Proteomic studies on protein N-terminus and peptide ion mobility by nano-scale liquid chromatography/tandem mass spectrometry

ナノスケール液体クロマトグラフィー/タンデム質量分析によるタンパク質 N 末端およびペプチドイオンモビリティーに関するプロテオミクス研究

2020

Chih-Hsiang Chang

CONTENT

ABBREVIATION

AA, ammonium acetate

ACN, acetonitrile

ANN, artificial neural network

BCA, bicinchoninic acid

BSA, bovine serum albumin

CAA, 2-chloroacetamide

CanNt-pepts, canonical protein N-terminal peptides

ChaFRADIC, charge-based fractional diagonal chromatography

CCS, collision cross section

COFRADIC, combined fractional diagonal chromatography

CZE, capillary zone electrophoresis

DMED, dulbecco's modified eagle medium

ESI, electrospray ionization

FDR, false discovery rate

GO, gene ontology

GOBP, gene ontology biological process

GOCC, gene ontology cellular component

GOMF, gene ontology molecular function

HPG-ALD, highly-branched polyglycerol polymer

HPLC, high performance liquid chromatography

HYTANE, hydrophobic tagging-assisted N-termini enrichment

IMS, ion mobility spectrometry

internal-pepts, internal peptides

ISP, intrinsic size parameter

KEGG, Kyoto encyclopedia of genes and genomes

LB, luria-bertani

LC, liquid chromatography

LS-MLR, length-specific multiple linear regression

MS, mass spectrometry

MW, molecular weight

NeoNt-pepts, neo-N-terminal peptides

PASEF, parallel accumulation–serial fragmentation

PTM, post-translational modification

PTS, phase-transfer surfactants

RP, reversed-phase

SCX, strong cation exchange chromatography

SDC, sodium deoxycholate

SDB-XC, styrene divinylbenzene

SSICalc, sequence-specific ion mobility calculator

SLS, sodium N-lauroylsarcosinate

SSRCalc, sequence-specific retention calculator

StageTip, stop and go extraction tip

T3, 3,3’,5-triiodo-L-thyronine

TAILS, terminal amine isotopic labeling of substrates

TCEP, tris(2-carboxyethyl)phosphine

TFA, trifluoroacetic acid

TIMS, trapped ion mobility spectrometry

TOF, time-of-flight

TrEMBL, translated EMBL nucleotide sequence data library

Tris-HCl, tris(hydroxymethyl)aminomethane hydrochloride

UniProtKB, universal protein resource knowledgebase

PREFACE

Proteomics provides insight into protein abundance, time-dependent expression patterns, post-translational modifications (PTMs), and protein-protein interactions, which can only be captured from proteins.[1,2] The combination of mass spectrometry and capillary liquid chromatography has revolutionized proteomics research, allowing for large-scale proteome profiling.[3] Despite the fact that LC/MS/MS analysis can identify thousands of proteins, the wide dynamic range of proteins and the high complexity of digested peptides make capturing the entire protein extremely challenging. In order to reduce the complexity of peptides, separation of peptides becomes a critical technique.[4] In general, there are two different approaches to reduce the complexity of peptides, the first is to reduce the complexity comprehensively to achieve overall proteome analysis,[4] and the second is to isolate the target, such as phosphorylated peptides,[5,6] glycosylated peptides.[7] In addition, improving the confidence of MS/MS-based identification can enhance proteome profiling by matching orthogonal features of peptide such as predicted retention time prediction[8] and predicted MSMS spectra,[9] to achieve a deeper analysis.

During my Ph D studies, I focused on the development of new approaches to achieve deep proteome profiling. In Chapter 1, the development of a new approach to enrich protein N-terminal peptides is described. This method is distinct from the past and without the complicated procedures and chemical reactions. By strong cation exchange chromatography combined with the new enzyme TrypN for digestion to isolate protein N-terminal peptides. This method was further applied to profiling the proteolytic procession during beige adipocytes maturation. In Chapter 2, deep proteome profiling by trapped ion mobility spectrometry (TIMS) is described. TIMS is a new type of ion mobility-based separation technology that can be integrated with LC/MS/MS to provide a new dimension of separation within the same analysis time. Using this comprehensive proteome dataset, I developed a predictive model for the collision cross section values of peptides and showed that the peptide structure in the gas phase is determined by the amino acid sequence of the peptide.

CHAPTER 1

Isolation of acetylated and unmodified protein N-terminal peptides by strong cation exchange chromatographic separation of TrypN-digested peptides

INTRODUCTION

Characterizing protein N-termini is essential to understand how the entire proteome is generated through biological processes such as translational initiation,[10-12] post-translational modifications[13,14] and proteolytic cleavages.[15,16] In order to perform N-terminomics using mass spectrometry (MS), peptides derived from protein N-termini must be selectively enriched, and many methods have been developed for this purpose.[17,18] Some of them use "positive selection" approaches in which chemically labeled protein N-terminal peptides are enriched by affinity purification.[15,19] However, these approaches are not applicable to proteins with *in vivo* N-terminal modifications. In contrast, "negative selection" approaches to isolate protein N-terminal peptides by depleting internal peptides have been used to comprehensively identify protein N-terminal peptides, including N-terminal modifications such as methylation, acetylation, and lipidation.[20,21] Gevaert et al. pioneered combined fractional diagonal chromatography (COFRADIC),[22] and this was followed by other negative selection approaches such as terminal amine isotopic labeling of substrates (TAILS),[23] the variant of COFRADIC called charge-based fractional diagonal chromatography (ChaFRADIC),[24] and hydrophobic tagging-assisted N-termini enrichment (HYTANE).[25] All of them require blocking of the primary amines at the protein level and depletion of digested internal peptides by means of chemical tagging-based separation. Thus, relatively large amounts of samples (~5 to 10 mg) are generally required to increase the identification number of protein N-terminal peptides. This limits the usefulness of these approaches in the case of hard-to-obtain biological samples.[26,27] Furthermore, limitations in the efficiency and specificity of the chemical derivatizations compromise the confidence of peptide identification. Therefore, a simple and sensitive approach to enrich protein N-terminal peptides is still needed for MS-based proteomics.

Strong cation exchange (SCX) chromatography, employing Coulombic interactions to separate peptides based on their charge at acidic pH, has been widely applied for deep proteome profiling.[28,29] Alpert et al. reported that the peptide retention in SCX is affected by charge and orientation.[30] In SCX separation of tryptic peptides, acetylated protein N-terminal peptides and protein C-terminal peptides are eluted first. Monophosphorylated peptides with +1 charge are then eluted, followed by peptides with +2 or more charge, such as unmodified protein N-terminal peptides, internal peptides and peptides containing missed cleavages.[31,32] Thus, it is impossible to isolate protein N-terminal peptides from tryptic peptides by SCX chromatography. This is also the case where LysN was used with SCX chromatography, based on the charge/orientation retention model.[30] To overcome this issue, we focused on TrypN, also known as LysargiNase, a metalloprotease that cleaves peptide chains mainly at the N-terminal side of Lys/Arg even in the case of Pro-Lys and Pro-Arg bonds, generating peptides with N-terminal Lys/Arg and yielding protein N-terminal peptides that do not contain Lys/Arg.[33] Unlike other kinds of LysargiNase such as ulilysin[34,35] and mirolysin,[36] which preferentially cleave the N-terminal side of either Lys or Arg, TrypN cleaves the N-terminal side of Lys and Arg equally at pH 6~8. Moreover, the peptide identification performance for N-terminal Lys/Arg peptides is comparable to that for tryptic peptides.[37]

In this chapter, we developed a new method to enrich protein N-terminal peptides without the need for chemical derivatization or complex procedures, taking advantage of the combination of proteinase TrypN-mediated protein cleavage and SCX separation of N-terminal peptides based on the extended charge/orientation retention model. We show that this rapid and simple approach to enrich protein N-terminal peptides enables comprehensive, high-throughput analysis of the human and bacteria N-terminomes. Finally, we applied this approach to the temporal N-terminome profiling during beige adipocyte maturation.

RESULTS AND DISCUSSION

Retention behavior of TrypN-digested peptides in SCX chromatography

Proteolysis with TrypN yields peptides with at least a +2 charge with Lys or Arg and an α-amino group at the peptide N-terminus at the acidic pH. On the other hand, peptides derived from protein N-termini have neither Lys nor Arg and are often acetylated at the protein N-terminus, so that most of them have a 0 or +1 charge, and only His-containing peptides with an unmodified protein N-terminus have a +2 charge (Figure 1-1). In this study, we focused on the fact that SCX chromatography under the acidic conditions might be able to separate peptides based on the number of positive charges as well as the localization of the charges according to the charge/orientation retention model,[30] and we attempted to separate protein N-terminal peptides from internal peptides among TrypN-digested peptides.

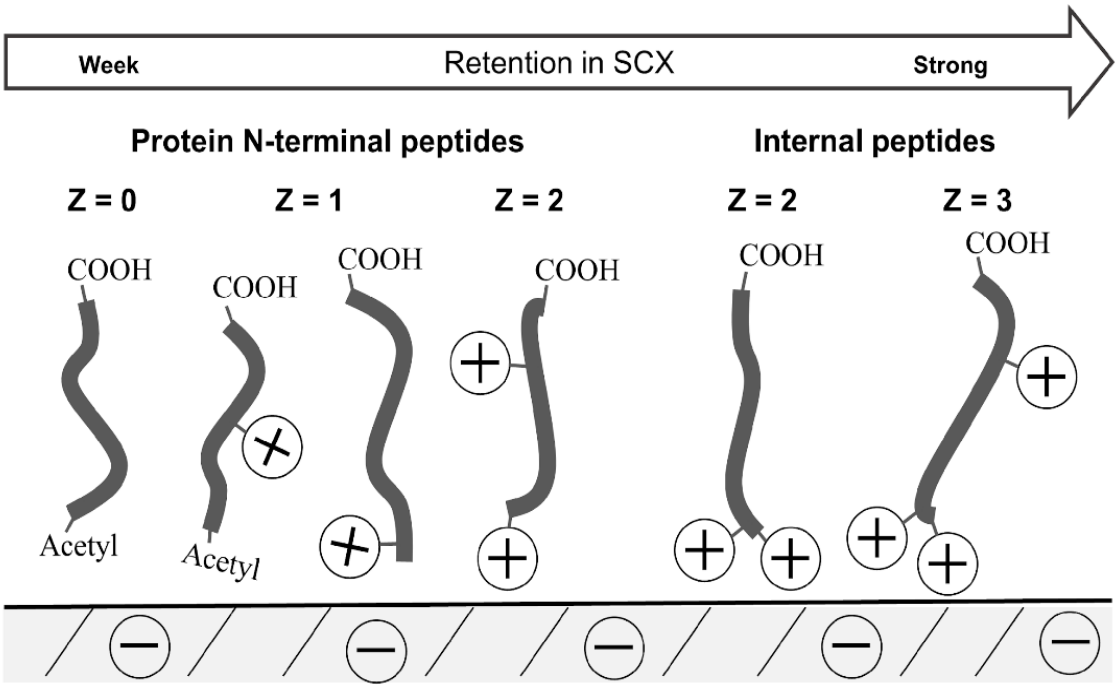

Figure 1-1. Charge/orientation-based SCX retention model for TrypN-digested peptides. Six types of TrypN-digested peptides are shown. Z is the charge number at acidic pH, which is based on the number of basic residues per peptide, such as unmodified N-terminus, Lys, Arg and His.

We first examined the number of missed cleavages in TrypN digestion. When digestion was performed in 0.1% RapiGest according to the manufacturer's protocol, the missed cleavage rate (the content of peptides with two or more missed cleavage sites) was 14%, almost equal to the value in the condition without addition of RapiGest (16%). On the other hand, when 1%

SDC was added instead of RapiGest, the missed cleavage rate was dramatically reduced to 5.8%. Thus, TrypN digestion was performed according to the PTS protocol in this study.[38]

Next, keeping in mind the need to separate protein N-terminal-derived peptides with both His residues and unmodified N-termini from TrypN-digested internal peptides, we investigated whether the peptides could be separated based on the position of the positive charge, in addition to the number of positive charges, by SCX chromatography. Studies with proteases that cleave either Lys/Arg, such as LysC and LysN or trypsin and TrypN, have indicated that the position of the positive charge affects the outcome in shotgun proteomics.[35,39] For example, it has been reported that peptides with N-terminal Lys or Lys/Arg are more strongly retained than peptides with a C-terminal Lys or Lys/Arg in reversed-phase liquid chromatography (RP-LC).[37] To determine how the Lys/Arg position of peptides affects their retention behavior in SCX chromatography at acidic pH, we examined a mixture of TrypN- and trypsin-digested peptides using the SCX HPLC system, followed by nanoLC/MS/MS. The 19,853 unique tryptic peptides generally showed weaker retention than the 11,334 unique TrypN peptides with the same charge states (Figure 1-2A, Supplemental Table 1). To characterize the SCX elution profiles in more detail, we compared the retention time in SCX HPLC for approximately 4,000 peptide pairs having sequences that differ only in the position of terminal Lys/Arg (Figure 1-2B). As expected, TrypN-digested peptides exhibit stronger SCX retention compared to Typsin analogs. This would be because the TrypN peptides carry two positively charged groups at the N-terminus, due to the α-amino group of the N-terminal Lys/Arg and the side chain ε-amino or guanidino group, whereas the positive charge of the C-terminal Lys/Arg of trypsin peptides was partially neutralized by the α-carboxy group under acidic condition (Figure 1-1). Alpert et al. and Gauci et al. reported that LysN-digested phosphopeptides with two basic moieties in close proximity tend to be more strongly retained on an SCX column than tryptic phosphopeptides.[30,32] Gussakovsky et al. reported a retention model for predicting the retention times in SCX chromatography of tryptic peptides, in which the position-dependent coefficient of basic amino acids is higher near the N-terminus.[40] We also found that the TrypN-digested peptides eluted in a narrower SCX fraction range than the tryptic peptides (Figure 1-2A). This may be due to the fact that the distance between N-terminal positive charge in the tryptic peptides differs depending upon the length of peptide, whereas the TrypN-digested peptides have the N-terminal Lys/Arg that minimizes the distance between the two positive charges. To our knowledge, the present work is the first to validate the peptide charge/orientation retention model in SCX using thousands of identical sequence pairs.

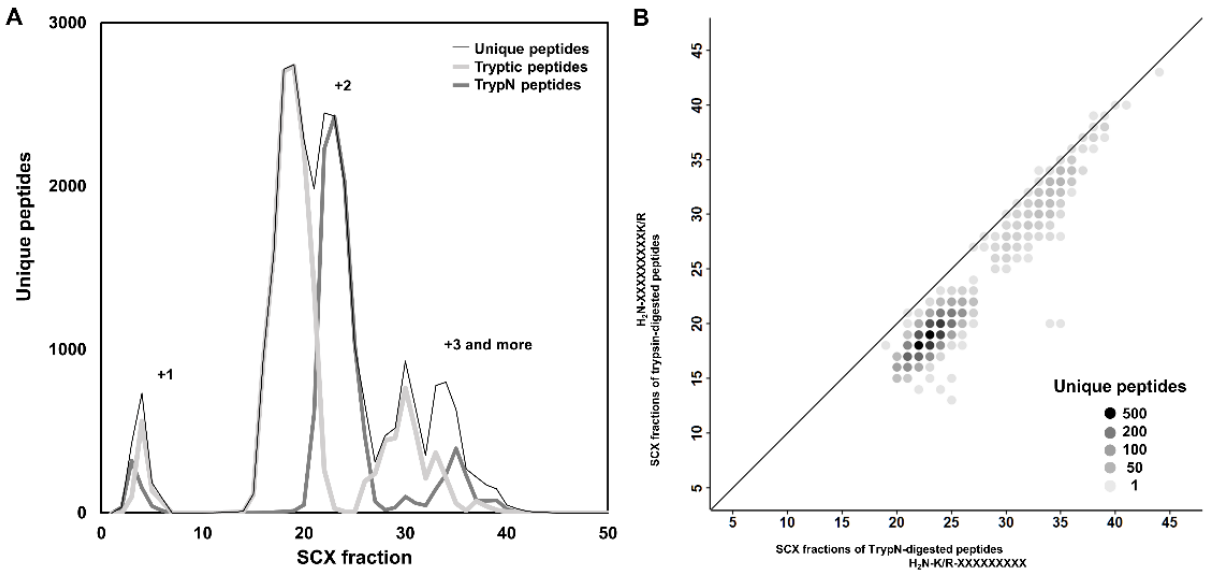

Figure 1-2. SCX separation of TrypN-digested and trypsin-digested peptides. (A) SCX separation of different types of peptides in digested HEK293T cell lysate. The peptide sample was prepared by mixing TrypN- and trypsin-digested HEK293T peptides with 1:1. The black, light gray and dark gray curves represent unique peptides, trypsin-digested peptides and TrypN-digested peptides, respectively. The charge number at acidic pH was labeled. (B) Comparison of SCX elution profiles of TrypN- and trypsin-digested peptides. Peptides with the same sequence except for the termini were selected (K/R-XXXXXX and XXXXXX-K/R for TrypN- and trypsin-digested peptides, respectively). The shade of the circle color indicates the number of peptides.

SCX HPLC separation of TrypN-digested HEK293T peptides

The HPLC system used in this study was equipped with a nonporous hydrophilic SCX column having a separation efficiency equivalent to that of a typical reversed-phase column (the peak width at half height was 12.4 ± 4.2 seconds and the peak capacity was 122; see Figure 1-3). As already shown in Figure 1-2A, this SCX HPLC system was able to separate TrypN-digested peptides with +1 and +2 charges from each other. Comprehensive SCX fractionation of TrypN-digested peptides derived from HEK293T cells was performed with a KCl salt gradient elution at pH 2.6, and peptide identification for each fraction was performed by nanoLC/MS/MS (Supplemental Table 2). As shown in Figure 1-4, nearly all of the protein N-terminal-derived peptides were clearly separated from the internal peptides, regardless of whether their N-termini were acetylated or unmodified. The fractions from 2 to 11 min contained mainly 0 and +1 peptides, including 2,207 acetylated protein N-terminal peptides, 345 His-containing acetylated N-terminal peptides, and 262 unmodified N-terminal peptides. The 12-18 min fractions contained +2 peptides, i.e., unmodified protein N-terminal peptides containing one His, Lys or Arg and acetylated protein N-terminal peptides containing two basic amino acids. The next fractions from 19 to 30 min also contained +2 peptides, but most of them were internal peptides based on the orientation effect, i.e., retention was stronger due to the high density of positive charge at the N-terminus of the peptides (Figure 1-2B). Thus, the protein N-terminal peptides can be easily isolated. Peptides with a charge greater than 2+ were sequentially eluted in the fractions after 31 min. These included protein N-terminal peptides containing missed cleavage sites, but their number was small due to the high efficiency of TrypN digestion by the PTS method. Up to 90% of non-redundant protein N-terminal peptides could be recovered in fractions up to 18 min by this approach (Figure 1-5), demonstrating that the combination of TrypN digestion with SCX HPLC enables simple and rapid protein N-terminal peptide enrichment. In addition, unlike trypsin, which is unable to cleave Lys-Pro and Arg-Pro bonds, TrypN can cleave Pro-Lys and Pro-Arg bonds, generating protein N-terminal peptides with Pro at the C-termini, and thus improving the coverage in the N-terminome analysis.

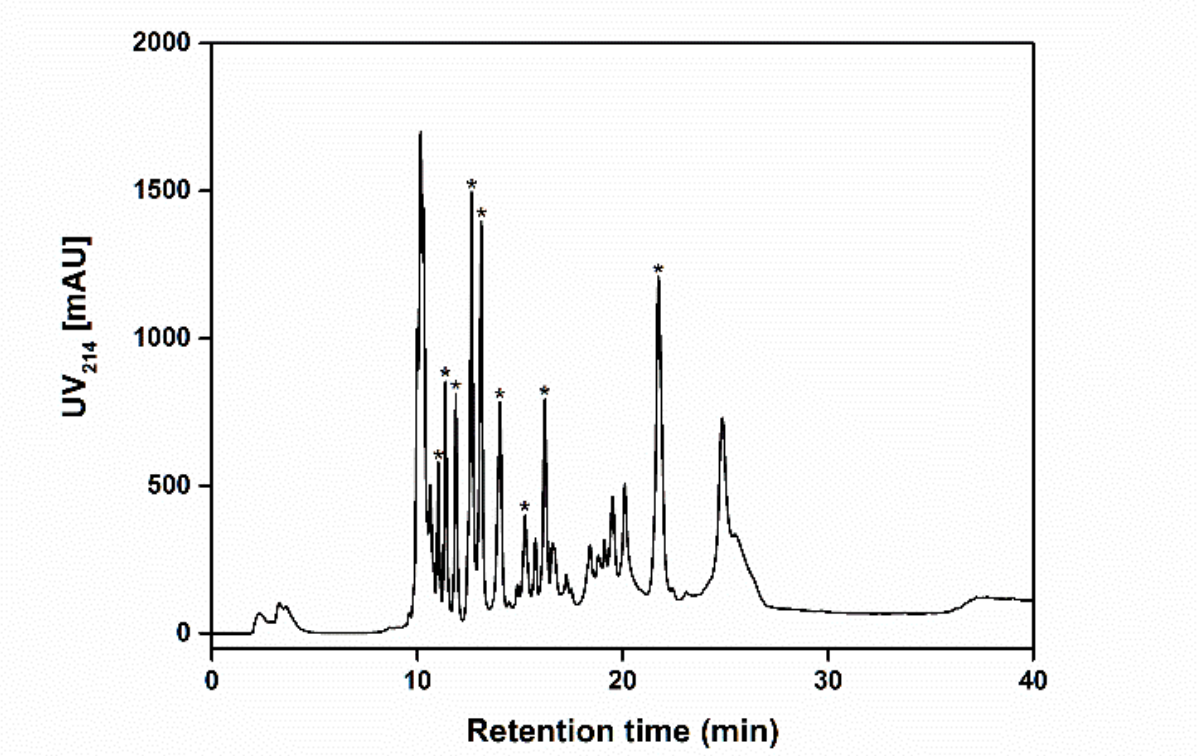

Figure 1-3. Typical chromatogram of high resolution SCX chromatography employed in this study. Tryptic HSA peptides (500 µg) were analyzed by using the Shimadzu Prominence HPLC system with Agilent BioIEX SCX column (250 mm × 4.6 mm inner diameter, 5 µm, non-porous beads made of poly(styrene-divinylbenzene) modified with sulfonate groups). A mixture of 7 mM sodium phosphate (pH 2.6) and ACN (7:3) was used for SCX buffer A, and 1 M NaCl was added to buffer A for SCX buffer B. The two-step linear gradient was performed as follows: 0% B for 5 min, 0-40% in 30 min, 40-99% in 5 min and 100% B for 10 min. Peak capacity was calculated using the following equation: PC = 1 + tg/W0.5. (W0.5 = 12.4±4.2 sec, PC = 122.) Peaks with asterisk were selected for measuring the half peak width (W0.5) to calculate peak capacity.

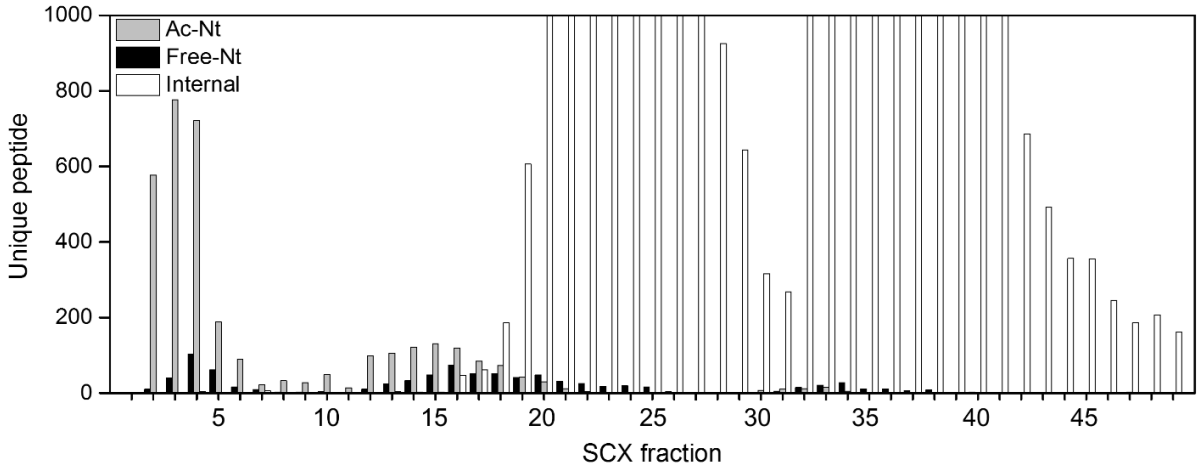

Figure 1-4. SCX elution profiles of TrypN-digested peptides. SCX HPLC fractionation of different types of peptides in TrypN-digested HEK293T cell lysates using KCl salt gradient elution. In fractions 2-6, protein N-terminal peptides with Z = 0 and 1 were observed, such as acetylated protein N-terminal peptides with or without one basic amino acid and unmodified protein N-terminal peptides without basic amino acid. Z is the charge number at acidic pH, which is based on the number of basic residues per peptide, such as unmodified N-terminus, Lys, Arg and His.

Since the charge number of the protein N-terminal peptides at acidic pH is generally smaller than that of internal peptides, we examined whether the identification efficiency of protein N-terminal peptides is affected by the low positive charge number. Since peptide identification is influenced by several steps, including ionization, ion transmission from MS1 to MS2, and fragmentation, four parameters such as the UV absorbance at 214 nm in SCX, the total ion currents in MS and MS/MS scans, and the Mascot peptide score distribution were measured for SCX 1-18 min fraction (protein N-terminal peptides were enriched) and 19-50 min fraction (internal peptides were enriched), respectively (

Table **1-1**). Considering that the UV absorbance ratio of the 1-18 min fraction to the 19-50 min fraction was smaller than the ratio of the average total ion current per MS scan, the ionization efficiency of protein N-terminal peptides was better than that of the internal peptides due to the lower sample complexity of the 1-18 min fraction. For ion transmission efficiency from MS1 to MS2, we did not see any difference between protein N-terminal peptides and internal peptides. As for fragmentation, the profiles of charge number distribution at acidic pH were significantly different between the protein N-terminal and internal peptides, but the obtained profiles of the score distribution were almost identical. This could be due to the similar distribution profiles of the charge states of the precursor ions. These results indicate that there is no clear disadvantage of using TrypN for the identification of the protein N-terminal peptides.

