications.

5.4 Introduction

The innovation of novel chromatographic methods for proteomics has stagnated in recent years while the repurposing of existing peptide separation modes has been a primary area for development. So far, mass spectrometry (MS) has been a major driving force in the ever-increasing capabilities of liquid chromatography tandem mass spectrometry (LC-MS/MS) in bottom-up proteomics. This calls for accelerated innovation and developments in peptide separation sciences moving forward.[1] The vast majority of bottom-up proteomics studies use specific proteases to generate proteolytic peptides and hyphenated LC-MS/MS technologies to analyse them. The gold standard of peptide separations is the use of C18 reversed-phase liquid chromatography (RPLC) as it affords sufficient distribution of peptides across commonly used acetonitrile gradients. RPLC has been utilized for peptide separations since the late 1970s[2], where elution occurs in the order of increasing hydrophobicity. While a specific mechanism can be debated, current models for features which govern peptide retention in RPLC[3] have supported this claim (reviewed[4]). Nonetheless, the longstanding history of RPLC is reflected in its robustness, feasibility of implementation, and amenability to soft ionization techniques which have played a foundational role in its use for proteomics studies.

Hydrophilic interaction chromatography (HILIC), a complementary choice to RPLC, has been reported as early as 1975[5] but was given its name in 1990 by Alpert.[6] It has been another important tool which gained importance for proteomics analyses[7,8] due to the ability to retain and separate hydrophilic peptides. Conversely to RPLC, HILIC is conventionally depicted with the opposite retention order, separating peptides by their hydrophilic character[6], but it exhibits effects of many modes such as reversed-phase (RP), normal phase, and ion-exchange (reviewed[9]). Significant efforts have been invested to determine the characteristics of peptides retention[10], including effects depending on the choice of sorbent[11,12] and eluent component.[13] Other modes, such as ion-exchange[14,15] and capillary electrophoresis[16] have been developed for proteomics but have not been implemented broadly due to a variety of drawbacks, including compatibility to mass-spectrometry, infeasibility to be implemented, issues with reproducibility, and low high-throughput capabilities.

In 2008, a novel HILIC-type separation mode, electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) was introduced by Alpert, incorporating features of weak anion-exchange (WAX).[17] In turn, analytes in ERLIC experience electrostatic attraction/repulsion

with an anion-exchange sorbent which can be modulated by the pH and buffer components, while simultaneous retention is conferred following a HILIC mechanism. For tryptic peptides, the application of low pH conditions sufficient to remove a negative charge from acidic amino acids reduces their excessively high retention, typically observed in anion exchange modes. Moreover, the cationic surface of the sorbent also exudes a repulsive effect on the positively charged termini of tryptic peptides, circumventing the high retention of basic peptides usually observed in HILIC. However, the use of high acetonitrile conditions compensates for the repulsive effect and confers retention in ERLIC, permitting the separation of tryptic peptides.

Another serendipitous discovery when investigating ERLIC was the selectivity of tryptic phosphopeptides based on the position of the phosphorylated residue in the peptide backbone. Alpert et al. noticed that the closer the modified residue was to the C-terminus of the peptide, the greater the retention of the respective peptide species.[18] This initially contradicted the presumed mechanism of tryptic peptides in anion-exchange where it was believed that the C-terminus and N-terminus are repelled by the cationic surface of the sorbent, leaving the internal regions of the peptide to govern retention. Based on the old model, peptides carrying phosphate groups near the internal regions of a peptide should be retained to a greater extent. The adapted mechanism for ERLIC by Alpert et al.[18] would correct this feature by having the C-terminus facing towards the sorbent while the positively charged N-terminus would be facing away due to electrostatic repulsion. In this scenario, the C-terminus of tryptic peptides forms a zwitterionic pair between the lysine or arginine groups and their corresponding carboxylic acid, thereby reducing the regions basicity and permitting the observed orientation-dependent selectivity. This was supported by observing the elution orders of synthetic peptides with asparagine permuted throughout the backbone. The closer the asparagine residue was to the C-terminus, the longer the peptide was retained on the column.

In 2010, ERLIC was used as a first-dimension in an ERLIC-RP fractionation scheme to profile the rat kidney proteome using a nonlinear gradient of ACN → water with a change in ion-pairing reagent from acetic acid → formic acid.[19] Under these conditions, the authors found that ERLIC fractionations resulted in a ~31% increase in protein/peptide identifications over strong cation exchange (SCX) while also resulting in a more even distribution of peptides across fractions. In 2013, this same group showed that a concatenated ERLIC fractionation had comparable results to concatenated high pH reversed-phase fractionation.[20] Moreover, ERLIC was found to be useful

for the separation of peptides carrying post translational modifications (PTMs). An RP-ERLIC scheme (ERLIC in the second dimension) was used for the characterization of deamidation products[21] and ERLIC was also found to enrich phosphorylated and glycosylated peptides[22,23] where recent efforts were made to optimize mobile phase[24] and ion-pairing conditions[25] for fractionation. While a plethora of studies have focused on employing ERLIC for PTM analyses, the applications for wider-coverage shotgun analyses have been scarce. As a result, the selectivity and mechanism of separation in ERLIC for large scale peptide analyses have yet to be comprehensive.

The goal of this study was to continue Dr. Andrew Alpert's leading studies by unravelling the retentive properties of peptides in ERLIC based on their positional-dependent amino acid composition. A proper description of peptide retention properties includes the breadth of physiochemical properties afforded from the molecular diversity of amino acid building blocks. For ERLIC, this reflects the sequence-specific features corresponding to peptide orientation previously mentioned. Our group has long contributed to peptide retention models which consider the effects of neighbouring amino acids, notably on the C- and N- termini. These have been applied across a plethora of separation modes and added to the Sequence-Specific Retention Calculator (SSRCalc, https://proteome.ad.umanitoba.ca/ssrcalc) family of prediction algorithms which can explain the retention of peptides from a chemical perspective.[3,10–12,26–28] Ultimately, these models can be applied for retention time prediction which aims to increase the confidence in peptide identification[29] by providing a chromatographic metric to filter out false positive hits. Moreover, retention time prediction can be used in the *a-priori* development of quantitative LC-MS analyses[30] and multidimensional applications.[31]

In a previous ERLIC study, fractionation using a gradient of ACN → water with ethylenediamine (EDA)-trifluoroacetic acid (TFA) → NH_4-TFA (reference[25], Figure S4 – F) appeared to have a good separation of the tryptic peptides while also resulting in the highest number of total peptide identifications. By adapting this method for wider coverage of the proteome, we determined that this ERLIC fractionation method outperformed other modes in terms of its orthogonality to reversed-phase with formic acid as an ion-pairing modifier and has a high potential in 2D LC-MS applications. Moreover, we note that sequence specific effects derived from our model are consistent with mechanisms previously described in the groundbreaking ERLIC studies by Dr. Andrew Alpert and co-workers.

5.5 Materials and Methods

5.5.1 Materials and Methods for Buffers and Sample Preparation

Deionized water and LC-MS grade acetonitrile were used to prepare all buffers for ERLIC. LC-MS grade water and acetonitrile were used for RPLC-MS analysis. Chemicals were sourced from Millipore Sigma (St. Louis, MO) unless otherwise noted. Sequencing grade modified trypsin (Promega, Madison, WI) was used for digestion. Siliconized 1.5 mL tubes (BioPlas, San Rafael, CA) were used for fraction collection. The custom designed standard peptides P1−P6[32] were synthesized by BioSynthesis Inc. (Lewisville, TX).

A tryptic digest of human Jurkat cells was used for 2D (ERLIC-RP) LC-MS. Jurkat cell lysate was generated from cell pellets by lysis with SDS-containing buffer, reduction and alkylation of cysteine residues with dithiothreitol and iodoacetamide, respectively (with quenching). Proteins were purified utilizing the SP3 method described elsewhere[33], followed by trypsin on-bead digestion. Tryptic Jurkat peptides were purified using C18 RPLC, aliquoted into vials with 100 µg each, and lyophilized prior to first-dimension ERLIC fractionation. The selection of ERLIC separation conditions was performed using our standard peptides (P1-P6) and a tryptic digest of bovine serum albumin (BSA). Tryptic BSA peptides were generated by a standard in-solution digestion following reduction and alkylation steps.[10]

5.5.2 First Dimension ERLIC Conditions

An Agilent 1100 series HPLC with a UV detector (214 nm) and 100 uL injection loop was used for ERLIC separations. A PolyWAX LP 2.1x200mm 5µm 300Å column (PolyLC, Maryland) was used with a flow rate was 250 µL/min. Eluent A (300mM NH_4-TFA, pH 2.5 in water) and eluent B (20 mM EDA-TFA, pH 2.5 in 90:10 ACN:water) were prepared according to a previous report[25] with minor adjustments. The optimized gradient method used during fraction collection followed a linear 0.45%/min water gradient with a 1mM salt gradient over the same window. Starting conditions were 99.9% eluent B, followed by a linear decrease 99.9-79.8% B over 40 minutes. Subsequently the slope was ramped to a linear 79.8-11.1% B decrease over 1 min and 11.1% B for 4 mins for column washing. Re-equilibration of the column post-run required a flow rate of 500 µL/min at 99.9% eluent B for 90 minutes. One-minute fractions were collected during the 40-minute linear gradient and desalted using Evotips prior to analysis. Approximately 200 ng

of peptides from each first-dimension fraction were spiked with our 6-peptide standard[32], and prepared for LC-MS on Evotips following manufacturers specifications.

