# Peak identification and quantification in proteomic mass spectrograms using non-negative matrix factorization

# プロテオミクスにおける非負値行列因子分解法による マススペクトログラムピークの同定および定量

**2020**

**Pasrawin Taechawattananant**

# Contents

# Abbreviations

- 3D : 3-Dimensional
- ACN : Acetonitrile
- DDA : Data-Dependent Acquisition
- FDR : False Discovery Rate
- bNMF : Basic Non-negative Matrix Factorization
- idotp : Skyline Isotope Dot Product
- KL Divergence: Kullback–Leibler Divergence
- LC : Liquid Chromatography
- MCR-ALS : Multivariate Curve Resolution-Alternating Least Squares
- pNMF : Proteomics Non-negative Matrix Factorization
- MS : Mass Spectrometry
- NMF : Non-negative Matrix Factorization
- PCA : Principal Component Analysis
- PeptScore : Mascot Peptide Score
- PSM : Peptide-Spectrum Match
- S/N : Signal-to-Noise Ratio
- TFA : Trifluoroacetic Acid
- UPS : Universal Proteomics Standard
- VQ : Vector Quantization

# Preface

I have always been interested in healthcare, probably, because I was often sick as a kid. I want to save people' lives by being a pharmacist and doing research in the field of pharmaceutical sciences. Since I was born in the era of the Human Genome Project, I saw many health-related questions that the knowledge of genome had been expected to provide answers. However, they could not. Proteins are the main workhorse of the cell and perform most of functions to support life. I have believed that the study about proteins can yield new cellular biology and the understanding of health. Therefore, I decided to study proteomics in graduate school.

The term "proteome" was invented to be similar to "genome", which refers to a whole set of proteins expressed at a specific time in a cell, tissue, or an organism[1]. The study of proteome or proteomics offers a comprehensive view of biological processes and networks in the cell. Mass spectrometry(MS)-based proteomics has become a universal analytical approach for large-scale protein identification and quantification due to its high-throughput and high sensitivity[2]. Since advancements in instrumentation have pressured on the downstream data processing, computational proteomics has been increasingly challenged to extract the full information from the mass spectrograms[3].

Beside my interest in healthcare, I love mathematics. I am very fortunate to be able to combine my interests and start working on computational proteomics in Ph.D. I had never coded and had few mathematical backgrounds from mathematical Olympiad camp. However, with strong supports from my professor and co-advisor, I was introduced into the machine learning world, particularly in audio processing. Non-negative matrix factorization (NMF) is a well-used technique that is applicable for audio source separation[4–6]. The audio data can be viewed as 3-dimensional (3D) space of frequency, time, and intensity. Surprisingly, it is similar to mass spectrograms which can be described as 3D space of mass-to-charge ratio (*m/z*), time, and intensity. It is fascinating to explore NMF computation for proteomics mass spectrogram analysis.

The challenging part is mass spectrograms of proteome samples have much higher complexity than spectrograms of audios. The current conventional proteomics software relies on picking only intense peaks from LC/MS mass spectrograms for quantification[7,8], and determines LC/MS/MS mass spectrograms using over-simplified models for identification[9]. I would like to show the potential of NMF for proteome analysis in benchmarking with the software that the proteome research community depends on now.

Peak identification and quantification in proteomics mass spectrograms using NMF is described herein. In chapter 1, a novel approach based on NMF was explored and developed to identify and quantify peaks in proteomics mass spectrograms using precursor ion information. In chapter 2, both precursor and product ion mass spectrograms were studied in order to design an integrated NMF framework for

proteomics applications. The overall goals are to improve miss-identification and miss-quantification in LC/MS/MS proteome analysis by using NMF.

# Chapter 1
# LC/MS-based Proteome Identification and Quantification using Non-negative Matrix Factorization

## Introduction

Recent advances in LC/MS technologies have permitted detection of a wide range of signal parameters in a three-dimensional mass spectrogram, including $m/z$, retention time, and intensity. However, despite the high resolving power of modern LC/MS instruments, a large part of the mass spectrogram remains unresolved[10,11]. A new approach based on mathematical and statistical methods is required to interpret mass spectrograms accurately and with higher sensitivity.

Various algorithms have been proposed for measuring peaks in mass spectrograms, but they often start by screening the peak shape or defining a threshold to remove presumed non-target ions. However, these preprocessing steps inevitably eliminate target ions with either low abundance or irregular peak shapes[7,8]. An alternative algorithm, multivariate curve resolution-alternating least squares (MCR-ALS) was recently proposed for application to high-resolution LC/MS data without the need for peak shape assumption or imposition of a prior threshold[12]. Since the mass spectrograms generated from LC/MS are sparse, meaning that a relatively small proportion of targeting signals are over a large space on the mass spectrogram, the algorithm incorporates a simple sparsity constraint. However, only eight resolved peak profiles from eight chemicals and four peak profiles from bacterial cell extracts were shown. The preprocessing requires manual selection of chromatographic windows to reduce complexity before data analysis. Moreover, the computation cannot remain non-negative, which conflicts with the non-negative nature of the mass spectrogram.

Non-negative matrix factorization (NMF) is an unsupervised machine-learning technique that is well applicable for matrix-format data source separation. NMF guarantees non-negative computation and effectively controls its sparse approximation with a sparsity constraint[13]. Unlike other factorizations, NMF learns a part-based representation of data input. According to the famous example of NMF application, NMF can recognize face image input from various components of face features while principal component analysis (PCA) and vector quantization (VQ) need similar whole face for recognition[14]. Therefore, NMF should be suitable for analyzing various signal components of non-negative $m/z$, retention time, and intensity in sparse proteomic mass spectrograms.

It was reported that application of NMF to LC/MS was feasible, and enabled annotation of 11 out of 18 mixed chemicals[15]. However, the simple algorithm employed

in that work has substantial limitations, separating the data of a small number of chemical compounds with few common or overlapped peaks. It thus cannot be applied to proteomic mass spectrograms. So far, application of NMF for peak identification and quantification in proteomics has not been explored.

In this study, I propose a new application of NMF-based approach, i.e., proteomics NMF (pNMF) for identifying and quantifying peaks in proteomics mass spectrograms. pNMF incorporates the isotopic *m/z* distribution, the predicted retention time of peptides, and the noise. I add a protein-peptide hierarchical relationship as a group information for a group sparsity constraint. pNMF shows a good performance without the need of thresholding, peak-picking, or complicated preprocessing that conventional approaches for proteomics require

## Theory and Design

NMF aims to decompose mixture data containing non-negative signals in a matrix $V \in \mathbb{R}^{m \times t}$ by finding two non-negative matrices, $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times t}$, whose product approximates the mixture data. $W$ is a collection of signal components called as a dictionary, and $H$ is an activation matrix where its column $H_t$ represents the decomposition as activation coefficients that approximate each $V_t$ onto the dictionary.

$$V \simeq WH$$

The basic NMF (bNMF)[16] finds $W$ and $H$ by minimizing the difference between $V$ and $WH$. The iterative approach of bNMF shows it is possible to update the initial values of both $W$ and $H$, but many applications is developed by updating only $H$ if it has a fixed dictionary for $W$[17]. The dictionary can be either preset or trained by many ways. The difference $D(V||WH)$ can be quantified by the cost function $D$, such as the Euclidean distance or the Kullback–Leibler divergence (KL divergence).

$$\underset{W \geq 0, H \geq 0}{\text{minimize}}\, D(V||WH)$$

I first examined the potential of bNMF for representing an observed mass spectrogram as matrix $V$ ($m/z$ × retention time) by using a fixed dictionary of peptide signals with monoisotopic $m/z$ of all theoretical peptide ions as matrix $W$ ($m/z$ × theoretical peptide ion). I employed $D(V||WH)$ as the KL divergence due to its superior performance in source separations[18]. The matrix $H$ (theoretical peptide ion × retention time) was updated for the activation of each theoretical peptide ion along the retention time profile.

To improve bNMF representation, I then designed pNMF (Figure 1). Two subspaces, i.e., the peptide signal subspace $W_S$ and the noise subspace $W_N$, were incorporated into the dictionary. This subspace partitioning prevents noise from interfering with interesting signal approximations[19]. $W_S$ was supervised by isotopic $m/z$ of theoretical peptides. $W_N$ was obtained by using learned noise patterns obtained from NMF with a blank region of the mass spectrogram. Since modern MS instruments can manage the fluctuation of background and lock certain chemical noise $m/z$ before reporting the results[2020], $W_N$ can help removing other remaining noises such as unexpected chemical noises from solvents in LC mobile phase, plasticizers from plastic devices, and metal adduct ions[21]. For the resulting matrix $H$, instead of allowing the algorithm to update freely, I scoped the activation region of the peptide subspace $H_S$ using predicted retention time.

Proteomic mass spectrograms have a unique group sparsity structure arising from the protein and peptide hierarchy, and it is appropriate to apply a group sparsity constraint Ω[22]. I imposed Ω on each protein using $\log/L1$ due to its simplicity and good

performance[5]. Sparseness measure was invented and applied for controlling the level of sparsity as a weight factor in NMF[13,23]. Since the number of members in the group affects the algorithm result, I applied sparseness measure as weight $\lambda$ and normalized it by the number of theoretical peptide ions per protein. $\lambda$ was then incorporated into Ω. pNMF aims to solve:

$$\underset{W\geq 0, H\geq 0}{\text{minimize}}\, D(V||WH) + \ \Omega(H_S)$$

In detail, $D$ is the Kullback-Leibler divergence and Ω is $\log/L1$ weighted with λ

$$D(V \parallel WH) = \sum_{mt}(V_{mt}\log\frac{V_{mt}}{(WH)_{mt}} - V_{mt} + (WH)_{mt})$$

$$\Omega = \log/L1 = \sum_{g=1}^{G}\lambda_g \log\left(||\, H_g||_1 + \varepsilon\right)$$

where $H_g$ is a group $(g)$ of peptide ions digested from the same protein. The majorizing function of $D(V||WH) + \Omega(H_S)$ can be derived for efficient multiplicative updates. The convergence of algorithm is shown in the previous study[5].

I presented two designs of pNMF for protein and peptide identifications. The most frequently used concept to identify peptides is database searching[24], in which all protein sequences provided in the database are in silico digested for reference. I first designed pNMF using a database-based dictionary and evaluated the performances with two standard samples. An alternative to database searching is library searching[25], in which the reference is constructed using a comprehensive collection of confident identification from previous experiments. With the reference as a strong prior knowledge, a library searching performs relatively quickly and provides a better identification, particularly for short-gradient and low-resolution experiments. I also designed pNMF using a library-based dictionary and applied it to a proteome-scale mass spectrogram.

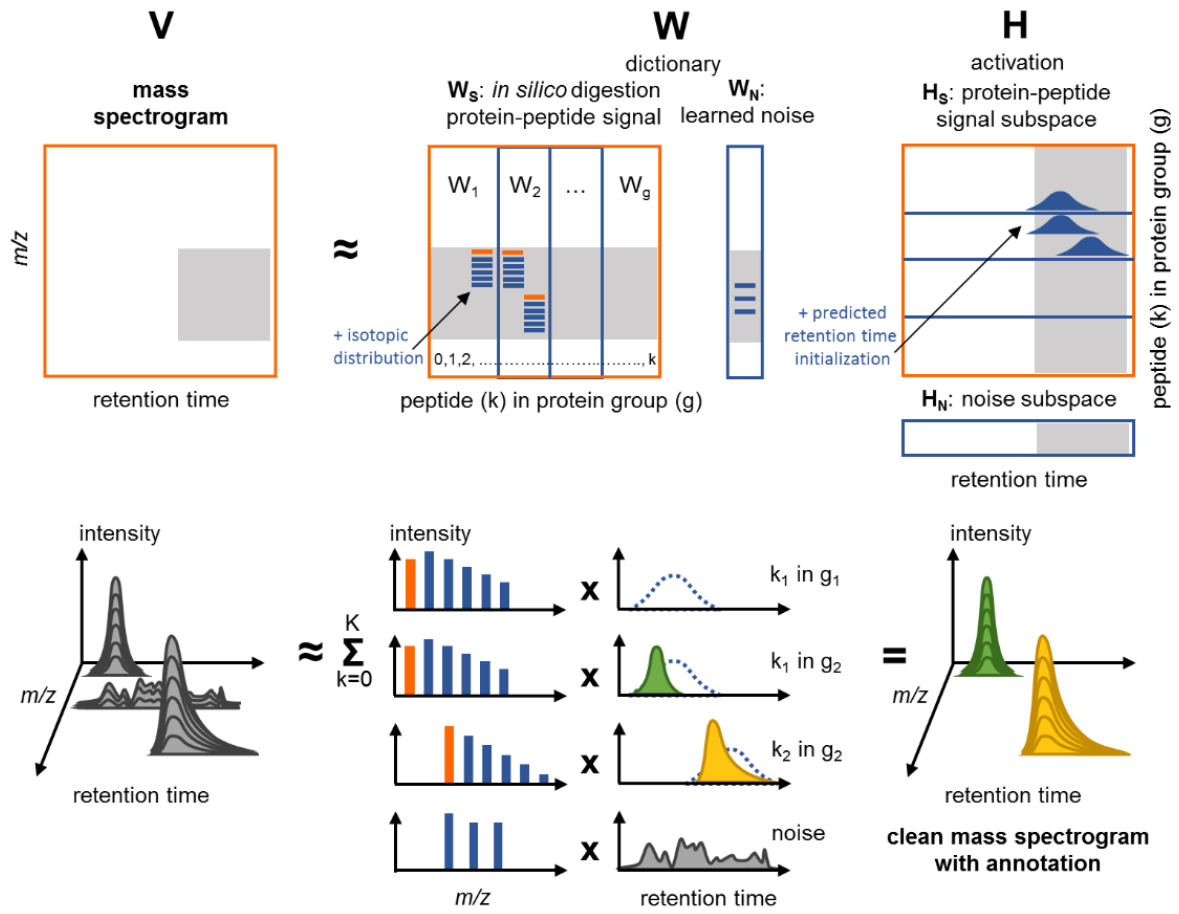


**Figure 1. Overview of pNMF.**

In bNMF method, three matrices are designed to correspond to an observed mass spectrogram $V$ ($m/z$ × retention time), a supervised dictionary of theoretical peptide monoisotopic $m/z$ $W$ ($m/z$ × theoretical peptide ion), and an activation matrix $H$ for updating the results (shown in orange). For pNMF, the supervised dictionary incorporates two subspaces: the peptide subspace $W_S$ and the noise subspace $W_N$. The $W_S$ column includes monoisotopic $m/z$ and subsequent isotopic $m/z$, and $W_N$ contains learned noise patterns. For the activation matrix at $H_S$, instead of allowing the algorithm to update freely, the activation region is scoped with predicted retention time windows (shown in blue) (top).