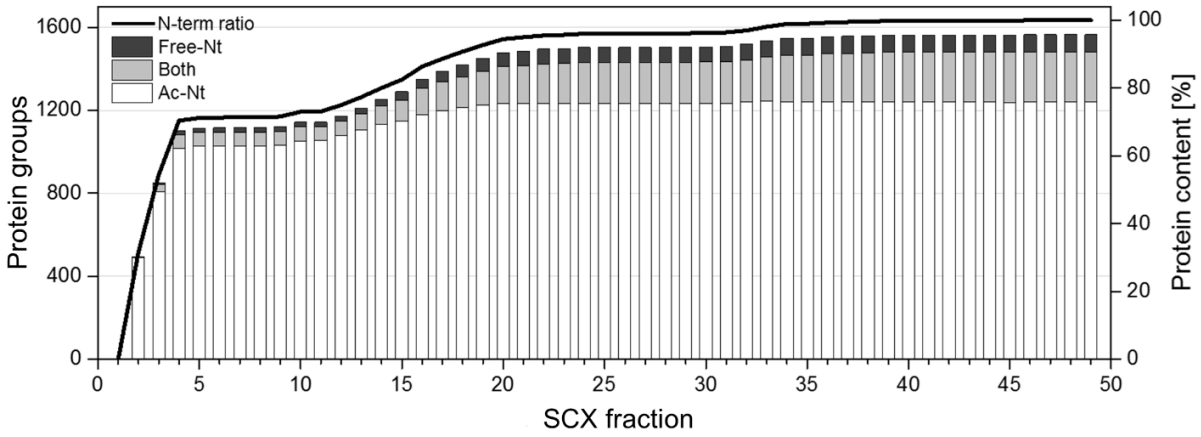

Figure 1-5. Accumulation of non-redundant protein groups based on protein N-terminal peptides during SCX elution. Sample: TrypN-digested HEK293T cell lysate. Free-Nt (black), both (gray) and Ac-Nt (white) squares in the bar graph represent unmodified, partially acetylated and acetylated protein N-termini, respectively. The curve represents the accumulated content of protein groups.

Table 1-1. Evaluation of the identification efficiency of TrypN-digested HEK293T peptides used in Figure 1-4.

	Total peak area in SCX (UV 214 nm)	Average total ion current per MS1 scan	Average total ion current per MS2 scan	Mascot peptide score (average ± SD)	Precursor ion charge state (1+/2+/3+/>3+)	Charge number at acidic pH (0/+1/+2/+3/>+3)
Protein N-terminal peptides (fr1-18)	5.8E+04	1.9E+08	3.0E+06	51 ± 26	1/86/11/2	21/41/37/1/0
Internal peptides (fr19-50)	3.5E+05	9.4E+08	1.5E+07	52 ± 24	0/62/33/4	0/0/52/28/20
Protein N-term/Internal ratio	0.17	0.20	0.20	0.98	-	-

Optimization of SCX separation using TrypN-digested *E. coli* peptides

To optimize the elution conditions for isolation of protein N-terminal peptides, we employed *E. coli* TrypN-digested peptides. Because bacterial proteins have less N-terminal modification than mammalian proteins, the bacterial sample was considered preferable to optimize the conditions for separating the protein N-terminal peptides with +2 charge (peptides with an unmodified N-terminus and one His residue) from the internal peptides (Figure 1-1). Three SCX buffers with different KCl concentrations were used for isocratic elution for 30 min, and the enrichment efficiencies for protein N-terminal peptides were compared (Supplemental Table 3). An enrichment specificity of more than 97% was obtained with 10 mM KCl (Table 1-2). When buffers with higher KCl concentrations were used, more internal peptides were identified (Figure 1-6A). In the case of 10 mM KCI buffer, we identified 53 His-containing protein N-terminal peptides out of 270 non-redundant protein N-terminal peptides without missed cleavage from 20 μg of *E. coli* lysate (19.6%, Figure 1-6B). Among *in silico* TrypN-digested peptides from the *E. coli* proteome, 20% of the protein N-terminal peptides contain one His, suggesting that our enrichment conditions have no bias in identifying His-containing protein N-terminal peptides. In other words, this SCX chromatography was able to isolate the protein N-terminal peptides from TrypN-digested *E. coli* peptides even in the most difficult cases where the unmodified protein N-terminal peptides contain an additional basic amino acid such as His, Lys or Arg near the N terminus (Figure 1-7). Although this SCX separation can be explained by the charge/orientation model, it is the first report to apply the retention model to N-unmodified protein N-terminal peptides.

Table 1-2. Enrichment of *E. coli* protein N-terminal peptides by SCX HPLC with isocratic elution at three different KCl concentrations.

Salt concentration	10 mM	12.5 mM	15 mM
Unique peptides	432	1,669	3,416
Unmodified protein N-terminal peptides	326	444	387
Acetylated protein N-terminal peptides	31	25	22
Protein N-terminal peptides (%) based on peptide counts	82.6	28.1	12.0
Protein N-terminal peptides (%) based on peak area	98.2	49.2	18.5

The enrichment specificity of protein N-terminal peptides is obtained by calculating the number and LC/MS peak area of protein N-terminal peptides among all identified peptides.

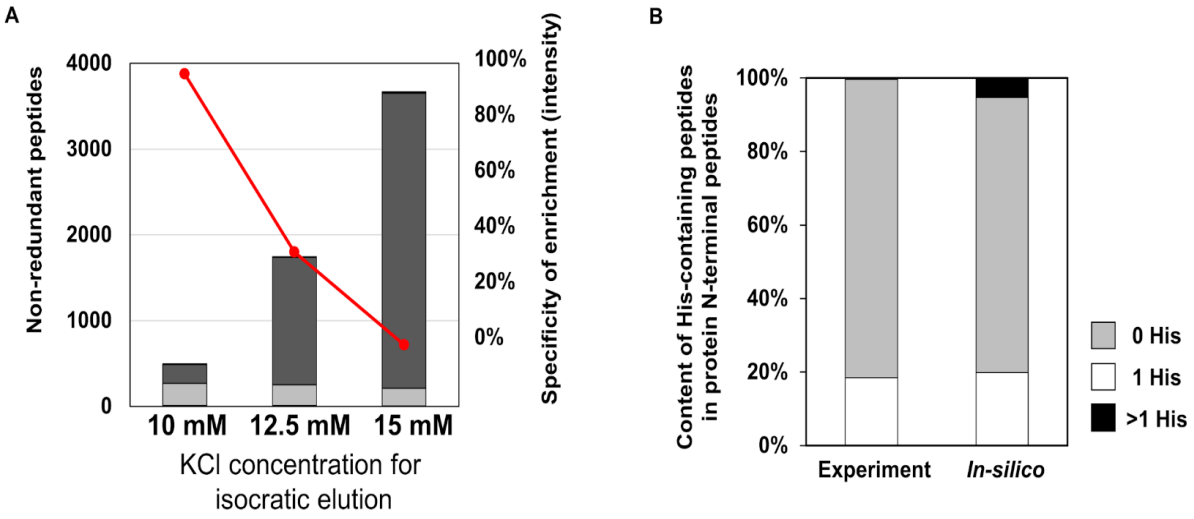

Figure 1-6. SCX elution profiles of TrypN-digested peptides. (A) SCX HPLC separation of TrypN-digested *E. coli* peptides under isocratic conditions with different KCl concentrations. The specificity in enriching protein N-terminal peptides and the numbers of peptides with different Z values are shown as a red curve. (B) The content of His-containing peptides in protein N-terminal peptides. The experimental results were those obtained with 10 mM KCl isocratic elution, and the *in-silico* results were calculated from the *E. coli* K-12 MG1665 protein sequence database. Details of *in-silico* digestion are described in the experimental section.

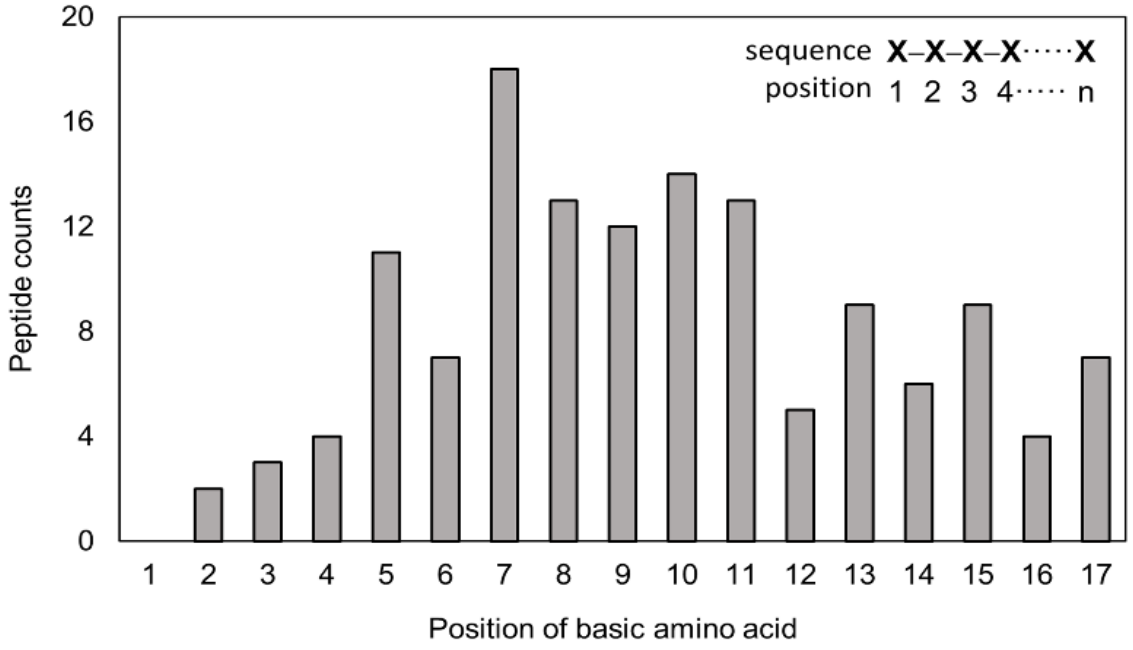

Figure 1-7. Dependence of the localization of basic amino acids (K, R, or H) on isolating protein N-terminal peptides by SCX with 10 mM KCl isocratic elution. Unmodified protein N-terminal peptides with +2 charge (137 peptides) isolated from TrypN-digested *E. coli* cell lysate were plotted.

HEK293T protein N-terminal peptide enrichment by TrypN-SCX approach

The N-terminal peptides of His-containing proteins could be successfully separated from the internal peptides of human and bacterial samples by SCX HPLC under optimized elution conditions. To validate the applicability of this method to large-scale N-terminal proteomics, we performed triplicate analyses using HEK293T cells, which have been widely used in N-terminal proteomics.[26,27] Triplicate SCX HPLC fractionations using 10 mM KCl isocratic elution were done for TrypN-digested HEK293T peptides (80 μg each time), and we subjected one-fourth of the isolated peptides to nanoLC/MS/MS in triplicate (9 runs in total). Default parameters, such as the Swiss-Prot human protein sequence database, specific TrypN cleavage, and minimum peptide length of 7 amino acids, were applied for peptide identification by database search (Figure 1-8A). The results are shown in Figure 1-8B, Table 1-3 and Supplemental Table 4. High correlations of peak areas of identified peptides were observed for intra- and interday preparation samples, (R^2 = 0.96 and R^2 = 0.75, 0.80, respectively). On average, we identified 1,550 unique acetylated and 200 unmodified protein N-terminal peptides from 20 μg of TrypN-digested HEK293T peptides in a single LC/MS/MS analysis. Contamination by internal peptides amounted to only 3% and 9% in peptide peak area and peptide number, respectively (Figure 1-8C, Table 1-3). Protein N-terminal peptides with missed cleavage were also enriched in the same elution, and 850 (~50%) miscleaved unique N-terminal peptides were identified on average, improving the coverage of the N-terminome. We identified 1,640 acetylated, 106 partially acetylated and 167 unmodified non-redundant protein N-termini. Note that 1,600 additional neo-N-terminal peptides were identified when semi-specific cleavage at the N-terminus was allowed in the data processing, although our purpose in this study was not to find novel proteoforms, but to establish a novel approach for N-terminomics. Furthermore, to compare our results with two published N-terminome datasets for HEK293T human cells,[17,18] we re-analyzed those datasets under the same conditions without the use of their original customized database or non-specific cleavage. In terms of the contents of acetylated and unmodified protein N-terminal peptides, all three datasets provided identical results, whereas the content of internal peptides as well as the number of unique peptides varied depending on the approach and the sample amount (Figure 1-9).

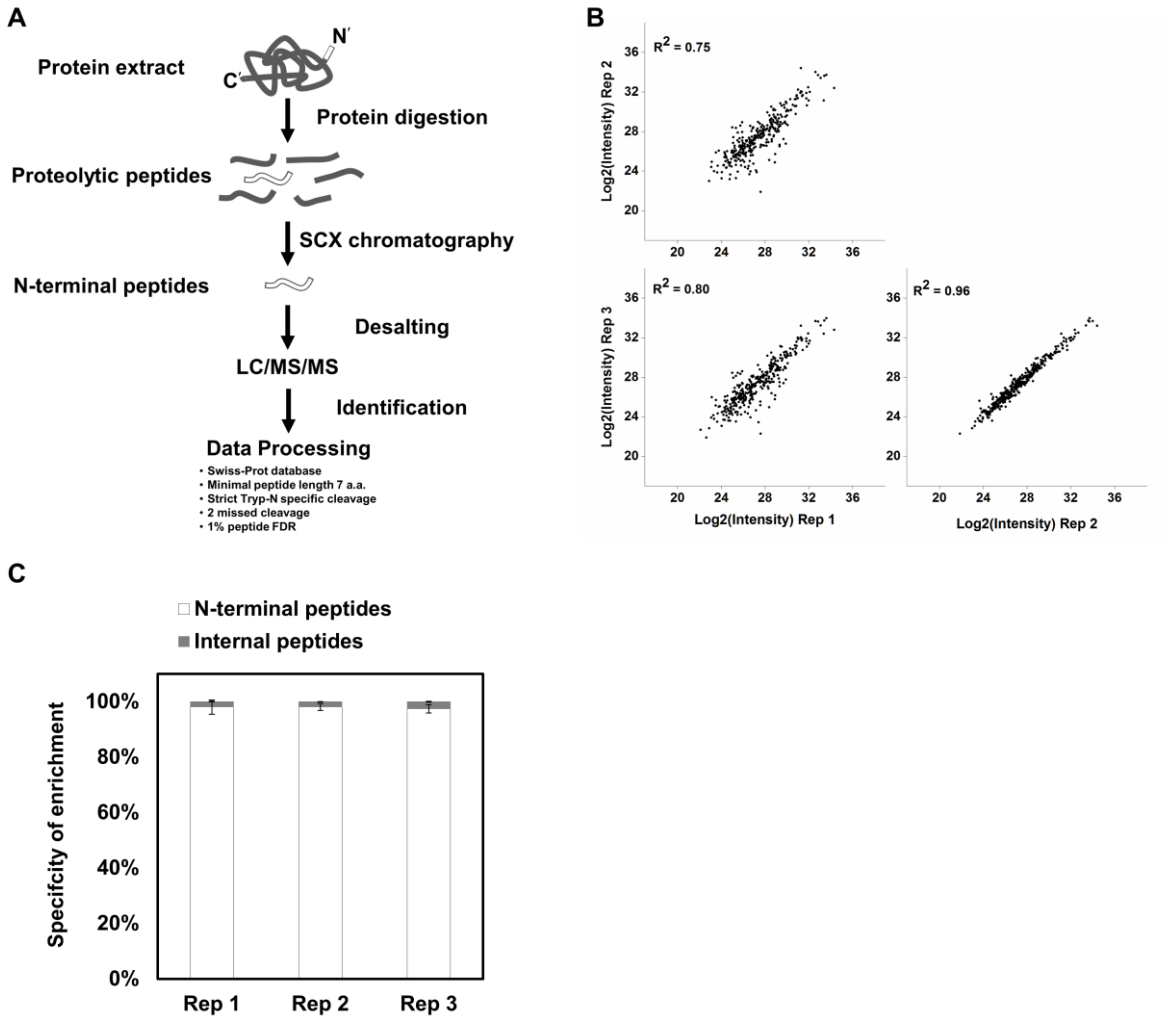

Figure 1-8. N-terminal proteomics using SCX chromatography. (A) Workflow of TrypN/SCX-based N-terminal proteomics. For details, see Materials and Methods. Three technical replicates were conducted on Day 1 (Rep 1) and Day 2 (Rep 2 and Rep 3) to evaluate inter- and intraday reproducibility. (B) Reproducibility in quantifying peptide peak areas between three technical replicates. Rep 1-Rep 2 and Rep 1-Rep 3 correlation represent interday reproducibility whereas Rep 2-Rep 3 correlation shows intraday reproducibility. (C) Enrichment specificity for protein N-terminal peptides in three technical replicates.

Table 1-3. Identification of protein N-terminal peptides from TrypN-digested HEK293T cells.

	Replicate 1	Replicate 2	Replicate 3	Total
Unmodified protein N-terminal peptides	199 (±3)	187(±5)	197 (±3)	352
Acetylated protein N-terminal peptides	1,854 (±13)	1,301 (±18)	1,509 (±15)	2,666
Internal peptides	160 (±8)	147 (±3)	232 (±6)	433
N-term ratio (%, peptide counts)	92.8 (±0.3)	91.0 (±0.2)	88.0 (±0.3)	
N-term ratio (%, peptide area)	97.4 (±0.3)	98.0 (±0.5)	97.4 (±0.5)	
Unmodified protein groups	115(±3)	100 (±4)	116 (±3)	167
Partially acetylated protein groups	36 (±2)	60 (±3)	50 (±2)	106
Acetylated protein groups	1,223 (±7)	1,000 (±9)	1,187 (±16)	1,640

Samples were prepared in triplicate (Replicates 1-3) and nanoLC/MS/MS of each sample was conducted in triplicate. Each number in the table is the average of triplicate measurements, and the total number is calculated after merging all results (n=9) and removing redundancy. The enrichment specificity of protein N-terminal peptides is obtained by calculating the number and peak area of protein N-terminal peptides among all identified peptides.

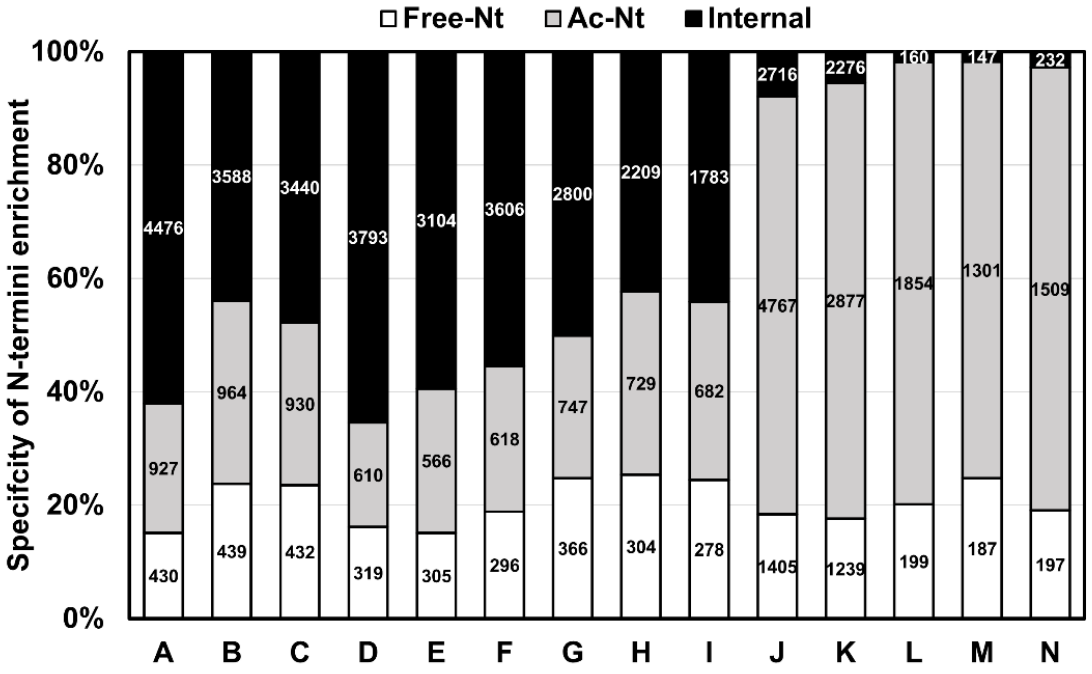

Figure 1-9. Comparison of the enrichment efficiency for protein N-terminal peptides from HEK293T cell lysate between published datasets and this study. The published raw files were downloaded from PRIDE, reanalysis by Mascot 2.6.1 with Swiss-Prot database (version 2017_04, 20,199), and peptides were considered identified if the false discovery rate is set to be less than 1% at peptide level. Specificity was calculated by the peptides area, and the numbers in the bar chart show the number of non-redundant peptides. Entry A-I depleted internal peptides by NHS-beads.[27] Entry J and K depleted internal peptides by HPG-ALD.[26] Entry L, M and N are the triplicate results in this article. Ac-Nt, Free-Nt and Internal represent acetylated protein N-terminal peptides, unmodified protein N-terminal peptides and internal peptides, respectively.

N-terminome profiling of beige adipocyte maturation

Beige adipocytes have the reversible thermogenic capacity with increasing Ucp1 expression in response to environmental stimuli.[41-43] Beige adipocytes acquire thermogenic function during maturation. And their maturation, including changes in morphology and functions, remodeling of the extracellular matrix and organelles turnover, are associated with the proteolytic process.[44] To inspect the post-translational events related to beige adipocyte maturation, we analyze the global proteome and N-terminome to profile proteases and proteolysis products. We induced the differentiation of the preadipocyte cells isolated from the inguinal white adipose tissue into beige adipocytes and harvested the cells every two days. For bottom-up proteome and N-terminome profiling, cell lysates from five time points were subjected to TrypN digestion. N-terminal peptides of the proteolysis products were isolated using the N-terminomics approach based on our TrypN-SCX strategy.[45] In this case, StageTip-based SCX chromatography was used for this large-scale analysis (Figure 1-10A).

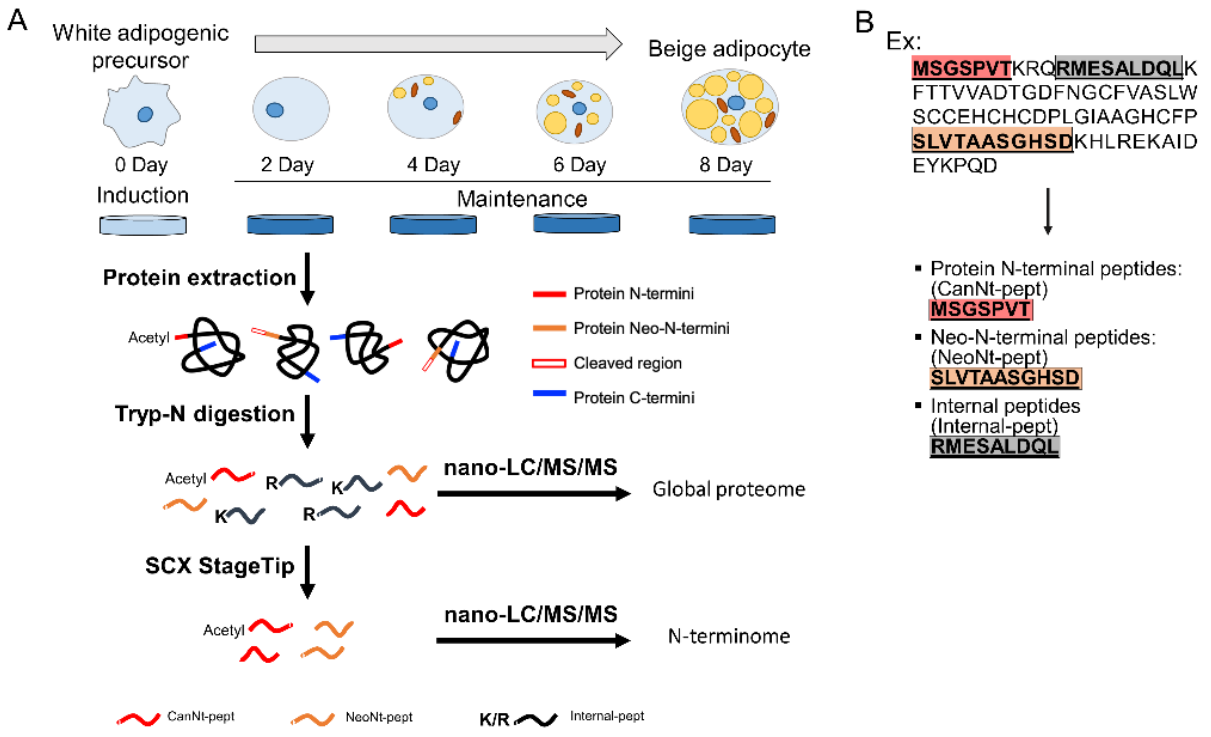

Figure 1-10. (A) Workflow of large-scale N-terminome profiling in beige adipocytes maturation. (B) Peptides annotation in N-terminome dataset.