5.5.3 Second Dimension RPLC-MS/MS

RPLC-MS/MS fraction analyses were performed using an Evosep 1 (Odense, Denmark) coupled to a Bruker timsTOF Pro2 mass spectrometer (Bruker, Bremen, Germany) by data-dependent acquisition (DDA) operating in PASEF (parallel accumulation-serial fragmentation) mode. The preset 30 samples per day gradient method and 500 µL/min flowrate was used for peptide separation, requiring a EV1137 performance column (1.5 µm particle, 150 µm ID × 150 mm, EvoSep) with mobile phase A (0.1% formic acid in water) and mobile phase B (0.1% formic acid in acetonitrile). Nanoflow emitters (20 µm) were charged at 1700 V. Linear calibration of the TIMS funnel voltage and inverse reduced ion mobility (1/*K*0) values were performed using three selected ions (*m*/*z* 622, 922, 1222) from the Agilent ESI-L Tuning Mix (Agilent, Santa Clara, CA, US). A 0.74 s acquisition cycle consisted of one full TIMS-MS scan followed by six DDA-PASEF MS/MS scans where singly charged ions were excluded using a standard polygon filter. The ion mobility was scanned from 0.7 to 1.5 V s/cm^2 at a ramp time of 100 ms, while ramp and accumulation times are locked to a 100% duty cycle. The collision energy was ramped linearly as a function of mobility from 59 eV at 1/*K*0 = 1.6 V s cm^{-2} to 20 eV at 1/*K*0 = 0.6 V s cm^{-2}. Dynamic exclusion was set to 0.4 min except for low-abundance precursors that have an intensity between 2000 and 12,500 counts which were scheduled for a repeat analysis before being excluded.

5.5.4 Data Analysis, Peptide Identification, and Retention Time Assignment

The Fragpipe software suite[34–36], v.22.0, was used to search individual fraction files for peptide identifications with the human swissprot database (downloaded October 16, 2022; 20,328 target sequences). Precusor and fragment mass tolerances were each set at 20 ppm with mass calibration and parameter optimization. Strictly tryptic cleavage rules were used with maximum possible missed cleavages set at 3. Peptide length was set from 7 to 50 amino acids with a corresponding mass range of 500 to 5000 Da. Carbamidomethyl cysteine (+57.02146 Da) was set as a fixed modification. A standard set of modifications were searched as variable mass offsets: methionine/tryptophan oxidation (+15.9949 and +31.98983 Da), asparagine/glutamine deamidation (+0.984016 Da), N-terminal acetylation (+42.0106 Da), glutamine/cysteine N-

terminal cyclization (-17.0265 Da), serine/threonine/tyrosine phosphorylation (+79.996635 Da), and serine/threonine O-linked HexNAcylation (+203.0793 Da). All other settings were default. Validation tools were used with default settings: MSBooster[37], Predict RT, Predict Spectra, Percolator[38], ProteinProphet, and reports were filtered at a 1% false discovery rate (FDR). Retention values 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.

5.6 Results and Discussion

5.6.1 Optimization of Gradient Conditions for ERLIC Separation of Nonmodified Peptides

The authors of the original publication from which we sourced our chromatographic conditions presented a gradient method which was intended to enrich phosphorylated and glycosylated peptides.[25] As a result, nonmodified peptides in this procedure were believed to elute in or near the void volume. However, we noticed that the author's mobile-phase 3 appeared to separate a tryptic peptide mixture with a relatively good distribution of analytes throughout the chromatogram. Although this was a UV trace from a first-dimension fractionation, the resulting high number of peptide identifications (reference[25], Figure 3-A), supported this claim. Our goal was to optimize the gradient parameters using this system to afford sufficient separation of nonmodified peptides for wider proteome coverage.

While "saltless" ion-pairing-based gradients have been used prior for full proteome characterization[19,20,39], we wanted to explore the use of salts as these likely enhance ERLICs electrostatic effects. We determined our optimized conditions similarly to previous experiments[10] by decreasing the gradient slope until our elution profile spanned a desired 40-minute range (**Figure 5.1**). Since the optimal conditions of ERLIC had yet to be comprehensively characterized, our standard peptides P1-P6, spanning a typical hydrophobicity range for tryptic peptides[32], were used for the evaluation of an approximate elution window for real digests (**Figure 5.1**, blue trace). Since the breadth of possible peptide species is much larger than the 6 standard peptides we used, a BSA digest was used as another surrogate sample to confirm the expected elution window of a tryptic digest (**Figure 5.1,** orange trace), resulting in a relatively good distribution of peaks across the chromatogram. Fractionation and 1-minute fraction collection of a Jurkat digest was then performed (**Figure 5.1,** green trace). It is important to note that column re-equilibration in our hands was considerably longer (90 minutes) to afford reproducible chromatograms compared to other studies with shorter times (5 minutes[19,20], 15 minutes[25]). However, this result was consistent with our previous studies and routine analyses using HILIC-type sorbents.[10,11]

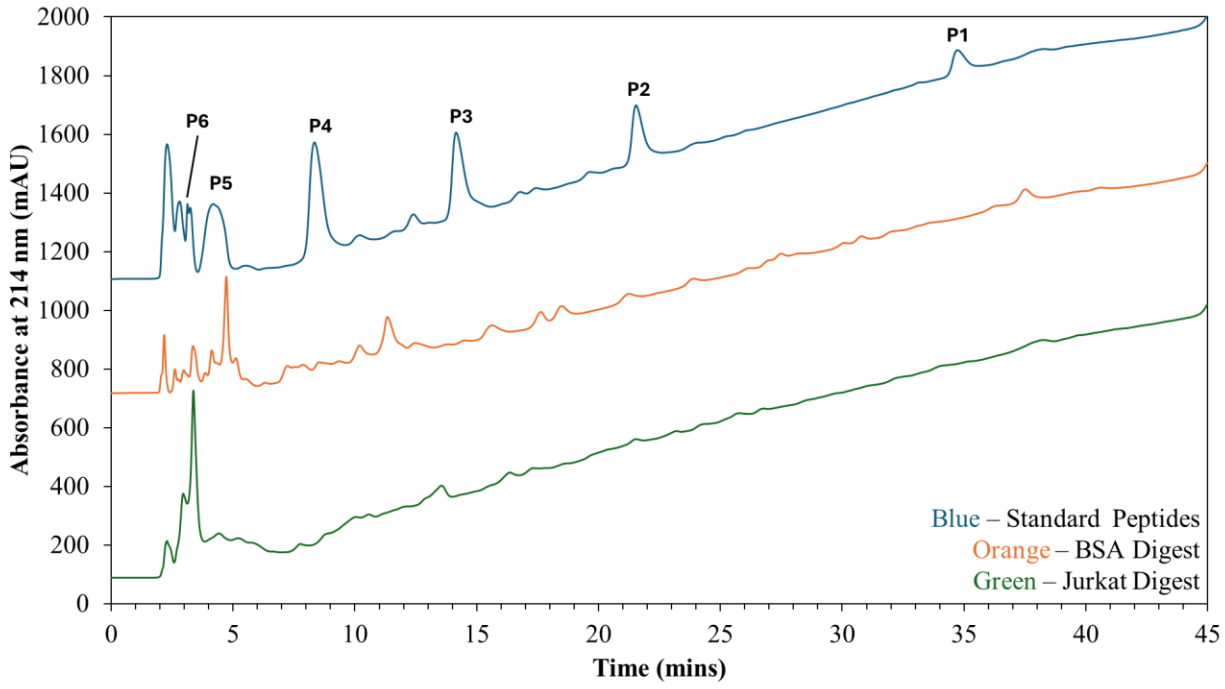


Figure 5.1 Offset chromatograms from experiments used for the optimization of gradient conditions and fraction collection. Our standard peptides (P1-P6) were used to determine a fraction collection window (blue). Confirmation of the elution window was performed with a BSA digest (orange). Jurkat digest was fractionated at 1-minute intervals between 1 to 45 minutes.

5.6.2 Peptide Retention Time Prediction for ERLIC under Optimized Conditions

Our laboratory has established a workflow for retention time modeling for use in a variety of chromatographic modes for proteomics applications.[40] Retention prediction for separation conditions compatible with ESI-MS are done by using a respective column in the second dimension of 2D LC-MS. The first dimension can use any separation mode regardless of eluent compatibility with MS detection. The fraction number in which a correspondent peptide is found is used as the experimental retention value for modeling.

Following our procedure for 2D ERLIC-RP LC-MS acquisition, identified peptides were filtered by their predicted retention times in the second-dimension RP separation mode (**Figure S5.1**). By comparing predicted to observed retention times, we filtered out outliers such that potential false positives were removed. In practice, this increased the confidence in the peptide identification output such that true positive identifications were used to train the ERLIC model for peptide retention. The first model generated from ~170,000 peptides was our standard SSRCalc model (**Figure 5.2A**).[26] In this model, peptide retention is based on the additive contribution of

each amino acid to a peptide's retention including the positional dependence of retention coefficients (relative to peptide termini). The optimization procedure generated retention coefficients for amino acids defined between internal residues (**Figure 5.3B**) and various positions near the C- and N-termini (**Figure 5.3, A and C**) which will be discussed in the next section. Ultimately, these values tell us how a peptide behaves based on ion-pairing, peptide-sorbent interactions, and can potentially elucidate mechanistic insights for driving factors in peptide retention. Our SSRCalc prediction model for ERLIC showed an R^2=0.9335 correlation for the dependence of experimental vs. predicted fraction number (**Figure 5.2A**) with a standard deviation of 2.65 fractions.