Specifically, the observed mass spectrogram $V$ can be viewed as the linear sum of the isotopic distribution profile of each peptide $k$ in the dictionary $W_S$ and noise profile in $W_N$ scaled by retention time-varying activations. The group sparsity constraint uses protein and peptide selection to identify a few highly related patterns to approximate the mass spectrogram (bottom).

**The workflow of pNMF**

The workflow comprises the following four steps.

**Step 1. Matrix construction**

**1.1 The observed mass spectrogram $V$:**

Either an mzML or a text file of a raw file from LC/MS converted by ProteoWizard MSconvert[26] can be processed by pNMF. The intensity values are normalized by mean and assigned to $V$ at appropriate $m/z$ and retention time bins. Thresholding is not performed, so that low-abundance peptide ions are included in the computation.

**1.2 The dictionary $W$:**

$W$ has two subspaces for protein-peptide signals and noise signals, respectively

- $\boldsymbol{W_S}$**:** *In silico* digestion was performed for the database-based dictionary with trypsin up to two missed cleavage sites. Peptide ions ranging in length from seven to 50 amino acids with charges ranging from +2 to +4 were included. Cysteine carbamidomethylation was set as a fixed modification. Methionine oxidation and N-terminal cysteine ammonia loss were set as variable modifications. Features of six isotopic peaks including the monoisotopic peak are calculated, normalized, and assigned to columns of $W_S$.

  The library-based dictionary had the peptide ions identified by Mascot[27] and the additional peptide ions from other possible charges of Mascot-identified peptides within $m/z$ range. Cysteine carbamidomethylation and methionine oxidation were set as fixed and variable modifications, respectively. Features of six isotopic peaks including the monoisotopic peak are calculated, normalized, and assigned to columns of $W_S$.
- $\boldsymbol{W_N}$**:** Noise was learned from a blank region of the mass spectrogram, i.e., the last 10 min of gradient elution, which was free from peptides, by using scikit-learn[2] NMF with 2 components. The obtained patterns were assigned to two columns of $W_N$.

**1.3 The activation $H$:**

- $\boldsymbol{H_S}$**:** The activations of each peptide are restricted to zero at retention time frames in which the peptide peaks are highly unlikely to be detected.
- $\boldsymbol{H_N}$**:** The points in the peptide subspace were fully initialized with positive values.

**Step 2. Sparsity constraint weight calculation**

The sparseness measure[13] with protein size $\boldsymbol{\lambda} = [\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \dots, \boldsymbol{\lambda_G}]^T$ can be calculated as follows

$$\boldsymbol{\lambda} = \frac{1}{\sqrt{M}} \sum_{m}^{M} \frac{\sqrt{T} - ||V_m||_1 / ||V_m||_2}{\sqrt{T} - 1} \boldsymbol{p}$$

where the $V$ matrix has $M$ rows of $m/z$ and $T$ columns of retention time ($m \times t$), and the $\boldsymbol{p}$ vector represents the number of peptide ions for each protein group as $\boldsymbol{p} = [\boldsymbol{p_1}, \boldsymbol{p_2}, \dots, \boldsymbol{p_G}]^T$. In the case of pNMF with the library-based dictioanry, since the update regions are constrained with narrow predicted time windows, a window-gradient ratio is incorporated into λ by multiplying $window\ (minute)/gradient\ (minute)$.

**Step 3. pNMF algorithm**

Since pNMF has a fixed dictionary, the algorithm aims to update $H$ by iterating between:

1. Updating $H$
2. Updating $H_g$
3. Checking $D(V||WH) + \Omega(H_S)$. If the change since the last iteration is less than the tolerance, the algorithm declares convergence.

The algorithm is summarized below:

**Algorithm 1: pNMF**

**input** : $V$, $W = [W_S \ W_N]$
**initialize** : $H$
**repeat**
    # updating $H$
    $H \leftarrow H .* (W^T(V./WH))$;
    # updating $H_g$
    **for** $g = 1{:}G$ **do**
        $H_g \leftarrow \frac{1}{1+ \lambda_g/(||H_g||_1 + \varepsilon)} H_g$;
    **end**
**until** *convergence*;
**return** $H$

where $.*$ and $./$ are component-wise multiplication and division, and $(.)^T$ is matrix transposition.

**Step 4. Post-processing for interpretation**

The resulting $H$ is multiplied by the mean to reverse the normalization in 1.1. The most confident peak in $H_k$, which has an activated intensity with a signal-to-noise ratio (S/N) > 3, is reported as a peptide candidate for identification and the criterion of S/N > 10 is used for quantification. Protein identification is reported if there are at least two identified peptides, and quantification is calculated based on the sum of activated intensities of the reported peptides.

For the detail of the most confident peak selection, the highest activation peak with cosine similarity more than 0.90 was selected as a default candidate. In the design of using the database-based dictionary, the additional post-processing was performed by

evaluating the other two highest annotated peaks for each peptide ion. If one of them has cosine similarity more than 0.97 and 20% better than the others, the default candidate will be replaced.

All computations were performed on a desktop computer with a 3.30 GHz E3-1226 v3 4-core processor and 32 GB main memory. The raw MS data were deposited at the ProteomeXchange Consortium via jPOST partner repository[28] with identifier JPST000663 for two standard proteins and JPST000765 for E. coli lysate. pNMF is available in Python 2.7 at GitHub https://github.com/pasrawin/ProteomicNMF and https://github.com/pasrawin/LibraryProteomicNMF.

# Experimental Section

Mass spectrograms with three different levels of complexity were used in this study. Two standard protein mixtures were prepared in-house from four proteins and Universal Proteomics Standard (UPS1) containing 48 proteins. *Escherichia coli* data sets were collected from the previously described method[29].

**Proteomics experiment setup and LC/MS/MS analysis**

**Materials:**

Recombinant human albumin, catalase from erythrocytes, recombinant human epidermal growth factor, recombinant human leptin, and UPS1 were obtained from Sigma-Aldrich (St. Louis, MO, USA). Dithiothreitol, iodoacetamide, Lys-C, trifluoroacetic acid (TFA), acetonitrile (ACN), and piperidine were obtained from Wako (Osaka, Japan). Trypsin was obtained from Promega (Madison, WI, USA).

**Sample preparation:**

For four-protein standard: 25 ng of albumin, catalase, EGF, and leptin were mixed. The sample was reduced with dithiothreitol (10% v/v) and alkylated with iodoacetamide (10% v/v) before Lys-C and trypsin digestion (1% w/w) for 3 h and overnight, respectively. Peptides were desalted with StageTips[30] and suspended in the loading buffer (0.5% TFA and 4% ACN) for subsequent LC/MS analysis.

For forty-eight-protein standard: a commercial UPS1 was premixed with 5 pmol of 48 proteins ranging in molecular mass from 6000 to 83,000 Da. The processes of reduction, alkylation, digestion, and desalting for subsequent LC/MS analysis were the same as described for the previous sample.

For *E. coli* sample: *E. coli* strain BW25113 cells grown in Luria–Bertani culture were digested according to the PTS protocol[31]. The resulting peptides were desalted with StageTips.

**LC/MS/MS analysis:**

For four-protein standard and forty-eight-protein standard, a self-pulled analytical column was prepared with ReproSil-Pur C18-AQ materials (Dr. Maisch, Ammerbuch, Germany). The mobile phases consisted of (A) 0.5% acetic acid and (B) 0.5% acetic acid in 80% ACN. A gradient condition with flow rate 500 nL/min was set: 5–10% B in 5 min, 10–40% B in 60 min, 40–100% B in 5 min, 100% B for 10 min, and 5% B for 30 min. An Ultimate 3000 pump (Thermo Fisher Scientific, Germering, Germany) and an HTC-PAL autosampler (CTC Analytics, Zwingen, Switzerland) were used coupled to a Q-Exactive hybrid quadrupole-orbitrap mass spectrometer (Thermo-Fisher Scientific). Tandem mass spectra were acquired using data-dependent acquisition. The MS1 survey scan range was scanned at a resolution of 70,000 for $m/z$ 300–1500, an AGC level of $3e$+6, and a maximum ion accumulation time of 100 ms. The top 10 ions having charges from +2 to +8 were selected in a scan range of $m/z$ 200–2000 with an isolation window of $m/z$ 2. The AGC target was $1e$+5, and the maximum ion accumulation time was 100 ms. A dynamic exclusion with a duration of 30 s was used to reduce repeated ions.

For *E. coli* sample, a self-pulled analytical column was prepared with ReproSil-Pur C18-AQ materials. A two-step linear gradient of 5 % to 40 % B was set for 70 and 550 minutes. An Ultimate 3000 pump and an HTC-PAL autosampler were used coupled to an LTQ-Orbitrap XL. The MS1 survey scan range was scanned for $m/z$ 300–1500 and an AGC level of $5e$+5. The top 10 ions were selected and the AGC target at $1e$+4 was set for the MS2 scan.

**pNMF setup**

The further information of matrix inputs are described as follows.

***V* input:**

For each axis of $V$, the full scan range of $m/z$ and the peptide retention time were divided into fine equal-width bins of 0.01 and 0.005-min, respectively. The retention time was smoothed by a moving average technique using a 12-s time window and 6-s time shift forward.

***H* input:**

The activations in peptide subspace are restricted by prior knowledge. For pNMF with the database-based dictionary, two initializations were combined using the predicted retention times and the similarities to the theoretical isotopic distribution, respectively. The predicted retention times were obtained from the algorithm, achrom3, which could be calibrated by using an in-house data set. The points within the possible retention times frame ($pRT$) were initialized using Gaussian distribution ($\mu$ = the predicted retention time, $\sigma$ = the average peak width/4). The similarities to the theoretical isotopic distribution were calculated for peptide ions by using cosine similarity ($cos$) at time t as

$$cos = \frac{\mathcal{W}_x \cdot \mathcal{V}_x}{\|\mathcal{W}_x\| \, \|\mathcal{V}_x\|}$$

where $\mathcal{W}_x$ is peaks in $W_k$ and $\mathcal{V}_x$ is peaks at $x$ in $V_t$

In order to include the surrounding distributions, $cos$ were calculated at $t-1$, $t+1$, and $x'$ which included the previous isotopic index into $x$. All $cos$ were summed and calculated for average of median and mean. In this work, the points outside 25-min $pRT$ and the $cos$ less than 0.5 were restricted to zero. Wide $pRT$ and low $cos$ were set to exclude only highly unreasonable retention times, and they could be adjusted according to the experiment.

For pNMF with the library-based dictionary, the predicted retention times were collected from references by combining between Mascot peak retention times from 70-min gradient experiment and converted retention times of 550-min gradient experiment using linear regression. Although Mascot-identified retention times are triggered by MS2, they are very dependable. $pRT$ were set for 5 minutes and $cos$ was not used.

# Results and Discussion

## 1. Four-protein standard

### 1.1 Effects of the NMF modifications

I applied bNMF method to the four-protein mixture to separate 1057 peptide ions. The correlation between the observed total ion currents (TICs) and the total NMF-activated signals at each time point was examined (Figure 2). I observed a proportional relationship with a correlation coefficient of 0.846. On average, 20% of TICs were accounted for. To improve the performance of the algorithm, I developed pNMF to incorporate precursor ion isotopic distributions, noise models, sparsity constraints, and other factors as described in the Theory and Design section, and evaluated the correlation to the observed TICs. With pNMF, I obtained a correlation coefficient of 0.982 and accounted for 70% of the TICs. As expected, I observed a significant improvement in terms of the extraction of peptide ion chromatograms from the mass spectrogram. The unaccounted-for signals were from peptides in unsupervised charged states and modifications, contaminants, and non-electronic noise outside the pNMF dictionary, since TICs included all signal contributions observed during LC/MS.

Next, I compared the peptide retention times approximated by the NMFs with those determined by Mascot and Skyline[32], two conventional software programs used in proteomics. I imported Mascot search results into Skyline MS1-Full Scan Filtering to obtain MS1 parameters. For the peptide ions commonly annotated by NMFs and Mascot/Skyline, the retention times by bNMF agreed with those determined by Mascot/Skyline, with $y = 1.003x$, $R = 0.902$ (Figure 3, left), while pNMF provided better accuracy in measuring the retention times of peptide ions, with $y = 1.000x$, $R = 0.999$ (Figure 3, right). These results indicate that our modifications to bNMF were effective for annotating peptide ions in the mass spectrogram.

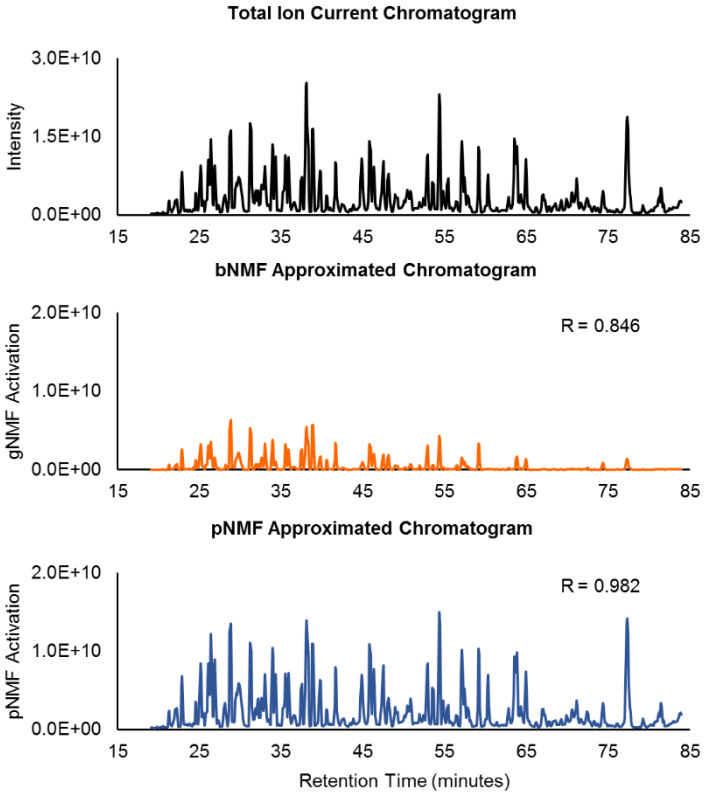


**Figure 2. Effects of NMF modifications on chromatogram extraction.**
The observed total ion current chromatogram (top). bNMF approximated chromatogram (middle). pNMF approximated chromatogram (bottom). Correlation coefficients were calculated between the observed total ion currents and total activated intensities of each NMF at the same retention time points.