The N-terminal peptides were identified with semi-specific cleavage at both N- and C-termini by Mascot or only at N-termini by MaxQuant with a threshold of 1% sequence FDR. N-terminome dataset was mainly classified into three categories: the canonical protein N-terminal peptides (CanNt-pepts), the neo-N-terminal peptides (NeoNt-pepts), and the internal peptides (internal-pepts). CanNt-pepts are defined as peptides in which the N-terminus matches the first or second amino acid residue of the N-terminus of the canonical protein. NeoNt-pepts are defined as peptides with a semicleaved N-terminus that may be generated from endopeptidases or exopeptidases. In NeoNt-pepts, several peptides were proximally truncated. This means that they have the same sequence at the C-terminus, but sequentially lost amino acids from the N-terminus. These peptides were most likely cleaved by endopeptidases and then by exopeptidases (proteolytic peptides). The third category was internal peptides with specific TrypN cleavage sites at both the N- and C-termini, producing peptides containing N-terminal lysine or arginine (Figure 1-10B). In the N-terminome dataset, 3,016 CanNt-pepts, 4,225 NeoNt-pepts and 381 internal peptides were identified (Supplemental Table 5). For peptide identification, we adopted a conservative criterion of only accepting peptides commonly identified by Mascot and MaxQuant searches, resulting in 2,124 CanNt-pepts and 1,301 NeoNt-pepts for quantitative analysis (Figure 1-11). Note that the number of NeoNt-peptides in quantitative analysis was without proximal truncation. In the proteome dataset, we identified 3,081 proteins, of which 3,032 were quantifiable (Figure 1-11, Supplemental Table 6). Gene Ontology (GO) and KEGG pathway enrichment analyses of CanNt-pepts, NeoNt-pepts, and proteins with the different temporal profiles (Figure 1-11) were performed using DAVID Bioinformatics Resources.[46,47] As a result, mitochondria, extracellular matrix, and metabolic functions were enriched in the increased cluster (Figure 1-12B, D, F), and extracellular matrix and cell adhesion were enriched in the decreased cluster (Figure 1-12A, C, E). These results indicate that the remodeling of extracellular matrix and mitochondrial biogenesis were the crucial signatures for beige adipocytes maturation.[48-50]

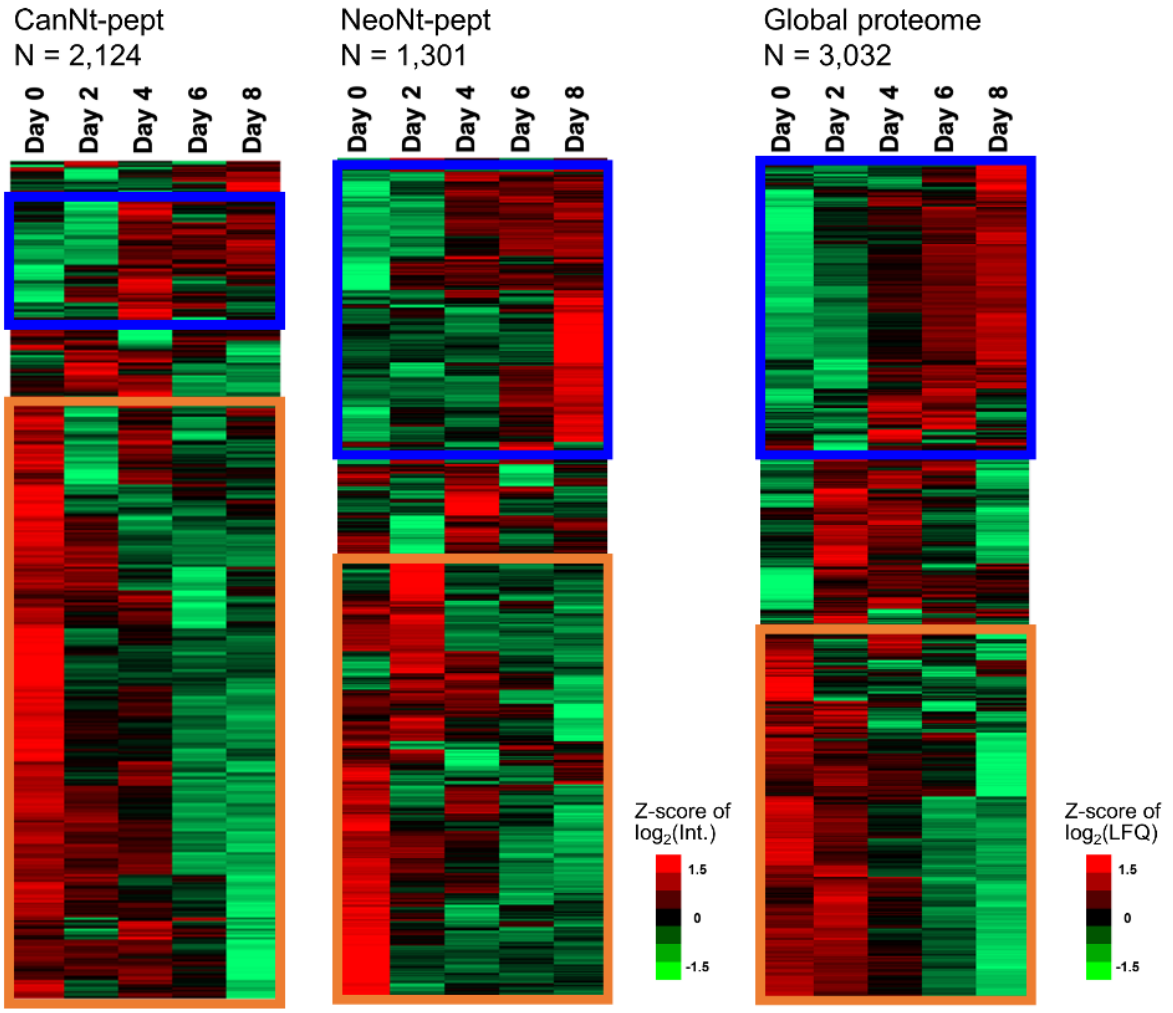

Figure 1-11. Heatmap of dynamic profiling in CanNt-pepts, NeoNt-pepts and global proteome. Increasing peptides and proteins were marked by blue box and orange showed decreasing peptides and proteins. Two clusters were functional annotation by DAVID.

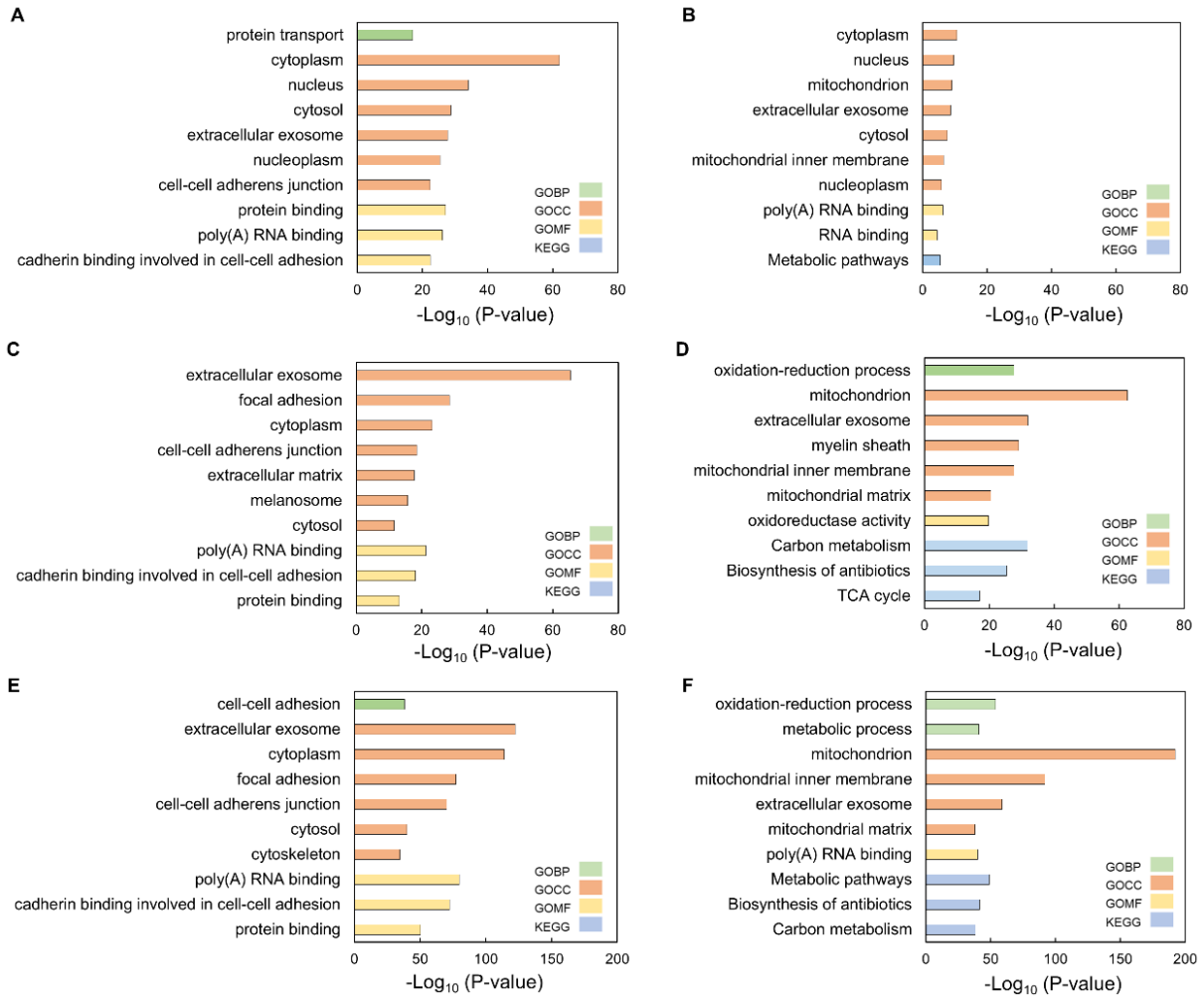

Figure 1-12. Gene Ontology (GO) and KEGG enrichment analysis in N-terminome and proteome. (A, B) Enriched functions and pathways for decreasing (A, orange box in heatmap) and increasing (B, blue box in heatmap) CanNt-pepts. (C, D) Enriched functions and pathways for decreasing (C) and increasing (D) NeoNt-pepts. (E, F) Enriched functions and pathways for decreasing (E) and increasing (F) proteins from global proteome.

Four clusters were selected from temporal profiles of beige adipocytes during maturation of NeoNt-pepts, which are considered to be the products of proteolysis by endogenous proteases (Figure 1-13A). In each cluster, eight terminal amino acids (four amino acids each at the N- and C-termini of the cleavage site) were selected as the cleavage sequence to construct a cleavage motif to compare with the cleavage recognition sequence of a specific protease. In the first cluster, indicated by the brown box in Fig. 1-13A, NeoNt-pepts increased after day 2 and remained at a high level. The cleavage preference showed a preference for basic amino acids at the -3 and -2 positions. In the cluster 2 (Figure 1-13A, blue), NeoNt-pepts increased after day 6 and had a trypsin-like cleavage preference. On the other hand, 15 proteases with known cleavage preference in the peptidase database MEROPS[51] were included in the quantitative

dataset of the global proteome (Figure 1-13B). By matching of temporal profiles between the cluster 2 and these proteases, cathepsin D (Ctsd), neprilysin, mitochondrial-processing peptidase subunit beta (Pmpcb) and plasmin (Plg) were selected as candidates as responsible proteases for NeoNt-pepts in the cluster 2. The cleavage preference of Pmpcb and Plg also supported the prediction. To confirm the prediction, Pmpcb, Ctsd and Plg were selected and siRNA experiments were conducted during beige adipocyte differentiation through the collaboration with Dr. Hsin-Yi Chang (Taipei Medical University). As shown in Figure 1-13C, by knocking down the genes of these proteases, adipogenic makers were enhanced and thermogenic makers were diminished, as expected.[52]

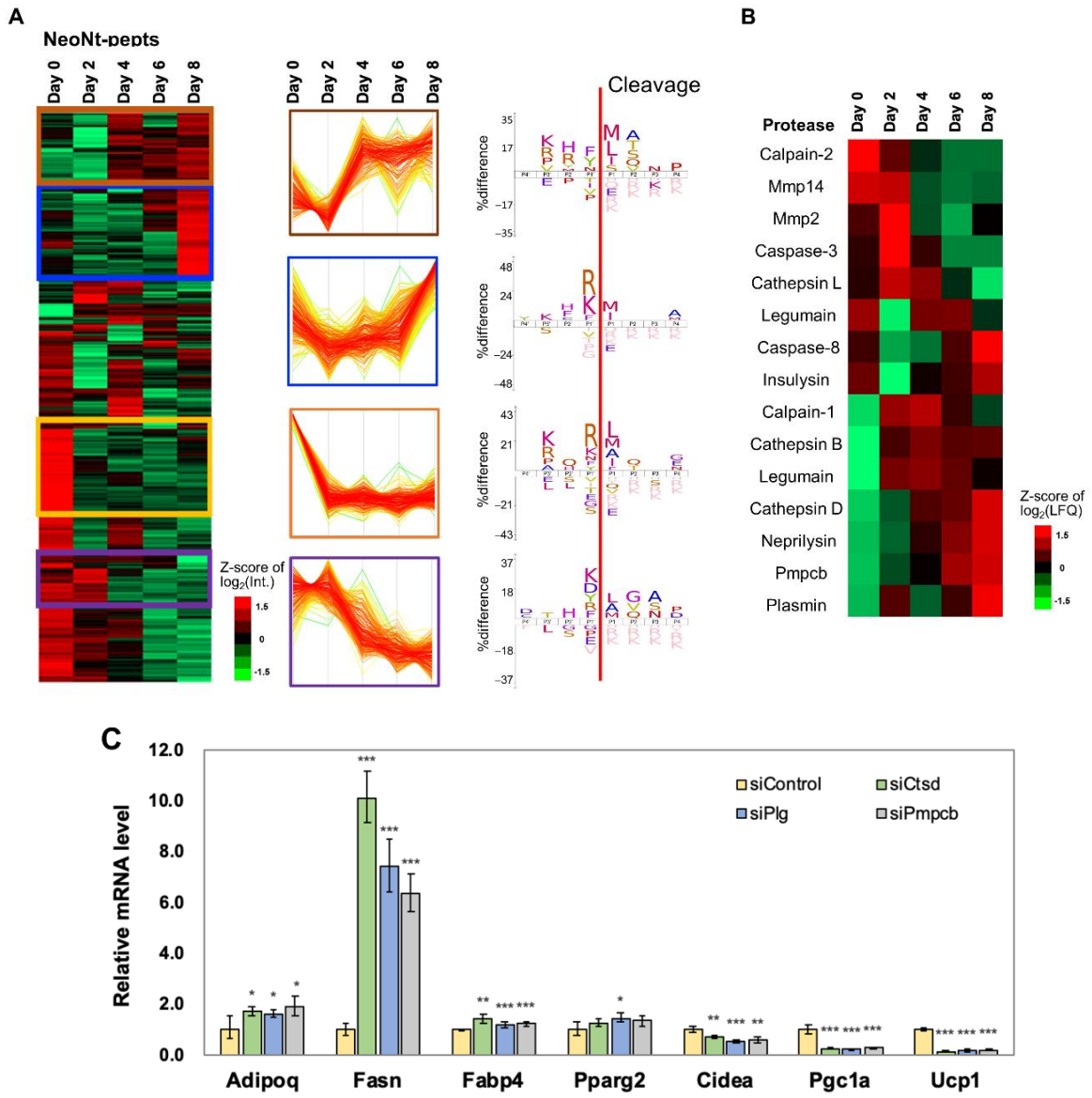

Figure 1-13. (A) Cleaved sequence logo for the selected cluster. The sequence logos were generated by iceLogo[53] and *Mus musculus* protein sequences were selected as background. (B) Dynamics of proteases were profiled from global proteome. (C) The mRNA expression of selected adipogenic makers and thermogenic makers by siRNA knockdown (*P<0.05; **P<0.005; ***P<0.001).

CONCLUSION

We have succeeded in developing a new N-terminomics method that does not require chemical reactions. This simple and rapid approach is suitable for high-throughput screening with minimal sample amounts. Our TrypN-SCX N-terminomics can enrich protein N-terminal peptides without bias, including peptides containing basic amino acids, with or without N-terminal modifications. This TrypN-SCX approach has great potential for expanding N-terminomics.

We also reported one application example to analyze the temporal N-terminome profiles during beige adipocytes maturation. Using peptide identification search engines with semi-cleavage specificity, we successfully identified peptides cleaved by endogenous proteases. By integrating the temporal profiles of the proteolytic peptides with the temporal profiles of proteases and their cleavage specificity, we can predict the proteases responsible for the proteolytic peptides. Knocking down the genes of candidate proteases resulted in the interruption of the beige adipocyte maturation, indicating that this prediction approach successfully identified the proteases required for thermogenesis in beige adipocytes maturation.

EXPERIMENTAL SECTION

Materials

Ammonium bicarbonate, tris(hydroxymethyl)aminomethane hydrochloride (Tris-HCl), sodium deoxycholate (SDC), sodium N-lauroylsarcosinate (SLS), ammonium bicarbonate, tris(2-carboxyethyl)phosphine (TCEP), 2-chloroacetamide (CAA), calcium chloride, ethyl acetate, acetonitrile (ACN), acetic acid, trifluoroacetic acid (TFA) and other chemicals were purchased from Fujifilm Wako (Osaka, Japan). RapiGest was purchased from Waters (Milford, MA). Dulbecco's Modified Eagle Medium (DMEM) from Gibco (Carlsbad, CA). Rosiglitazone, indomethacin, dexamethasone, 3-Isobutyl-1-methylxanthine and 3,3',5-triiodo-L-thyronine (T3) were purchased from Sigma-aldrich (St. Louis, MO). Modified trypsin was from Promega (Madison, MA). TrypN was from Protifi (Huntington, NY). Styrene divinylbenzene (SDB-XC) EmporeTM disk was purchased from GL Sciences (Tokyo, Japan). Water was purified by a Millipore Milli-Q system (Bedfold, MA).

Cell culture and protein extraction

HEK293T (human embryonic kidney) cells were cultured to 80% confluence in 10-cm diameter dishes. *Escherichia coli* K-12 BW25113 cells were grown to mid-log phase in LB broth with vigorous shaking at 37°C. To differentiate beige adipocytes, the preadipocytes from the inguinal white adipose tissue were incubated in induction medium consisting of 0.5 μM rosiglitazone, 125 μM indomethacin, 2 μg/ml dexamethasone, 0.5 mM 3-Isobutyl-1-methylxanthine, 5 μg/ml insulin, and 1 nM T3 in DMEM. After 2 days of induction, medium was renewed with maintain medium (DMEM containing 0.5 μM rosiglitazone, 5 μg/ml insulin, and 1 nM T3) every 2 days. Cells were fully differentiated into mature fat cells about 6 days after adding the induction medium. The preadipocytes were cultured as the controlled time point: after reaching to 80% confluence (day 0), after two-day induction (day 2), every two days during the maturation under the maintain medium (day 4, 6 and 8). These cells were collected by centrifugation and resuspended in the PTS lysis buffer containing protease inhibitors (Sigma), 12 mM SDC, 12 mM SLS, 10 mM TCEP, 40 mM CAA in 100 mM Tris buffer (pH 8.5).[38,54] The lysate was vortexed and sonicated on ice for 20 min. The concentration of protein crude extract was determined by means of bicinchoninic acid (BCA) protein assay (Thermofisher Scientific, Rockford, IL).

Protein Digestion

For optimization of TrypN digestion conditions, protein pellets were prepared by methanol/chloroform precipitation as described previously,[32] and were dissolved with 0.1% RaipGest in the buffer consisting of 25 mM trimethylammonium acetate, 2 mM $CaCl_2$ and 0.1 mM $MnCl_2$ at pH 7.4, followed by TrypN digestion overnight at 55°C according to the manufacturer's protocol. The PTS buffer[38] or the urea buffer consisting of 1 M urea, 25 mM trimethylammonium acetate, 2 mM $CaCl_2$ and 0.1 mM $MnCl_2$ at pH 7.4 instead of the RapiGest buffer was also used for the TrypN digestion.

For TrypN digestion after optimization, the cell lysate in the PTS buffer was diluted 10-fold with 10 mM $CaCl_2$ and digested with TrypN (1: 50 w/w) overnight at 37 °C. Note that TrypN can be replaced with LysargiNase (Merck Millipore, Darmstadt, Germany). In the case of tryptic digestion, the protein solution was digested with LysC (1:50 w/w) for 3 h at 37 °C, followed by 5-fold dilution with 50 mM ammonium bicarbonate and trypsin digestion (1:50 w/w) overnight at 37 °C. After enzymatic digestion, an equal volume of ethyl acetate was added to each sample solution, and the mixture was acidified with 0.5% trifluoroacetic acid (TFA, final concentration) according to the PTS protocol reported previously.[38] The resulting mixture was shaken for 1 min and centrifuged at 15,700 g for 2 min to separate the ethyl acetate layer. The aqueous layer was collected and desalted by using StageTips with SDB-XC disk membranes (SDB-StageTip).[55] The peptides were quantified by LC-UV at 214 nm using BSA digest as a standard and kept in 80% ACN and 0.5% TFA at −20 °C until use.

Peptide fractionation by SCX HPLC

SCX chromatography was performed using a Prominence HPLC system (Shimadzu, Kyoto, Japan) with a BioIEX SCX column (250 mm × 4.6 mm inner diameter, 5 μm non-porous beads made of poly(styrene-divinylbenzene) modified with sulfonate groups (Agilent, Santa Clara, CA).

For examination of the SCX separation characteristics, 75 μg each of trypsin- and TrypN-digested HEK293T peptides were mixed and directly loaded onto the SCX column at 0.8 mL/min. A mixture of 5 mM potassium phosphate (pH 3.0) and ACN (7:3) was used as SCX buffer A, and a mixture of 500 mM potassium phosphate (pH 3.0) and ACN (7:3) was used as SCX buffer B. Gradient elution was performed as follows: 0% B for 5 min, 0-10% in 20 min, 10-50% in 10 min, 50-100% in 5 min and 100% B for 4 min. Fractions were manually collected at one min intervals for 45 min. After evaporation of the solvent in a SpeedVac SPD121P (Thermofisher Scientific), fractionated peptides were resuspended in 50 μL of 0.1% TFA and

desalted by using SDB-StageTips. One-fourth of each fraction was analyzed by nanoLC/MS/MS using a TripleTOF 5600 (SCIEX, Foster City, CA) as described below.

For gradient SCX fractionation of TrypN-digested HEK293T peptides, 80 μg of digested peptides were analyzed using the SCX HPLC system described above. A mixture of 7.5 mM potassium phosphate (pH 2.6) and ACN (7:3) was used as SCX buffer A, and 350 mM KCl was added to buffer A for SCX buffer B. Gradient elution was performed as follows: 0.5% B for 15 min, 0.5-1% B in 10 min, 1-4% B in 10 min, 4-10% B in 3 min, 10-100% B in 3 min, and 100% B for 5 min. Fractions were manually collected at one min intervals for 50 min. The fractionated peptides desalted by using SDB-StageTips as described above. One-fourth of each fraction for Fr.1-43 and one-tenth of each fraction for Fr.44-50 were analyzed by nanoLC/MS/MS using an Orbitrap Fusion Lumos mass spectrometer (Thermofisher Scientific) as described below.

Enrichment of protein N-terminal peptides by SCX HPLC with isocratic elution

Enrichment of protein N-terminal peptides from 30 μg of TrypN-digested *E. coli* peptides was performed using the SCX HPLC system under the following isocratic conditions: SCX buffer A was a mixture of 7.5 mM potassium phosphate solution (pH 2.2) containing 10, 12.5 or 15 mM KCl and ACN (7:3), and buffer B was a mixture of 7.5 mM potassium phosphate solution (pH 2.2) containing 500 mM KCl and ACN (7:3). Isocratic elution was performed with 100% A for 30 min and then the system was washed with 100% B. The collected fractions were lyophilized, re-suspended in 50 μL of 0.1% TFA and desalted using SDB-StageTips. Two-thirds of the enriched peptides were analyzed by nanoLC/MS/MS using the Orbitrap Fusion Lumos.

To isolate protein N-terminal peptides from TrypN-digested HEK293T peptides, the digested peptides (80 μg) were analyzed by the SCX HPLC system under isocratic conditions, eluting with a mixture of 7.5 mM potassium phosphate solution (pH 2.2) containing 10 mM KCl and ACN (7:3) for 30 min to collect the desired fraction and desalted using SDB-StageTips as described above. We analyzed one-fourth of the enriched peptides by nanoLC/MS/MS in triplicate using the Orbitrap Fusion Lumos.

N-terminal peptides enrichment by SCX separation for beige adipocytes

N-terminal peptides were enriched as described previously.[45] The SCX-StageTip was prepared as following. SCX-StageTip packed with one-layer SCX disk membrane into 200-μL tips. The active and eluted buffers were prepared as follows: 7.5 Mm phosphate buffer 500 mM KCl, pH 2.2 and 30% ACN; 7.5 Mm phosphate buffer 12.5 mM KCl, pH 2.2 and 30% ACN. Conditioning and equilibration were done through sequential passing 100 μL buffer and

centrifugation at 1000 × g for 1 min of the following reagent: methanol, active buffer and elution buffer. 20 ug of digested peptides were loaded into SCX-StageTips and through centrifugation at 1000 × g for 1 min, collecting as flowtrough, and sequential eluted bound peptides with 100 µL of eluted buffer by centrifugation at 1000 × g for 1 min, collecting the eluate into the same tube as enriched N-terminal peptides. The eluted fraction was evaporated by SpeedVac SPD121P (Thermo Scientific), resuspended in 50 uL of 0.1% TFA and desalted by SDB-StageTip. The desalted samples were concentrated in a vacuum evaporator followed by the addition of 0.1% TFA in 4% ACN for subsequent nanoLC-MS/MS analysis.

NanoLC/MS/MS analysis

NanoLC/MS/MS analyses were performed on a TripleTOF 5600 mass spectrometer or an Orbitrap Fusion LUMOS mass spectrometer, connected to a Thermo Ultimate 3000 pump and an HTC-PAL autosampler (CTC Analytics, Zwingen, Switzerland). Peptides were separated on self-pulled needle columns (150 mm length × 100 µm ID, 6 µm opening) packed with Reprosil-Pur 120 C18-AQ 3 µm reversed-phase material (Dr. Maisch, Ammerbuch, Germany). The injection volume was 5 µL and the flow rate was 500 nL/min. The mobile phases were (A) 0.5% acetic acid and (B) 0.5% acetic acid and 80% ACN. For TripleTOF 5600 analysis, gradient elution was performed as follows: 12−40% B in 20 min, 45-100% B in 1 min, 100% B for 5 min. For Orbitrap analysis, gradient elution of fractionated samples was performed as follows: 12−40% B in 15 min, 40-100% B in 1 min, 100% B for 5 min. For protein N-terminal peptide-enriched samples, gradient elution was performed as follows: 10−40% B in 100 min, 40-100% B in 10 min, 100% B for 10 min. For proteome samples, gradient elution was performed as follows: 8−35% B in 100 min, 35-50% B in 10 min, 5-100% B in 5 min, 100% B for 10 min. Spray voltages of 2300 V in the TripleTOF 5600 system and 2400 V in the Orbitrap system were applied. The mass scan range of the TripleTOF 5600 system was *m/z* 300−1500, and the top ten precursor ions were selected in each MS scan for subsequent MS/MS scans. The mass scan range for the Orbitrap system was *m/z* 300−1500, with an automatic gain control value of 1.00e + 06, a maximum injection time of 50 ms and detection at a mass resolution of 60,000 at *m/z* 200 in the orbitrap analyzer. The top ten precursor ions with +2, +3 or +4 charge were selected in each MS scan for subsequent MS/MS scans with an automatic gain control value of 5.00e + 04 and a maximum injection time of 300 ms. Dynamic exclusion was set for 25 s with a 10 ppm gate. The normalized HCD was set to be 30, with detection at a mass resolution of 15,000 at *m/z* 200 in the Orbitrap analyzer. A lock mass (445.1200025) function was used to obtain constant mass accuracy during the gradient.