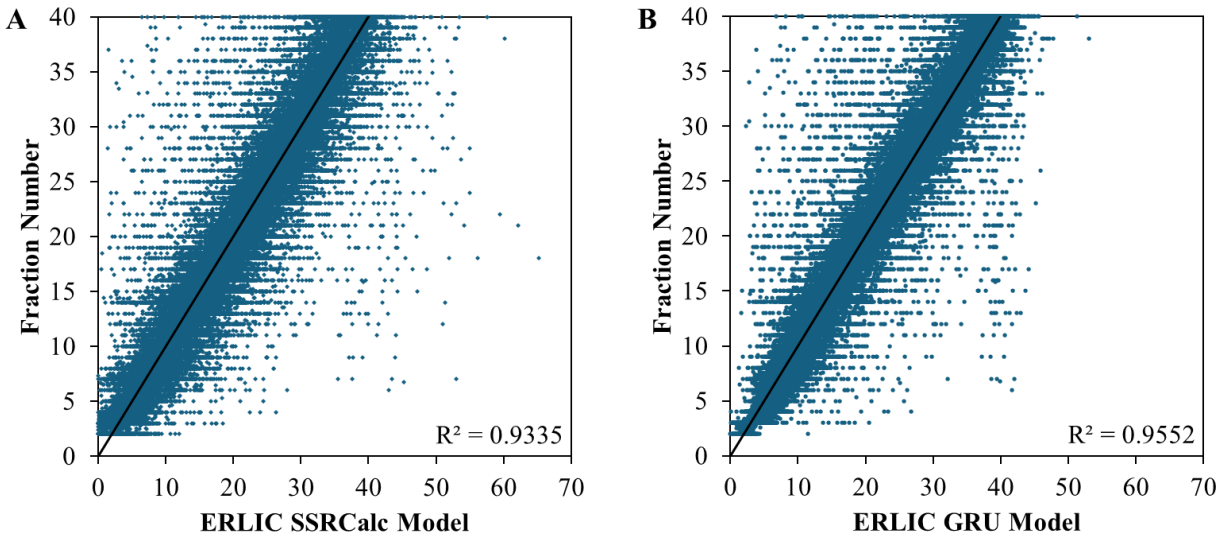


Figure 5.2 Retention prediction for ERLIC using a ~170,000 unique peptides retention dataset. Plots represent the observed fraction number of each individual peptide against the models trained and used for prediction. A – ERLIC SSRCalc our standard model with position-dependent retention coefficients with a standard deviation of 2.65 fractions; B – ERLIC GRU model using machine learning approach with a standard deviation of 2.17 fractions.

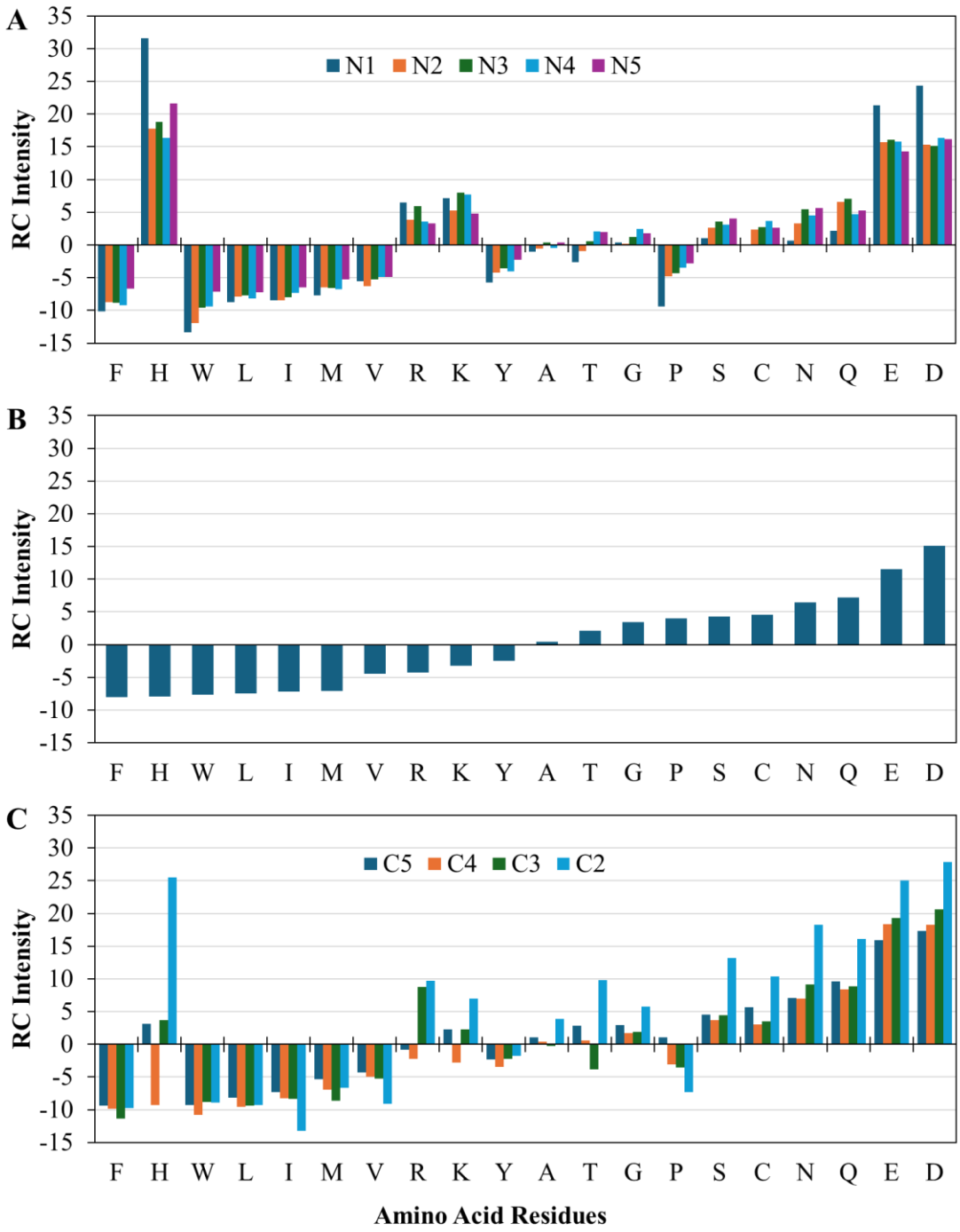


Figure 5.3 Retention coefficients (RCs) from the ERLIC SSRCalc Model for N-terminal (A), internal (B) and C-terminal (C) positions of amino acids. Terminal RCs are represented by the distance from the termini e.g. N1 = N-terminal position, N2 = second position in from the N-terminus etc. Since tryptic peptides only contain K and R on the C-terminus, C1 coefficients are omitted due to lack of representation for all other amino acids.

ERLIC presented an interesting case for retention modelling where sequence-specific effects for peptide retention were based on two orthogonal mechanisms: electrostatic effects and polarity. This is particularly interesting considering other separation modes generally have one major mechanism, while non-specific and alternative sorbent-analyte interactions become nuanced in the bigger picture. As an example, periodic distribution of hydrophobic residues (promoting amphipathic helices) contributes to the broadening of linear regressions for predicted versus observed retention times in RPLC. In all modes, a larger number of nonspecific or alternative mechanisms generally decreases prediction accuracy in simplistic models using a limited number of variables, such as the SSRCalc model. As many features are speculated to drive retention under HILIC modes[9] and have been subsequently modelled by our group[10,11], it is unsurprising that adding WAX electrostatic effects complicates retention modelling for ERLIC, reducing the prediction accuracy with SSRCalc. As an alternative to limited variable models, retention time prediction was improved in this study by training our shallow machine learning model[41], hereafter referred to as the ERLIC gated recurrent unit (GRU) model. This afforded increased prediction accuracy (**Figure 5.2B**) with R^2=0.9552 and standard deviation of 2.17 fractions. While it provided a better prediction accuracy, this approach represented a typical machine learning "black box" without mechanistic insights available for analysis. Interestingly, many of the outliers in this model were the same as in the SSRCalc model but appeared to have a narrower distribution in the linear regression for predicted versus observed fractions.

The negative outliers in our prediction model consisted mostly of peptides carrying long strings of acidic residues and acidic residues next to positively charged groups such as basic residues or the N-terminus. Positive outliers from our prediction model consisted of peptides carrying larger numbers of cysteine, asparagine, and glutamine residues, and proline in strings or on the C- and N- termini. These comprise potential directions for future improvements to the ERLIC prediction model that would require consideration of special sequences, and manual validation using synthetic peptides.

5.6.3 Sequence-Specific Effects in Peptide ERLIC

It was cathartic to find that the early work on describing retention mechanisms for ERLIC, which found tryptic peptide retention selectivity dependent on orientation with a limited number of peptides[18], was consistent with the output of our advanced retention model based on a 170,000-

analyte retention dataset. However, our observations appeared to support this initial discovery and expanded into the affects of other amino acids. The first of many of these supportive observations in our study was that C- and N-terminal distance appeared to be a major driving factor in peptide retention (**Figure 5.3**). At the C2-position for all hydrophilic amino acids, a dramatic increase in retention coefficients was observed (**Figure S5.2C**). Moreover, as the position of amino acids approaches C3 to C5, the contribution resembled that of internal amino acids. Approaching the N-terminus (N5 to N1), the contribution of the hydrophilic amino acids decreased with the lowest point on the N-terminus. These results reflected a similar observation by Alpert et al. that hydrophilic amino acids approaching the C-terminus increased peptide retention in ERLIC, based on the observation of permuting asparagine throughout the peptide backbone.[18]

These observations were consistent for acidic amino acids with the exception of the N-terminus. Retention coefficients for acidic amino acids appeared to increase in magnitude as they approach both termini (**Figure S5.2F**). On the C-terminus, this was likely the result of the same distance dependent peptide-sorbent interaction explained previously for hydrophilic amino acids. On the N-terminus, the presence of an acidic amino acid would reduce the basicity of the amino group[28] and therefore reduce ion-pairing with the hydrophobic ion-pairing reagent, TFA. A reduction in ion-pairing with TFA would thus increase the effective hydrophilicity of peptides and increased their observed retention in ERLIC.