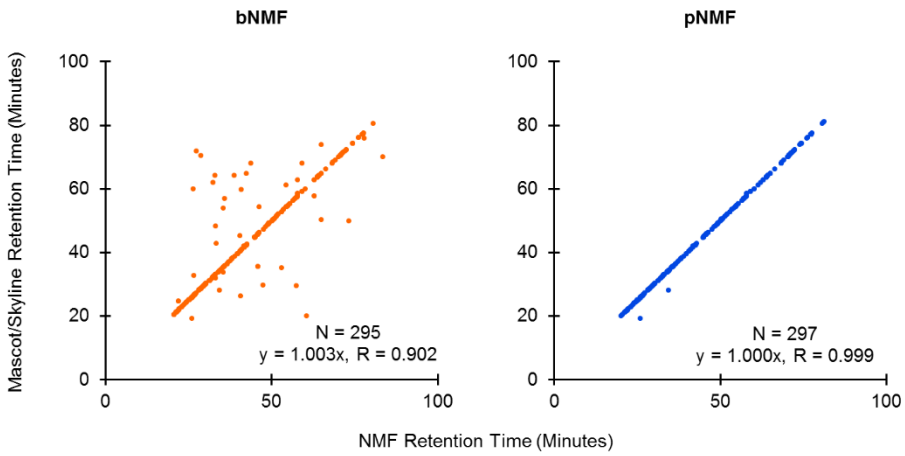


**Figure 3. Effects of NMF modifications on the accuracy of peptide retention time annotation.**

Proportional relationship between bNMF (x-axis) and the conventional method (y-axis) for commonly annotated peptide ions (left). Proportional relationship between pNMF (x-axis) and the conventional method (y-axis) for commonly annotated peptide ions (right).

### 1.2 Peptide identification

Peptide identification by current conventional approaches, including Mascot/Skyline, is based on the use of MS/MS spectra to acquire sequence-related information, meaning that the success of peptide identification depends to a large extent on the quality of the MS/MS spectra, whereas NMF-based approaches require the signal in MS1. pNMF identified 297 out of 303 (98.02%) of Mascot-identified peptides (Figure 4, left), and the remaining six peptides had 52.29 times lower intensity, 20.78 lower Mascot Peptide Score (PeptScore), and 4.12% lower Skyline isotope dot product (idotp) than the peptides identified in common, calculated by median (Figure 5). In other words, these six peptides have limited intensities in MS1.

Moreover, pNMF identified an additional 299 peptide ions, and provided 100% sequence coverage for the four proteins, whereas the conventional method gave only 94.83% coverage on average (Figure 6). Generally, MS1-based methods possess higher sensitivity to detect a larger number of peptides regardless of MS2 information, although MS1 information alone is insufficient for identifying peptides in some cases when the poor profiles, such as distorted isotopic patterns, are obtained.

### 1.3 Protein and peptide quantification

I compared the peak intensities obtained by pNMF with those obtained by Skyline for commonly identified ions at the peptide level. The comparison at the protein level was done using the sum of the intensities of the common peptide ions in both methods. The quantification results for both methods agreed well with each other. Correlation coefficients of 0.997 and 0.914 were obtained at the protein and peptide levels, respectively. I observed four peptide ion outliers calculated from Tukey's fences, or 1.5 times the interquartile range above the third quartile of the ratio of absolute intensity difference (Figure 4, right).

I identified two causes of the four outliers, exemplifying the advantage and disadvantage of the algorithms. For the outlier 1, Skyline and pNMF reported uncorrelated intensities of KLVAASQAALGL with +2 charge for $1.41e{+}05$ at 42.1 min and $5.12e{+}09$ at 41.60 min. According to Skyline data visualization, I found that this peptide ion had three significant hits at 41.47, 42.10, and 43.13 min by Mascot, but the maximum MS1 intensity was located at 42.10 min, within a dynamic exclusion time. Since Skyline, which uses an MS2-triggered MS1-peak integrated approach, was able to find an appropriate peak close to the significant hit at 42.10 min, it missed the true maximum intensity at 41.61 min because of the distance. Consequently, Skyline underestimated the intensity from the subordinated peak due to dynamic exclusion (Figure 7). In contrast, pNMF is a completely MS1-dependent approach, and can report the maximum MS1 peak intensity as an alternative peak quantification tool. The outliers 2, 3, and 4 resulted from HPYFYAPELLFFAKR with charges of +2, +3, and +4. This mixture contains RHPYFYAPELLFFAK, which shares the same precursor isotopic distribution profile as HPYFYAPELLFFAKR; an MS1-based algorithm such as pNMF therefore cannot distinguish between these two peptides. More features are necessary to quantify peptide isomers exclusively.

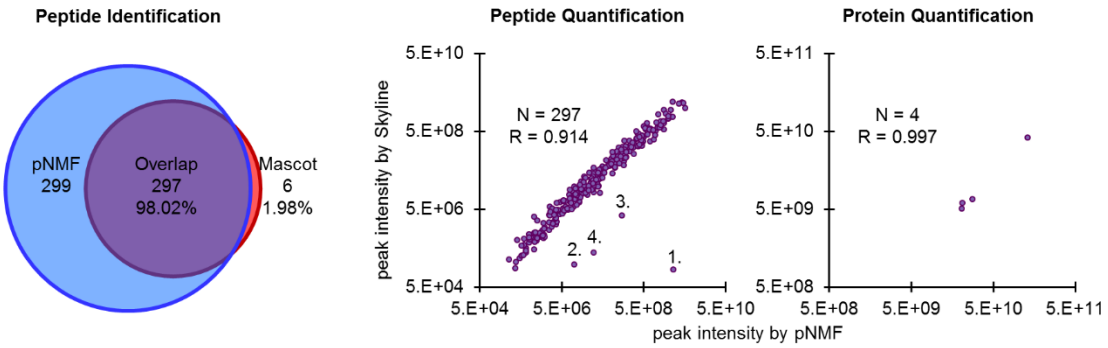


**Figure 4. pNMF for four-protein standard.**

The Venn diagram represents the peptide ions identified using pNMF and Mascot (left). The peak intensities of peptide ions commonly identified by pNMF (x-axis) and Skyline (y-axis) show correlation coefficients of 0.916, with four outliers at the peptide level, and 0.999 at the protein level (right).

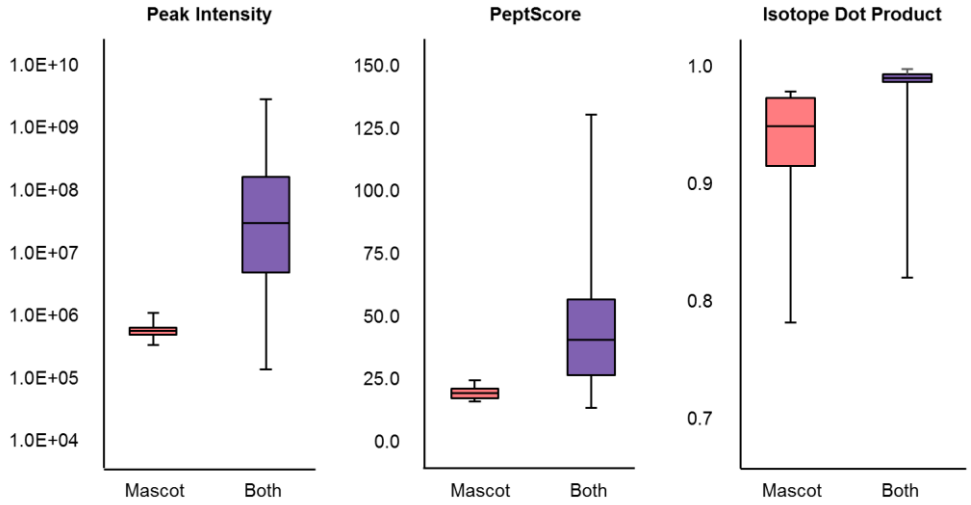


**Figure 5. Profile of peptides exclusively identified by Mascot in four-protein standard.**

The box plot shows that the six peptides exclusively identified by Mascot have 52.29 times lower intensity, 20.78 lower Mascot Peptide Score (PeptScore), and 4.12% lower Skyline isotope dot product (idotp) than the 297 peptides commonly identified by both Mascot and pNMF.

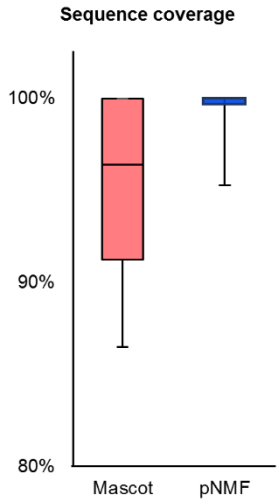


**Figure 6. Sequence coverage comparison between 303 peptides identified by Mascot and 596 peptides identified by pNMF.**
Mascot and pNMF yielded medians of 94.83% and 100.00% coverage, respectively.

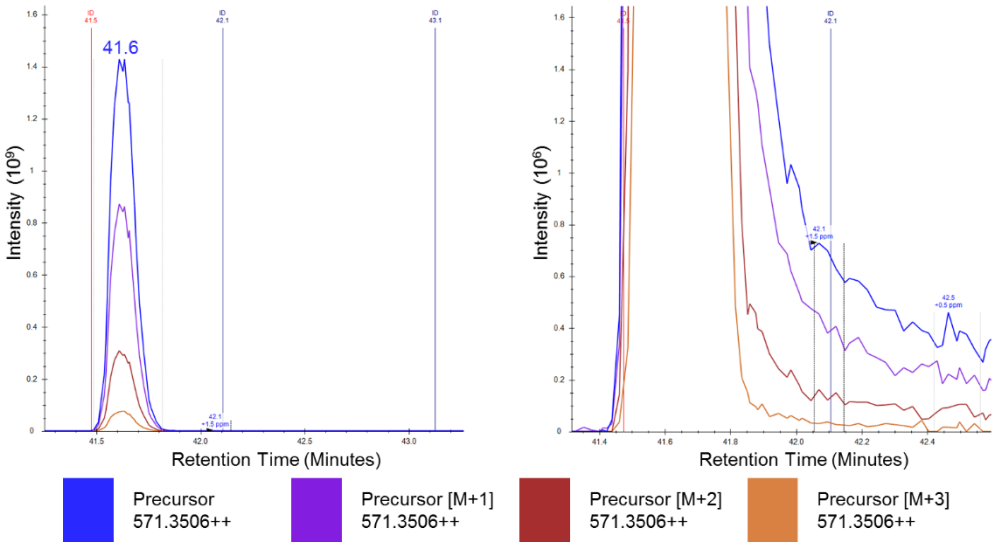


**Figure 7. Outlier investigation for KLVAASQAALGL++ of *m/z* 571.35.**
Skyline graphic interface shows the extracted chromatogram with peak retention time at 41.60 min (left). Detailed examination shows that the Skyline peak integration boundary selecting the peak closest to the significant hit at 42.10 min results in incorrect quantification.

## 2. Forty-Eight-Protein Standard

### 2.1 Protein and peptide identification

Universal Proteomics Standard (UPS1) is a premixed standard containing 5 pmol of 48 proteins ranging in molecular mass from 6000 to 83,000 Da. It represents a simple proteome complexity with 9319 *in silico* peptide ions from 4045 unique peptides. In benchmarking with the conventional method, pNMF reported 1322 out of 1481 peptide ions or 89.26% of Mascot identification. In terms of accuracy in measuring the retention times of commonly identified peptide ions, the correlation between the two methods was excellent, with $y = 0.999x$, $R = 0.986$ (Figure 8).

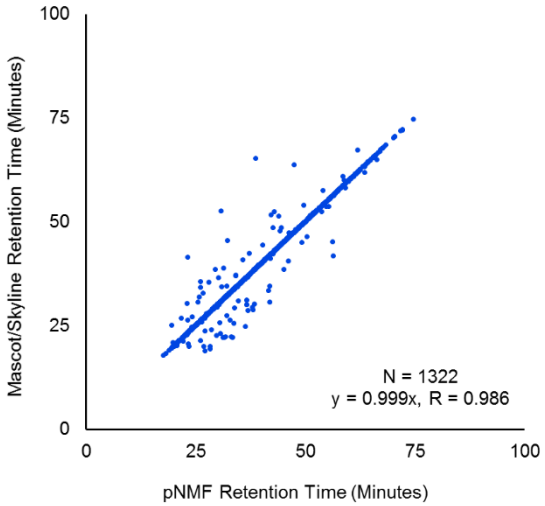


**Figure 8. Accuracy of peptide retention time annotation in forty-eight-protein standard.**

Proportional relationship between pNMF (x-axis) and the conventional method (y-axis) for commonly annotated peptide ions.

The conventional method and pNMF exclusively identified 159 and 2103 peptide ions, respectively (Figure 9, left). The complexity of the proteomics standard increased both exclusive identifications, presumably because of the larger proportion of low-intensity peptide ions and poor or overlapped MS1 peaks. I observed a 3.45 times lower intensity, 8.59 times lower PeptScore, and 3.37% lower idotp values for the 159 peptides exclusively identified by Mascot, similar to the case of four-protein standard (Figure 10). The conventional method and pNMF identified all proteins present in the sample, and yielded median sequence coverages of 62.02% and 87.29%, respectively (Figure 11).