Proteomics Data Processing for HEK293T and *E. coli*

Two peak lists in ".mgf" and ".apl" formats were generated from the MS/MS spectra by MaxQuant 1.5.8.0 33. The peptides and proteins were identified by Mascot v2.6.1 (Matrix Science, London, U.K.) against the Swiss-Prot database (version 2017_4, 20,199 sequences) or the *E. coli* K-12 MG1665 protein sequence database 34 with a precursor mass tolerance of 20 ppm (TripleTOF 5600) or 10 ppm (Orbitrap), a fragment ion mass tolerance of 0.1 Da (TripleTOF 5600) or 20 ppm (Orbitrap), TrypN/trypsin specificity allowing for up to 4 missed cleavages for TrypN/trypsin mixed proteolytic peptides and strict TrypN specificity allowing for up to 2 missed cleavages for TrypN-digested peptides. Carbamidomethylation of cysteine was set as a fixed modification, and methionine oxidation and protein N-terminal acetylation were allowed as variable modifications. False discovery rates at a peptide level of less than 1% were applied for peptide identification based on a target-decoy approach.

Proteomics Data Processing for beige adipocytes

For N-terminome dataset, the peak list ".mgf" were generated from the MS/MS spectra by MaxQuant. The peptides and proteins were identified by Mascot v2.6.1 and MaxQuant v1.6.2.10 against the Universal Protein Resource Knowledgebase (UniProtKB, version 2019_2, 93,387 sequences; 25,228 entries form Swiss-Prot including isoforms and 68,159 entries from TrEMBL (Translated EMBL Nucleotide Sequence Data Library)). In MaxQuant searching were performed with the following parameters: both precursor and fragment set 20 ppm for mass tolerance, N-terminally semi-TrypN specificity allowing for up to 2 missed cleavage, 7 amino acids were the required minimum peptide sequence length for specific cleavage, 7 amino acids for semi-specific cleavage, the maximum peptide mass was limited to 4,600 Da and peptides were quantitated through match-between-runs. In Mascot searching, standard settings with the additional options were selected as following: a precursor mass tolerance of 10 ppm, a fragment ion mass tolerance of 20 ppm, semi-TrypN specificity allowing for up to 2 missed cleavage, the required minimum peptide sequence length was 7 amino acids. Both search engines applied the setting as follows: carbamidomethylation of cysteine was set as a fixed modification, methionine oxidation and N-terminal acetylation was allowed as a variable modification. A reversed sequence library was employed to control the false discovery rate (FDR) less than 1% in sequence level for both MaxQuant and Mascot.

For the global proteome dataset, the peptides and proteins were identified by MaxQuant against UniProtKB and the standard settings as shown above. Additional options were selected as following: strict TrypN specificity allowing for up to 2 missed cleavage in mono proteolytic

peptides, the FDR was less than 1% for peptide spectrum matches and protein group identifications. The master proteins were selected based on the maximum overlap with the N-teminome dataset. Match-between-runs and label-free protein quantification were performed with the MaxLFQ algorithm. Carbamidomethylation of cysteine was set as a fixed modification, methionine oxidation and protein N-terminal acetylation was allowed as a variable modification.

Preparation of *in-silico* digested *E. coli Protein* N-terminal peptide list

The list of protein N-terminal peptides obtained by *in silico* TrypN-digestion of *E. coli* proteome was prepared using *E. coli* K-12 MG1665 protein sequence database (4,316 sequences).[56] First, a strict TrypN-specific cleavage at protein N-terminus was performed and missed cleavage was prohibited. Second, peptides shorter than 7 amino acids in length and the redundant sequences were removed. A total of 1,494 protein N-termini were obtained.

Bioinformatics of proteome and N-terminome dataset

The statistical analysis was performed using Perseus v.1.6.14.0. The peptide intensities and protein MaxLFQ values were used for quantification in N-terminomics and global proteomics and filtered based on two non-zero values in the three technical replicates at least one time-point. The values were log2 transformed and replaced the missing values using the heuristic random-tail method.[57] The replicates in each time-point were averaged and normalized by z-score independently and subject to hierarchical clustering analysis with following option: Euclidean distance with option of preprocessed with k-means clustering in 300 number of clusters and 10 maximal number of iterations. The GOBP, GOCC, GOMP, and KEGG were performed using DAVID 6.8 for functional annotation.[46,47]

CHAPTER 2

Sequence-specific model for predicting peptide collision cross-section values in proteomic ion mobility spectrometry

INTRODUCTION

Ion mobility spectrometry (IMS) has been long considered as a promising tool for many applications in structural biology,[58] proteomics[59] and many other analytical applications.[60,61] Separation of isobaric peptides,[62] improving signal-to-noise ratio in bottom-up approaches,[63] studying protein conformation and protein assemblies[64] represents an incomplete list of its proteomic applications.[65-70] One of the attractive options of IMS is the possibility to model gas-phase peptide behaviour in ion-mobility based separation. Building a comprehensive collisional cross section (CCS) prediction model for peptides will allow not only the direct application to improve confidence of MS/MS-based identification[71] by providing the orthogonal property for machine learning approaches such as Percolator[72] but will help better the current understanding of underlying mechanisms for ion mobility-based separations, resulting in improving MS/MS-based quantitation by reducing the complexity of peptide ions prior to tandem mass spectrometry.[73,74]

In the past, CCS measurements have been used to understand the effect of sequence motifs and charge on structure[75-81] and for determining gas-phase conformation of protein complexes.[82,83] The effects of structural features on peptide separation have been studied in both, gas phase with IMS,[75,76,84-86] and liquid phase separations with reversed-phase high performance liquid chromatography (RP-HPLC).[87-89] Considering the history in the progress of this field, it is easy to notice striking similarity between IMS and RP-HPLC. Both were conceived long before arrival of MS based proteomics.[90-92] Both techniques are used as front-end devices to improve delivery of separated compounds into the mass spectrometer and are amenable to modeling of the separation processes – driven by peptides' size/shape in the gas phase and hydrophobicity, respectively.[71,93]

Due to its preparative capability, RP-HPLC became one of the most important techniques in protein/peptide analytical chemistry long before the proteomic era. Initial peptide retention time prediction models aimed at improving separation method development during early 1980s.[93,94] These early models were based on an additive principle, which considers that the hydrophobicity for a peptide is equal the sum of its constituent residues' hydrophobicities. The effect of peptide sequence in addition to its composition in RP-HPLC has been reported in 1987[95] and first sequence specific model has been developed in 2004 based on the collection of just 346 tryptic peptides.[96] The authors suggested using position-dependent hydrophobicity coefficients for individual amino acids to compensate for unique features of peptide N-termini observed due to ion-pairing interactions; however, sufficient data density is required to model this concept. Development of mass spectrometry and proteomics brought increased throughput and confidence in peptide identifications, thus increasing the size of high-quality datasets available for prediction modeling.[97] Proteomic peptide datasets have allowed the implementation of more complex prediction algorithms and opportunities to study a variety of structural features in peptides. In mid 2000s, Petritis et al. described retention modeling via an artificial neural network (ANN) approach using datasets of ~6,000[98] and later ~300,000 peptides.[99] The increasing size of proteomic datasets near the 2010s opened the opportunity to study the effects of structural motifs such as N-cap helical stabilization on peptide retention in RP-HPLC. Given only a small portion of peptides exhibit amphipathic helicity, such study required a collection of ~280,000 peptides.[89] Continuous efforts in standardization of RP HPLC separation in proteomics[100,101] and progressive growth in MS productivity in the past decade has allowed for the collection of high quality RP-HPLC data in the size of hundreds of thousand to over a million peptides.[102-105] This paved the possibility for wider application of high data density machine learning techniques to address peptide retention time prediction problems.[105-107]

Clemmer and co-workers led the way in the development of IMS technology for proteomic applications,[108] peptide IMS data collection, and modeling peptide ion mobility.[71,109] Valentine et al. used 660 peptides to derive the intrinsic size parameter (ISP) coefficients, which multiplicatively scales the mass of individual residues used in CCS additive prediction models.[109] The same group of authors expanded this algorithm to the collection of 2,094 tryptic peptides 5-15 amino acids long.[71] In a different approach, Shah et al. built a machine-learning based model attempting to introduce additional features including but not limited to: normalized retention time in RP-HPLC, peptide length, gas phase basicity, and number of negatively/positively charged groups.[110] However, the size of this dataset, which contained

3933 (2+), 3916 (3+), 717 (4+ peptides), was not sufficient to define sequence-specific features. Peptide structural properties are of ultimate importance for IMS. The same peptide species can assume different conformations with drastically different CCS values.[111] This feature is the most obvious for 3+ peptide populations, which exhibit significant split in CCS versus molecular weight plots – designated as compact and extended structure populations.[110,112,113] Another argument confirming sequence dependent character of peptide IMS was provided by Lietz et al.[112] The authors used LysC and LysN digests of K562 to show that N-terminal location of Lys results in lower CCS values compared to the same sequences with C-terminal Lys for 14 peptide pairs.

Overall, there is ample evidence of sequence-dependent character of IMS separation. Yet, compared to RP HPLC, there are no CCS prediction models incorporating these features. One of the problems is a significant advantage of chromatographic separations in terms of data availability: hundreds of thousand data points[101-103,105] vs. thousands.[71,110] The timsTOF Pro, a quadrupole/time-of-flight (Q/TOF) mass spectrometer coupled with trapped ion mobility spectrometry (TIMS) cells, achieves a resolving power of over 220K and the scan speed (100 ms per scan) between LC and Q/TOF mass analyzer, showed a lot of promise in this regard.[113-115] Similar to chromatographic applications, measuring CCS values for few hundred thousand peptides may provide sufficient data for application of machine learning approaches. At the time of our work, Meier et al. concurrently developed a deep learning CCS prediction model using 570,000 unique combinations of sequence, charge state (2+, 3+ and 4+), including peptides with oxidized methionine.[116] However, many machine learning approaches often operate as "black boxes", providing limited information on the underlying separation mechanisms. Meier et al. have demonstrated the contributions of the 20 amino acid residues and the qualitative trends for hydrophobicity, Pro content, and position of His.[38] They highlighted the difficulty to model the observed physicochemical properties along with sequence dependent features directly with the linear sequences and our work here is able to address such difficulties as well as investigate finer composition and position dependent features that are novel to our approach. Therefore, a semi-empirical Sequence-Specific Retention Calculator (SSRCalc) approach based on position-dependent correction coefficients was applied in this work to establish a Sequence-Specific Ion mobility Calculator (SSICalc).[117] The SSRCalc has been used successfully in the past for modeling various modes of peptide HPLC,[117-119] and capillary zone electrophoresis.[120] The dataset for CCS modeling was obtained by the 2D LC (SCX/RP)/ESI/TIMS/Q/TOF analysis of multiple alternative proteases digests using timsTOF Pro.

RESULTS AND DISCUSSION

Data selection for model optimization.

In this work, seven protease-digested (trypsin, LysargiNase, LysC, LysN, GluC, AspN and chymotrypsin) cell lysates have been analysed using SCX-StageTip fractionation applied prior to RP-LC/TIMS/Q/TOF analysis. SCX chromatography was chosen aiming to improve representation of peptides in different charge states. The selection of peptides for model optimization are crucial for generating a representative high-quality dataset. NanoLC/TIMS/Q/TOF measurements provided the reduced ion mobility coefficients ($1/K_0$) and retention time for all identified peptides (Figure 2-1). These identifications have been additionally filtered using the SSRCalc peptide retention time prediction model as shown in Figure 2-2 (Supplemental Table 1). Less than 1% of identified peptides were removed based on large retention prediction errors or low confidence score of $-3 < \log(e) < -1$. Moreover, since IMS separates peptides based on their conformations similar to previous studies,[110,112] multiple peptide conformations were detected in some instances (Figure 2-3). The $1/K_0$ values for model optimization were then selected corresponding to the most intense peptide MS/MS hit on each mobilogram (Figure 2-3) followed by the removal of redundant identifications in order to merge the dataset into 133,946 entries. There were 14,482, 86,268, 27,463 and 5,733 peptides belonging to the 1+, 2+, 3+ and 4+ populations, respectively (Supplemental Table 8). Peptides contained 1-11 positively charged residues (Lys, Arg, His and unmodified N-termini) and were 5-50 amino acid long (560-5245 Da): 14.7 residues on average. This represents a typical size/charge distribution of peptides encountered in bottom-up proteomics experiments.

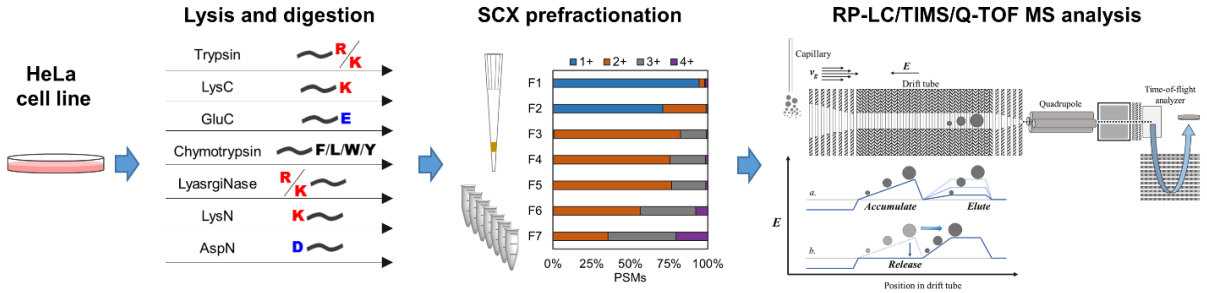

Figure 2-1. Workflow of experimental data collection.

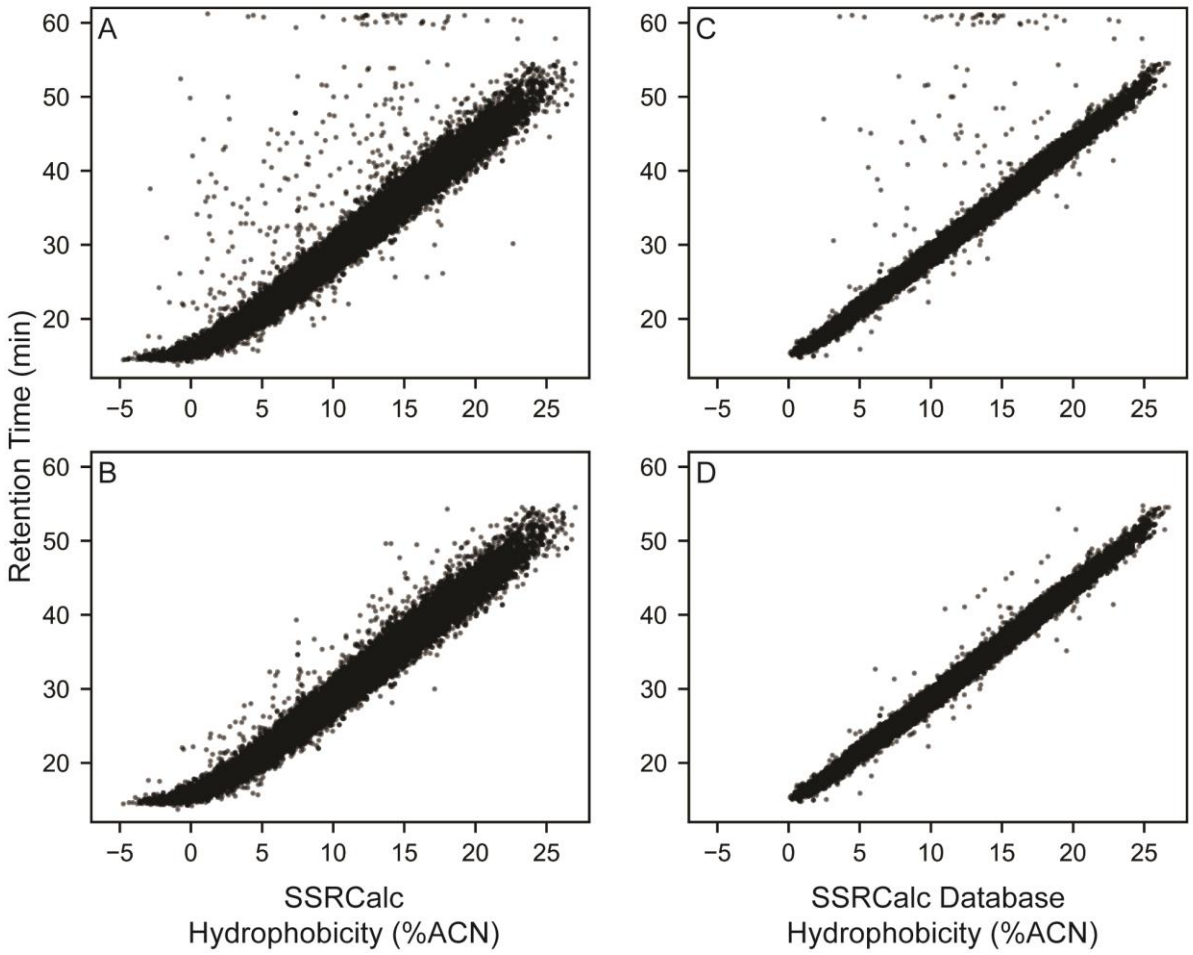

Figure 2-2. Elimination of false-positive identifications using peptide retention time prediction with de-novo SSRCalc model (A, B) or SSRCalc retention Database (C, D). All peptides with retention time prediction error of more than ±6 min and low confidence identification scores (-3 < log (e) <-1) have been removed. Plots before (A, C) and after (B, D) the removal of the outliers with lower confidence scores are shown.

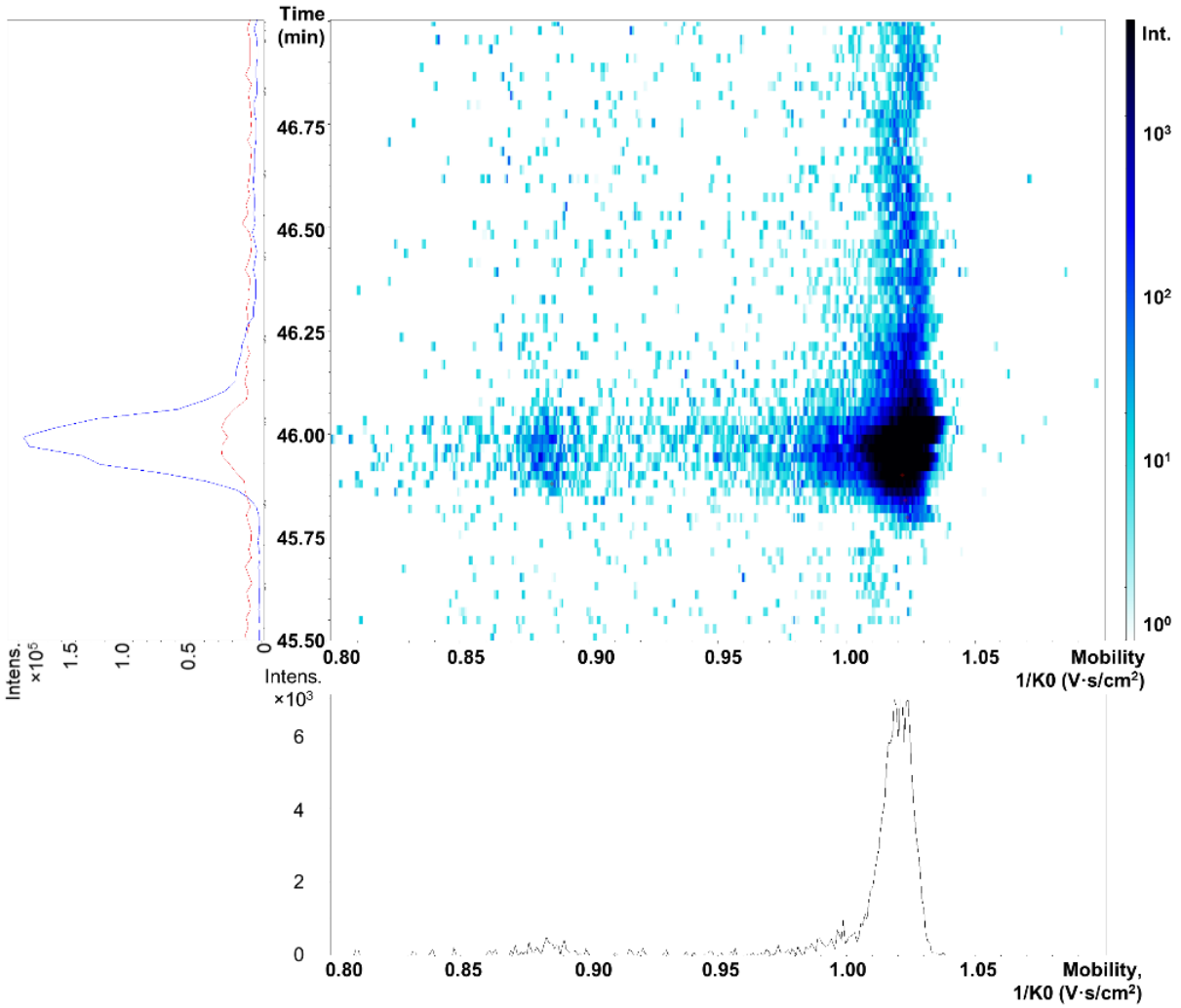

Figure 2-3. Distribution of extracted ion (*m/z* 754.04-754.06) in chromatogram and mobilogram. The gradient blue on the right shows the intensity scale in MS1, retention time on y-axis, and 1/K0 on x-axis. Extracted ion chromatogram and extracted ion mobilogram are projected on the left and bottom axes, respectively.

Evaluation of peptide bulk properties affecting CCS.

Similar to prior work,[110,113] plotting the dependence of CCS values on *m/z* resulted in definitive trend lines corresponding to four individual charges (Figure 2-4A). We used the characteristic shapes of these plots and properties of 100 peptides that are the most significant positive/negative outliers (Table 2-1, Figure 2-5, for each charge state to assess the effects of peptide bulk properties.

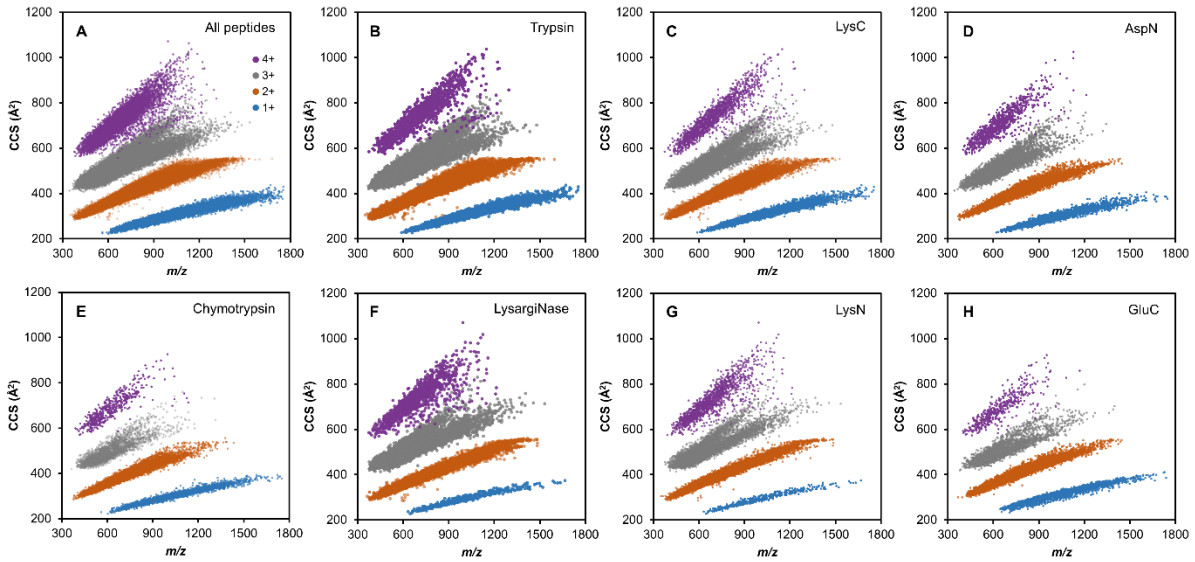

Figure 2-4. CCS versus *m/z* plots for 133,946 peptides from HeLa cell digests for total (A) and protease-specific populations (B-H). Individual charge states are color coded as: blue, 1+; orange, 2+; gray, 3+ and purple, 4+.

Table 2-1. Average bulk properties of top 100 positive and negative outliers in charge specific CCS vs. *m/z* plots.

Charge/ prediction error	Tryptic/non-tryptic peptides*	Peptide length	Agadir α-helicity[121]	SSRCalc Hydrophobicity (% ACN)	pI[122]	# of basic residues
1+/pos	93/7	12.60	0.31	11.88	7.13	2.00
1+/neg	6/94	13.10	0.11	7.92	3.55	1.23
2+/pos	70/30	18.56	1.41	15.35	6.88	2.93
2+/neg	44/56	21.52	0.16	12.09	4.44	2.19
3+/pos	60/40	29.44	1.35	18.43	5.97	3.41
3+/neg	27/73	28.87	0.23	12.96	4.30	2.81
4+/pos	42/58	33.77	0.68	16.01	6.42	4.34
4+/neg	40/60	36.27	0.32	15.67	4.44	4.12

* - independently of the protease used tryptic peptides correspond to Lys/Arg terminated species.