For other groups of amino acids, their retention coefficients lead to alternative conclusions. For aromatic and hydrophobic amino acids (**Figure S5.2, A and B**) the retention coefficients did not change largely with respect to their position in the peptide, with the exception of proline on internal positions. Proline is usually an outlier in our retention modelling as it carries unique hydrophobic characteristics[40], likely due to conformational effects. The consistent effect for retention across positions would suggest that the major contribution of aromatic and hydrophobic amino acids was in the HILIC-type mechanism, altering the affinity of a peptide into the immobilized water-layer of the sorbent rather than affecting the peptide-sorbent interactions.

Alanine and glycine in this study were referred to as neutral amino acids, as their contributions to peptide retention were overall small compared to other amino acids (**Figure 5.3**). Similarly to hydrophilic amino acids, they displayed an increase in retention contribution when in a C2-position (**Figure S5.2D**) but to a lesser extent. This could be the result of facilitating free

rotation of the C-terminus as they are comparatively less conformationally rigid than other amino acids, however a definitive mechanism was not elucidated in this study for these amino acids.

The anomalous behaviour for positional dependent retention coefficients for basic amino acids (**Figure S5.2E**) did not have a clear explanation. Since positional dependent retention coefficients depended on many instances where these amino acids were present (i.e. at N1, N2, N3, etc.), the low number of instances for basic amino acids in these positions was likely the reason behind these results and were thus not representative of true values. This is not surprising for lysine and arginine since tryptic digests generally produce the majority of these residues on the C-terminus. Synthetic peptides would be a future consideration to properly address the positional-dependent retention characteristics of peptides carrying basic residues but were not conducted for this study. Conversely, the internal retention coefficients for basic amino acids followed expected outcomes with negative retention coefficients. As the PolyWAX LP sorbent acts as an anion exchange sorbent, one would have expected electrostatic repulsion with the positively charged basic residues.

5.6.4 Comparison of ERLIC to Other HILIC Separation Modes

Figure 5.4 shows the dependence of normalized internal retention coefficients of ERLIC model against other SSRCalc algorithms developed in our laboratory such as strong anion exchange (SAX) and other HILIC modes under acidic conditions (pH 2.7): XBridge Amide, HILIC, ZIC-HILIC, and ZIC-cHILIC (manuscript submitted to the Journal of Chromatography A). The most comparable to ERLIC HILIC mode was Zic-cHILIC (**Figure 5.4D**), followed by XBridge Amide, Zic-HILIC. This behaviour was expected since ZIC-HILIC carries distal negative charge and ZIC-cHILIC a distal positive charge - the most comparable surface chemistry to PolyWAX LP. It is interesting to note that the majority of amino acid contributions in HILIC modes were largely in agreement for ERLIC with the exception of charged residues. This reflects that ERLIC is comparable to other modes, simply adding additional selectivity to charged groups. For SAX, retention coefficients bared nearly no resemblance to ERLIC aside from acidic residues due to significant differences in eluent pH and composition.

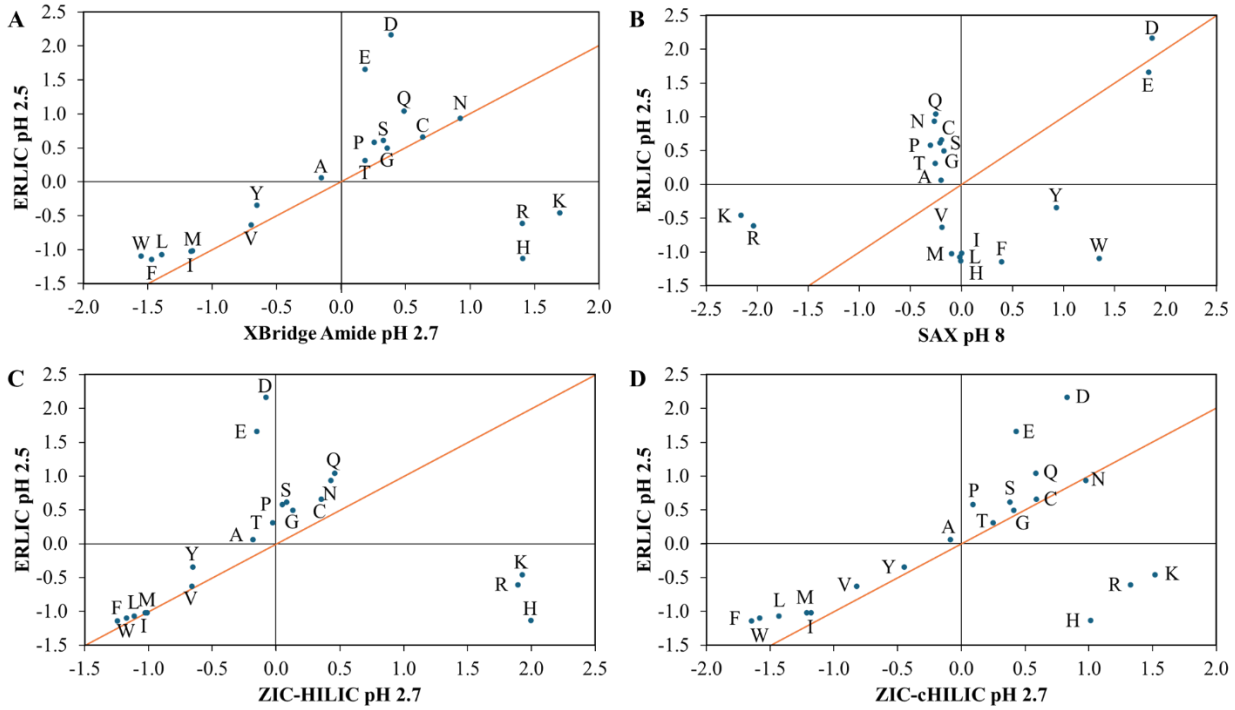


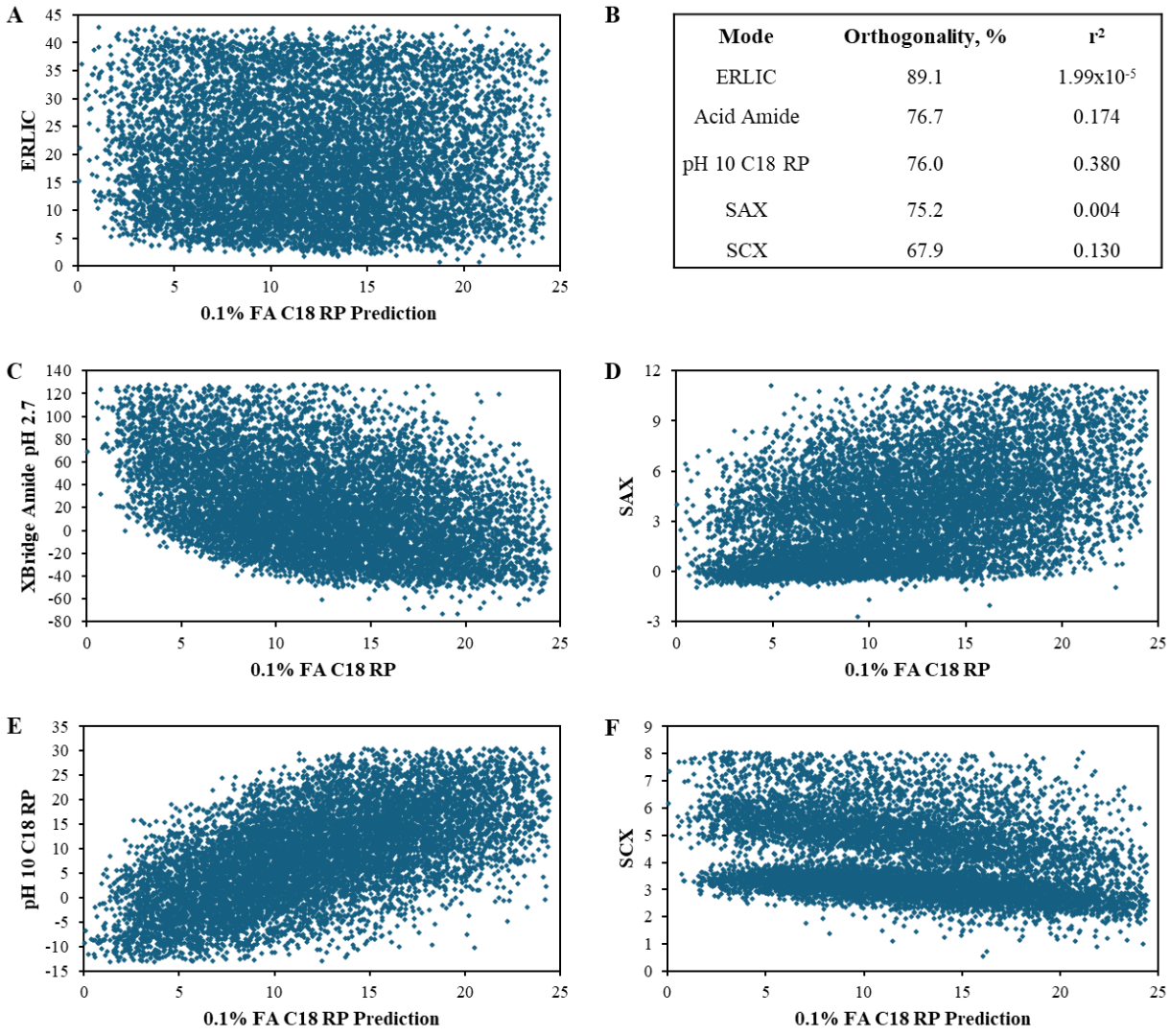
Figure 5.4 Plots of normalized internal retention coefficients for ERLIC against other modes. A – ERLIC vs XBridge Amide (HILIC) at pH 2.7. B – ERLIC vs SAX pH 8. C – ERLIC vs ZIC-HILIC at pH 2.7. D – ERLIC vs ZIC-HILIC at pH 2.7. Orange represents the function y = x (Mode 1 = Mode 2) which aids in illustrating the difference in amino acid dependence for peptide retention in each mode: the closer the values to the line, the more similar two modes are for any amino acid.