## 2.2 Protein and peptide quantification

I compared the peak intensities of 1322 commonly identified peptides obtained by the conventional method with those obtained by pNMF. For 1319 peptides, after eliminating three peptide ions with S/N ≤ 10, the quantification results from both methods agreed with each other. Correlation coefficients of 0.976 and 0.909 were obtained at the protein and peptide levels, respectively. I observed 39 outlier peptides, or 2.96% of the total, calculated from Tukey's fences (Figure 9, right). I investigated the reasons for uncorrelated quantification. It is unclear to judge whether which algorithm presented the most accurate amounts among 34 cases but 14 of them have poor isotopic patterns or limited MS1 intensities. pNMF was unable to distinguish exactly overlapped MS1 patterns for 3 cases, and Skyline missed the right peak top for 2 cases.

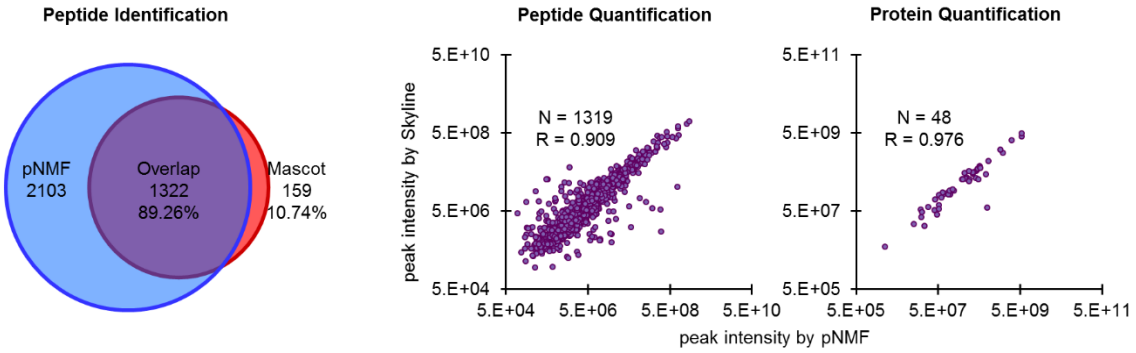


**Figure 9. pNMF for forty-eighty-protein standard.**
The Venn diagram represents the peptides identified using pNMF and Mascot (left). The peak intensities of peptide ions commonly identified by pNMF (*x*-axis) and Skyline (*y*-axis) show correlation coefficients of 0.909 at the peptide level, and 0.976 at the protein level (right).

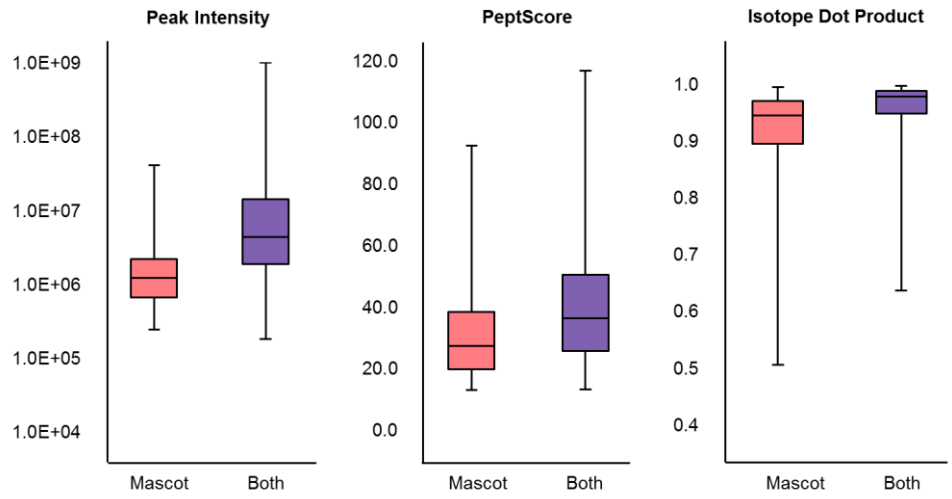


**Figure 10. Profile of peptides exclusively identified by Mascot in forty-eight-protein standard.**

The box plot shows that the 159 peptides exclusively identified by Mascot have 3.45 times lower intensity, 8.59 lower PeptScore, and 3.37% lower idotp than the 1322 peptides commonly identified by both Mascot and pNMF.

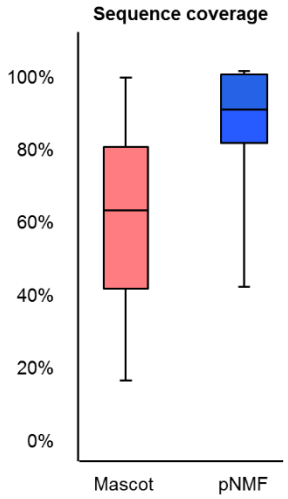


**Figure 11. Sequence coverage comparison between 1481 peptides identified by Mascot and 3425 peptides identified by pNMF.**

Mascot and pNMF yielded medians of 62.02% and 87.29% coverage, respectively.

## 3. Proteome-scale application

Library searching has been continuously developed to improve the identification speed, accuracy, and sensitivity[33–35]. I applied pNMF with a library-based dictionary for proteome-scale data using whole E. coli cell lysates with 70-min gradient. The reference library was the combination of the detected peptides analyzed from E. coli cell lysates with 70-min gradient and 550-min gradient by Mascot at 1% false discovery rate (FDR) confidence. The dictionary was then constructed to cover all library peptides and extended to include every charge states possible within *m/z* range. The result from the longer gradient separation provides a larger number of peptides for creating a comprehensive dictionary and offers retention times for initializing an activation matrix. I converted observed retention times to reference retention times by linear regression with ±2.5 min window.

### 3.1 Protein and peptide identification

In benchmarking with the conventional method, pNMF identified 3286 out of 3509 or 93.64% of Mascot-identified peptides. The correlation between two methods for measuring the retention times of commonly identified peptide ions was very well agreed, with y = 0.999x, R = 0.996 (Figure 12). In protein level, pNMF identified more than twice of Mascot identification number and covered 497 out of 502 or 99.00% of Mascot identification (Figure 13, top). The conventional method and pNMF identified provided median sequence coverages of 14.80% and 25.12%, respectively (Figure 14).

### 3.2 Protein and peptide quantification

Finally, I quantified the peak intensities of commonly identified peptide ions and proteins. After eliminating peaks with S/N ≤ 10, the correlation coefficient of was observed at 0.982 and 0.941 at the protein and peptide levels, respectively (Figure 13, bottom).

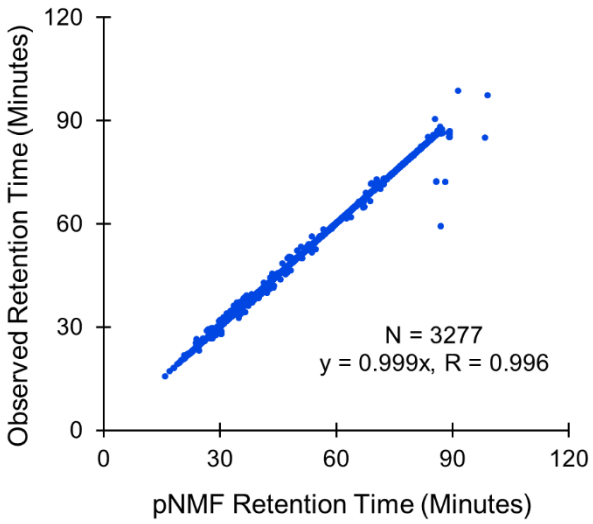


**Figure 12. Accuracy of peptide retention time annotation in proteome-scale application.**

Proportional relationship between pNMF (x-axis) and the conventional method (y-axis) for commonly annotated *E. coli* peptide ions.

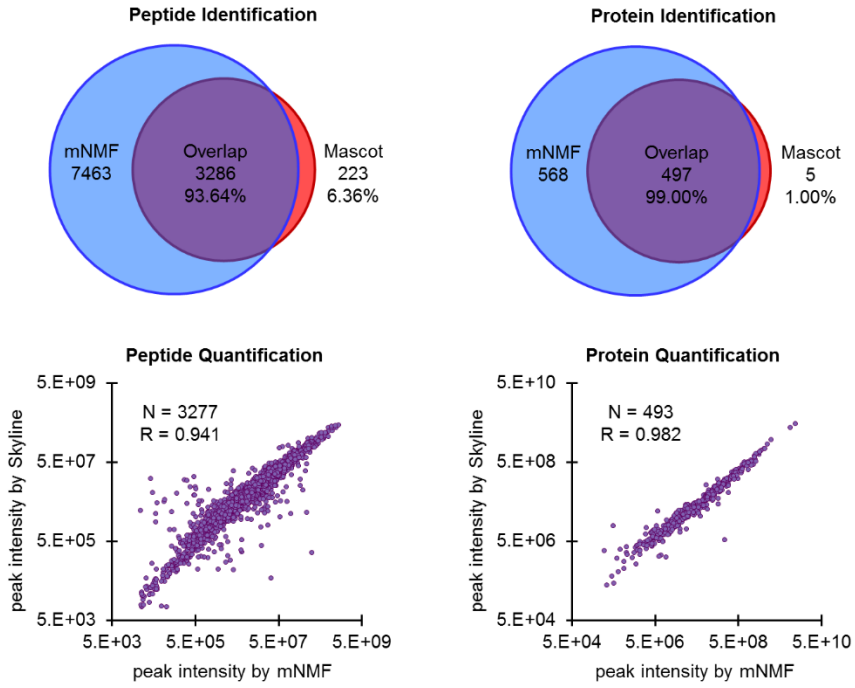


**Figure 13. pNMF for proteomics application.**

The Venn diagram represents the *E. coli* peptide ions and proteins identified using pNMF and Mascot (top). The peak intensities of peptide ions commonly identified by pNMF (x-axis) and Skyline (y-axis) show correlation coefficients of 0.941 at the peptide level, and 0.982 at the protein level (bottom).

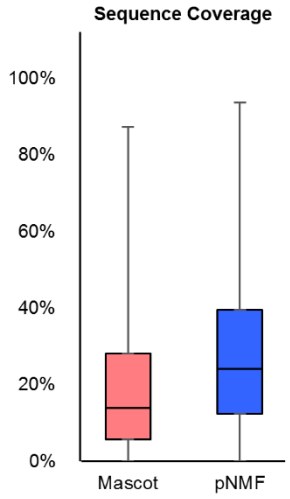


**Figure 14. Sequence coverage comparison between 3774 peptides identified by Mascot and 10749 peptides identified by pNMF.**

Mascot and pNMF yielded medians of 14.80% and 25.12% coverage, respectively.

## 4. Computational performance

The computational cost of pNMF depends on the sizes of the matrices. For four-protein standard, pNMF required approximately 7.30 min on the system I used. Specifically, reading the mzML file, running *in silico* digestion, learning noise, constructing $V$ and $W$ matrices, initializing $H$ matrix, updating pNMF, and printing report took approximately 1.11, 0.38, 1.92, 2.41, 0.95, 0.35 and 0.18 min, respectively. The forty-eight-protein standard by pNMF took approximately 45.78 min. Despite the size of the computational tasks, which involve several million elements, the computation time and space required by pNMF are manageable on a common desktop computer, based on code optimization and appropriate choices of libraries for sparse computation, particularly, scipy-sparse[36]. For proteome-scale application using *E. coli* sample, pNMF could reduce running time due to precise predicted retention times to narrow down update regions. The algorithm took 24.17 min for all processing. The memory footprints for all sample processing were less than 2 GB.

# Conclusion

I have proposed NMF-based approach to specifically analyze three-dimensional mass spectrograms obtained from proteomics studies efficiently. Our pNMF incorporates isotopic distribution, learns noise, exploits a protein-peptide hierarchical relationship by using a group sparsity constraint, and configures a reasonable initialization by using predicted retention times. The proposed update rule with the constraint of choice guarantees convergence to a local optimum, meaning that with appropriate initialization, pNMF guarantees convergence to the right solution.

In the case of four-protein standard, pNMF gave a better resolved chromatogram with a higher accuracy of peptide retention times than bNMF. For forty-eight-protein standard and *E. coli* samples, pNMF provided the results with excellent correlations to the conventional methods, Mascot/Skyline for both identification and quantification without the need for preprocessing. Additionally, pNMF increased number of identified peptides and enabled quantification of more than a thousand proteins. In this study, I have focused on protein identification and quantification, but the same approach should be readily applicable for various purposes related to the interpretation of mass spectrograms.

The present results indicate that the NMF algorithm-based approach is very effective for mass spectrometry-based proteome analysis, particularly by using the library-based dictionary. Further improvements should also be possible by incorporating other significant features such as product ions from proteomic mass spectrograms, which are presented in the next chapter.

# Chapter 2
# An Integrated Non-negative Matrix Factorization Framework to Analyze LC/MS/MS Spectrograms in Proteomics

## Introduction

Popular conventional approaches for peptide identification require LC/MS/MS information, since precursor *m/z*, retention times, and intensities from LC/MS are insufficient dimensions to separate peptide ion peaks in mass spectrograms derived from highly complex proteome, i.e., human proteome. Product ion patterns from the peptide fragmentation process provide another 3D information of product ion *m/z*, times, and intensities. The typical algorithm relies on peptide-spectrum match (PSM) where the peptide answer of an experimental product ion patterns are concluded by the similarity to the reference product ion patterns from other experiments or the theoretical pattern simulations[9,34]. However, the lack of reference patterns restricts the discovery of new peptides, and the perfect theoretical simulations are not established due to the incomplete knowledge in fragmentation process, particularly the intensity dimension. As a result, the experimental patterns cannot resemble the theoretical ones which leads to miss-identification. One of the current trends for the theoretical product pattern generation is predicting product ion intensities by sophisticated probabilistic models or machine learning methods for the subsequent PSM[37–39].

In this study, I propose an extension of NMF-based approach for identifying product ion mass spectrograms. The previous chapter pNMF approach is modified for precursor mass spectrogram inputs. *V*' is input with top product ion peak picking with special inclusion for y1- and b2-ion peaks. *W*' dictionary is then constructed for the precursor candidates using precursor *m/z* filter from *V*'. Lastly, *H*' is updated with the group sparsity constraint using a precursor-product ion hierarchical relationship. The results of product ion pNMF can integrate to precursor pNMF by replacing the *in silico* digestion dictionary of precursor pNMF. Our new computational method can identify product ion mass spectrogram without the need of intensity prediction algorithms or similarity measurement functions for PSM. The integrated information of precursor and product ion mass spectrograms in statistical framework is prospective to identify more peptides and enhance the accuracy in protein quantification.