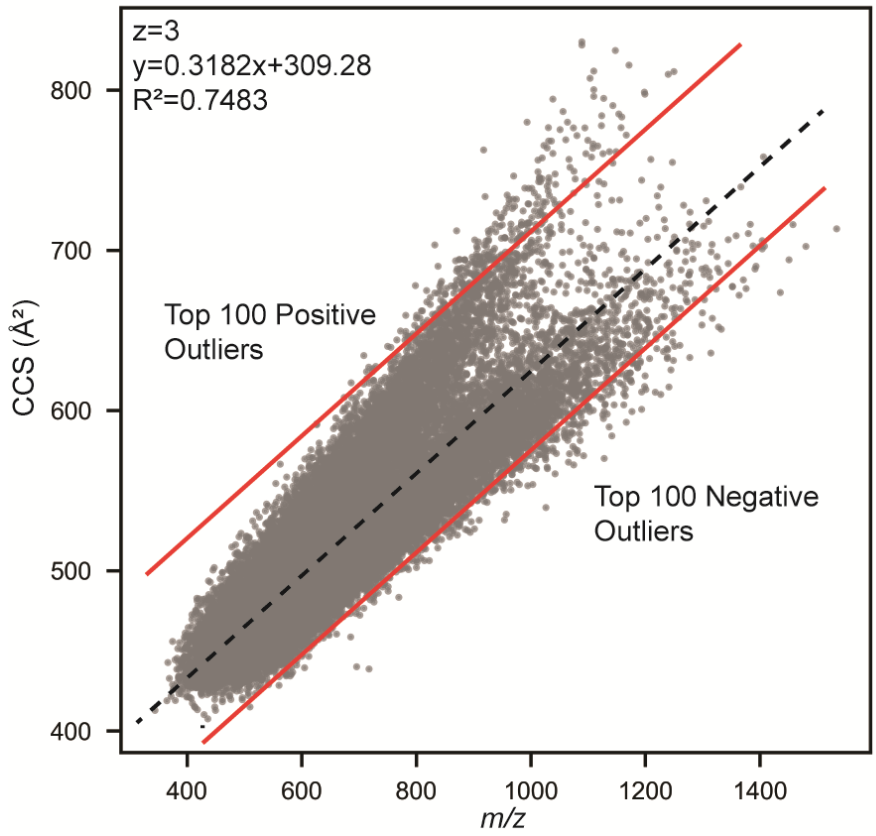

Figure 2-5. Graphical representation of the selection process for the Top 100 Positive Outliers and Top 100 Negative Outliers shown in Table 2-1. The outlier selection lines (red) are parallel to the trendline of the CCS vs *m/z* correlation (black dotted) as we are picking the outliers with the largest positive difference for the Top 100 Positive Outlier set and the largest negative difference for the Top 100 Negative Outlier set.

The CCS versus *m/z* correlation plots for singly and doubly charged peptides are slightly concave, indicating the preference of longer peptides to be in more compact conformation. For longer highly charged (3+, 4+) peptides, the CCS trends became dispersed and a clear split-population appeared in 3+ species, corresponding to compact (low CCS) and extended (high CCS) features observed previously.[110,112,113] In addition, we found that the distributions between compact and extended conformation are protease-specific, e.g., 3+ LysN digested peptides containing two positive charges at N-termini predominantly assume compact conformation (Figure 2-4G). The non-tryptic (terminated by any amino acid other than Lys or Arg) peptide populations of all charges exhibited lower CCS values compared to tryptic ones. This observation was also confirmed by analyzing the population of the outliers (Table 2-1). For example, 93% of 1+ peptides with largest positive deviations are tryptic, while 94% non-tryptic species were found among 100 most negative outliers. Similar finding has been reported by Lietz et al., who compared CCS values for 3+ peptides from LysC and LysN digest. The authors explained this behaviour by the electrostatic interaction of N-terminal/C-terminal Lys with

peptide macro-dipole, which should destabilize/stabilize peptide's helical conformation, respectively. Similarly, 3+ AspN peptides exhibit even distribution between compact and extended structures, while GluC generated species tend to be in the latter conformation Figure 2-4D, H). Overall, LysargiNase/LysN/GluC destabilize the helix favouring compact, whereas trypsin/LysC/AspN stabilize the helix inducing more extended form through interaction of terminal residues with peptide's macro-dipole.[112] Detailed analysis of 1+ and 2+ correlation plots (Figure 2-4) also shows that trypsin/LysC/AspN populations show some splitting between the dominant compact and extended subpopulations, although the CCS difference between two conformations was subtle. Similarly, LysargiNase/LysN/GluC exhibit more uniform distribution in the 1+ and 2+ peptide populations. Furthermore, for 4+ trypsin/LysC/AspN, more preference to the extended conformation was observed compared to LysargiNase/LysN/GluC. For each protease, increasing charge state of a peptide led to higher tendency to be in extended conformation.

Based on comparison of average peptide length for the entire population (Table 2-2) vs. the most significant outliers (Table 2-1), we find that shorter peptides are more consistent in conformational behaviour than the longer outlier peptides. In fact, for peptides with MW < 1600, 1600 to 2400, and >2400, the percent errors were 2.2%, 3.1%, and 4.7%, respectively. This behaviour is particularly obvious for 3+ peptides, which exhibit a split for more than 20-mer species. Peptides with large positive prediction errors exhibit higher helical content calculated using the Agadir algorithm.[121] Alpha-helical peptides are more linear in geometry and are unable to fold to smaller states thus will exhibit higher than expected CCS consistent with the positive prediction error we observed. Peptides with large positive deviations (Table 2-1) are more hydrophobic. This observation supports previous findings by Valentine et al.[109] and Shvartsburg et al.[123] reporting on high ISP values of hydrophobic residues.

Table 2-2. Accuracy of prediction model (R^2-value) for step-by-step optimization.

Datasets Average peptide length and range	**Step 0** No correction CCS~*m/z*	**Step 1** 20 ISP values for each: tryptic/ non-tryptic subsets 2x20 = 40 parameters	**Step 2** 20 ISP values for each charge state 2x20x4 =160 parameters	**Step 3** Position-dependent ISP for each charge 2x20x4x13=2080 parameters
1+ (14482) 9.7aa (5-19 aa)	0.912	0.925	0.952	0.977
2+ (86268) 13.7 aa (6-33 aa)	0.938	0.952	0.962	0.969
3+ (27463) 18.5 aa (8-49 aa)	0.763	0.815	0.826	0.864
4+ (5733) 24.5 aa (13-50 aa)	0.750	0.79	0.802	0.832
All (133946) aligned 14.7aa (5-50 aa)	R^2 = 0.966 δ95% 36.2 Å^2	R^2 = 0.973 δ95% 32.2 Å^2	R^2 = 0.976 δ95% 30.5 Å^2	R^2 = 0.981 δ95% 27.1 Å^2

δ95% - prediction error range excluding 5% of the most significant outliers

The peptide pI showed no obvious correlation with deviation from CCS vs. *m/z* plots, when entire peptide population was considered (Figure 2-6). However, when we isolate the top 100 outliers in the dataset, as shown in Table 2-1, the general trends amongst all the charges are that positive prediction errors are associated with higher average pI values and vice versa. We have an interest to investigate if there is a correlation between peptide CCS and electrophoretic mobility measured by capillary zone electrophoresis (CZE). Our advanced SSRCalc CZE model has $R^2 \sim 0.995$ correlation with experimental values[120] and should provide an accurate estimation of electrophoretic mobility in solution at acidic pH when compared to experimental CCS. However, Figure 2-7A shows poor correlation between these two systems. Important to note that this plot consists of multiple sub-populations corresponding to peptides carrying different number of charged residues versus their CCS values for a particular charge state. Peptide electrophoretic mobility at acidic pH depends mostly from peptide charge (number of basic residues) and mass. The sequence-specific features in CZE largely are limited by N-terminal position of Asp and Glu, which reduces N-terminal charge/basicity. IMS separation is affected by many processes, including formation of helical structures, which results in poor correlation of CCS versus μ_{ef} plots even when peptides with identical number of charged residues are considered. As shown previously,[120] the semi-empirical model $\mu_{ef} = k(Z/M^X)$ (μ_{ef}: electrophoretic mobility, k: constant, M: molecular weight, Z: net charge) can be optimized by modulating X such as 1/3, 2/3, or 1/2. Note that if the molecular volume is proportional to M and the molecular shape is globular, M^X (X= 1/3, 2/3, 1/2) is proportional to the radius of the molecule, the cross-sectional area, and the van der Waals radius, respectively. As CCS is

proportional to mass (Figure 2-4A), we were able to linearize the CCS versus μ_{ef} correlation by modulating X to 1/3 with $R^2 = 0.849$ (Figure 2-7B).

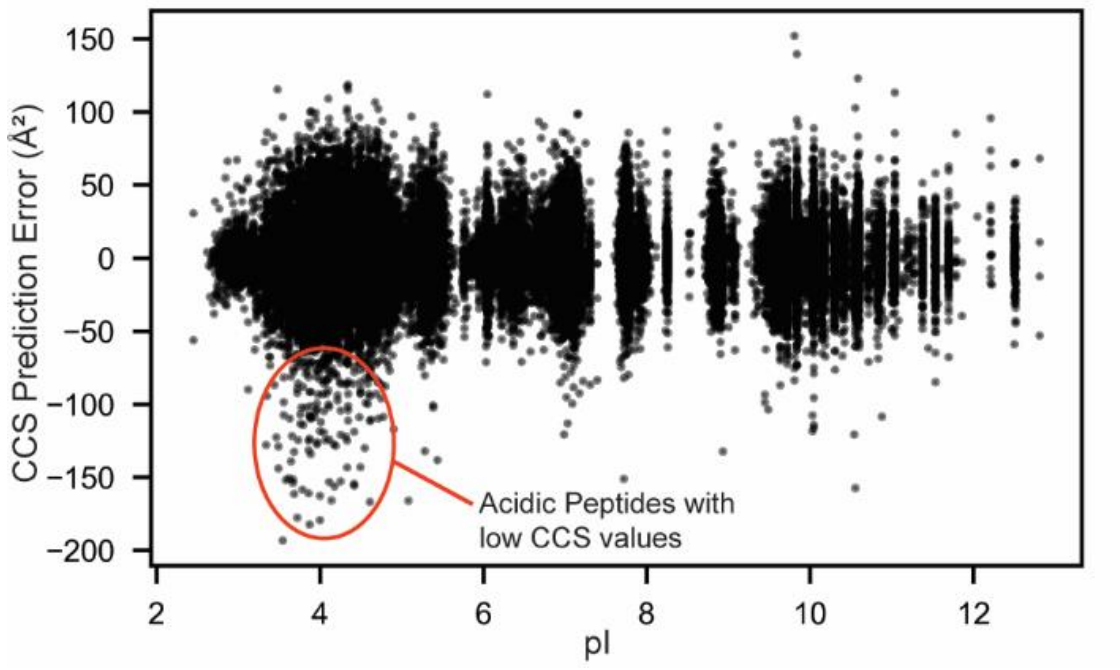

Figure 2-6. Correlation between CCS prediction error and peptide pI.[122] Prediction errors have been calculated for uncorrected CCS vs. *m/z* plots for each charge state.

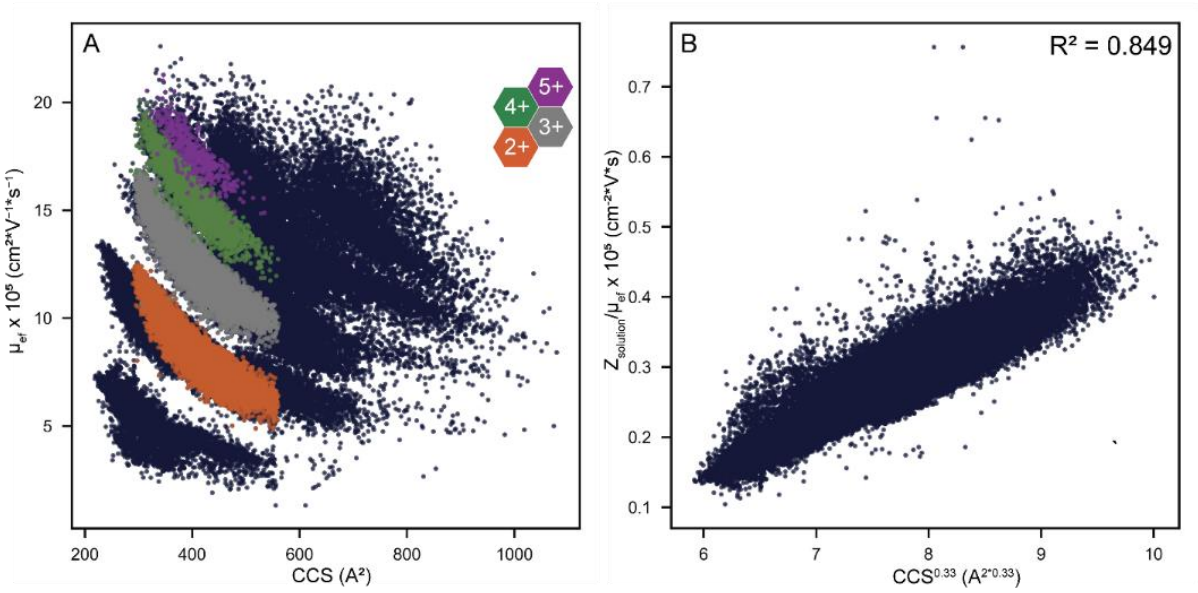

Figure 2-7. Correlation between predicted peptide electrophoretic mobility[120] and experimental CCS values. (A) Peptides detected in 2+ charge state by mass spectrometer, but carrying different charge (2+, 3+, 4+, 5+) in solution at acidic pH are highlighted, (B) the semi-empirical model $\mu_{ef} = k(Z/M^X)$ can be optimized by modulating X such as 1/3, 2/3, or 1/2, as shown in the paper.[2] As CCS $\propto$ M (in **Figure 2-4**A), we attempted the X modulation and 1/3 had the best correlation of R^2=0.849.

Length-specific multiple linear regression model.

To explore the properties involved in the ion mobility of peptides in IMS, an ensemble of length-specific multiple linear regression (LS-MLR) model and charge states was built to predict the CCS values of the peptides. In each model, the number of independent variables increases with peptide length, as there are 20 amino acids per position in the peptide sequence. However, due to the nature of enzymatic digestions, the number of experimental data per peptide length was only distributed over a narrow range, and in particular, the number of longer peptides was sparse (Figure 2-9A). To reduce the number of independent variables, in addition to the terminal cleaved site, the 20 amino acids were grouped into 7 by similarity and LS-MLR analyses were performed using the relative position coefficients (P_j) corrected for the mass (G_k) of each amino acid. Different proteases cleaved at N/C-terminal sides of different amino acids, resulting in the diverse CCS values observed in IMS (Figure 2-4). Therefore, individual LS-MLR models were built for peptides generated by different proteases and in different charged states, and the position coefficients were converted to coefficients for each amino acid based on mass as shown in Figure 2-8.

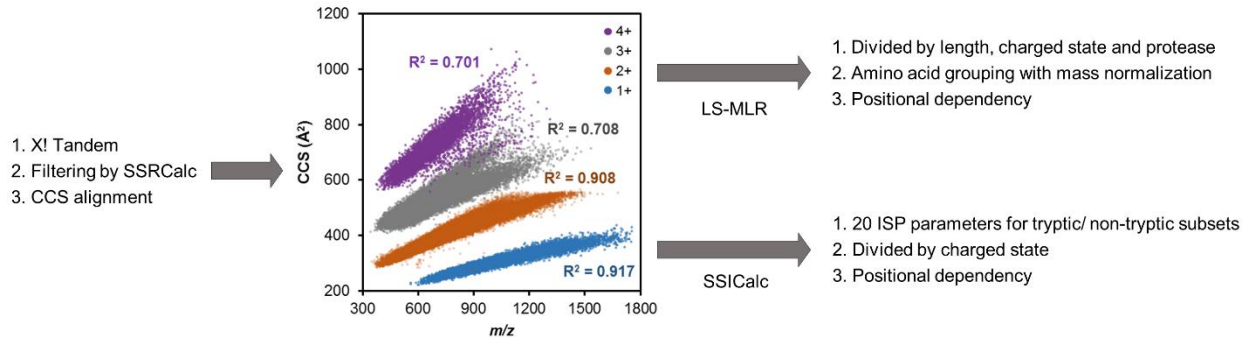

Figure 2-8. The workflow of prediction model optimization.

As a first step, LS-MLR was applied to trypsin- and LysargiNase-digested peptides that have opposing terminal Lys/Arg (Figure 2-9B). The trypsin-digested peptides have higher coefficients at the C-termini especially for aliphatic and positively charged amino acids, whereas positive contribution at P3 position was observed for LysargiNase-digested peptides. Second, both 2+ and 3+ tryptic had larger coefficients at the C-termini, and only the 3+ peptides had larger coefficients at the N-termini. Moreover, some polar amino acids such as D, E, S, T, N, and Q have lower coefficients in the internal region. Finally, LS-MLR was performed for each length of tryptic peptides. The bottom two heatmaps (Figure 2-9B) show the coefficients

in CCS prediction for 3+ tryptic peptides of different lengths. At a peptide length of 15, the coefficients are higher for both termini and lower for the polar amino acids in the central region. However, as the length of the peptide increases, the coefficients become more uniform in the central region.

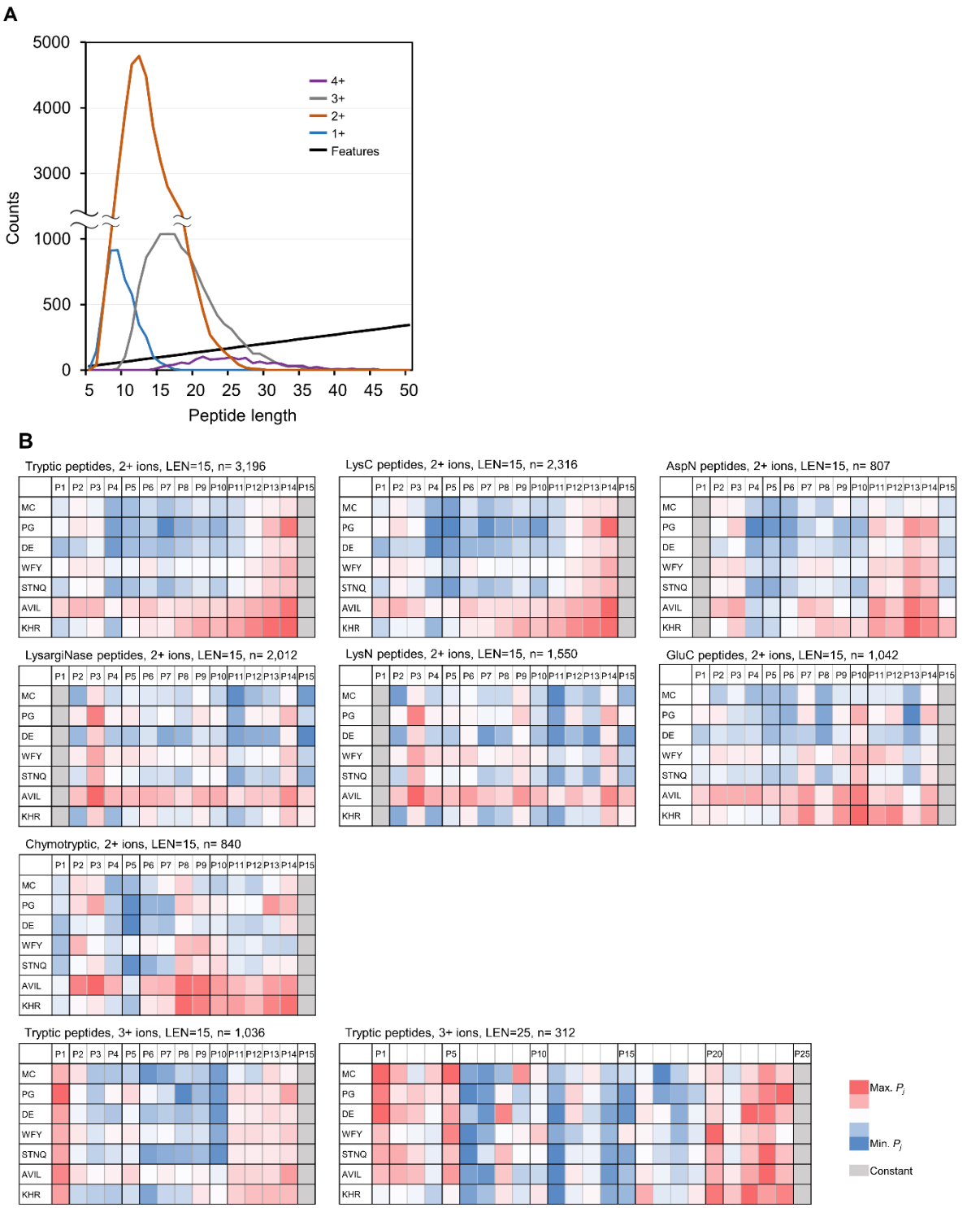

Figure 2-9. (A) The number of peptides for each charge state and length in tryptic dataset compared to the number of features for each length in the LS-MLR, (B) Heatmaps of position and group dependent coefficients obtained by LS-MLR model.

The LS-MLR model has been able to achieve an R^2 value of 0.977 for the CCS prediction derived from the tryptic peptides for specific charge and length (Figure 2- 10, 1+ peptides with 7-12 a.a., 2+ peptides with 8 -20 a.a. and 3+ peptides with 11-25 a.a.). While the LS-MLR models could produce good correlations for peptides of particular length, the overfitting still occurred due to the limited size of the dataset at each length. Therefore, it is necessary to apply an alternative predictive model that is not limited by the number of features to obtain a global prediction of CCS values. To compensate for these features, we employed the physicochemical properties of trypsin/non-trypsin peptides, charge states, and amino acid positions via a step-by-step charge and protease dependent ensemble linear regression model optimization, as shown in the next section.

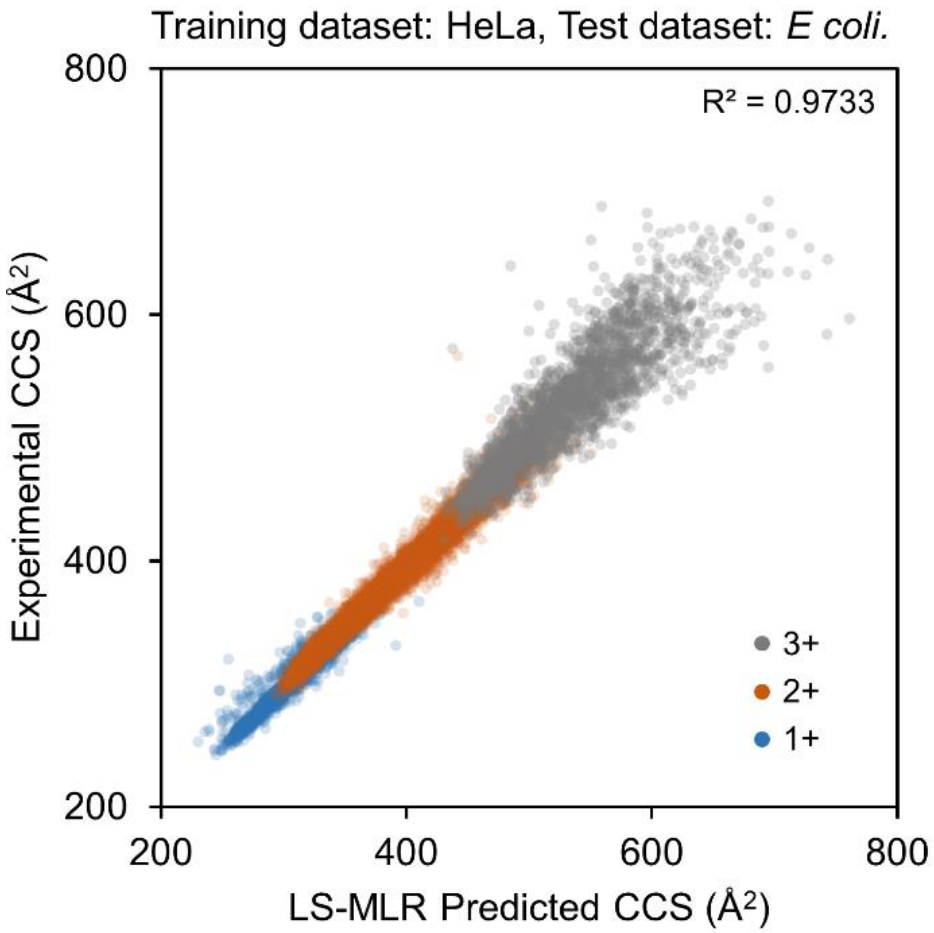

Figure 2- 10. Correlation of LS-MLR predicted versus experimental CCS values. The predicted CCS values derived from the tryptic peptides (12,347 *E. coli* peptides) for specific charges and length (1+ peptides with 7-12 a.a., 2+ peptides with 8 -20 a.a. and 3+ peptides with 11-25 a.a.) by LS-MLR. Note that peptides for each category in (B) were selected not by the actual proteases but by the types of peptides. For example, peptides with C-terminal R or K and 2 missed cleavage at maximum were selected as “tryptic peptides”.

Peptide length independent step-by-step optimization using Intrinsic Size Parameters approach as a starting point

Each optimization step has been followed by the alignment of eight peptide subsets: tryptic/non-tryptic in four different charge states to fit all eight correlation plots to CCS_{pred} versus CCS_{exp} with slope 1 and intercept 0 as shown in Table 2-2. In every step each of the eight independently optimized sub-models have their own CCS-aligning slope and intercept values that are allowed to vary from this initial *m/z* data alignment. It should be noted that the R^2 correlation for the combined collection of peptides is higher than individual subgroups due to wider range of the experimental CCS values.

Step 1: twenty ISP values optimized for two datasets (tryptic and non-tryptic). Figure 2-11A shows comparison of ISPs reported by Valentine et al.[71] and by Shvartsburg et al.[123] versus ours optimized for tryptic/non-tryptic datasets. Most of the hydrophobic residues' ISP values are larger compared to the ones reported previously. Conversely, polar residues showed lower ISP values, favouring more compact structures. These deviations are likely originated from the significant difference in charge and size distribution in these two datasets. ISP values for Lys and Arg, which have been found to be similar to His and in close agreement with a-priori predicted ISP values by Shvartsburg et al. using sum of the projection areas for constituent atoms.[123]

Step 2: charge dependent ISP values have been optimized for subsets of peptides and improved correlation values for all of the submodels (Table 2-2). Figure 2-11B shows these values for both tryptic and non-tryptic species. ISP values of hydrophobic residues increase for highly charged 3+ and 4+ peptides; the opposite is true for polar residues (D, E, N, Q). These trends follow the difference observed between Valentine et al. values and ours, indicating inclusion of highly charged longer peptides determined overall differences in ISP values in Figure 2-11A. Pro exhibits the lowest ISP values, favouring compact structures, for 3+ peptides, while Gly in 4+ species promote extended conformation.

Step 3: thirteen position dependent ISP coefficients have been introduced for each residue: six on both termini plus a general internal position – similar to all SSRCalc models for peptide HPLC. This led to further improvements in correlation values in all respective subsets shown in Table 2-2. Selected position dependent trends are shown in Figure 2-12 and the entire collection of optimized coefficients is provided in Table 2-3.

The hydrophobic residues (A, V, L, I, M, F, Y, W) show virtually no position dependence except for a small decrease in ISP for internal position, especially for aromatic Phe, Tyr, and

Trp (Figure 2-12A). Also, an evident decrease in the internal position for Pro. It is interesting to note that Pro ISP values in terminal positions are above 1, which corresponds to the value determined by Shvartsburg et al. based on atom projections.[77] In other words, the behavior of Pro (low ISP) is determined by its known property as helix breaker, rather than size of the side chain. While located inside of the peptide Pro tends to form kinks in the structure favouring sequence bending, resulting in reduced CCS. The decrease in internal Pro ISP is smaller for 2+ ions compared to 1+ and much more significant for 3+ (Figure 2-12B).