5.6.5 Peptide Separation Orthogonality for ERLIC-RP Compared to Other First Dimension Separation Modes

The orthogonality between two chosen dimensions for 2-dimensional liquid chromatography (2D LC) applications is another consideration which should be met tactfully. Ideally, completely orthogonal modes presents the best case for achieving the maximum number of identifications. This is achieved by reducing the number of coeluting species in a standard RP LC-MS run through implementation of a first-dimension fractionation, thereby also increasing the overall dynamic range (reviewed[8,42]).

Our group has previously covered peptide orthogonality in detail[43] but did not include ERLIC in this comparison. The orthogonality of ERLIC was hinted to in a previous study comparing high-pH RP to ERLIC as a first dimension in a 2D LC setup.[20] The authors remarked that concatenated ERLIC performed similarly to high-pH reversed phase, however, a direct comparison on non-concatenated systems was not reported. When plotting the retention in RP

against the observed fractions for the peptides used for retention modelling, we initially observed a high level of orthogonality (**Figure S5.3**), inspiring us to investigate this further. For the most part, SCX, high pH reversed-phase, and HILIC are all popular choices for 2D LC applications as they possess sufficient orthogonal separation selectivity to reversed-phase.[8,42] Thus, ERLIC was compared to these modes by randomly selecting a subset of 10,000 peptide sequences from our peptide library (adapted from[42]), running predictions from each of our models, then plotting the predicted retention in each mode against the predicted retention of 0.1% formic acid RP, summarized by orthogonality scores and R^2 values (**Figure 5.5**).



Mode	Orthogonality, %	r^2
ERLIC	89.1	1.99×10^{-5}
Acid Amide	76.7	0.174
pH 10 C18 RP	76.0	0.380
SAX	75.2	0.004
SCX	67.9	0.130

Figure 5.5 Orthogonality plots for various separation modes based on 10,000 randomly selected peptides against reversed-phase (RP) with formic acid (FA) as an ion-pairing modifier.

Orthogonality is scored following the same procedure outlined in our previous publication [43]. A – ERLIC vs 0.1% FA C18 RP. B – Summary of Orthogonality scores and r^2 values. C – XBridge Amide HILIC vs 0.1% FA C18 RP. D – SAX vs 0.1% FA C18 RP. E – high pH reversed phase vs 0.1% FA C18 RP. F – SCX vs 0.1% FA C18 RP.

Orthogonality scores were explained in our previous study[43], representing the proportion from possible random distribution across both dimensions.[42] ERLIC was found to have an orthogonality score of 89.1%, the highest recorded from our datasets and an increase of ~12% from the next closest comparisons: HILIC, then high pH reversed phase, SAX, and SCX. This suggested that ERLIC is a highly promising candidate for 2D LC applications. Considering that ERLIC also affords enrichment of glycosylated and phosphorylated peptides, future development for large scale proteomics applications using ERLIC could be highly valuable for a comprehensive overview of proteomes.

5.6.6 Behaviour of Common Post-Translational Modifications in ERLIC.

It is important to understand the roles of spontaneous and endogenous post-translational modifications (PTMs) when assessing the results of proteomics analyses. The former requires careful consideration to their effect on identification outputs while the latter is often a focal point in understanding their roles in the biological context. Thus, we have summarized the retention shifts of various PTMs in ERLIC, represented by the average and standard deviations of fraction shifts between modified and nonmodified sequence counterparts (**Figure 5.6**). It is important to note that the reported average shift values represent general effects of the modifications on peptide retention in ERLIC, while the standard deviations represent the chemical diversity within a modification that is influenced by a peptides primary structure. While a common definition of the standard deviation relates to the multiple measurements of the same parameter (e.g. retention shifts for the same peptide), standard deviations in **Figure 5.6** are given for multiple measurements of the shifts for different peptides.

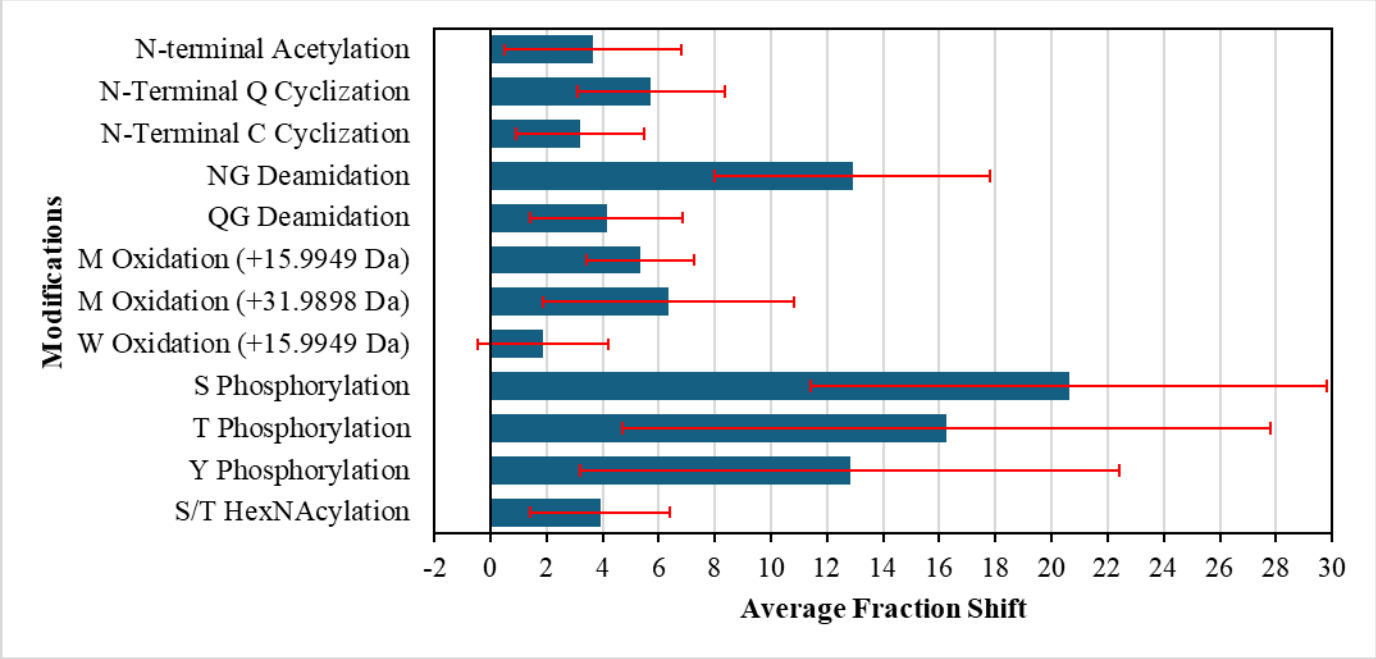


Figure 5.6 Effects of various PTMs on peptide retention in ERLIC. Average fraction shifts were calculated by generating pairs of modified and nonmodified sequence counterparts carrying only one modification. Error bars represent the standard deviation of all fraction shifts for each respective modification.

N-terminal cyclization of glutamine and carbamidomethyl-cysteine, deamidation of asparagine and glutamine (limited to NG and QG sequences), and oxidation of methionine (+15.9949 Da and +31.9989 Da) and tryptophan (+15.9949 Da) are some of the most common spontaneous modifications to affect proteomics outputs and thus were suitable candidates for evaluation of their behaviour in ERLIC. N-terminal acetylation, Serine and threonine O-linked N-acetylhexosamine (O-HexNAc) and phosphorylation, which also includes tyrosine, are suitable choices for evaluating ERLIC characteristics for studying biologically relevant modifications. We have found that in each of these cases, positive retention shifts were observed.