## Design

I designed an extension of pNMF for analyzing product ion mass spectrograms. First, a vector is used for representing a mass spectrogram of each time as $V'$ ($m/z$ × 1). A supervised dictionary presents peptide candidates. According to precursor $m/z$ information of $V'$, it can be used as a filter to select a list of candidates from *in silico* digestion database. The theoretical $m/z$ with a single charge for all possible b-ions and y-ion types are calculated and designated to each column as $W'$ ($m/z$ × ion type). All ion types from the same peptide candidates provide a group structure as $g'$. An activation vector $H'$ (ion type × 1) is allowed to update with a group sparsity constraint. The highest sum of activation of $H'_g$, in $H'$ provides the precursor candidate for integrating to precursor pNMF.

The precursor candidates from the product ion level of pNMF can integrate to precursor pNMF by replacing *in silico* digestion dictionary. Features of six isotopic peaks of activated precursor candidates $H'_g$, are calculated and multiplied by its original intensity from $V'$. Precursor candidates from the same parent peptide are stacked, normalized, and assigned to columns of the precursor dictionary $W'_k$ in $W'_S$. This particular stacked design exploits the nature of precursor information, since all peptide ion precursors of the same peptide share the same retention time profile. In other words, the algorithm activates the corresponding row $H'_k$ in $H'_S$ along retention time for all stacked precursors with the same profile instantaneously. Furthermore, instead of depending on retention time prediction algorithm, time $t_x$ of each product ion mass spectrogram is useful information for initializing the precursor activation matrix. The pNMF for precursor mass spectrogram is then computed for peptide identification and quantification.

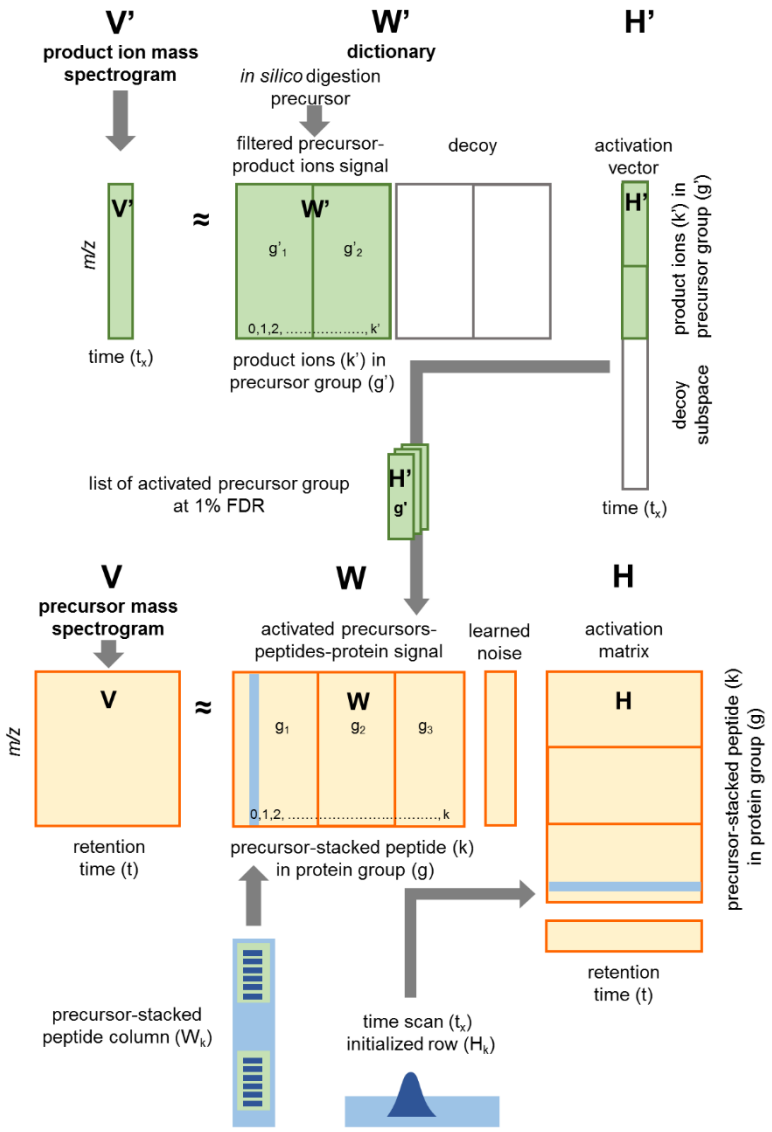


**Figure 1. Overview of integrated pNMF for product ion and precursor mass spectrograms.**

**The workflow of pNMF**

The workflow comprises the following four steps.

**Step 1. Matrix construction**

**1.1 The observed mass spectrogram $V'$:**

A vector is used to represent mass spectrogram of product ion information for each time. Top 6 in window ± 30 approach is applied to select top intensities due to its best performance among preprocessing techniques[40]. According to theoretical calculation, possible y1- and b2-ion peak intensities are included regardless their intensities. The intensity values are transformed by log and assigned to $V$ at appropriate $m/z$ bins.

**1.2 The dictionary $W'$:**

Peptide precursor candidates are selected into the dictionary from *in silico* digestion. For *in silico* digestion, peptide ions ranging in length from seven to 50 amino acids with charges ranging from +2 to +8 were included. Cysteine carbamidomethylation was set as a fixed modification. Methionine oxidation and N-terminal methionine excision were set as variable modifications. For selection, according to precursor $m/z$ information of $V$, it can be used as a filter to exclude peptides candidates unless its $m/z$ is in the range of ±0.01 $m/z$. In case of no peptide candidates are selected, the range up to ±0.03 $m/z$ is applied with maximum number of two candidates. Except b1-ion, all theoretical masses of b-ions and y-ions are calculated and input into each column with intensity as 1. Decoy subspace is constructed by the concept of a concatenated target-decoy approach using a shuffle sequence[41,42].

**1.3 The activation $H'$:**

$H'$ is allowed to be updated with a group sparsity constraint according to precursor-product ion group structure.

**Step 2. Sparsity constraint weight calculation**

Since $V'$ is extremely sparse, a maximum sparseness measure[13] for vector is used as 1. The weight is then multiplied with square root peptide length for standardization.

**Step 3. pNMF algorithm**

pNMF algorithm is as same as the previous chapter where $D$ is the Kullback-Leibler divergence and $\Omega$ is $\log/L1$.

**Step 4. Post-processing for interpretation**

The resulting $H'_g$, is summed and divided by square root of its length to refer the activation of each peptide candidate. For statistical evaluation, 1% FDR is used for calculating minimum activation cut-off.

All computations were performed on a desktop computer with a 3.30 GHz E3-1226 v3 4-core processor and 32 GB main memory.

# Experimental Section

### Proteomic experiment setup and LC/MS/MS analysis

Mass spectrogram obtained from four-protein standard was prepared as described in the previous chapter.

# Results and Discussion

## 1. The proteome complexity in precursor mass spectrograms

I first investigated the ambiguities of precursor mass spectrograms. Precursor *m/z* and retention times provide sufficient information to separate peaks of peptide ions in mass spectrograms derived from moderate complex samples. However, a highly complex proteomics sample, such as human proteome, increases substantially the number of peptide ions with both similar *m/z* and retention times.

In previous pNMF design, I represented peaks with 0.01 *m/z* bin and 0.2 min retention time bin. According to this setting criteria, I quantified the numbers of overlapped experimental peptide precursor peaks in human proteome based on the Mascot-identified human peptide ions of cytoplasmic, organelle, and nuclear fractions in 60-min gradient-experiment[43]. Using ±0.01 monoisotopic *m/z* bin and ±0.2 min retention time bin, 11896 of 11961 peaks or 99.46% were distinguishable from other precursor peaks without overlaps (Figure 2a). However, precursor peaks cannot be described by only monoisotopic *m/z* peaks, the isotopic peaks are usually presented in mass spectrograms. I included the first monoisotopic peaks into calculation and found that the number of non-overlap peaks decreased to 10116 of 11961 peaks or 84.57%. Number of more than one overlap peaks were also discovered at 2.85% (Figure 2b). The overlap peaks harden NMF to identify peaks since the isotopic distribution pattern of each peptide becomes unclear and different from theoretical isotopic distribution. The quantification is also worsened by peaks unresolved in time. These results indicate that while precursor *m/z* and retention times are useful information, more parameters are necessary to distinguish peptide ions for identification and quantification in the large proteome-scale sample.

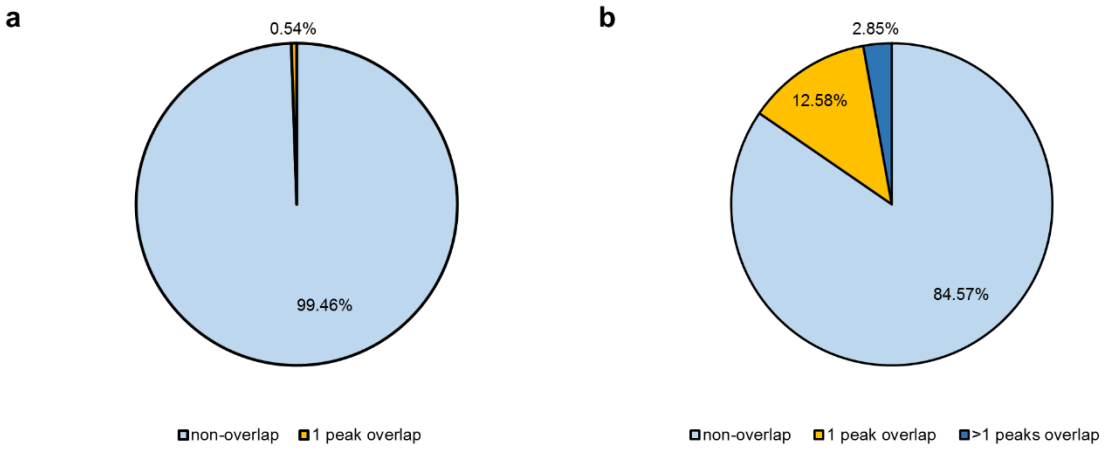


**Figure 2. The complexity of experimental human proteome precursor mass spectrogram.**

(a) The overlap of monoisotopic *m/z* peaks with similar retention times.

(b) The overlap of monoisotopic and first isotopic *m/z* peaks with similar retention times.

## 2. Prior knowledge analysis of product ion mass spectrogram

The process of fragmentation using low-energy collision produces a predominant series of b- and y-ions. Unlike precursor isotopic distributions, fragment peptide ions do not possess theoretical patterns. The product ion mass spectrograms cannot be predicted. In order to provide prior knowledge for NMF, I explored the types and numbers of ions used for identification.

I applied web scraping to extract Mascot-detected and identified fragment peptide ions from ion matches used for scoring and all possible ion matches in detection, respectively. Collecting 3818 high-quality product ion spectra with Mascot-identification from *E. coli* sample, I found b6-, b5-, and b7-ions were top three ion types used for scoring and in overall profile. Mascot unconsidered approximately 30% of overall detection for scoring (Figure 3a). In case of y-ions, y6-, y5-, and y4-ions were top three for scoring and in overall profile (Figure 3b).

Here, peptide precursors were fragmented and identified with the minimum length at seven amino acids. The small ion types from three to eight were likely to be reported more than the large ion types. I normalized each type of ion with the number of precursors possibly possessing the ion. In other words, the numbers of spectra with the possibility to possess each ion, i.e. the length of peptide precursors is longer than the type of product ion, were used as 100%. Surprisingly, the fragmentation provided up to 94.53% in case of b6-ion and 97.43% in case of y6-ion. Both ion types from three to eight were detected more than 80% in all spectra possibly possessing them. These high percentages of ions could be a result of analyzing high-quality spectra collection. Mascot discarded approximately less than 50% of each ion type for scoring. Assignments of small N- and C-terminal ions, such as y1- and b2-ions, provide very useful information for peptide sequence. I found that although only 27.74% and 53.30% of all spectra possibly having y1- and b2-ions respectively, but 19.17% and 42.98% of y1- and b2-ions were identified and used for scoring, which emphasized their importance for peptide sequencing process (Figure 3c-3d).

For numbers of ions used for scoring, Mascot mainly took both y- and b-ions less than ten ions. However, plenty of ions detected were not considered. I found that more than 80% of spectra had more than ten y- and b-ions (Figure 4).