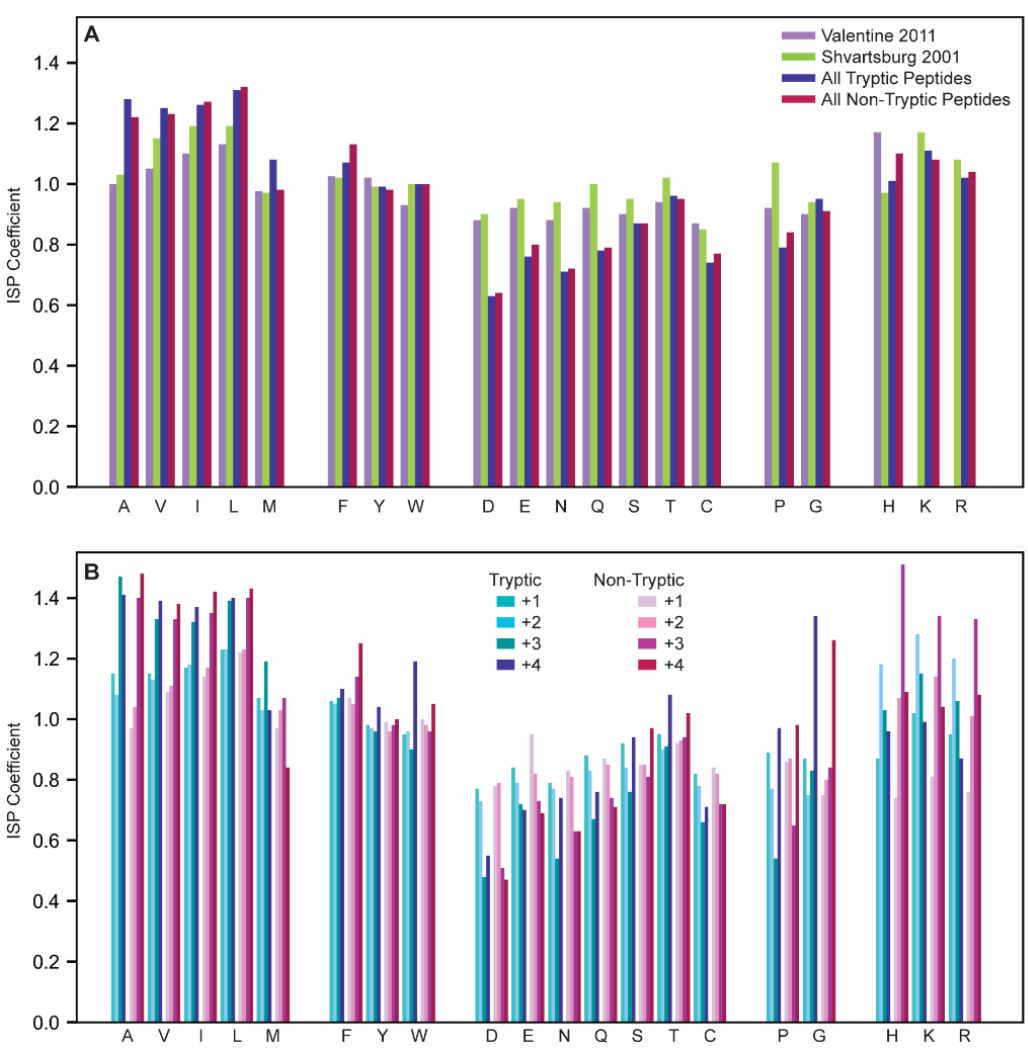

Figure 2-11. Comparison of previously reported ISP values with ones obtained for our dataset. (A) ISP values for tryptic and non-tryptic peptides vs. Valentine et al.15 and Shvartsburg et al.16 data (Model Step 1), (B) charge specific ISP values (Step 2).

The position dependence for basic Arg, His, Lys is also unique (Figure 2-12C-F). Generally, ISP values increase slightly from N- to C-terminus corresponding to their increasing interactions with the helix macro-dipole near C-terminus. This trend is more visible for 2+ and

3+ peptides compared to 1+. Polar acidic residues exhibit lower CCS values for 1+ internal position (Figure 2-12C), which is similar to Pro. Asp, and Glu ISP values for 3+ peptides showed the effect of interaction with macro-dipole opposite to the basic residues as ISPs decrease from N to C-terminus.

While we do have a diverse set of peptides derived from different protease cleavages, there could be some issues with representation of amino acids at particular positions which will result in model over-fitting. For example, position dependent ISP values for charged residues showed significant variation due to their small representation in the 1+ subset. Position dependent ISP values for 4+ charge state also vary significantly, making it hard to extract consistent ISP trends. In other words, CCS prediction model for the 5,733 4+ peptides with 520 parameters is over-fitted. Additive retention time prediction models in peptide RP-HPLC show representative results starting at a 1:5 parameter to peptide ratio,[124] suggesting significant variation in peptide conformation in IMS separation. Nonetheless, overall model was able to achieve an R^2 value of 0.981 and demonstrated robustness in predicting CCS values consistent with experimental trends.

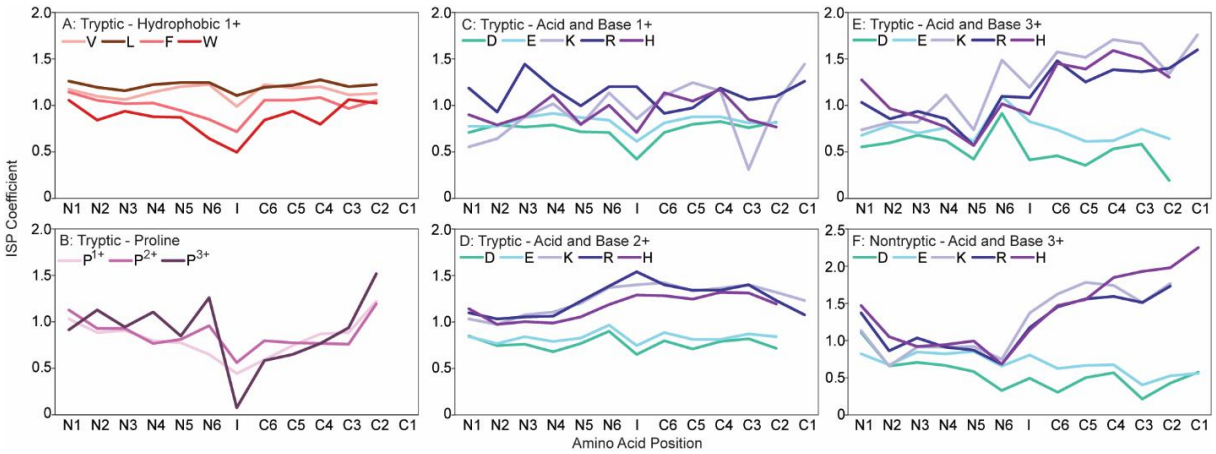

Figure 2-12. Selected examples of ISP positional trends: for hydrophobic amino acids (A), Pro in different charged peptides (B), acidic/basic amino acids among different charged peptides (C-F).

Table 2-3. Intrinsic size parameter values obtained by SSICalc model.

1+ tryptic peptides

	N1	N2	N3	N4	N5	N6	INT	C6	C5	C4	C3	C2	C1
A	1.13	1.04	1.06	1.27	1.28	1.37	1.15	1.27	1.27	1.25	1.12	1.18	-
V	1.19	1.11	1.08	1.16	1.22	1.24	1.00	1.24	1.20	1.22	1.13	1.14	-
I	1.22	1.16	1.14	1.17	1.16	1.06	1.12	1.12	1.19	1.24	1.14	1.17	-
L	1.28	1.21	1.17	1.24	1.26	1.26	1.12	1.21	1.23	1.29	1.22	1.24	-
M	1.07	1.01	1.00	1.06	1.05	0.95	1.01	1.11	1.16	1.09	1.02	1.08	-
F	1.16	1.07	1.03	1.04	0.96	0.86	0.73	1.07	1.07	1.10	0.98	1.07	-
Y	0.99	1.01	0.99	1.03	0.94	0.97	0.69	0.86	0.97	1.07	0.93	0.97	-
W	1.07	0.85	0.95	0.89	0.88	0.65	0.50	0.85	0.95	0.81	1.08	1.04	-
D	0.72	0.80	0.78	0.80	0.73	0.72	0.43	0.72	0.81	0.84	0.77	0.83	-
E	0.79	0.79	0.88	0.93	0.88	0.85	0.62	0.82	0.89	0.89	0.82	0.82	-
N	0.85	0.81	0.84	0.78	0.77	0.62	0.51	0.71	0.79	0.90	0.82	0.75	-
Q	0.92	0.88	0.92	0.93	0.86	0.79	0.66	0.75	0.90	0.91	0.85	0.82	-
S	0.93	0.84	0.86	0.99	0.94	0.86	0.73	1.01	1.06	1.00	0.91	0.92	-
T	1.00	0.95	0.93	1.00	1.04	0.91	0.72	0.96	1.03	1.01	0.91	0.94	-
C	0.90	0.83	0.88	0.82	0.74	0.54	0.63	0.68	0.88	0.90	0.79	0.76	-
P	1.05	0.90	0.92	0.81	0.79	0.66	0.45	0.60	0.76	0.88	0.90	1.24	-
G	1.00	0.75	0.85	1.00	0.94	0.90	0.56	1.03	0.95	0.95	0.80	0.95	-
H	0.91	0.80	0.90	1.13	0.81	1.02	0.72	1.15	1.06	1.19	0.86	0.78	-
K	0.56	0.65	0.89	1.03	0.80	1.15	0.87	1.12	1.26	1.17	0.32	1.03	1.46
R	1.20	0.94	1.46	1.20	1.01	1.22	1.22	0.93	0.99	1.20	1.08	1.11	1.28

1+ non-tryptic peptides

	N1	N2	N3	N4	N5	N6	INT	C6	C5	C4	C3	C2	C1
A	1.44	1.00	0.93	1.19	1.12	1.15	0.88	1.05	1.13	1.13	0.95	1.00	0.90
V	1.34	1.11	1.08	1.23	1.10	1.12	0.98	1.05	1.11	1.20	1.07	1.09	1.14
I	1.30	1.17	1.15	1.24	1.15	1.15	0.94	1.15	1.17	1.26	1.15	1.15	1.12
L	1.39	1.25	1.24	1.33	1.24	1.24	1.13	1.21	1.27	1.34	1.20	1.23	1.12
M	1.18	1.05	1.05	1.14	1.08	0.98	0.90	1.12	1.08	1.11	1.01	1.04	0.96
F	1.21	1.13	1.10	1.19	1.12	1.09	1.00	1.08	1.08	1.15	1.08	1.11	0.97
Y	1.16	1.08	1.03	1.12	1.01	0.90	0.85	1.02	1.05	1.12	1.02	1.10	0.92
W	1.11	1.08	1.07	1.10	1.05	0.89	0.72	1.07	1.03	1.09	1.06	1.03	0.87
D	0.75	0.89	0.86	0.94	0.84	0.83	0.72	0.83	0.91	0.94	0.80	0.86	1.19
E	0.93	0.84	0.86	0.94	0.83	0.83	0.66	0.80	0.88	0.93	0.82	0.88	1.35
N	1.12	0.89	0.89	0.95	0.90	0.82	0.62	0.86	0.96	0.94	0.84	0.91	0.80
Q	1.07	0.94	0.95	1.02	0.87	0.82	0.79	0.91	0.94	1.00	0.89	0.95	0.88
S	1.18	0.90	0.89	1.03	0.91	0.87	0.77	0.92	0.98	1.03	0.84	0.92	0.87
T	1.17	0.97	0.96	1.08	0.95	0.99	0.87	0.95	1.03	1.07	0.93	0.96	0.89
C	1.04	0.92	0.88	0.91	0.81	0.82	0.76	0.84	0.86	0.90	0.89	0.86	0.83
P	1.18	0.90	0.95	1.02	0.94	0.90	0.77	0.86	0.97	1.05	0.88	0.95	1.14
G	1.46	0.82	0.70	1.05	0.92	0.86	0.63	0.83	0.92	0.98	0.70	0.81	0.83
H	1.03	0.97	0.91	1.05	0.90	0.99	0.78	1.13	1.06	1.09	1.00	1.09	1.02
K	0.93	1.17	1.09	1.22	1.21	1.29	0.99	1.22	1.29	1.16	1.05	1.14	-
R	0.83	0.93	0.99	1.16	1.14	1.22	1.22	1.06	1.32	1.26	0.97	1.08	-

2+ tryptic peptides

	N1	N2	N3	N4	N5	N6	INT	C6	C5	C4	C3	C2	C1
A	1.25	1.06	1.04	1.06	1.20	1.53	1.14	1.23	1.01	0.96	0.94	1.11	-
V	1.26	1.17	1.15	1.13	1.22	1.40	1.11	1.20	1.12	1.05	1.01	1.14	-
I	1.26	1.26	1.22	1.17	1.25	1.44	1.16	1.28	1.16	1.14	1.09	1.22	-
L	1.32	1.31	1.30	1.22	1.30	1.47	1.21	1.31	1.21	1.19	1.16	1.27	-
M	1.13	1.08	1.02	1.03	1.04	1.23	1.00	1.06	0.99	0.99	1.01	1.02	-
F	1.16	1.12	1.08	1.07	1.08	1.20	1.01	1.10	1.04	1.02	1.03	1.06	-
Y	1.03	1.07	1.02	1.00	1.02	1.12	0.92	1.03	0.97	0.97	0.99	0.96	-
W	1.06	1.09	1.07	1.01	1.02	1.10	0.86	1.03	0.95	0.85	0.97	0.91	-
D	0.86	0.76	0.77	0.69	0.78	0.91	0.66	0.81	0.72	0.80	0.83	0.73	-
E	0.85	0.78	0.85	0.80	0.84	0.98	0.76	0.90	0.82	0.82	0.88	0.85	-
N	0.93	0.80	0.80	0.75	0.81	0.91	0.67	0.85	0.79	0.78	0.83	0.71	-

Q	1.00	0.90	0.93	0.83	0.84	0.98	0.74	0.95	0.82	0.83	0.87	0.86	-
S	0.99	0.88	0.87	0.86	0.93	1.14	0.82	0.97	0.84	0.80	0.89	0.76	-
T	1.04	0.99	0.92	0.91	1.00	1.17	0.91	0.99	0.88	0.87	0.87	0.87	-
C	0.95	0.87	0.86	0.78	0.82	0.87	0.67	0.86	0.78	0.76	0.82	0.79	-
P	1.14	0.94	0.94	0.78	0.82	0.97	0.57	0.81	0.79	0.78	0.77	1.21	-
G	1.15	0.83	0.87	0.82	0.97	1.31	0.69	0.94	0.70	0.63	0.64	0.59	-
H	1.16	0.99	1.02	1.00	1.07	1.20	1.31	1.30	1.26	1.34	1.33	1.21	-
K	1.05	0.99	1.09	1.12	1.22	1.39	1.42	1.44	1.35	1.38	1.42	1.34	1.25
R	1.11	1.05	1.07	1.08	1.24	1.40	1.56	1.42	1.36	1.36	1.42	1.25	1.09

2+ non-tryptic peptides

	N1	N2	N3	N4	N5	N6	INT	C6	C5	C4	C3	C2	C1
A	1.53	1.07	0.99	1.06	1.24	1.49	1.09	1.11	1.09	0.95	0.99	0.92	0.56
V	1.47	1.17	1.14	1.12	1.26	1.44	1.11	1.17	1.16	1.06	1.07	1.03	0.84
I	1.44	1.22	1.21	1.18	1.29	1.47	1.17	1.23	1.22	1.13	1.14	1.15	0.94
L	1.49	1.30	1.29	1.25	1.37	1.50	1.22	1.29	1.30	1.21	1.22	1.20	0.92
M	1.31	1.09	1.02	1.03	1.13	1.28	1.06	1.08	1.03	1.07	0.96	0.99	0.71
F	1.25	1.10	1.09	1.09	1.17	1.26	1.05	1.12	1.11	1.04	1.07	1.06	0.76
Y	1.18	1.04	1.04	1.02	1.08	1.17	0.92	1.07	1.06	1.00	1.01	0.99	0.69
W	1.15	1.02	1.08	1.03	1.03	1.07	0.94	1.10	1.08	0.93	1.00	1.00	0.74
D	1.12	0.83	0.79	0.74	0.86	0.98	0.74	0.87	0.89	0.78	0.79	0.79	0.48
E	1.05	0.86	0.85	0.83	0.92	1.04	0.81	0.91	0.88	0.84	0.81	0.80	0.60
N	1.11	0.86	0.83	0.75	0.89	1.02	0.79	0.92	0.89	0.82	0.84	0.78	0.53
Q	1.10	0.93	0.92	0.88	0.94	1.05	0.83	0.93	0.92	0.84	0.85	0.85	0.65
S	1.20	0.92	0.87	0.87	0.99	1.16	0.85	0.98	0.94	0.83	0.83	0.80	0.48
T	1.24	0.98	0.94	0.93	1.06	1.22	0.95	1.03	1.00	0.91	0.90	0.86	0.60
C	1.05	0.86	0.89	0.82	0.92	1.02	0.79	0.89	0.89	0.80	0.83	0.81	0.67
P	1.43	1.02	0.93	0.84	0.98	1.12	0.82	0.94	0.93	0.85	0.97	0.89	0.81
G	1.53	0.88	0.79	0.80	1.03	1.28	0.74	0.95	0.90	0.75	0.75	0.65	0.28
H	1.16	0.97	0.94	0.96	1.01	1.17	1.09	1.19	1.12	1.09	1.23	1.22	0.92
K	1.36	1.01	1.03	1.06	1.15	1.35	1.23	1.35	1.28	1.18	1.17	1.18	-
R	1.12	0.97	0.96	0.95	1.05	1.17	1.08	1.19	1.15	1.13	1.12	1.09	-

3+ tryptic peptides

	N1	N2	N3	N4	N5	N6	INT	C6	C5	C4	C3	C2	C1
A	0.77	1.58	1.21	1.40	1.18	2.14	1.83	1.53	1.14	1.06	0.86	1.19	-
V	1.08	1.71	1.19	1.32	1.12	1.86	1.59	1.40	1.13	1.06	1.03	1.26	-
I	1.09	1.60	1.38	1.46	1.21	1.84	1.46	1.29	1.26	1.14	1.20	1.42	-
L	1.20	1.69	1.50	1.60	1.35	1.91	1.56	1.39	1.32	1.10	1.18	1.27	-
M	0.86	1.57	0.92	1.44	1.12	1.49	1.28	1.16	1.06	1.04	1.24	1.23	-
F	1.14	1.48	1.22	1.16	1.02	1.47	1.10	1.12	0.90	0.99	1.04	1.12	-
Y	0.93	1.03	1.05	1.16	0.81	1.40	0.99	1.05	1.05	1.11	0.92	0.86	-
W	0.88	0.90	1.06	1.09	1.41	1.25	0.90	0.90	0.52	0.71	0.88	0.91	-
D	0.56	0.61	0.69	0.63	0.43	0.93	0.42	0.47	0.36	0.54	0.59	0.20	-
E	0.69	0.80	0.71	0.77	0.62	1.11	0.84	0.75	0.62	0.63	0.76	0.65	-
N	0.76	0.74	0.73	0.62	0.48	1.09	0.36	0.66	0.48	0.76	0.92	0.24	-
Q	0.81	0.98	0.82	0.95	0.59	1.19	0.56	0.57	0.63	0.70	0.88	0.85	-
S	0.52	0.94	0.89	0.90	0.64	1.49	0.83	0.96	0.53	0.77	0.85	0.22	-
T	0.78	1.21	0.90	1.14	0.74	1.45	1.03	0.91	0.63	0.71	0.98	0.59	-
C	1.17	1.04	0.84	0.87	0.87	1.01	0.52	0.56	0.68	0.59	0.76	0.63	-
P	0.93	1.14	0.96	1.12	0.86	1.28	0.08	0.59	0.66	0.78	0.95	1.54	-
G	0.54	1.11	1.30	1.17	0.42	1.95	0.89	0.93	0.80	0.82	0.49	0.11	-
H	1.29	0.98	0.89	0.78	0.58	1.03	0.92	1.47	1.41	1.61	1.52	1.32	-
K	0.75	0.83	0.83	1.13	0.75	1.51	1.21	1.60	1.54	1.73	1.69	1.36	1.78
R	1.05	0.87	0.95	0.87	0.58	1.11	1.10	1.50	1.27	1.40	1.38	1.42	1.62

3+ non-tryptic peptides

	N1	N2	N3	N4	N5	N6	INT	C6	C5	C4	C3	C2	C1
A	0.92	1.41	1.35	1.45	1.40	0.94	1.62	0.92	1.00	1.04	0.64	0.71	1.29
V	0.99	1.33	1.40	1.45	1.30	1.06	1.46	0.94	1.16	0.99	0.85	0.88	1.44

I	1.01	1.46	1.30	1.48	1.32	0.99	1.43	1.14	1.13	1.20	0.95	1.01	1.51
L	1.12	1.58	1.48	1.49	1.37	1.12	1.51	1.11	1.20	1.25	0.91	1.06	1.26
M	1.19	1.00	1.50	1.01	1.63	0.43	1.17	0.95	0.89	1.01	1.07	0.59	0.88
F	1.25	1.32	1.30	1.31	1.05	1.00	1.19	0.89	0.92	0.98	0.78	0.94	1.00
Y	0.93	1.15	1.12	1.14	1.10	1.01	1.05	0.82	0.89	0.93	0.75	0.80	0.76
W	1.13	1.32	1.20	1.00	1.04	0.88	0.92	0.72	0.74	0.95	0.91	0.70	0.94
D	1.12	0.66	0.71	0.67	0.59	0.33	0.50	0.31	0.51	0.57	0.22	0.43	0.58
E	0.83	0.68	0.85	0.83	0.86	0.66	0.81	0.63	0.67	0.68	0.41	0.53	0.56
N	0.85	0.88	0.77	0.82	0.72	0.44	0.56	0.50	0.44	0.60	0.52	0.61	0.72
Q	1.17	0.94	1.01	0.94	0.86	0.59	0.67	0.57	0.65	0.78	0.59	0.61	0.95
S	0.58	0.86	0.88	0.77	0.77	0.48	0.85	0.69	0.65	0.78	0.53	0.35	1.03
T	0.96	1.05	0.98	0.99	1.03	0.74	1.03	0.71	0.63	0.88	0.55	0.58	1.05
C	1.19	0.95	1.02	0.87	0.80	0.64	0.63	0.28	0.78	0.77	0.47	0.49	1.02
P	1.12	1.17	1.13	1.19	1.01	0.58	0.38	0.46	0.61	0.94	0.96	0.99	1.76
G	0.42	1.12	0.99	1.23	0.91	0.32	0.78	0.63	0.66	0.92	0.11	0.37	1.33
H	1.48	1.06	0.93	0.95	1.00	0.69	1.15	1.48	1.56	1.86	1.94	1.99	2.26
K	1.14	0.66	0.92	0.92	0.93	0.75	1.38	1.64	1.79	1.75	1.52	1.78	-
R	1.38	0.87	1.04	0.92	0.88	0.69	1.18	1.46	1.57	1.60	1.52	1.74	-

4+ tryptic peptides

	N1	N2	N3	N4	N5	N6	INT	C6	C5	C4	C3	C2	C1
A	-0.70	1.20	1.13	0.85	0.87	1.34	1.48	1.67	1.71	1.79	1.41	0.79	-
V	-0.13	1.17	1.38	2.01	1.25	1.21	1.40	1.08	1.24	1.43	1.17	1.47	-
I	0.12	1.53	0.64	1.22	1.48	1.57	1.44	1.17	1.51	1.10	1.74	1.47	-
L	0.86	1.59	1.63	1.60	1.34	1.24	1.36	1.50	1.16	1.22	0.87	1.40	-
M	-2.42	1.20	2.90	1.46	1.23	0.90	1.09	0.75	0.60	0.63	2.30	0.71	-
F	1.09	0.88	1.04	1.38	1.45	0.85	1.09	1.88	1.29	0.80	0.54	0.85	-
Y	0.14	1.30	0.65	0.75	1.53	0.79	1.10	1.08	1.29	1.47	0.49	1.31	-
W	0.39	1.58	2.12	1.99	0.69	0.33	1.01	2.12	0.48	1.82	0.18	-1.57	-
D	-0.49	0.76	0.70	0.57	0.42	0.96	0.45	0.90	0.78	0.96	0.79	0.55	-
E	0.29	0.40	0.75	0.92	0.81	0.82	0.68	0.84	0.62	0.88	1.08	0.93	-
N	-0.46	0.75	1.15	0.88	0.37	0.61	0.61	1.09	0.64	0.43	1.12	0.82	-
Q	0.45	0.84	0.80	0.70	1.18	1.25	0.67	1.06	0.79	0.78	0.88	0.38	-
S	-0.23	0.77	1.11	1.10	0.98	0.94	0.95	1.38	0.90	1.29	1.00	0.78	-
T	-0.94	0.89	0.73	1.38	1.07	1.22	1.19	1.06	0.83	1.01	0.98	0.88	-
C	0.24	1.21	0.52	1.32	0.29	0.25	0.59	0.98	1.26	1.26	0.57	1.22	-
P	0.50	1.47	0.79	1.64	1.07	1.48	0.69	0.87	0.80	1.76	0.93	1.69	-
G	-0.60	1.31	1.30	1.00	1.87	1.73	1.35	2.01	1.09	1.25	0.70	0.37	-
H	0.32	1.14	0.94	0.90	0.99	1.09	0.84	1.34	0.82	0.62	0.84	1.03	-
K	0.04	1.09	1.33	0.89	1.02	1.36	0.98	0.96	1.08	1.08	0.79	0.44	1.63
R	0.33	0.89	0.96	0.81	1.13	0.64	0.73	0.75	0.82	0.59	0.52	0.70	1.56

4+ non-tryptic peptides

	N1	N2	N3	N4	N5	N6	INT	C6	C5	C4	C3	C2	C1
A	-0.37	1.76	1.09	1.74	1.41	1.98	1.41	0.98	1.47	1.36	0.64	0.97	1.37
V	0.14	1.37	1.24	1.50	1.37	1.28	1.33	1.43	1.20	1.39	1.31	1.09	1.62
I	0.92	1.26	1.03	1.68	1.44	1.41	1.34	1.39	1.26	1.23	0.86	1.48	0.84
L	0.61	1.58	1.47	1.57	1.51	1.89	1.24	1.42	1.19	1.36	1.12	1.09	1.26
M	1.29	1.67	0.58	1.18	1.23	0.89	1.07	1.09	0.56	1.07	1.00	1.20	0.50
F	0.32	1.16	1.36	1.44	1.39	1.43	1.15	1.11	0.74	0.97	1.20	0.87	0.91
Y	-0.09	1.03	0.79	1.48	1.27	1.04	0.84	1.15	0.99	0.51	1.04	1.09	0.71
W	0.40	-0.39	2.66	1.23	1.64	1.94	1.00	2.02	0.91	1.50	0.51	0.11	0.74
D	0.65	0.51	0.55	0.63	0.52	0.62	0.45	0.51	0.88	0.91	0.67	-0.51	0.22
E	0.38	0.67	0.86	0.88	0.60	1.10	0.63	0.82	0.64	0.70	0.61	0.59	0.19
N	-0.03	0.62	1.13	0.65	0.85	0.82	0.46	0.68	1.34	0.75	0.50	0.23	0.38
Q	0.43	1.01	0.95	0.78	0.99	0.86	0.55	0.97	0.73	0.70	0.53	0.97	0.85
S	0.39	1.04	1.06	1.67	1.12	1.22	0.83	0.58	1.15	0.80	0.47	0.74	0.62
T	-0.20	1.30	0.97	1.13	0.85	1.22	0.92	0.60	0.75	0.84	1.27	1.05	0.95
C	0.52	0.87	0.91	1.30	0.85	0.67	0.56	1.18	0.84	0.32	0.85	0.70	0.46
P	0.33	0.79	1.14	1.17	1.46	1.37	0.77	0.99	0.75	1.59	1.31	1.43	1.41
G	0.18	1.42	1.57	1.64	1.66	2.69	1.14	0.64	1.39	1.14	0.57	0.84	1.09
H	0.40	0.85	1.02	0.99	1.15	1.23	0.67	1.35	1.03	1.38	1.13	1.40	1.17

K	1.00	0.90	1.15	1.03	0.79	1.12	0.80	1.15	1.29	1.37	0.82	1.14	-
R	1.20	1.05	0.81	0.98	0.65	0.71	0.66	1.18	1.22	1.15	1.08	1.39	-

Tryptic peptides are terminated by Lys/Arg and non-tryptic by any other amino acid, independent of the enzyme used.