As N-terminal cyclizations, N-terminal acetylation, and deamidation of asparagine and glutamine reduce the basicity of a peptide, these positive retention shifts in ERLIC are unsurprising. N-terminal cyclization removes the N-terminal amine's positive charge while simultaneously reducing ion-pairing with TFA, thereby decreasing electrostatic repulsion and making the peptide more hydrophilic. Deamidation on the other hand, transforms asparagine and glutamine residues into aspartic acid/β-aspartic acid and glutamic acid/γ-glutamic acid pairs, respectively.[44,45] The conversion of amide to carboxylic acid functionality increases the overall

acidity of peptides, thereby increasing their interaction with the anion-exchange sorbent. The larger effect observed for asparagine over glutamine can be rationalized by its proximity to the backbone, asparagine having one less methylene group in the side chain. In our previous CZE study, we observed a 3-fold larger inductive effect for asparagine than glutamine affecting the basicity of the N-terminus.[28] It is likely that this is consistent across for different positions in the peptide, which would explain our observations and is further supported by the increased retention coefficients for aspartic acid compared to glutamic acid (**Figure 5.3B**).

A similar feature was observed for phosphorylation, but with larger retention shifts than most other modifications. Phosphate groups are more acidic than other peptide functionalities and likely retained a negative charge at pH 2.5 which resulted in their characteristics ERLIC retention. In our case, most phosphorylated peptides resided in later fractions, indicating that a wider collection window and increased gradient length would be required to model retention more accurately.

Oxidation, and serine/threonine O-HexNAc modifications were not expected to change the overall charge of a peptide but do represent hydrophilic modifications which increase retention in HILIC modes.[46,47] Therefore, the positive retention shifts of these modifications can be explained by increasing the hydrophilicity of peptides consistent with HILIC characteristics in previous reports.

5.7 Conclusion

ERLIC was initially presented as a unique separation mode for analytes, incorporating both HILIC and WAX properties with the goal of overcoming the drawbacks in each of these individual separation modes.[17] Further investigation revealed that ERLIC had an orientation-specific selectivity for tryptic peptides which orient with their C-terminus towards the sorbent.[18] This innovative development and interesting feature for peptides presented an attractive avenue for method development targeting unique separations. Our laboratory has a long history of evaluating the chromatographic properties of peptides, with a focus on elucidating retention mechanisms for the development of separation sciences in proteomics applications. This work continued our long-standing experience in collaborating with Dr. Andrew Alpert and PolyLC, thus we undertook the development of a retention model for ERLIC using a very large proteomics-derived dataset.

The first retention time prediction model for ERLIC was generated with our standard SSRCalc approach using 170,000 peptides, affording R^2=0.9335. Improvement of this model to R^2=0.9552 was achieved by using our shallow machine learning model[41] but removed the ability to extract any relevant information for elucidating a retention mechanism, therefore this information was processed using the SSRCalc model. The extracted retention coefficients from the standard SSRCalc model presented similar features to the early assumptions of ERLICs retention mechanism, reflecting a dependence of retention on the position of charged and hydrophilic residues in the backbone of a peptide. Generally, hydrophilic residues near or on the C-terminus had a larger influence on increasing retention than on the N-terminus. Similarly, acidic residues presented the same feature on the C-terminus, however, also increased retention on the N-terminus as this reduced the basicity of N-terminal amino group and decreased ion-pairing with the hydrophobic TFA ion-pairing modifier.

By comparing the retention coefficients in ERLIC to similar modes such as HILIC, Zic-HILIC, Zic-cHILIC (all three at acidic pH), and SAX, we were able to demonstrate that ERLIC carried a higher selectivity to charged amino acids than in any other case, with the most similar mode being Zic-cHILIC. Comparisons to other modes popularly employed in 2D LC formats were able to demonstrate that ERLIC is more orthogonal, presenting the most orthogonal mode studied in our laboratory to date. The general characteristics of common PTMs in ERLIC were also characterized; all of them resulting in increased peptide retention.

To rival RPLC in application and feasibility, future development in ERLIC should emphasize the application towards total-proteome coverage and usage of higher quality matrices, including a focus on optimization of eluent components and corresponding concentrations to improve peak shape and reduce equilibration time between runs. On-line ERLIC-MS approaches will likely suffer from a lack of amenability to salt matrices, thus limiting the scope of ERLIC in such formats. On the other hand, application as a first separation dimension in large-scale proteomics is very promising and would benefit further from the development of high-efficiency stationary phases for ERLIC. Further improvements in ERLIC separation efficiency and robustness would potentially allow it to challenge the current method of choice for the first-dimension peptide separation in 2D HPLC – high pH reversed phase HPLC.

5.8 Acknowledgements

This work was supported by the grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2022-05015 O.V.K.). The authors thank the Department of Internal Medicine, Shared Health Manitoba and Canada Foundation for Innovation for continuous support of Manitoba Centre for Proteomics and Systems Biology. The authors also thank Dr. Christopher J. Wike (PolyLC) for providing the columns for experiments and advice during the early stages of this project.

5.9 Data Availability

The raw LC-MS files for this experiment were uploaded to the MassIVE repository under the accession number MSV000095965 (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp) for public access. Contact authors for further information regarding data analysis.

5.10 Supplementary Information

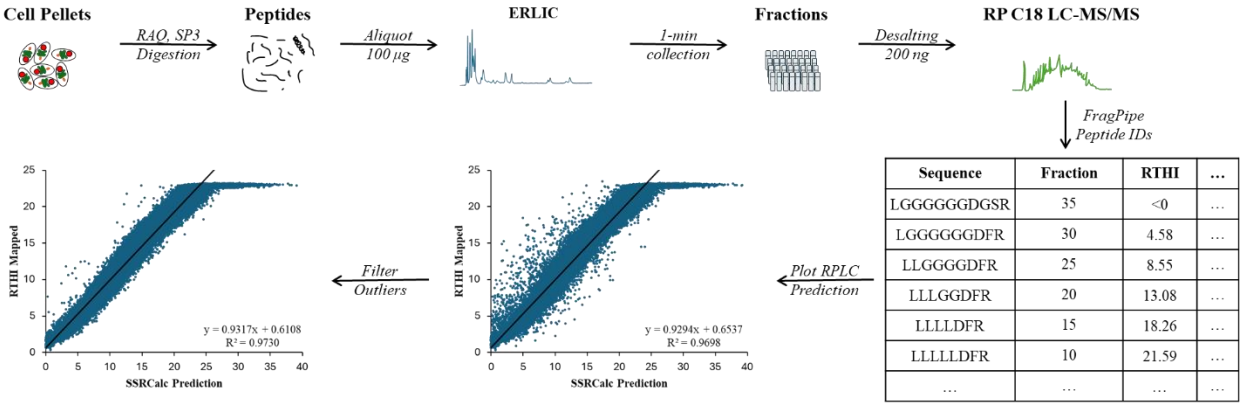


Figure S5.1 Experimental Workflow sample preparation, peptide identification, and RPLC filtering.

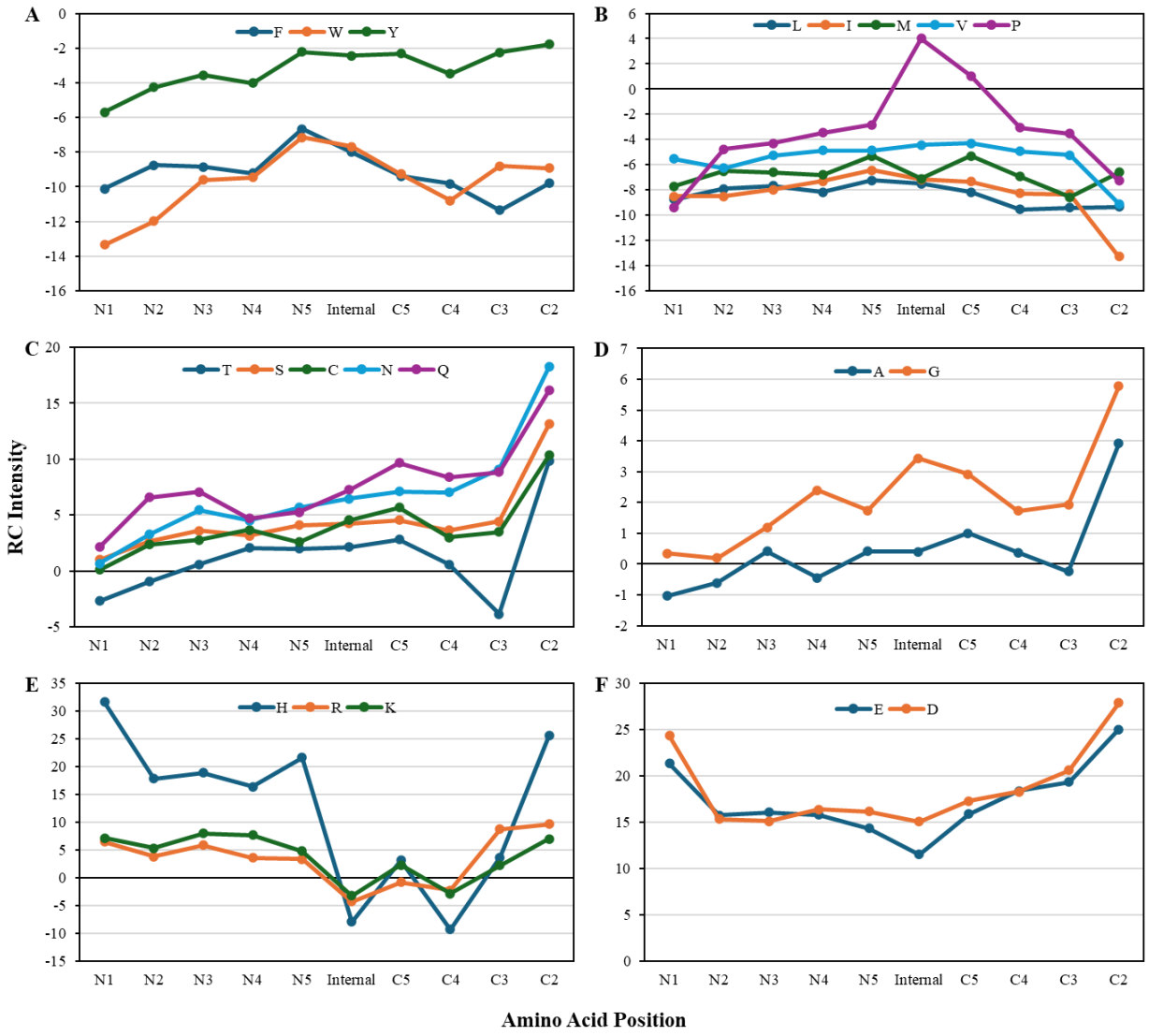


Figure S5.2 Trends for retention coefficients based on the position of amino acids. Amino acids are represented by their one letter codes. A – aromatic amino acids. B – Hydrophobic amino acids. C – Hydrophilic amino acids. D – Neutral amino acids. E – Basic amino acids. F – Acidic amino acids.