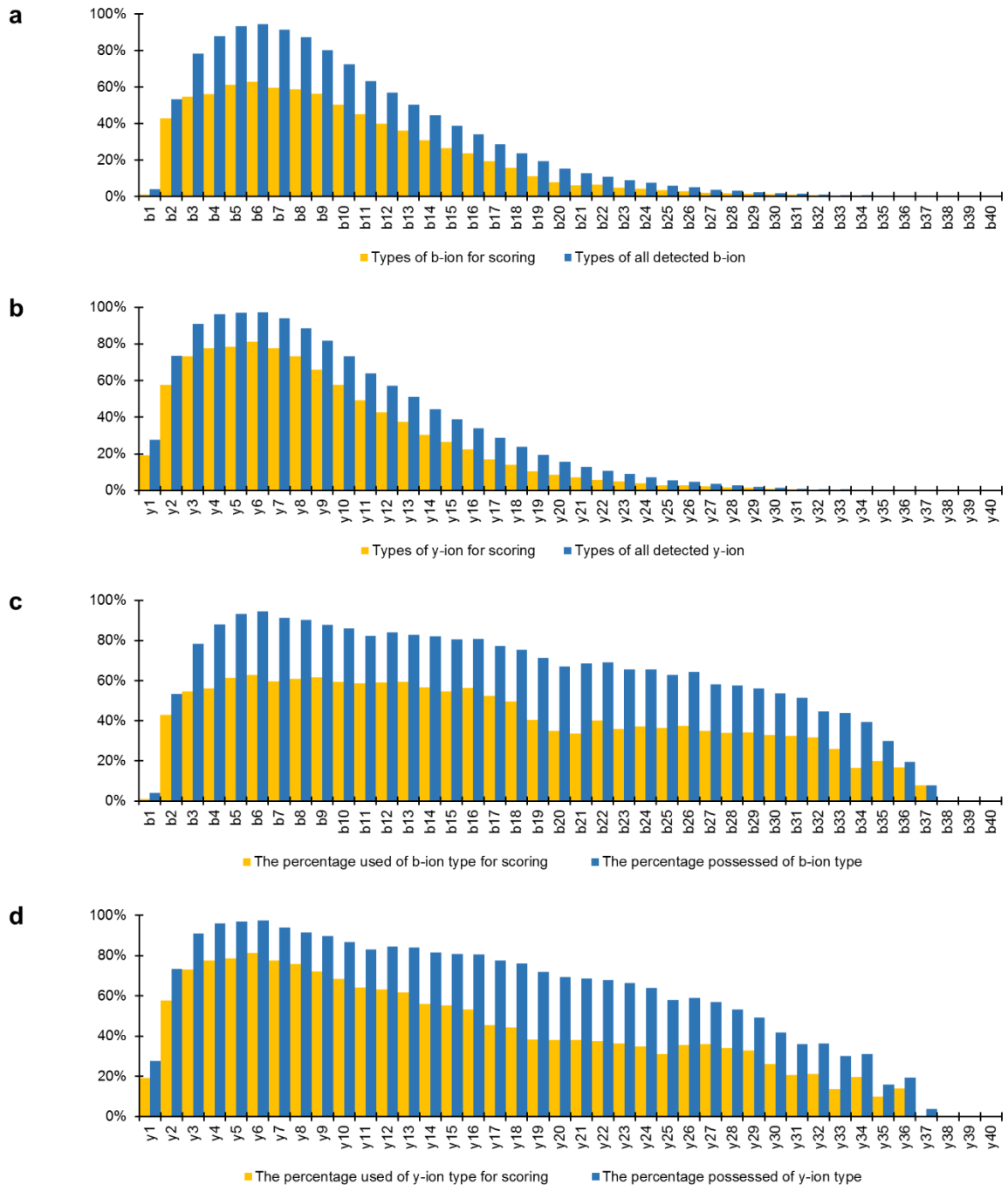


**Figure 3. The b-ion and y-ion distribution profiles in Mascot-scoring and overall detection.**

(a) Types of b-ions in Mascot-scoring (shown in yellow) and overall detection (shown in blue).

(b) Types of y-ions in Mascot-scoring (shown in yellow) and overall detection (shown in blue).

(c) The percentage used of b-ions in Mascot-scoring (shown in yellow) and overall detection (shown in blue).

(d) The percentage used of y-ions in Mascot-scoring (shown in yellow) and overall detection (shown in blue).

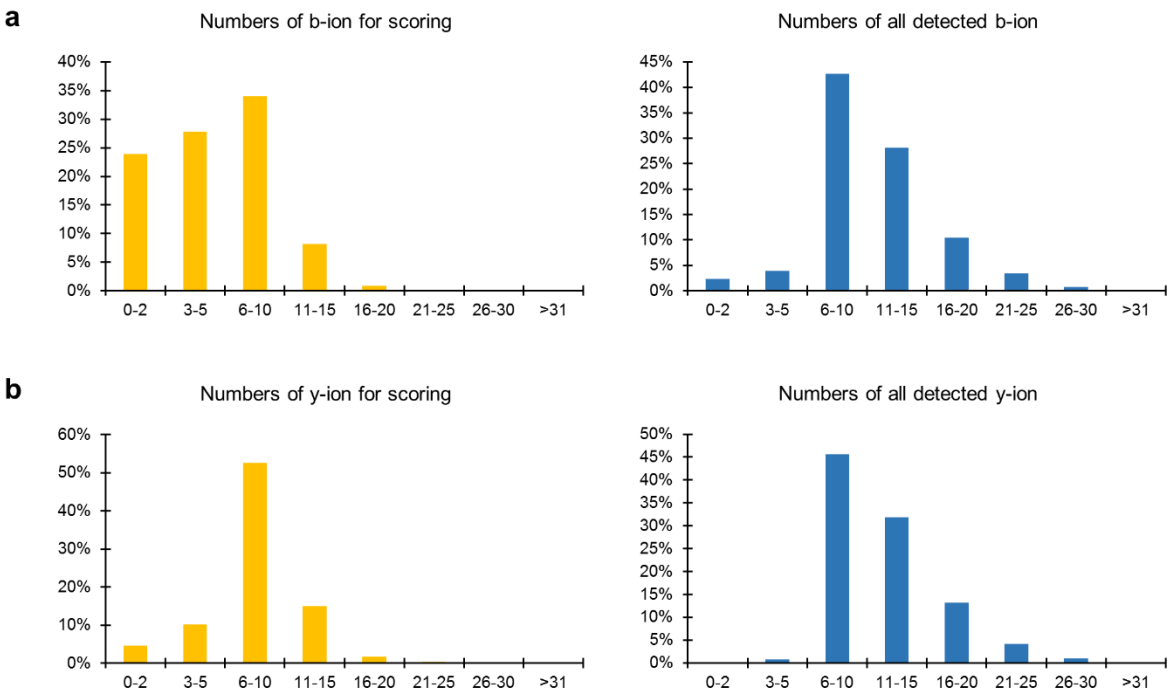


**Figure 4. The number of b-ions and y-ions per each product ion mass spectrum in Mascot-scoring and overall detection.**

(a) Numbers of b-ion for scoring (shown in yellow) are generally less than in the overall detection profile (shown in blue).

(b) Numbers of y-ion for scoring (shown in yellow) are generally less than in the overall detection profile (shown in blue).

### 3. pNMF for product ion identification

I first examined the possibility of using NMF for identifying product ion spectrum without integrating with precursor level. I collected 346 high quality spectra with Mascot identification results of four-protein standard and applied pNMF. I obtained all accurate identified peptide sequences for 346 spectra (100%).

I then tested integrated pNMF for all 21865 spectra, regardless of qualities and identification results, of four-protein standard. pNMF identified 707 out of 730 (96.85%) of Mascot-identified spectra (Figure 5a). I referred to identified peptides from spectrum identification results. pNMF identified 341 out of 342 (99.71%) of Mascot-identified spectra (Figure 5b). The remaining one peptide belonged to an unidentified spectrum with few singly charged production ions as illustrated above. For the peptides commonly identified by pNMF and Mascot, the retention time correlation was excellent with $y$ = 1.000x, $R$ = 0.997 (Figure 5c). pNMF identified an additional 126 peptides, and provided 97.59% sequence coverage for the four proteins, whereas the conventional method gave 96.59% coverage on average.

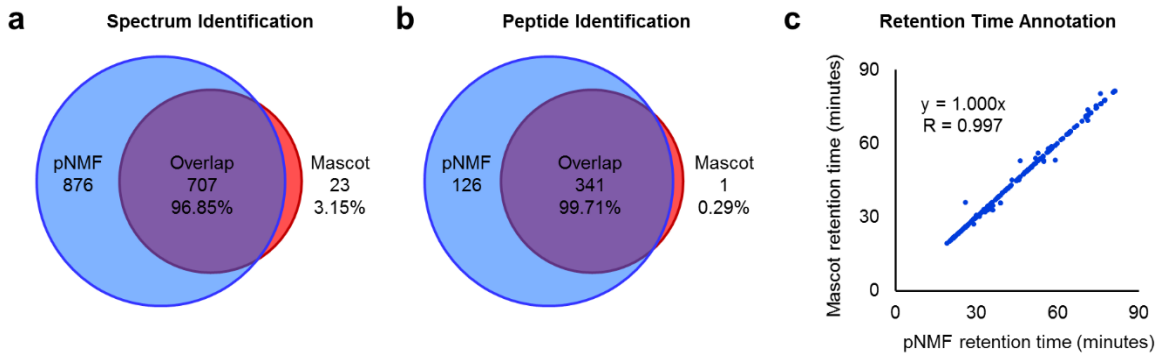


**Figure 5. Product ion identification.**

(a) The Venn diagram represents the common spectra identified using pNMF and Mascot.

(b) The Venn diagram represents the common peptide ions identified using pNMF and Mascot.

(c) Accuracy of peptide retention time annotation in four protein standard shows proportional relationship between pNMF (x-axis) and the conventional method (y-axis) for commonly annotated peptide ions.

I investigated the causes for the remaining 23 spectra that were exclusively identified by only Mascot. Actually, pNMF identified 19 of them but their activations were below 1% FDR cut-off. These 19 spectra also had low PeptScore, indicating of poor peak profiles. Other four spectra were not identified due to very low product ion intensities (Figure 6a) or few singly charged product ions (Figure 6b).

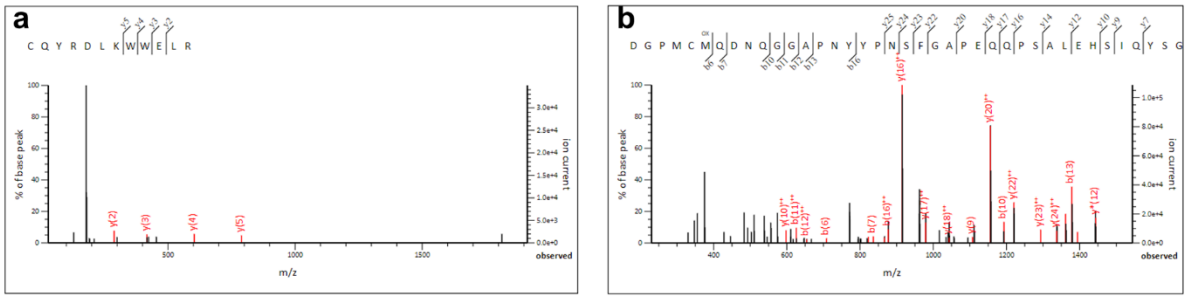


**Figure 6. Example cases of unidentified product ion spectra.**

(a) Very low product ion intensities of CQYRDLKWWELR at 438.97, 4+

(b) Few singly-charged product ion peaks of DGPMCMQDNQGGAPNYYPNSFGAPEQQPSALEHSIQYSGEVRR at 960.63, 5+

I compared the peaks of precursor ions (Figure 7a), peptides (Figure 7b), and proteins (Figure 7c) commonly quantified by pNMF and Skyline. The correlations in precursor ion level is less than peptide level and protein level. The correlations calculated by peak intensities at the top are less than peak areas in all levels. These results suggest that pNMF is effective for calculating peptide and protein quantification. However, the design of the algorithm can be improved with refinements to address the peak tops and to model each precursor ratios more precisely.

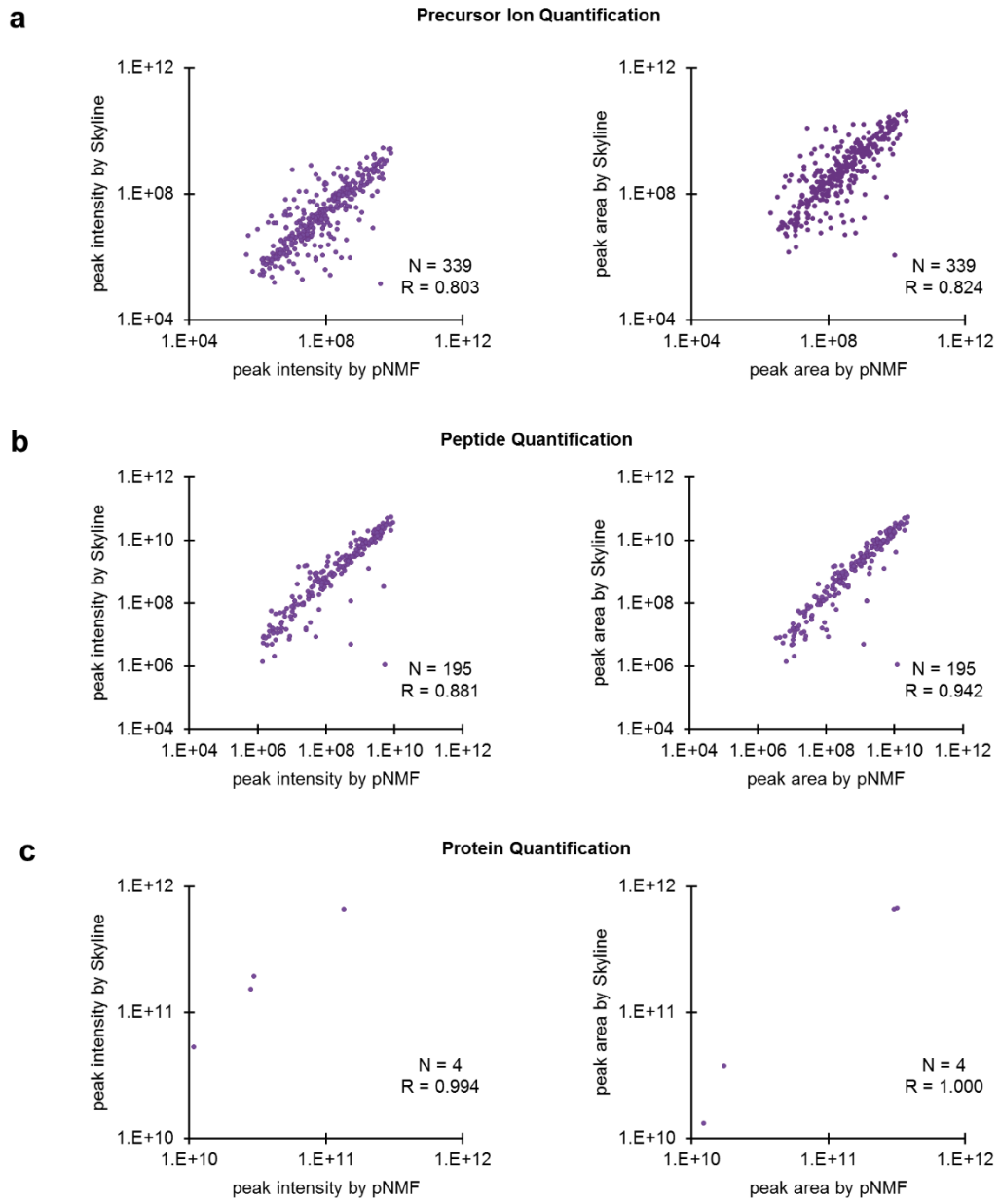


**Figure 7. The peak intensities and areas of commonly quantification by pNMF (x-axis) and Skyline (y-axis).**

(a) Commonly quantified precursor ions show correlation coefficients of 0.803 and 0.824 calculated from peak intensities and peak areas, respectively.