Composition and sequence-specific features driving peptide IMS separation.

The original work on incorporating ISP concept has been done using collections of structurally similar 1+ and 2+ tryptic peptides without internal Arg/Lys residues.[109] The largest dataset used by Valentine et al. consisted of 2,094 peptides, 10.7 residues long on average.[71] We anticipated that inclusion of the entire population of peptides without restriction on protease type, number of basic residues, charges and peptide length will complicate model optimization. At the same time, it has provided additional information on the mechanism of ion mobility separation. Due to the increased size of the dataset, we were able to elucidate position-dependent ISP and found significant effect of the structural features rather than geometric size of individual residues.

The geometry of peptides in gas phase are strongly affected by the charge of the peptides. As seen in the plot of CCS versus molecular weight (Figure 2-13A), increasing in peptide charge leads to higher CCS values. To explain our findings, we use Counterman & Clemmer's approach[111] that have described the notion of exposed cationic charges being solvated by the backbone carbonyls of the peptide leading to the compact globular structures (Figure 2-13B). From our results in proteomic IMS separations, it suggests that ability of a peptide to solvate the charge as a globular peptide is based on the peptide flexibility and availability of polar groups as elaborated below in the instances of 1+ peptides. As charge density increases, the repulsive effects of cationic charges in proximity starts to be become an issue in charge solvation. The repulsion reduces the stability of globular structures and starts to approach other stable conformations that are more linear in orientation. The different stages of peptide structures will be described as: closed globular, open globular, hinged-helix, alpha-helix, and linear listed in order of increasing CCS.

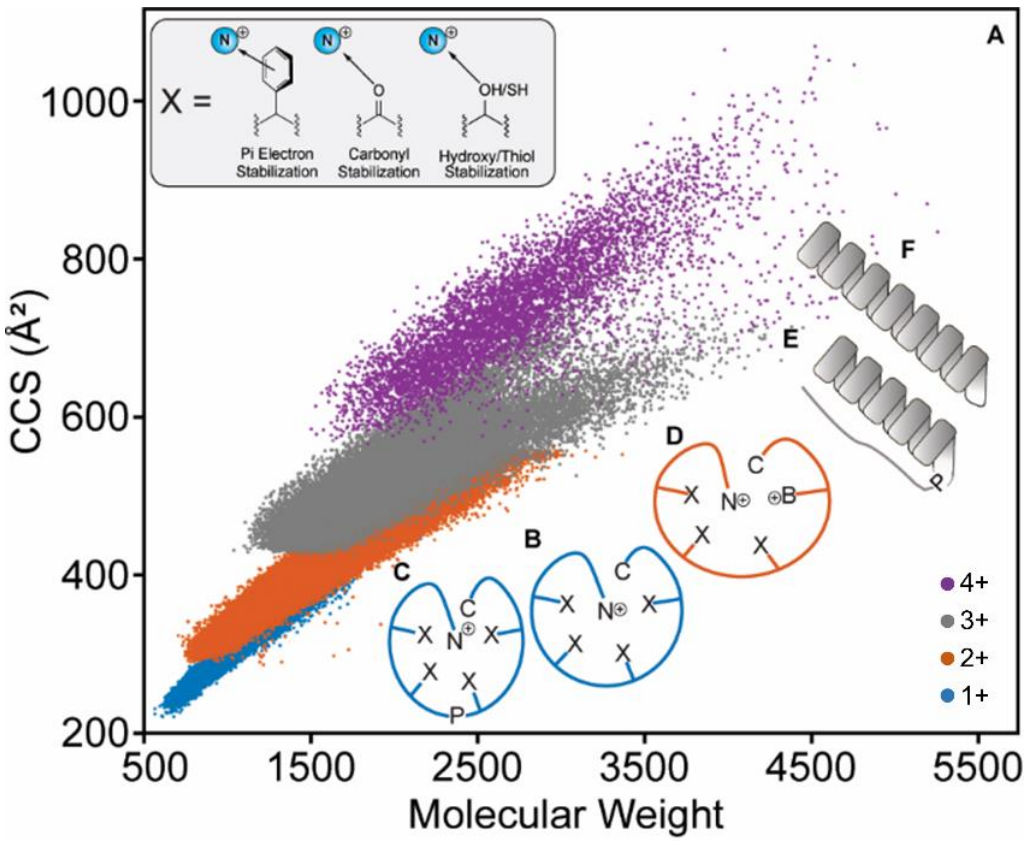

Figure 2-13. Compositional and sequence specific features driving the separatory behaviour observed in CCS vs. Molecular Weight plot (A). The geometry of the peptides in gas phase (C, B, D, E, F) ordered from lowest to highest CCS. N+ corresponds to the N-terminal, C to the C-terminal, B+ is an internal basic residue, and P is Pro.

In singly charged peptides (1+), the predominant geometry will be closed globular allowing this group to have the lowest CCS. The amino acid side chain structure will influence the size of the globular peptide based on steric interactions and electronic effects. The solvation of the exposed cation will be enhanced in the presence of partially negative functional groups as shown in Figure 2-13B. Acidic Asp and Glu show low ISP coefficients as they improve the solvation with their carboxylate side chains, which assist in compacting the globular structure. Asn and Gln also follow a similar trend where the carbonyl on the amide also assist in compacting the globule. Cys, Ser and Thr have polar thiol or hydroxyl groups that can stabilize the cation therefore exhibiting low ISPs. Aromatic amino acids stabilize the peptide electronically via the pi orbitals on the aromatic rings and are able to condense the peptide structure. In the case of aliphatic amino acids, their side groups do not contribute electronically to the cation stabilization but rather add steric bulk to the peptide leading to the observed increase in CCS. The flexible Gly and Pro do not contribute electronically or sterically but rather their flexibility allows for tighter peptide solvation to the cationic core allowing lower CCS conformations (Figure 2-13C).

For the case of doubly charged peptides (2+), they follow a similar trend in internal ISP values with +1 peptides; however, exhibiting smaller change in ISP values between terminal and internal positions. This suggests the 2+ peptides are also globular but given the electrostatic repulsion between the two positive charges, the peptide will not be able to fold as tightly (Figure 2-13D). This effect forces 2+ peptides to the open globular conformation consistent with the higher CCS than 1+ peptides found in our experimental data. This observation is supported by the divergence in acidic and basic amino acid ISP values shown for 1+ and 2+ peptides in Figure 2-12C and D where the mediation of an acidic side chain assists in lowering the repulsive effects therefore decreasing CCS and vice versa for basic amino acids.

The triply charged peptides (3+) exhibit a divergent pattern when CCS is plotted against molecular weight. Prior work in the field has demonstrated that the pattern can be attributed to two main peptide geometries:[111] a fast hinged-helix orientation and a slow alpha-helix orientation as displayed in Figure 3E, F. As peptides are now able to take helical conformations in this charge state, the ISP values of aliphatic amino acids are increased from their 2+ counter parts. The helix-breaker Pro exhibits the lowest ISP value for +3 charge state (especially for internal positions as shown in Figure 2-12B) due to their ability to bend the peptide to favour the hinged-helix orientation allowing the peptide to have lower CCS.[77,125] Similarly, acidic and polar amino acids also decrease in CCS from 2+ peptides as the effect of the cationic solvation is more drastic in larger ions found in the 3+ sets (Figure 2-11B).

Our findings on splitting population of 3+ peptides were confirmed by protease-specific features of CCS vs. *m/z* plots driven by interaction of acidic/basic residues with peptide macro-dipole. Peptides featuring acidic residues at C-termini and basic ones at N-termini (LysargiNase/LysN/GluC) tend to be in a compact conformation. Meanwhile trypsin/LysC/AspN peptides show more even distribution between two conformational states. Surprisingly, ions in the other charged states (1+, 2+ and 4+) also showed similar specificity, albeit with uneven distribution between conformational states. Lesser number of 1+ and 2+ peptides assume extended and 4+ compact conformation, respectively. Compared to previous studies, we can identify this novel finding due to the diversity of proteases employed.

Quadruply charged peptides (4+) in the past have not been well characterized due to their limited representation in the optimization datasets.[126] Based on our novel CCS information we conclude that the geometry of the peptides are generally more linear and helical than 3+ peptides. In terms of the aliphatic amino acids, the ISP values are largely similar to 3+ peptides (Figure 2-11B) supporting our notion that helicity is still a strong contributor in the 4+ charge state. Interestingly, Pro and Gly increase in ISP values (Figure 2-11B). Electrostatic contributions

from acidic amino acids in quadruply charged peptides are analogous to the triply charged peptides, whereas basic amino acids experience a decrease in ISP values (Figure 2-11B).

CONCLUSIONS

Through pairing high-throughput proteomics with IMS, we were able to collect a high-quality CCS database of ~134,000 peptides and establish the first sequence-specific model to predict peptide CCS. Our collections and resultant prediction model are detailed for each charge state, in enabling expansion for the current observations of 3+ and 4+ peptides in finer detail, and in attaining an R^2 value of 0.981 for the entire dataset. The gas phase peptide geometry dictates the CCS of the peptide and the conformations are heavily influenced by charge, sterics, and helical propensity of the constituent amino acids. Singly charged peptides have the lowest CCS in the entire dataset as it maintains a small profile in a closed globular conformation with the cation stabilized by backbone carbonyls or polar side groups. The globular structure can be further stabilized and condensed in the presence of Pro. For doubly charged peptides, the geometric behaviours are similar to 1+ peptides; however, the two cations experience electrostatic repulsion causing the structure to expand to an open globular conformation. Triply charged peptides establish two main conformations, a fast hinged-helix or a slow alpha helix structure. The ISP contributions of hydrophobic amino acids increase compared to the previous two charge states as these amino acids have high helical propensity favouring the alpha-helix conformation. Pro also exhibits the lowest ISP in 3+ peptides as its ability to bend the peptide favour the formation of the hinged-helix structure. We observe a divergent trend between acidic and basic amino acids' position dependent ISPs in triply charged peptides due to the macro-dipole interaction, which is also characteristic for helical structures. For the first time, 2+ peptides as well as 1+ and 4+ peptides were identified to exhibit similar splitting behaviour, due to the position of acidic/basic residues that favour helical stabilization via interaction with the peptide macro-dipole. Quadruply charged peptides maintain similar ISP values and trends as 3+ peptides with the exception of Pro and Gly increasing drastically in ISP. Other structural outliers have been observed for long and highly charged peptides with multiple proline residues. These motifs are one of the reasons for the high prediction errors observed in 3+ and 4+ peptides as the interactions of adjacent prolines may result in the formation of left-hand helices, which extends the peptide conformation. There is an active effort to understand the behaviour of different polyproline isomers; however, with current literature it is difficult to definitively align our diverse observations for such species. To fully elucidate the nature of our prediction errors, molecular dynamics paired with hydrogen-deuterium exchange experiments for a majority of the peptides will be needed to understand the true diversity of gas-phase peptide conformations.

Despite the difficulties in ascertaining outlier behaviours in our dataset, we are able to provide a variety of novel insights for the influence of peptide properties in real world CCS prediction.

EXPERIMENTAL SECTION

Materials

Ammonium bicarbonate (ABC), 2-amino-2-(hydroxymethyl)-1,3-propanediol hydrochloride (Tris-HCl), sodium deoxycholate (SDC), ammonium acetate (AA), sodium N-lauroylsarcosinate (SLS), tris(2-carboxyethyl)phosphine (TCEP), 2-chloroacetamide (CAA), calcium chloride, ethyl acetate, acetonitrile, acetic acid, trifluoroacetic acid, V8 protease (GluC), lysyl endopeptidase (LysC) and other chemicals were purchased from Fujifilm Wako (Osaka, Japan). Modified trypsin, chymotrypsin and AspN/LysN/LysargiNase were procured from Promega (Madison, WI)/Thermo Fisher Scientific (Waltham, MA)/Merck (Darmstadt, Germany), respectively. Polystyrene-divinylbenzene (SDB) and cation exchange-SR (SCX) Empore^{TM} disks were purchased from GL Sciences (Tokyo, Japan). Water was purified by a Millipore Milli-Q system (Bedford, MA).

HeLa cell culture and protein extraction

HeLa S3 (human cervical adenocarcinoma) cells were cultured to 80% confluence in 10-cm diameter dishes then harvested in lysis buffer containing protease inhibitors (Sigma-Aldrich, St. Louis, MO), 12 mM SDC, 12 mM SLS, 10 mM TCEP, 40 mM CAA in 100 mM Tris buffer (pH 8.5). The lysate was vortexed and sonicated on ice for 20 min. The final protein concentration of the sample was determined using the bicinchoninic acid (BCA) protein assay (Thermo Fisher Scientific).

E. coli culture and protein extraction

E. coli K12 strain BW25113 cells grown in Luria-Bertani (LB) cultures at 37 °C were used in this study. The cell pellet was prepared by centrifugation at 4500g for 10 min and was resuspended in 10 mL of ice-cold 1 M KCl, 15 mM Tris (pH 7.4). A protease inhibitor AEBSF was added to the final concentration of 10 mM. Proteins were extracted with 12 mM SDC, 12 mM SLS, and 50 mM ammonium bicarbonate; reduced with 10 mM dithiothreitol at room temperature for 30 min; and alkylated with 55 mM iodoacetamide in the dark at room temperature for 30 min.

Protein Digestion

The proteins were digested using previously described phase-transfer surfactants (PTS) method.[38] For LysargiNase digestion, protein extract was diluted 10-fold by using 10 mM $CaCl_2$ and digested with LysargiNase (1: 40 w/w) overnight at 37 °C. For other proteases, extracts

were diluted 5-fold by with 50 mM ABC and digested overnight at 37 °C using trypsin (1: 40 w/w), LysC (1: 20 w/w), LysN (1: 50 w/w), GluC (1: 20 w/w), AspN (1: 40 w/w), chymotrypsin (1: 50 w/w) protease/substrate ratios. After enzymatic digestion, an equal volume of ethyl acetate was added, and the mixture was acidified with 0.5% trifluoroacetic acid (final concentration) according to the PTS protocol. The mixture was shaken for 1 min and centrifuged at 15,700 g for 2 min to separate ethyl acetate phase from the aqueous phase. The latter was collected and desalted by using SDB-StageTips.[127] The amount of peptides was quantified by LC-UV at 214 nm relative to standard BSA tryptic digests and kept in 80% ACN and 0.5% TFA at −20 °C until use.

Peptide fractionation by Strong Cation Exchange StageTip

The preparation of SCX-StageTips were performed in 200-µL tips format as described previously.[128] SCX buffers were made in 15% acetonitrile with stepwise increase of elution buffer strength: F1 - 0.1% TFA; F2 - 1.0% TFA; F3 - 2.0% TFA; F4 - 3.0% TFA; F5 - 3.0% TFA and 100 mM AA; F6 - 3.0% TFA and 500 mM AA and; F7 - 0.1% TFA and 500 mM AA, as described previously.[28] Two technical replicate SCX separations have been done for each digest. Conditioning and equilibration were done through sequential passing 100 µL buffer and centrifugation at 1000 × g for 1 min of the following buffers: methanol, F7, F5 and F1. 20 µg of digests from HeLa cell lysate were loaded into the SCX-StageTip, spun at 1000 × g for 1 min and the eluate was collected as flow-through (FT). The bound peptides eluted with 100 µL of F1 by centrifugation at 1000 × g for 1 min. Subsequent fractions were collected using 100 µL of SCX buffers F2 to F7. F5-F7 were lyophilized, resuspended in 50 µL of 0.1% TFA and desalted by SDB-StageTips.

NanoLC/TIMS/Q/TOF analysis

NanoLC/MS/MS analyses were performed using a hybrid ESI/TIMS/Q/TOF mass spectrometer (timsTOF Pro, Bruker, Bremen, Germany), which was connected to an Ultimate 3000 pump (Thermo Fisher Scientific, Germering, Germany) and an HTC-PAL autosampler (CTC Analytics, Zwingen, Switzerland). Peptides were separated at 50 °C using 150 mm length × 100 µm ID capillary column with 6 µm ID ESI tip, packed with Reprosil-Pur 120 C18-AQ 3 µm particles (Dr. Maisch, Ammerbuch, Germany). The injection volume was 5 µL and the flow rate was 500 nL/min. The mobile phases consisted of (A) 0.5% acetic acid and (B) 0.5% acetic acid and 80% ACN. A two-step linear gradient of 5−40% B in 45 min, 40−99% B in 1 min, keeping at 99% B for 5 min was employed.

The timsTOF Pro mass spectrometer was operated in parallel accumulation–serial fragmentation (PASEF) mode.[129] Two methods were applied in IMS separation. Method 1 was applied for covering singly and multiply charged ions and method 2 was mainly used for depleting the contaminants usually singly charged background ions, respectively. The setting parameters are described in Table 2-4.

Table 2-4. Parameter settings of timsTOF Pro mass spectrometer in PASEF analysis.

Spray voltage	4000 V	
RF potential on electrodynamic funnel	350 Vpp	
Mode of IMS Separation	Method 1	Method 2
PASEF number	5	10
Potential for ramp start	180 V	130 v
Ramp time	250 ms	100 ms
Scan range of 1/K0	0.65 – 2.27 V·s/cm^2	0.7 – 1.40 V·s/cm^2
Scan range of *m/z* for MS and MS/MS	100 – 1750 *m/z*	100 – 1700 *m/z*
Target value for PASEF-MS/MS scan	2.40e + 04	
dynamic exclusion	25 s	
Step-up in collision energy		
Ramp time	collision energy	
0 –19%	52 eV	
19 –38%	47 eV	
38–57%	42 eV	
57–76%	37 eV	
76–100%	32 eV	
Isolation width of quadrupole (Th)		
m/z range	Charge dependent isolation width (1+/2+/3+)	
<200	4.0/3.0/3.0	2.0/2.0/2.0
200-700	5.0/4.0/4.0	2.0/2.0/2.0
700-800	5.0/4.0/4.0	3.0/3.0/3.0
800-1500	6.0/5.0/4.0	3.0/3.0/3.0
>1500	7.0/6.0/5.0	3.0/3.0/3.0

TIMS funnel's voltages were linearly calibrated using Agilent ESI-L Tuning Mix to obtain reduced ion mobility coefficients ($1/K_0$) for three selected ions (*m/z* 622, 922, 1222).[130] The $1/K_0$ was converted to CCS using the Mason-Schamp equation (Eq.1).[131]

$$\mathrm{CCS} = \frac{3ze}{16n_0}\sqrt{\frac{2\pi}{\mu k_B \mathrm{T}}}\frac{1}{K_0} \qquad \text{(Eq.1)}$$

The z is the charge of the ions, e is the elemental charge (1.602×10^{-19} A·s), n_0 is Loschmidt constant (2.686×1025 m^{-3}), k_b is Boltzman's constant (1.380×10^{-23} $kg \cdot m^2 \cdot K^{-1} \cdot s^{-2}$), μ is the reduced mass ($m_i m_g / (m_i + m_g)$, m_i is the mass of ion; m_g is the mass of N_2, 1 Da = 1.660×10^{-27} kg), K_0 is the reduced mobility, (10^{-4} $cm^2 \cdot V^{-1} \cdot s^{-1}$) and T is the temperature (305 K). For the CCS calculation, pure N_2 is assumed as the drift gas.

Peptide identification and retention time prediction data filtering

The peak list in mascot generic format ".mgf" was generated by MaxQuant v1.6.7.0,[132] encoding information on both retention time and $1/K_0$ for each spectrum. The peptides were identified using X!Tandem Cyclone (12.10.01.1)[133] against human subset of the Swiss-Prot database (July 2016 extraction) with 20 ppm mass tolerance for both precursor and product ions. Carbamidomethyl of cysteine was set as a fixed modification. Oxidation of methionine and tryptophan, deamidation of glutamine and asparagine, cyclization of N-terminal glutamine and cysteine and protein N-terminal acetylation were allowed as variable modifications, and strict enzymatic specificity allowing for up to 2 missed cleavages as search parameters. Redundant peptide identifications have been removed leaving the most intense peptide MS/MS hits with their correspondent $1/K_0$ and retention time values. Peptides with variable modifications were also removed for CCS prediction. All peptides with confidence score log (e) < -1 or better were additionally filtered using latest version of SSRCalc retention time prediction model.[117] All peptides with retention time prediction error of more than ±6 min and low confidence score (-3 < log (e) <-1) have been removed as shown in Figure 2-2.

Model optimization

The preliminary length-specific ensemble of multiple linear regression (LS-MLR) models by R package[134] used to explore the variable space in CCS prediction has been derived for peptides with the selected charges and length (Eq.2):

$$\mathrm{CCS} = \sum_{j=1}^{25} \sum_{k=1}^{7} P_j G_k + b_0 \qquad \text{(Eq.2)}$$

where, P_j is the position and group dependent coefficients, G_k is the mass of each amino acid and b_0 is a constant. Amino acids have been grouped in seven categories based on their physicochemical properties as follows: basic K, R and H; acidic D and E; polar S, T, N and Q; aliphatic A, V, I and L; aromatic F, W and Y; aliphatic/polar side chains M and carbamidomethyl-Cys; P and G as amino acids with low helical propensity.

On the other hand, the final algorithm of the Sequence-Specific Ion Mobility Calculator (SSICalc) encodes 13 position-dependent ISP values (*j*) for each amino acid (*i*) in a charge (*z*) and protease (*e*) tryptic/non-tryptic dependence: six on each terminus plus internal position. Our equation for the SSICalc model is shown in Eq.3 as the summation of a coefficient (*P*) multiplied by the number of amino acids (*AA*) in the peptide with the corresponding *e*, *z*, *i*, *j* state listed above and mass of the amino acid (G_i) along with a constant b_0 term for the combined model:

$$\mathrm{CCS} = \sum_{e=1}^{2} \sum_{z=1}^{4} \sum_{i=1}^{20} \sum_{j=1}^{13} \left(P_{e,z,i,j} * AA_{e,z,i,j} * G_i \right) + b_0 \qquad \text{(Eq.3)}$$

Optimization of the charge sub-divided models followed a simple stochastic hill-climbing approach maximizing to the highest R^2 correlation. In each iteration of the optimization, a randomly selected parameter was adjusted along a shift value until the prediction versus observed CCS stopped improving until which a subsequent parameter is selected for optimization. The initial variable-space parameters were set to a matrix of ones and the signed shift value was randomly selected.

SUMMARY

During my PhD study, I developed two approaches, a novel enrichment method for protein N-terminal peptides and a novel model to predict peptide CCS values. The former one combines the new protease, TrypN, and SCX chromatography to achieve a simple and rapid isolation. For the latter one, position-dependent correction coefficients are applied to establish a Sequence-Specific Ion mobility Calculator (SSICalc).

In Chapter 1, I described a simple and rapid method to enrich protein N-terminal peptides by strong cation exchange chromatography according to a retention model based on the charge/orientation of peptides. This approach was applied to 20 μg of human HEK293T cell lysate proteins to profile the N-terminal proteome. On average, 1,550 acetylated and 200 unmodified protein N-terminal peptides were successfully identified in a single LC/MS/MS run with less than 3% contamination with internal peptides. The method was further applied to beige adipocytes for large-scale N-terminome profiling. In total, 3,016 CanNt-pepts, 4,225 NeoNt-pepts were identified with 2,124 quantitative CanNt-pepts and 1,301 NeoNt-pepts. Integrating the temporal profiling and cleavage site preferences of NeoNt-pepts and 15 proteases, Pmpcb and Plg showed high activity in the late stages of beige maturation and were verified by protease knockdown.

In Chapter 2, more than 134,000 peptides of four different charge states were acquired using a two-dimensional LC/trapped ion mobility spectrometry/quadrupole/time-of-flight MS analysis of HeLa cell digests created using 7 different proteases and was converted to CCS values. Position dependent ISPs were independently optimized, resulting in prediction accuracy of ~0.981 for the entire population of peptides. Overall, the N-terminal peptide enrichment method shows highly specific and sensitive and can be used for large-scale profiling of proteolytic processes. High precision CCS predictions have been established and can be further applied.

In my perspective, reducing the sample complexity is a crucial to enhance the sensitivity of the shotgun proteome, both in terms of the complexity of the peptide in the sample and the confidence of the peptide identification. Appropriate methods to reduce complexity should attempt to keep the procedures simple and prevent instabilities such as chemical reactions or contaminants. The N-terminal peptide enriched method developed in this thesis is a perfect example to show that about 2000 protein N-terminal peptides can be identified with only 20 μg of starting material without chemical reactions or tedious steps of interference. In the near future,

I expect to see more separation methods for peptides, as well as improvements in proteome coverage.