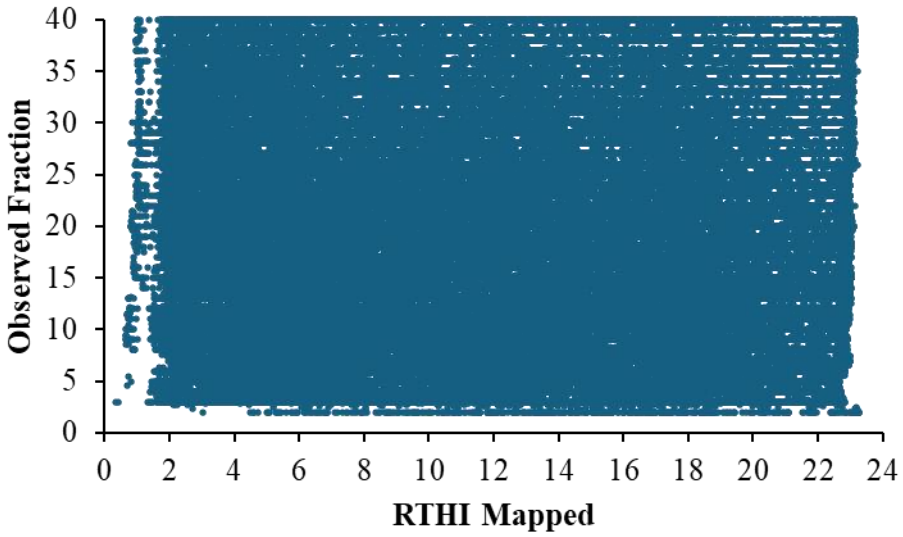


Figure S5.3 Orthogonality plot of peptides identification outputs from ERLIC fractionation. RTHI is values were mapped from retention times using our 6-peptide standard. Corresponding fractions were used to visualize orthogonality.

Table S5.1 All amino acid retention coefficients from the SSRCalc ERLIC model.

Amino Acid	N1	N2	N3	N4	N5	Internal	C5	C4	C3	C2	C1
F	-10.11	-8.75	-8.86	-9.21	-6.66	-8.01	-9.39	-9.83	-11.34	-9.8	9.66
H	31.59	17.82	18.85	16.39	21.63	-7.92	3.14	-9.26	3.68	25.51	79.19
W	-13.34	-11.99	-9.6	-9.45	-7.16	-7.68	-9.27	-10.8	-8.81	-8.92	0.25
L	-8.74	-7.92	-7.68	-8.16	-7.23	-7.51	-8.18	-9.55	-9.42	-9.33	4.7
I	-8.5	-8.5	-7.96	-7.3	-6.45	-7.16	-7.34	-8.28	-8.35	-13.28	8.67
M	-7.69	-6.51	-6.62	-6.81	-5.29	-7.12	-5.31	-6.93	-8.59	-6.62	14.92
V	-5.53	-6.26	-5.26	-4.86	-4.88	-4.43	-4.31	-4.94	-5.24	-9.12	8.13
R	6.47	3.84	5.89	3.58	3.33	-4.29	-0.78	-2.2	8.75	9.67	12.67
K	7.13	5.3	7.98	7.68	4.81	-3.22	2.27	-2.81	2.25	6.99	8.98
Y	-5.7	-4.27	-3.56	-4.01	-2.23	-2.44	-2.31	-3.49	-2.25	-1.79	18.38
A	-1.03	-0.61	0.42	-0.45	0.42	0.4	1.01	0.37	-0.24	3.92	36.15
T	-2.67	-0.94	0.61	2.07	1.99	2.15	2.83	0.6	-3.84	9.82	87.41
G	0.35	0.2	1.2	2.4	1.74	3.44	2.92	1.73	1.93	5.77	35.49
P	-9.37	-4.78	-4.3	-3.48	-2.84	4.03	1.05	-3.06	-3.52	-7.28	23.26
S	1.02	2.62	3.62	3.15	4.08	4.25	4.56	3.65	4.43	13.14	38.06
C	0.11	2.37	2.77	3.7	2.6	4.56	5.68	3.02	3.5	10.36	87.45
N	0.69	3.27	5.44	4.5	5.67	6.48	7.09	7.02	9.1	18.23	51.36
Q	2.14	6.58	7.07	4.68	5.28	7.24	9.65	8.38	8.82	16.13	59.62
E	21.34	15.75	16.07	15.78	14.32	11.53	15.87	18.37	19.32	25	53.36
D	24.32	15.32	15.12	16.38	16.16	15.06	17.28	18.28	20.6	27.87	42.04

5.11 References

1. 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.
2. Bennett, H. P. J.; Hudson, A. M.; McMartin, C.; Purdon, G. E. Use of Octadecasilyl-Silica for the Extraction and Purification of Peptides in Biological Samples. Application to the Identification of Circulating Metabolites of Corticotropin-(1-24)-Tetracosapeptide and Somatostatin *in Vivo*. *Biochemical Journal* **1977**, *168* (1), 9–13. https://doi.org/10.1042/bj1680009.
3. 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.
4. 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.
5. Linden, J. C.; Lawhead, C. L. Liquid Chromatography of Saccharides. *J Chromatogr A* **1975**, *105* (1), 125–133. https://doi.org/10.1016/S0021-9673(01)81096-7.
6. 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.
7. 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.
8. Di Palma, S.; Hennrich, M. L.; Heck, A. J. R.; Mohammed, S. Recent Advances in Peptide Separation by Multidimensional Liquid Chromatography for Proteome Analysis. *J Proteomics* **2012**, *75* (13), 3791–3813. https://doi.org/10.1016/j.jprot.2012.04.033.
9. Buszewski, B.; Noga, S. Hydrophilic Interaction Liquid Chromatography (HILIC)—a Powerful Separation Technique. *Anal Bioanal Chem* **2012**, *402* (1), 231–247. https://doi.org/10.1007/s00216-011-5308-5.
10. 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.

11. 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. https://doi.org/10.1016/j.chroma.2017.12.046.
12. Yeung, D.; Klaassen, N.; Mizero, B.; Spicer, V.; Krokhin, O. V. Peptide Retention Time Prediction in Hydrophilic Interaction Liquid Chromatography: Zwitter-Ionic Sulfoalkylbetaine and Phosphorylcholine Stationary Phases. *J Chromatogr A* **2020**, *1619*. https://doi.org/10.1016/j.chroma.2020.460909.
13. Alpert, A. J. Effect of Salts on Retention in Hydrophilic Interaction Chromatography. *J Chromatogr A* **2018**, *1538*, 45–53. https://doi.org/10.1016/j.chroma.2018.01.038.
14. Davis, M. T.; Beierle, J.; Bures, E. T.; McGinley, M. D.; Mort, J.; Robinson, J. H.; Spahr, C. S.; Yu, W.; Luethy, R.; Patterson, S. D. Automated LC–LC–MS–MS Platform Using Binary Ion-Exchange and Gradient Reversed-Phase Chromatography for Improved Proteomic Analyses. *J Chromatogr B Biomed Sci Appl* **2001**, *752* (2), 281–291. https://doi.org/10.1016/S0378-4347(00)00547-8.
15. Dai, J.; Wang, L.-S.; Wu, Y.-B.; Sheng, Q.-H.; Wu, J.-R.; Shieh, C.-H.; Zeng, R. Fully Automatic Separation and Identification of Phosphopeptides by Continuous PH-Gradient Anion Exchange Online Coupled with Reversed-Phase Liquid Chromatography Mass Spectrometry. *J Proteome Res* **2009**, *8* (1), 133–141. https://doi.org/10.1021/pr800381w.
16. Fonslow, B. R.; Yates III, J. R. Capillary Electrophoresis Applied to Proteomic Analysis. *J Sep Sci* **2009**, *32* (8), 1175–1188. https://doi.org/10.1002/jssc.200800592.
17. Alpert, A. J. Electrostatic Repulsion Hydrophilic Interaction Chromatography for Isocratic Separation of Charged Solutes and Selective Isolation of Phosphopeptides. *Anal Chem* **2008**, *80* (1), 62–76. https://doi.org/10.1021/ac070997p.
18. Alpert, A. J.; Petritis, K.; Kangas, L.; Smith, R. D.; Mechtler, K.; Mitulović, G.; Mohammed, S.; Heck, A. J. R. Peptide Orientation Affects Selectivity in Ion-Exchange Chromatography. *Anal Chem* **2010**, *82* (12), 5253–5259. https://doi.org/10.1021/ac100651k.
19. Hao, P.; Guo, T.; Li, X.; Adav, S. S.; Yang, J.; Wei, M.; Sze, S. K. Novel Application of Electrostatic Repulsion-Hydrophilic Interaction Chromatography (ERLIC) in Shotgun

Proteomics: Comprehensive Profiling of Rat Kidney Proteome. *J Proteome Res* **2010**, *9* (7), 3520–3526. https://doi.org/10.1021/pr100037h.