(b) Commonly quantified peptides show correlation coefficients of 0.881 and 0.942 calculated from peak intensities and peak areas, respectively.

(c) Commonly quantified proteins show correlation coefficients of 0.994 and 1.000 calculated from peak intensities and peak areas, respectively.

# Conclusion

I have applied the pNMF algorithm to analyze product ion mass spectrograms using prior knowledge. The new extension of pNMF exploits precursor *m/z* for filtering peptide candidates for the dictionary and specially includes y1- and b2-ion peaks for analysis. The search space, i.e., the dictionary, is reasonably reduced from the number of all precursors provided by *in silico* digestion to the number of activated precursor candidates provided by product ion pNMF. Instead of arranging these activated precursor candidates into the column of the dictionary one-by-one, they are configured in stack of the same parent peptide. This particular stacked design fits the nature of precursors which share the parent peptide will share the retention time profile. Furthermore, the integration of product ion pNMF improves an initialization of precursor pNMF by providing times. As a result, pNMF can identify more spectra and peptide ions than the conventional methods at 1% FDR.

The prior knowledge analysis of product ion mass spectrogram provides a number of possible ways to leverage the approach using more significant product ions for the dictionary. Future improvements should be possible by incorporating +2 and more charge states for b-ions and y-ions, and more ion types such as a-ions and immonium-ions. In the current design, y1- and b2-ions are promoted and b1-ions are suppressed according to the prior knowledge, but other ions can also be adjusted.

# Summary

In this thesis, I show that NMF is suitable to identify and quantify peaks in mass spectrograms obtained from LC/MS and LC/MS/MS analyses. Our major modifications to the bNMF are the incorporation of group sparsity constraints by exploiting the hierarchical relationships of protein-peptides and peptide precursor-product ions.

In chapter 1, our pNMF can perform chromatogram extraction, identify and quantify more peptides than the conventional approaches using the proteome-scale sample. The simultaneous identification and quantification are another strong point of pNMF, since the conventional methods only calculate the peaks after the identification in LC/MS/MS which leads to miss-quantification problem. In chapter 2, pNMF shows the capability to identify LC/MS/MS mass spectrograms with improved numbers in all levels of spectrum, peptide precursor ion, and peptide. The workflow of pNMF does not need the theoretical spectrum prediction which is one of the weak points in computational proteomics due to lack of the complete knowledge of fragmentation. In other words, I provide the evidence that the intensity prediction algorithm, which is a popular research at this moment, is not important. In overall, I show the new possibility to use NMF algorithm in mass spectrogram analysis and solve persistent problems of proteomics in miss-identification and miss-quantification.

In my perspective, machine learning will have become an important technique for computational proteomics. The appropriate mathematical and statistical models will solve the persistent problems such as identifying unidentified peptides in mass spectrograms, accelerating the process of analyzing large data, and improving the understandings of MS-based proteomics. In these few years, I expect to see more algorithms incorporated into the conventional pipeline and provide significant improvements. Interdisciplinary research, such as computational proteomics, is always not easy, since people tend to misunderstand the complexities of other fields. The requests for higher sophisticated algorithms or larger data are unavoidable from the professionals of each discipline. However, I think we should step out to show the possibilities of our inventions. In this way, we can fill the gap of knowledge and discover new solutions to improve people' lives.

# Supplementary Information

## An Introduction to Conventional Proteomics Mass Spectrogram Generations and Analyses

Proteomics is a large-scale study of proteins. Since proteome samples are often highly complex with numerous proteins involved, they need strong separation techniques and proper representations that ensure the feasibility of subsequent data interpretation. Shotgun proteomics[44] has become a universal method to identify and quantify peptides in a mixture by integrating two effective separation systems: LC and MS, and representing the analytical results in the form of mass spectrograms.

## Shotgun Proteomics for Mass Spectrogram Generations

Shotgun proteomics starts from protease digestion to cleave proteins into peptides. Peptides are comparatively simple and easy than proteins for separating by LC/MS system. LC separates peptides in the mixture, according to their physicochemical properties. Peptide samples are injected and passed through a stationary phase column with mobile phase. The time required to elute a particular peptide from LC system is called as retention time. The retention time becomes the first dimension of separation in the mass spectrograms. In MS, eluted peptides are ionized by electrospray, and separated based on *m/z* by mass analyzer as the second dimension. Lastly, separated peptide ions are counted by detector to record the signal intensity as the third dimension (Figure 1).

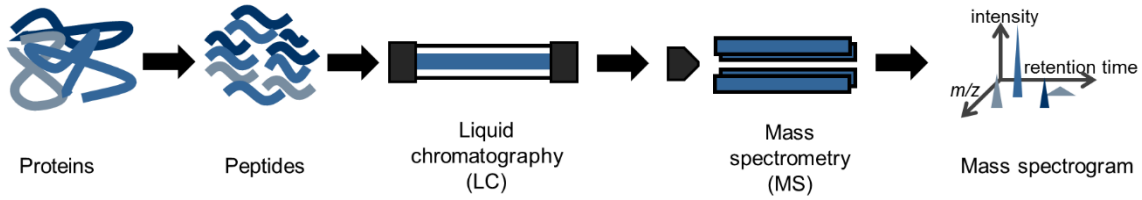


**Figure 1. An overview of common shotgun proteomics experiment using LC/MS.**

Modern MS has fast scan speed with high mass resolving power to resolve the adjacent *m/z*, isotopic patterns, and other fine details of peptide ions. Thus, all dimensions are outspread and peptide information is located sparsely due to the precise scales. The final mass spectrograms can be depicted as a profile of peptide peaks at very specific locations in large non-negative 3D space (Figure 2).

Currently, mass spectrometers are commonly operated in tandem. Tandem MS (MS/MS) employs two mass analyzers, MS1 and MS2. The configuration consists of a fragmentation region in between two stages. In MS1, peptide ions or precursor ions are analyzed, detected, and selectively passed to be fragmented with low-energy collision-

induced dissociation. Product ions resulting from fragmentation are scanned and recorded in MS2 as product ion mass spectrograms.

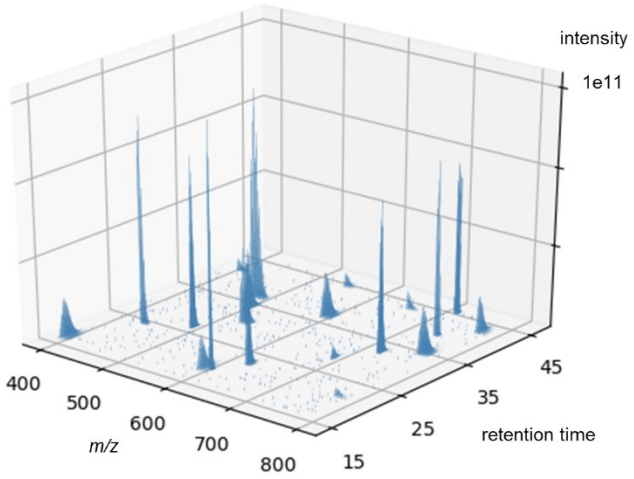


**Figure 2. Precursor mass spectrogram representation as a large sparse non-negative 3D space with *m/z*, retention time, and intensity.**

The first dimension of product ion analysis is time, which is correlated to retention time in precursor mass spectrogram. However, at a particular retention time, many precursors are possibly eluded from LC column. In order to separate the overlapped precursors, MS2 quickly performs scanning several times for each selected precursor found in a mass spectrum of MS1. The second dimension is *m/z* of product ions. Product ion *m/z* patterns are a result of low-energy collision. Low-energy collision dominantly breaks C–N bonds and produces b- and y-ions from N-terminal and C-terminal precursor ions, respectively. The last dimension is product ion signal intensity (Figure 3).

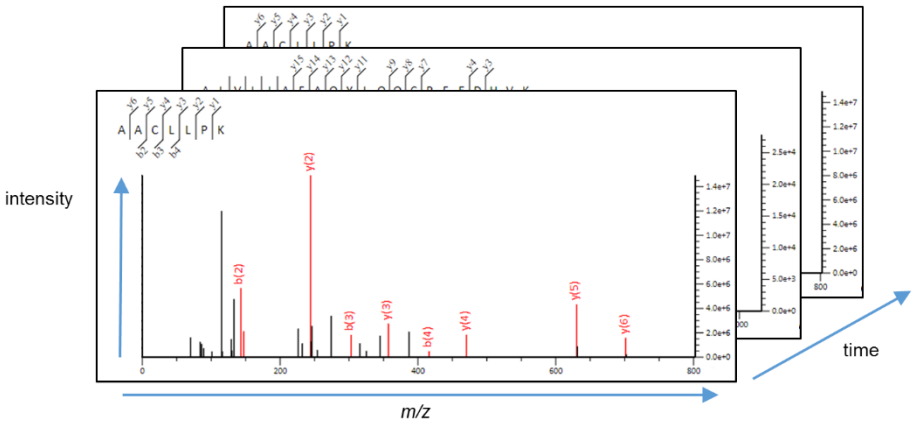


**Figure 3. Product ion mass spectrogram representation by *m/z*, time, and intensity.**

# Computational Proteomics for Mass Spectrogram Analyses

## Conventional approaches

The increasing size and complexity of mass spectrograms demand mathematics and statistics for data interpretation. The algorithms underlying computational proteomics platforms are a key factor to identify and quantify peptide peaks from the large dimensional data in accurate and efficient manners. Mostly, the conventional approaches remove the complexities in mass spectrogram by performing preprocessing prior to peptide identification.

## Preprocessing

Preprocessing demonstrates the improvements for analysis[45] by deciding which peaks to remove from precursor mass spectrograms in order to simplify big data into statistically manageable information. Several techniques are applied for recovering peptide peaks from baseline, noise, and contamination.

- **Deisotoping and deconvolution:** deisotoping and deconvolution simplify data by reducing the number of peaks originated from the same peptides. Elements naturally have multiple masses due to their stable isotopes. The combination of key elements of peptides: carbon (C), hydrogen (H), oxygen (O), and nitrogen (N), creates a series of *m/z* peaks with isotopic distribution. For deisotoping, it merges or collapses isotopic peaks of each peptide to a single peak. For deconvolution, it replaces peptides peaks at multiple charged states with one peak of singly charged peptide. The resulting mass spectrogram is thus cleaner and easier to interpret[46].
- **Peak picking:** Peak picking relies on feature detection techniques to obtain the information of peptide intensities with the least interferences from non-peptide compounds. Several theories have been proposed using signal-to-noise ratio, continuous wavelet transform, or Gaussian function model. However, completely accurate models of the peptide features do not exist. The resulting picked peaks are possibly included non-peptide peaks and excluded peptide peaks[47,48].
- **Noise:** Noises in mass spectrograms are inherent and ubiquitous. The noise peaks come from air particles, small molecules released from materials contacted in the experimental procedures, and protein-related contaminants introduced from human sources. The lists of common mass spectrometry contaminants have been reported[49]. Another type of noise is electronic noise. Appropriate techniques, such as baseline subtraction and smoothing, are applied to filter their uniformly distributed peaks throughout the profile. However, low-abundance peptides, which have weak signals in mass spectrogram, are likely discarded by these processes.

For product ion mass spectrograms, preprocessing methods mainly discard peaks with low intensities by selecting a limited number of top peaks. Three main categories of preprocessing include top X intensity approaches, top X intensity in Y regions, and top X intensity in a window of ±Z[40].

Although preprocessing eases the subsequent data analysis, it is arguable how many peptide signals are lost during the process, particularly low abundant peptides, the delicate part of proteomics study.

**Peptide identification**

Precursor mass spectrograms provide necessary information of separated *m/z* features of peptides in order to perform identification. The separated *m/z* can be observed in mass spectrum, a subspace of mass spectrogram with dimensions of *m/z* and intensity (Figure 4).

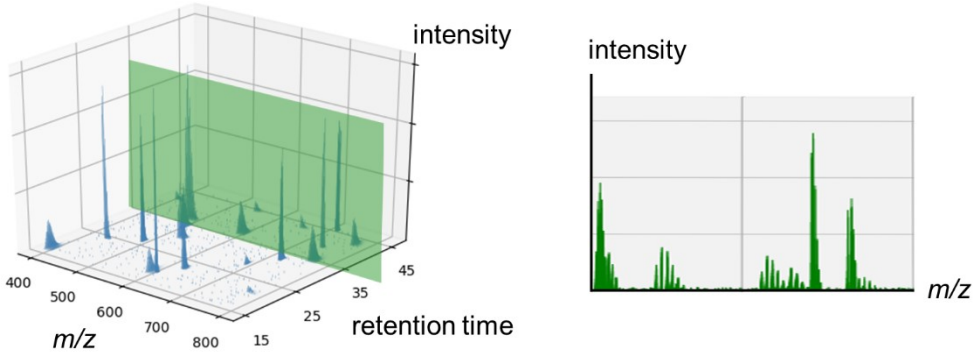


**Figure 4: Mass spectrum, the subspace of mass spectrogram with dimensions of *m/z* and intensity, for peptide identification**

Tandem mass spectrometry has become a routine task for conventional identification. Currently, the common computational methods perform peptide identification from product ion mass spectrograms from MS2 only with one of three strategies: database searching, spectral library searching, and *de novo* sequencing.

- **Database searching:** Database search is a process of statistical comparison between an experimental spectrum and a theoretical spectrum in the database and calculating a confidence of match from built-in score function. Database searching is currently the most popular approach with several well-used platforms, such as SEQUEST[24], Mascot[27], X! Tandem[50], and MS-GF+[51]. However, the variability of peptides, such as post-translational modifications, can weaken the search performance due to expanding the search space. Complicated fragmentation process is also not yet completely understood to design appropriate theoretical fragment spectra and suitable scoring functions for

comparison. The database search engines also require the expertise from users to define parameters, such as protein molecular mass range, mass tolerance, possible charges, and possible modifications, to obtain the correct output.