ACKNOWLEDGEMENT

I would like to thank my supervisor, Professor Yasushi Ishihama (Graduate School of Pharmaceutical Sciences, Kyoto University), for his guidance during my doctoral studies. I am grateful for your support in all aspects of my academic research and abroad life, and I could not have completed this difficult challenge without your kindness and continued support. In the final stages of my doctoral studies, I realized how fortunate I was to have been your student and to have completed my studies under your tutelage. I look forward to enjoying research with you again in the future.

I would also like to thank Professor Juri Rappsilber, Professor Oleg Krokhin, Professor Hsin-Yi Chang, Dr. Chia-Feng Tsai, Professor Naoyuki Sugiyama, Professor Norie Araki, Professor Koshi Imami, Professor Miao-Hsia Lin, Professor Kosuke Ogata, all past and current members of the lab for their help and kindness.

Finally, I would like to thank my family and friends, especially my parents, for their love and support.

Always.

SUPPLEMENTAL TABLES

Supplemental Table 1. Identification list of TrypN- and trypsin-mixed HEK293T peptides used in Figure 1-2.

Supplemental Table 2. Identification list of TrypN-digested HEK293T peptides used in Figure 1-4, 1-5.

Supplemental Table 3. Identification list of TrypN-digested *E. coli* peptides used in Table 1-1, Figure 1-6.

Supplemental Table 4. Identification list of TrypN-digested HEK293T peptides used in Table 1-3 and Figure 1-8.

Supplemental Table 5. Identified peptides in N-terminome of beige adipocytes used in Figure 1-11, 1-13A.

Supplemental Table 6. Identified protein groups in global proteome of beige adipocytes used in Figure 1-11, 1-13B.

Supplemental Table 7. Peptide retention lists used in Fig 2-2.

Supplemental Table 8. Identified peptides in RP-LC/TIMS/Q/TOF analysis used in Figure 2-4.

Supplemental Table 1-4 are available free of charge at ASBMB website (https://www.mcponline.org/cms/10.1074/mcp.TIR120.002148/attachment/143946ff-5f35-433e-8c14-c4e8069b9ea8/mmc1.xlsx).

Supplemental Table 5,6 have been deposited with the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the jPOST partner repository (http://jpostdb.org) with the data set identifier PXD024334/JPST001099.

Supplemental Table 7,8 have been deposited with the ProteomeXchange Consortium via the jPOST partner repository with the data set identifier PXD021440/JPST000959 and PXD022800/JPST001017.

REFERENCES

(1) Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D. N., et al. *Nature* **2014**, *509*, 575-581.

(2) Wilhelm, M.; Schlegl, J.; Hahne, H.; Gholami, A. M.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J. H.; Bantscheff, M.; Gerstmair, A., et al. *Nature* **2014**, *509*, 582-587.

(3) Iwasaki, M.; Sugiyama, N.; Tanaka, N.; Ishihama, Y. *J. Chromatogr. A* **2012**, *1228*, 292-297.

(4) Wolters, D. A.; Washburn, M. P.; Yates, J. R. *Anal. Chem.* **2001**, *73*, 5683-5690.

(5) Masuda, T.; Sugiyama, N.; Tomita, M.; Ishihama, Y. *Anal. Chem.* **2011**, *83*, 7698-7703.

(6) Thingholm, T. E.; Jorgensen, T. J. D.; Jensen, O. N.; Larsen, M. R. *Nat. Protoc.* **2006**, *1*, 1929-1935.

(7) Zhang, H. M.; Guo, T. N.; Li, X.; Datta, A.; Park, J. E.; Yang, J.; Lim, S. K.; Tam, J. P.; Sze, S. K. *Mol. Cell. Proteomics* **2010**, *9*, 635-647.

(8) Bouwmeester, R.; Gabriels, R.; Hulstaert, N.; Martens, L.; Degroeve, S. *bioRxiv* **2020**, 2020.2003.2028.013003.

(9) Gabriels, R.; Martens, L.; Degroeve, S. *Nucleic Acids Res.* **2019**, *47*, W295-W299.

(10) Nakahigashi, K.; Takai, Y.; Kimura, M.; Abe, N.; Nakayashiki, T.; Shiwa, Y.; Yoshikawa, H.; Wanner, B. L.; Ishihama, Y.; Mori, H. *DNA Res.* **2016**, *23*, 193-201.

(11) Ingolia, N. T. *Nat. Rev. Genet.* **2014**, *15*, 205-213.

(12) Van Damme, P.; Gawron, D.; Van Criekinge, W.; Menschaert, G. *Mol. Cell. Proteomics* **2014**, *13*, 1245-1261.

(13) Hwang, C. S.; Shemorry, A.; Varshavsky, A. *Science* **2010**, *327*, 973-977.

(14) Starheim, K. K.; Gevaert, K.; Arnesen, T. *Trends Biochem. Sci.* **2012**, *37*, 152-161.

(15) Mahrus, S.; Trinidad, J. C.; Barkan, D. T.; Sali, A.; Burlingame, A. L.; Wells, J. A. *Cell* **2008**, *134*, 866-876.

(16) McDonald, L.; Robertson, D. H. L.; Hurst, J. L.; Beynon, R. J. *Nat. Methods* **2005**, *2*, 955-957.

(17) Leitner, A. *Anal. Chim. Acta* **2018**, *1000*, 2-19.

(18) Klein, T.; Eckhard, U.; Dufour, A.; Solis, N.; Overall, C. M. *Chem. Rev.* **2018**, *118*, 1137-1168.

(19) Xu, G.; Shin, S. B.; Jaffrey, S. R. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 19310-19315.

(20) Varland, S.; Osberg, C.; Arnesen, T. *Proteomics* **2015**, *15*, 2385-2401.

(21) Lai, Z. W.; Petrera, A.; Schilling, O. *Curr. Opin. Chem. Biol.* **2015**, *24*, 71-79.

(22) Gevaert, K.; Goethals, M.; Martens, L.; Van Damme, J.; Staes, A.; Thomas, G. R.; Vandekerckhove, J. *Nat. Biotechnol.* **2003**, *21*, 566-569.

(23) Kleifeld, O.; Doucet, A.; auf dem Keller, U.; Prudova, A.; Schilling, O.; Kainthan, R. K.; Starr, A. E.; Foster, L. J.; Kizhakkedathu, J. N.; Overall, C. M. *Nat. Biotechnol.* **2010**, *28*, 281-288.

(24) Venne, A. S.; Solari, F. A.; Faden, F.; Paretti, T.; Dissmeyer, N.; Zahedi, R. P. *Proteomics* **2015**, *15*, 2458-2469.

(25) Chen, L. F.; Shan, Y. C.; Weng, Y. J.; Sui, Z. G.; Zhang, X. D.; Liang, Z.; Zhang, L. H.; Zhang, Y. K. *Anal. Chem.* **2016**, *88*, 8390-8395.

(26) Na, C. H.; Barbhuiya, M. A.; Kim, M. S.; Verbruggen, S.; Eacker, S. M.; Pletnikova, O.; Troncoso, J. C.; Halushka, M. K.; Menschaert, G.; Overall, C. M.; Pandey, A. *Genome Res.* **2018**, *28*, 25-36.

(27) Yeom, J.; Ju, S.; Choi, Y.; Paek, E.; Lee, C. *Sci. Rep-Uk* **2017**, *7*.

(28) Adachi, J.; Hashiguchi, K.; Nagano, M.; Sato, M.; Sato, A.; Fukamizu, K.; Ishihama, Y.; Tomonaga, T. *Anal. Chem.* **2016**, *88*, 7899-7903.

(29) Essader, A. S.; Cargile, B. J.; Bundy, J. L.; Stephenson, J. L., Jr. *Proteomics* **2005**, *5*, 24-34.

(30) Alpert, A. J.; Petritis, K.; Kangas, L.; Smith, R. D.; Mechtler, K.; Mitulovic, G.; Mohammed, S.; Heck, A. J. R. *Anal. Chem.* **2010**, *82*, 5253-5259.

(31) Helbig, A. O.; Gauci, S.; Raijmakers, R.; van Breukelen, B.; Slijper, M.; Mohammed, S.; Heck, A. J. R. *Mol. Cell. Proteomics* **2010**, *9*, 928-939.

(32) Gauci, S.; Helbig, A. O.; Slijper, M.; Krijgsveld, J.; Heck, A. J.; Mohammed, S. *Anal. Chem.* **2009**, *81*, 4493-4501.

(33) Wilson, J. P.; Ipsaro, J. J.; Del Giudice, S. N.; Turna, N. S.; Gauss, C. M.; Dusenbury, K. H.; Marquart, K.; Rivera, K. D.; Pappin, D. J. *J. Proteome Res.* **2020**, *19*, 1459-1469.

(34) Tallant, C.; Garcia-Castellanos, R.; Seco, J.; Baumann, U.; Gomis-Ruth, F. X. *J. Biol. Chem.* **2006**, *281*, 17920-17928.

(35) Huesgen, P. F.; Lange, P. F.; Rogers, L. D.; Solis, N.; Eckhard, U.; Kleifeld, O.; Goulas, T.; Gomis-Ruth, F. X.; Overall, C. M. *Nature Methods* **2015**, *12*, 55-58.

(36) Koneru, L.; Ksiazek, M.; Waligorska, I.; Straczek, A.; Lukasik, M.; Madej, M.; Thogersen, I. B.; Enghild, J. J.; Potempa, J. *Biol. Chem.* **2017**, *398*, 395-409.

(37) Tsiatsiani, L.; Giansanti, P.; Scheltema, R. A.; van den Toorn, H.; Overall, C. M.; Altelaar, A. F. M.; Heck, A. J. R. *J. Proteome Res.* **2017**, *16*, 852-861.

(38) Masuda, T.; Tomita, M.; Ishihama, Y. *J. Proteome Res.* **2008**, *7*, 731-740.

(39) Raijmakers, R.; Neerincx, P.; Mohammed, S.; Heck, A. J. *Chem Commun (Camb)* **2010**, *46*, 8827-8829.

(40) Gussakovsky, D.; Neustaeter, H.; Spicer, V.; Krokhin, O. V. *Anal. Chem. (Wash.)* **2017**, *89*, 11795-11802.

(41) Harms, M.; Seale, P. *Nat. Med.* **2013**, *19*, 1252-1263.

(42) Kajimura, S.; Spiegelman, B. M.; Seale, P. *Cell Metab* **2015**, *22*, 546-559.

(43) Giralt, M.; Villarroya, F. *Endocrinology* **2013**, *154*, 2992-3000.

(44) Wilson-Fritch, L.; Burkart, A.; Bell, G.; Mendelson, K.; Leszyk, J.; Nicoloro, S.; Czech, M.; Corvera, S. *Mol. Cell. Biol.* **2003**, *23*, 1085-1094.

(45) Chang, C.-H.; Chang, H.-Y.; Rappsilber, J.; Ishihama, Y. *Mol. Cell. Proteomics* **2021**, 100003.

(46) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. *Nucleic Acids Res.* **2009**, *37*, 1-13.

(47) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. *Nat. Protoc.* **2009**, *4*, 44-57.

(48) Ikeda, K.; Maretich, P.; Kajimura, S. *Trends Endocrinol. Metab.* **2018**, *29*, 191-200.

(49) Pope, B. D.; Warren, C. R.; Parker, K. K.; Cowan, C. A. *Trends Cell Biol.* **2016**, *26*, 745-755.

(50) Forner, F.; Kumar, C.; Luber, C. A.; Fromme, T.; Klingenspor, M.; Mann, M. *Cell Metab* **2009**, *10*, 324-335.

(51) Rawlings, N. D.; Alan, J.; Thomas, P. D.; Huang, X. D.; Bateman, A.; Finn, R. D. *Nucleic Acids Res.* **2018**, *46*, D624-D632.

(52) Wang, W.; Seale, P. *Nat. Rev. Mol. Cell Biol.* **2016**, *17*, 691-702.

(53) Colaert, N.; Helsens, K.; Martens, L.; Vandekerckhove, J.; Gevaert, K. *Nat. Methods* **2009**, *6*, 786-787.

(54) Humphrey, S. J.; Azimifar, S. B.; Mann, M. *Nat. Biotechnol.* **2015**, *33*, 990-U142.

(55) Rappsilber, J.; Ishihama, Y.; Mann, M. *Anal. Chem.* **2003**, *75*, 663-670.

(56) Riley, M.; Abe, T.; Arnaud, M. B.; Berlyn, M. K. B.; Blattner, F. R.; Chaudhuri, R. R.; Glasner, J. D.; Horiuchi, T.; Keseler, I. M.; Kosuge, T.; Mori, H.; Perna, N. T.; Plunkett, G.; Rudd, K. E.; Serres, M. H.; Thomas, G. H.; Thomson, N. R.; Wishart, D.; Wanner, B. L. *Nucleic Acids Res.* **2006**, *34*, 1-9.

(57) Robles, M. S.; Humphrey, S. J.; Mann, M. *Cell Metab* **2017**, *25*, 118-127.

(58) Konijnenberg, A.; Butterer, A.; Sobott, F. *Biochim. Biophys. Acta-Prroteins Proteom.* **2013**, *1834*, 1239-1256.

(59) Baker, E. S.; Burnum-Johnson, K. E.; Ibrahim, Y. M.; Orton, D. J.; Monroe, M. E.; Kelly, R. T.; Moore, R. J.; Zhang, X.; Theberge, R.; Costello, C. E.; Smith, R. D. *Proteomics* **2015**, *15*, 2766-2776.

(60) Kanu, A. B.; Dwivedi, P.; Tam, M.; Matz, L.; Hill, H. H. *J. Mass Spectrom.* **2008**, *43*, 1-22.

(61) Zhou, Z. W.; Shen, X. T.; Tu, J.; Zhu, Z. J. *Anal. Chem.* **2016**, *88*, 11084-11091.

(62) Winter, D. L.; Mastellone, J.; Kabir, K. M. M.; Wilkins, M. R.; Donald, W. A. *Anal. Chem.* **2019**, *91*, 11827-11833.

(63) Pfammatter, S.; Bonneil, E.; McManus, F. P.; Prasad, S.; Bailey, D. J.; Belford, M.; Dunyach, J. J.; Thibault, P. *Mol. Cell. Proteomics* **2018**, *17*, 2051-2067.

(64) Uetrecht, C.; Rose, R. J.; van Duijn, E.; Lorenzen, K.; Heck, A. J. R. *Chem. Soc. Rev.* **2010**, *39*, 1633-1655.

(65) Clowers, B. H.; Belov, M. E.; Prior, D. C.; William, F. D.; Ibrahim, Y.; Smith, R. D. *Anal. Chem.* **2008**, *80*, 2464-2473.

(66) Xuan, Y.; Creese, A. J.; Horner, J. A.; Cooper, H. J. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 1963-1969.

(67) Akashi, S.; Downard, K. M. *Anal. Bioanal. Chem.* **2016**, *408*, 6637-6648.

(68) Haynes, S. E.; Polasky, D. A.; Dixit, S. M.; Majmudar, J. D.; Neeson, K.; Ruotolo, B. T.; Martin, B. R. *Anal. Chem.* **2017**, *89*, 5670-5673.

(69) Chouinard, C. D.; Nagy, G.; Webb, I. K.; Shi, T. J.; Baker, E. S.; Prost, S. A.; Liu, T.; Ibrahim, Y. M.; Smith, R. D. *Anal. Chem.* **2018**, *90*, 10889-10896.

(70) Sisley, E. K.; Ujma, J.; Palmer, M.; Giles, K.; Fernandez-Lima, F. A.; Cooper, H. J. *Anal. Chem.* **2020**, *92*, 6321-6326.

(71) Valentine, S. J.; Ewing, M. A.; Dilger, J. M.; Glover, M. S.; Geromanos, S.; Hughes, C.; Clemmer, D. E. *J. Proteome Res.* **2011**, *10*, 2318-2329.

(72) Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. *Nat. Methods* **2007**, *4*, 923-925.

(73) Ogata, K.; Ishihama, Y. *Anal. Chem.* **2020**, *92*, 8037-8040.

(74) Meier, F.; Brunner, A.-D.; Frank, M.; Ha, A.; Bludau, I.; Voytik, E.; Kaspar-Schoenefeld, S.; Lubeck, M.; Raether, O.; Aebersold, R.; Collins, B. C.; Röst, H. L.; Mann, M. *bioRxiv* **2020**, 656207.

(75) Tao, L.; Dahl, D. B.; Perez, L. M.; Russell, D. H. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1593-1602.

(76) McLean, J. R.; McLean, J. A.; Wu, Z. X.; Becker, C.; Perez, L. M.; Pace, C. N.; Scholtz, J. M.; Russell, D. H. *J. Phys. Chem. B* **2010**, *114*, 809-816.

(77) Florance, H. V.; Stopford, A. P.; Kalapothakis, J. M.; McCullough, B. J.; Bretherick, A.; Barran, P. E. *Analyst* **2011**, *136*, 3446-3452.

(78) May, J. C.; McLean, J. A. *Proteomics* **2015**, *15*, 2862-2871.

(79) Xiao, C. Y.; Perez, L. M.; Russell, D. H. *Analyst* **2015**, *14*, 6933-6944.

(80) Stiving, A. Q.; Jones, B. J.; Ujma, J.; Giles, K.; Wysocki, V. H. *Anal. Chem.* **2020**, *92*, 4475-4483.

(81) Kinnear, B. S.; Kaleta, D. T.; Kohtani, M.; Hudgins, R. R.; Jarrold, M. F. *J. Am. Chem. Soc.* **2000**, *122*, 9243-9256.

(82) Chen, S. H.; Russell, D. H. *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 1433-1443.

(83) McCabe, J. W.; Mallis, C. S.; Kocurek, K. I.; Poltash, M. L.; Shirzadeh, M.; Hebert, M. J.; Fan, L. Q.; Walker, T. E.; Zheng, X. Y.; Jiang, T.; Dong, S. Y.; Lin, C. W.; Laganowsky, A.; Russell, D. H. *Anal. Chem.* **2020**, *92*, 11155-11163.

(84) Kinnear, B. S.; Hartings, M. R.; Jarrold, M. F. *J. Am. Chem. Soc.* **2001**, *123*, 5660-5667.

(85) Zilch, L. W.; Kaleta, D. T.; Kohtani, M.; Krishnan, R.; Jarrold, M. F. *J. Am. Soc. Mass Spectrom.* **2007**, *18*, 1239-1248.

(86) Jarrold, M. F. *PCCP* **2007**, *9*, 1659-1671.

(87) Buttner, K.; Blondelle, S. E.; Ostresh, J. M.; Houghten, R. A. *Biopolymers* **1992**, *32*, 575-583.

(88) Sereda, T. J.; Mant, C. T.; Hodges, R. S. *J. Chromatogr. A* **1995**, *695*, 205-221.

(89) Spicer, V.; Lao, Y. W.; Shamshurin, D.; Ezzati, P.; Wilkins, J. A.; Krokhin, O. V. *Anal. Chem.* **2014**, *86*, 11498-11502.

(90) Howard, G. A.; Martin, A. J. P. *Biochem. J.* **1950**, *46*, 532-538.

(91) McDaniel, E. W.; Barnes, W. S.; Martin, D. W. *Rev. Sci. Instrum.* **1962**, *33*, 2-7.

(92) Kirkland, J. J. *J. Chromatogr. Sci.* **1971**, *9*, 206-214.

(93) Meek, J. L. *Proc. Natl. Acad. Sci. USA* **1980**, *77*, 1632-1636.

(94) Parker, J. M. R.; Guo, D.; Hodges, R. S. *Biochemistry* **1986**, *25*, 5425-5432.

(95) Houghten, R. A.; Degraw, S. T. *J. Chromatogr.* **1987**, *386*, 223-228.

(96) Krokhin, O. V.; Craig, R.; Spicer, V.; Ens, W.; Standing, K. G.; Beavis, R. C.; Wilkins, J. A. *Mol. Cell. Proteomics* **2004**, *3*, 908-919.

(97) Moruz, L.; Kall, L. *Mass Spectrom. Rev.* **2017**, *36*, 615-623.

(98) Petritis, K.; Kangas, L. J.; Ferguson, P. L.; Anderson, G. A.; Pasa-Tolic, L.; Lipton, M. S.; Auberry, K. J.; Strittmatter, E. F.; Shen, Y. F.; Zhao, R.; Smith, R. D. *Anal. Chem.* **2003**, *75*, 1039-1048.

(99) Petritis, K.; Kangas, L. J.; Yan, B.; Monroe, M. E.; Strittmatter, E. F.; Qian, W. J.; Adkins, J. N.; Moore, R. J.; Xu, Y.; Lipton, M. S.; Ii, D. G. C.; Smith, R. D. *Anal. Chem.* **2006**, *78*, 5026-5039.

(100) Krokhin, O. V.; Spicer, V. *Anal. Chem.* **2009**, *81*, 9522-9530.

(101) Zolg, D. P.; Wilhelm, M.; Yu, P.; Knaute, T.; Zerweck, J.; Wenschuh, H.; Reimer, U.; Schnatbaum, K.; Kuster, B. *Proteomics* **2017**, *17*.

(102) Rosenberger, G.; Koh, C. C.; Guo, T. N.; Rost, H. L.; Kouvonen, P.; Collins, B.; Heusel, M.; Liu, Y. S.; Caron, E.; Vichalkovski, A.; Faini, M.; Schubert, O. T.; Faridi, P.; Ebhardt, H. A.; Matondo, M.; Lam, H.; Bader, S. L.; Campbell, D. S.; Deutsch, E. W.; Moritz, R. L., et al. *Sci. Data* **2014**, *1*.

(103) Spicer, V.; Ezzati, P.; Neustaeter, H.; Beavis, R. C.; Wilkins, J. A.; Krokhin, O. V. *Anal. Chem.* **2016**, *88*, 2847-2855.

(104) 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., et al. *Nat. Methods* **2017**, *14*, 259-262.

(105) Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; Reimer, U.; Ehrlich, H. C.; Aiche, S.; Kuster, B.; Wilhelm, M. *Nat. Methods* **2019**, *16*, 509-518.

(106) Ma, C. W.; Ren, Y.; Yang, J. R.; Ren, Z.; Yang, H. M.; Liu, S. Q. *Anal. Chem.* **2018**, *90*, 10881-10888.

(107) Müller, J. B.; Geyer, P. E.; Colaço, A. R.; Treit, P. V.; Strauss, M. T.; Oroshi, M.; Doll, S.; Virreira Winter, S.; Bader, J. M.; Köhler, N.; Theis, F.; Santos, A.; Mann, M. *Nature* **2020**, *582*, 592-596.

(108) Henderson, S. C.; Valentine, S. J.; Counterman, A. E.; Clemmer, D. E. *Anal. Chem.* **1999**, *71*, 291-301.

(109) Valentine, S. J.; Counterman, A. E.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 1188-1211.

(110) Shah, A. R.; Agarwal, K.; Baker, E. S.; Singhal, M.; Mayampurath, A. M.; Ibrahim, Y. M.; Kangas, L. J.; Monroe, M. E.; Zhao, R.; Belov, M. E.; Anderson, G. A.; Smith, R. D. *Bioinformatics* **2010**, *26*, 1601-1607.

(111) Counterman, A. E.; Clemmer, D. E. *J. Am. Chem. Soc.* **2001**, *123*, 1490-1498.

(112) Lietz, C. B.; Yu, Q.; Li, L. J. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 2009-2019.

(113) 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.; Rather, O.; Mann, M. *Mol. Cell. Proteomics* **2018**, *17*, 2534-2545.

(114) Michelmann, K.; Silveira, J. A.; Ridgeway, M. E.; Park, M. A. *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 14-24.

(115) Ridgeway, M. E.; Lubeck, M.; Jordens, J.; Mann, M.; Park, M. A. *Int. J. Mass spectrom.* **2018**, *425*, 22-35.

(116) Meier, F.; Köhler, N. D.; Brunner, A.-D.; Wanka, J.-M. H.; Voytik, E.; Strauss, M. T.; Theis, F. J.; Mann, M. *bioRxiv* **2020**, 2020.2005.2019.102285.

(117) Krokhin, O. V. *Anal. Chem.* **2006**, *78*, 7785-7795.

(118) Krokhin, O. V.; Ezzati, P.; Spicer, V. *Anal. Chem.* **2017**, *89*, 5526-5533.

(119) Gussakovsky, D.; Neustaeter, H.; Spicer, V.; Krokhin, O. V. *Anal. Chem.* **2017**, *89*, 11795-11802.

(120) Krokhin, O. V.; Anderson, G.; Spicer, V.; Sun, L. L.; Dovichi, N. J. *Anal. Chem.* **2017**, *89*, 2000-2008.

(121) Munoz, V.; Serrano, L. *J. Mol. Biol.* **1995**, *245*, 275-296.

(122) Kozlowski, L. P. *Biol. Direct* **2016**, *11*, 55.

(123) Shvartsburg, A. A.; Siu, K. W. M.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **2001**, *12*, 885-888.

(124) Shamshurin, D.; Spicer, V.; Krokhin, O. V. *J. Chromatogr. A* **2011**, *1218*, 6348-6355.

(125) Pierson, N. A.; Chen, L. X.; Russell, D. H.; Clemmer, D. E. *J. Am. Chem. Soc.* **2013**, *135*, 3186-3192.

(126) Kondalaji, S. G.; Khakinejad, M.; Tafreshian, A.; Valentine, S. J. *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 947-959.

(127) Rappsilber, J.; Mann, M.; Ishihama, Y. *Nat. Protoc.* **2007**, *2*, 1896-1906.

(128) Ishihama, Y.; Rappsilber, J.; Mann, M. *J. Proteome Res.* **2006**, *5*, 988-994.

(129) Meier, F.; Beck, S.; Grassl, N.; Lubeck, M.; Park, M. A.; Raether, O.; Mann, M. *J. Proteome Res.* **2015**, *14*, 5378-5387.

(130) Stow, S. M.; Causon, T. J.; Zheng, X. Y.; Kurulugama, R. T.; Mairinger, T.; May, J. C.; Rennie, E. E.; Baker, E. S.; Smith, R. D.; McLean, J. A.; Hann, S.; Fjeldsted, J. C. *Anal. Chem.* **2017**, *89*, 9048-9055.

(131) Mason, E. A.; McDaniel, E. W. *Transport Properties of Ions in Gases*; Wiley, New York, 1988.

(132) Prianichnikov, N.; Koch, H.; Koch, S.; Lubeck, M.; Heilig, R.; Brehmer, S.; Fischer, R.; Cox, J. *Mol. Cell. Proteomics* **2020**, *19*, 1058-1069.

(133) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466-1467.

(134) Wilkinson, G. N.; Rogers, C. E. *J. R. Stat. Soc. C-Appl.* **1973**, *22*, 392-399.