20. Hao, P.; Ren, Y.; Dutta, B.; Sze, S. K. Comparative Evaluation of Electrostatic Repulsion–Hydrophilic Interaction Chromatography (ERLIC) and High-PH Reversed Phase (Hp-RP) Chromatography in Profiling of Rat Kidney Proteome. *J Proteomics* **2013**, *82*, 254–262. https://doi.org/10.1016/j.jprot.2013.02.008.
21. Hao, P.; Qian, J.; Dutta, B.; Cheow, E. S. H.; Sim, K. H.; Meng, W.; Adav, S. S.; Alpert, A.; Sze, S. K. Enhanced Separation and Characterization of Deamidated Peptides with RP-ERLIC-Based Multidimensional Chromatography Coupled with Tandem Mass Spectrometry. *J Proteome Res* **2012**, *11* (3), 1804–1811. https://doi.org/10.1021/PR201048C.
22. Hao, P.; Guo, T.; Sze, S. K. Simultaneous Analysis of Proteome, Phospho- and Glycoproteome of Rat Kidney Tissue with Electrostatic Repulsion Hydrophilic Interaction Chromatography. *PLoS One* **2011**, *6* (2), e16884. https://doi.org/10.1371/journal.pone.0016884.
23. 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. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2015**, *1854* (5), 460–468. https://doi.org/10.1016/j.bbapap.2015.01.006.
24. Cui, Y.; Yang, K.; Tabang, D. N.; Huang, J.; Tang, W.; Li, L. Finding the Sweet Spot in ERLIC Mobile Phase for Simultaneous Enrichment of N-Glyco and Phosphopeptides. *J Am Soc Mass Spectrom* **2019**, *30* (12), 2491–2501. https://doi.org/10.1007/s13361-019-02230-6.
25. Cui, Y.; Tabang, D. N.; Zhang, Z.; Ma, M.; Alpert, A. J.; Li, L. Counterion Optimization Dramatically Improves Selectivity for Phosphopeptides and Glycopeptides in Electrostatic Repulsion-Hydrophilic Interaction Chromatography. *Anal Chem* **2021**, *93* (22), 7908–7916. https://doi.org/10.1021/acs.analchem.1c00615.
26. 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.

27. 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.
28. 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. https://doi.org/10.1021/acs.analchem.6b04544.
29. 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.
30. Lange, V.; Picotti, P.; Domon, B.; Aebersold, R. Selected Reaction Monitoring for Quantitative Proteomics: A Tutorial. *Mol Syst Biol* **2008**, *4* (1). https://doi.org/10.1038/msb.2008.61.
31. 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.
32. 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. https://doi.org/10.1021/ac9016693.
33. 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* (10), 757. https://doi.org/10.15252/MSB.20145625.
34. 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.
35. Teo, G. C.; Polasky, D. A.; Yu, F.; Nesvizhskii, A. I. Fast Deisotoping Algorithm and Its Implementation in the MSFragger Search Engine. *J Proteome Res* **2021**, *20* (1), 498–505. https://doi.org/10.1021/acs.jproteome.0c00544.

36. Yu, F.; Haynes, S. E.; Teo, G. C.; Avtonomov, D. M.; Polasky, D. A.; Nesvizhskii, A. I. Fast Quantitative Analysis of TimsTOF PASEF Data with MSFragger and IonQuant. *Molecular & Cellular Proteomics* **2020**, *19* (9), 1575–1585. https://doi.org/10.1074/mcp.TIR120.002048.
37. 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* (1), 4539. https://doi.org/10.1038/s41467-023-40129-9.
38. Käll, 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* (11), 923–925. https://doi.org/10.1038/nmeth1113.
39. de Jong, E. P.; Griffin, T. J. Online Nanoscale ERLIC-MS Outperforms RPLC–MS for Shotgun Proteomics in Complex Mixtures. *J Proteome Res* **2012**, *11* (10), 5059–5064. https://doi.org/10.1021/pr300638n.
40. 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* (37), 6348–6355. https://doi.org/10.1016/j.chroma.2011.06.092.
41. 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.
42. 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.
43. 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.
44. Robinson, A. B.; Rudd, C. J. Deamidation of Glutaminyl and Asparaginyl Residues in Peptides and Proteins; 1974; pp 247–295. https://doi.org/10.1016/B978-0-12-152808-9.50013-4.

45. Riggs, D. L.; Silzel, J. W.; Lyon, Y. A.; Kang, A. S.; Julian, R. R. Analysis of Glutamine Deamidation: Products, Pathways, and Kinetics. *Anal Chem* **2019**, *91* (20), 13032–13038. https://doi.org/10.1021/acs.analchem.9b03127.
46. Badgett, M. J.; Boyes, B.; Orlando, R. The Separation and Quantitation of Peptides with and without Oxidation of Methionine and Deamidation of Asparagine Using Hydrophilic Interaction Liquid Chromatography with Mass Spectrometry (HILIC-MS). *J Am Soc Mass Spectrom* **2017**, *28* (5), 818–826. https://doi.org/10.1007/s13361-016-1565-z.
47. Badgett, M. J.; Boyes, B.; Orlando, R. Predicting the Retention Behavior of Specific O-Linked Glycopeptides. *J Biomol Tech* **2017**, *28* (3), 122–126. https://doi.org/10.7171/jbt.17-2803-003.

6 Chapter 6: General Conclusions

Proteomics relies on the application of robust and consistent chromatographic separations to identify the highest number of peptides and proteins, thereby generating higher quality data with better sequence coverage of the proteome. One problem in this case is the prevalence of false-positive identifications from peptide identification algorithms. Currently, scrutinization of MS data is the major avenue for filtering identification outputs to produce lower false-positive rates. However, retention time prediction models can be folded into these workflows, permitting another dimension of filtering capabilities and thereby increasing the quality of identification outputs. For PTMs, their naturally low abundance often precludes large-scale analyses and therefore the generation of high-quality retention models.

This thesis presents several developments which have been made in outlining chromatographic features of PTM carrying peptides, including the generation of larger scale modified peptide datasets from proteomics studies. All PTMs studied so far by the Krokhin group were summarized in Chapter 2 with generalized models, largely focusing on RPLC applications. In-depth analyses were performed for select PTMs, deamidations (Chapter 3) and glycosylations (Chapter 4), and focused mainly on RPLC as well, however alternative separation modes were also explored for 2D LC applications. The detailed studies of deamidations and glycosylations explored the effects of having modified residues at different positions in the peptide backbone, i.e. sequence-specific effects. Alterations of amphipathic helicity, ion-pairing properties at the N-terminus, and the nearest amino acid neighbours comprised major contributors to a PTMs characteristic behaviour in such cases.

The generation of a large-scale dataset was also performed to generate a retention time prediction model for ERLIC in Chapter 5, a novel addition to the SSRCalc family of peptide retention time prediction algorithms. This model elucidated the retention features of ERLIC as a combination of HILIC and the directional effects observed when employing anion-exchange sorbents for separations of tryptic digests. This was also used to evaluate the general characteristics of select PTMs for ERLIC applications in proteomics.

Ultimately, these models can be applied to proteomics studies to increase the confidence in peptide identification outputs. The future of proteomics is likely to shift more towards studying PTMs, and in many ways it already has with the advent and development of enrichment protocols. The ability of MS technologies to detect PTMs on peptides and proteins as mass offsets is one of

its most powerful applications. However, identifying PTMs comes with challenges as they can have drastically different mass spectrometric properties, such is the case for labile modifications like glycosylations. Applying LC models which carefully characterize the sequence-specific effects in these cases would aim to supplement the information lost from MS applications, thereby recovering the confidence in identification outputs while simultaneously filtering false positives.

The overarching research objective of this thesis was to increase the confidence in peptide identification outputs using chromatographic information to generate models. While a concrete example of this was not directly provided, an application of LC models can be extracted, starting from Chapter 2 using modified vs nonmodified peptide regression plots. These would be examples of coarse models which do not account for sequence-specific effects i.e. peptide primary and secondary structure. To integrate this model into proteomics workflows, the hydrophobicity of the nonmodified peptide would need to be known or predicted. The expected retention shift for a given modification could then be calculated using the slopes and intercepts of the coarse model. This would utilize manually adjusted or calculated cutoff values for observed retention shifts to filter potential false positives. In some cases, by applying the sequence-specific information provided in Chapters 3 and 4, some of the outliers from the regression plots can be explained, reducing the number of false negatives.

Ideally, the best model would be one with a sufficient number of identifications for a given modification such that retention times can be analyzed to introduce a new amino acid variable (amino acid #21, which carries the modification). In this case, sequence-specific retention coefficients (as explained in chapter 5) can be used to calculate predicted retention times directly. Since the sequence-specific effects are modelled discretely in this case, this would also provide better prediction accuracy compared to coarse models and permit a better investigation into understanding the retention mechanisms.

For future studies, the generation of enrichment protocols and larger retention datasets would be beneficial to elucidating other variables which govern a peptide's characteristic behaviours ultimately increasing the prediction accuracy of retention models.