- **Spectral library searching:** Spectral library searching has emerged as an alternative of database search engines for peptide identification. The search utilizes a reference library of a set of confidently assigned spectra obtained from previous experiments as a template for comparison. The result of search heavily relies on the choice of library of template spectra which should be of high quality and similar in analytical process. Reliable sources of spectra are provided[52–54]. The major limitation of this method is the incapability to identify novel peptides or unexpected mutations and modifications due to lack of references.

- ***De novo* sequencing:** *De novo* sequencing identifies peptide without constructing database or library, and thus can discover novel identifications regardless of the reference. The process takes the mass difference between two neighboring peaks to compute the mass of an amino acid present in a candidate peptide. The popular *de novo* sequencing software are such as PEAKS[55], PepNovo[56] and NovoHMM[57]. However, the wrong identification can be caused by fragment peaks are unclear by overlaps, shifts, or losses. *De novo* peptide sequencing also generally requires more time and high precision data as an input.

The conventional algorithms have different strengths and weaknesses for peptide identification. However, it is unquestionable that all approaches abandon too many useful features of mass spectrograms. Several alternative platforms have proposed to exploit mass spectrograms obtained from both LC/MS and LC/MS/MS analyses[58,59]. Integrating information of precursor ions with product ions solves miss-identification caused by missing data in either one of two mass spectrograms and improves poor identification caused by ambiguities of both mass spectrograms.

**Protein identification**

One of the most common challenges in proteomics studies revolves around protein identification, i.e., protein inference, the process of inferring the accurate present proteins in the experimental peptide mixture generated by shotgun proteomics.

The ambiguities arise from too low dependencies or too high redundancies between peptide and protein evidence. A "one-hit wonder" refers to an insufficient dependency to draw a confident conclusion for a protein which has only one peptide identified. Generally, the identification with at least two peptides is the criteria to determine a particular protein identification, but the situation likely remains unresolved if identified peptides are shared products of multiple proteins[60]. Tracing peptides back to the parent protein is a complicated issue. The suitable statistical analysis that can incorporate the dependency between protein and peptide are necessary to solve these problems.

**Peptide and protein quantification**

Peptide and protein quantifications enable more meaningful interpretations for proteomics than solely depend on identification. The simple quantification strategy, namely label-free quantification, provides high reproducibility performance without extra modifications on peptide sample. Chromatogram, a subspace of precursor mass spectrogram with the dimensions of retention time and intensity, provides information for measuring peptide and protein abundances (Figure 5). A typical way of quantification integrates signal intensities of a particular identified peptide over the retention time profile for area under the curve. The difficulties in quantification present when different peptides have both similar *m/z* and retention time profiles. Without any other features of mass spectrogram, the issue becomes mathematically underdetermined system, where information or equation are insufficient to solve the variables.

The most common freely available platforms for proteome quantification are Skyline[32] and MaxQuantLFQ[58], based on different modified statistical models. Skyline can build its own spectral library or import identification results from other searches, such as Mascot search engine. The built-in algorithm includes peak picking, ion intensity trapezoidal summation of peptide isotopic peaks, background subtraction, and scoring system before reporting total area and height parameter for each peptide. Skyline platform is well-used because of the compatibility with mass spectrograms from four major vendors and the graphical interface that allows users to explore and modify the results manually. MaxQuantLFQ obtains the identification result from Andromeda search engine[61] to calculate the area. Peptide peaks are fitted into Gaussian shape, smoothed, and extracted by significant local minima for integration. The algorithm applies a sophisticated intensity normalization procedure to report precise answers.

The limitation of the conventional methods is due to its workflow where identification is needed before quantification. However, the identification can be missed by several reasons such as a bias in selecting high-abundance precursors for fragmentation, a low ionization efficiency or a poor fragmentation process. Without the

product ions, the conventional methods skip to quantify these peptides even their peaks are prominent in precursor mass spectrograms.

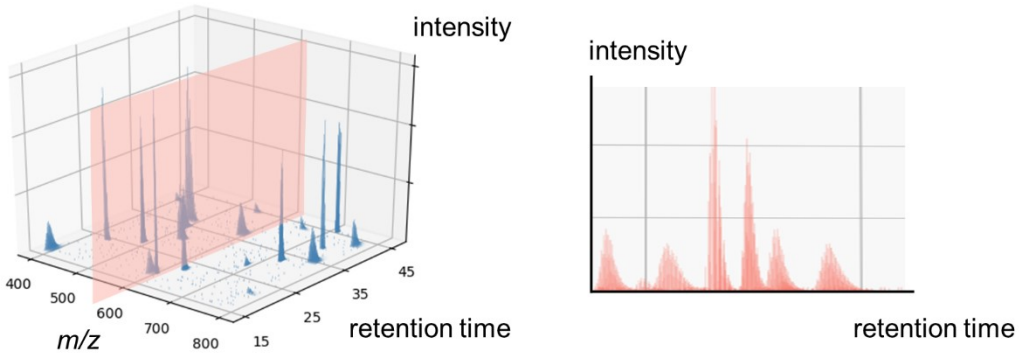


**Figure 5: Chromatogram, the subspace of mass spectrogram with dimensions of retention time and intensity, for peptide quantification**

**Machine learning approaches**

Machine learning has emerged as a tool to solve persistent problems in mass spectrogram analysis. Currently, a prominent implementation has been trending to the downstream analysis of proteomics data. Although, the gold-standard approach based on machine learning techniques for peak identification and quantification in proteomics mass spectrogram is not yet available, many new algorithms are investigated and applied to aid some steps in a pipeline of peptide identification and quantification. The goals of this section is not to present an exhaustive list of all available software tools but to focus on well-known applications for analyzing retention time, *m/z*, and intensity dimensions of mass spectrograms.

Retention time prediction algorithms has become a valuable source of prior knowledge. Predicted retention times can be applied as a filter to remove false positive and increase the accuracy of protein identification. The reliable model based on its extensive tests using a large number of real samples still depends on using properties of amino acid individually and collectively without machine learning techniques such as SSRCalc[62]. For machine learning, small-scale experiments under specific chromatographic conditions were studied using support vector machine[63,64] and neural network[65–67]

In order to reduce complexity of *m/z* dimension, isotopic clustering techniques are helpful as preprocessing step. There are many applications search the isotopic pattern of peptide peaks based on the rule of selected isotopic peptide peak properties[68–70]. Bayesian network was recently introduced to cluster of peaks into envelopes [71].

The prediction of intensity dimension of LC/MS/MS is highly challenging since the fragmentation process is not well understood. However, the predicted intensities can significantly improve peptide identification because they are required for generating the correct template for database search engine or defining the right rule for *de novo* sequencing. A probabilistic decision tree is the first successful technique to model the probability of observing a fragment intensity in mass spectrograms[72] and develop as a tool [73,39]. Deep learning has emerged because of the capability to understand complex data. The current applications are developed to increase the accuracy of prediction in various experimental conditions[74,75].

Another noteworthy implementation of machine learning is to enhance the quality of identification after database search. Percolator[76] is a well-used postprocessing algorithm aiming to validate the correct identification of a database search algorithm. A support vector machine is chosen to learn from correct and incorrect groups. The features, such as mass, length, and misclevage, of fragment mass spectra are evaluated to return a confident score of final identification results.

Challenges in mass spectrogram analysis have remained. Integrating appropriate machine learning techniques with maximizing the utilities of mass spectrogram features, i.e. both precursors and product ions, is promising to identify, quantify, and verify public, yet undiscovered, truth about proteome.

# Acknowledgement

I would like to thank Professor Yasushi Ishihama (Graduate School of Pharmaceutical Sciences, Kyoto University), my supervisor, for guiding me through these years. I appreciate you supporting me to do the project on machine learning, among the other things. I could not have come this far without your kindness and continued support. At the end, because of the extraordinary difficulties, I finally understand how extraordinarily fortunate I am to have been with the right supervisor for Ph.D.

I would like to thank Professor Kazuyoshi Yoshii (Graduate School of Informatics, Kyoto University), my co-advisor, for introducing me to machine learning and giving me constant advices. I appreciate your kind words and support very much.

I would also like to thank Professor Naoyuki Sugiyama, Professor Akiyasu Yoshizawa, Professor Koshi Imami, all past and current lab members for help and kindness.

A special thanks for Japanese Government Scholarship for supporting my Master and Ph.D.

Last but not the least, I would like to thank my friends and family, especially my parents and my sisters for love and support everything I have done in my life.

Always.

# References


(1) Blackstock, W. P.; Weir, M. P. Proteomics: Quantitative and Physical Mapping of Cellular Proteins. *Trends Biotechnol.* **1999**, *17* (3), 121–127. https://doi.org/10.1016/S0167-7799(98)01245-1.

(2) Nilsson, T.; Mann, M.; Aebersold, R.; Yates, J. R.; Bairoch, A.; Bergeron, J. J. M. Mass Spectrometry in High-Throughput Proteomics: Ready for the Big Time. *Nat. Methods* **2010**, *7* (9), 681–685. https://doi.org/10.1038/nmeth0910-681.

(3) Domon, B.; Aebersold, R. Challenges and Opportunities in Proteomics Data Analysis. *Mol. Cell. Proteomics MCP* **2006**, *5* (10), 1921–1926. https://doi.org/10.1074/mcp.R600012-MCP200.

(4) Smaragdis, P.; Brown, J. C. Non-Negative Matrix Factorization for Polyphonic Music Transcription. In *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*; 2003; pp 177–180. https://doi.org/10.1109/ASPAA.2003.1285860.

(5) Sun, D. L.; Mysore, G. J. Universal Speech Models for Speaker Independent Single Channel Source Separation. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*; 2013; pp 141–145. https://doi.org/10.1109/ICASSP.2013.6637625.

(6) Ueda, S.; Shibata, K.; Wada, Y.; Nishikimi, R.; Nakamura, E.; Yoshii, K. Bayesian Drum Transcription Based on Nonnegative Matrix Factor Decomposition with a Deep Score Prior. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; 2019; pp 456–460. https://doi.org/10.1109/ICASSP.2019.8683129.

(7) Nesvizhskii, A. I. Protein Identification by Tandem Mass Spectrometry and Sequence Database Searching. *Methods Mol. Biol. Clifton NJ* **2007**, *367*, 87–119. https://doi.org/10.1385/1-59745-275-0:87.

(8) Zhang, J.; Gonzalez, E.; Hestilow, T.; Haskins, W.; Huang, Y. Review of Peak Detection Algorithms in Liquid-Chromatography-Mass Spectrometry. *Curr. Genomics* **2009**, *10* (6), 388–401. https://doi.org/10.2174/138920209789177638.

(9) Liu, J.; Bell, A. W.; Bergeron, J. J. M.; Yanofsky, C. M.; Carrillo, B.; Beaudrie, C. E. H.; Kearney, R. E. Methods for Peptide Identification by Spectral Comparison. *Proteome Sci.* **2007**, *5*, 3. https://doi.org/10.1186/1477-5956-5-3.

(10) Michalski, A.; Cox, J.; Mann, M. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority Is Inaccessible to Data-Dependent LC-MS/MS. *J. Proteome Res.* **2011**, *10* (4), 1785–1793. https://doi.org/10.1021/pr101060v.

(11) Griss, J.; Perez-Riverol, Y.; Lewis, S.; Tabb, D. L.; Dianes, J. A.; del-Toro, N.; Rurik, M.; Walzer, M. W.; Kohlbacher, O.; Hermjakob, H.; Wang, R.; Vizcaíno, J. A. Recognizing Millions of Consistently Unidentified Spectra across Hundreds of Shotgun

Proteomics Datasets. *Nat. Methods* **2016**, *13* (8), 651–656. https://doi.org/10.1038/nmeth.3902.

(12) Cook, D. W.; Rutan, S. C. Analysis of Liquid Chromatography-Mass Spectrometry Data with an Elastic Net Multivariate Curve Resolution Strategy for Sparse Spectral Recovery. *Anal. Chem.* **2017**, *89* (16), 8405–8412. https://doi.org/10.1021/acs.analchem.7b01832.

(13) Hoyer, P. O. Non-Negative Matrix Factorization with Sparseness Constraints. *J Mach Learn Res* **2004**, *5*, 1457–1469.

(14) Lee, D. D.; Seung, H. S. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature* **1999**, *401* (6755), 788–791. https://doi.org/10.1038/44565.

(15) Rapin, J.; Souloumiac, A.; Bobin, J.; Larue, A.; Junot, C.; Ouethrani, M.; Starck, J.-L. Application of Non-Negative Matrix Factorization to LC/MS Data, 2015.

(16) Lee, D. D.; Seung, H. S. Algorithms for Non-Negative Matrix Factorization. *Adv. Neural Inf. Process. Syst.* 556–562.

(17) Févotte, C.; Vincent, E.; Ozerov, A. Single-Channel Audio Source Separation with NMF: Divergences, Constraints and Algorithms. In *Audio Source Separation*; Makino, S., Ed.; Springer International Publishing: Cham, 2018; pp 1–24. https://doi.org/10.1007/978-3-319-73031-8_1.

(18) Pauca, V. P.; Piper, J.; Plemmons, R. J. Nonnegative Matrix Factorization for Spectral Data Analysis. *Linear Algebra Its Appl.* **2006**, *416* (1), 29–47. https://doi.org/10.1016/j.laa.2005.06.025.

(19) Wilson, K. W.; Raj, B.; Smaragdis, P.; Divakaran, A. Speech Denoising Using Nonnegative Matrix Factorization with Priors. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*; 2008; pp 4029–4032. https://doi.org/10.1109/ICASSP.2008.4518538.

(20) Perry, R. H.; Cooks, R. G.; Noll, R. J. Orbitrap Mass Spectrometry: Instrumentation, Ion Motion and Applications. *Mass Spectrom. Rev.* **2008**, *27* (6), 661–699. https://doi.org/10.1002/mas.20186.

(21) Hodge, K.; Have, S. T.; Hutton, L.; Lamond, A. I. Cleaning up the Masses: Exclusion Lists to Reduce Contamination with HPLC-MS/MS. *J. Proteomics* **2013**, *88*, 92–103. https://doi.org/10.1016/j.jprot.2013.02.023.

(22) Kim, J.; Monteiro, R. D. C.; Park, H. Group Sparsity in Nonnegative Matrix Factorization. In *Proceedings of the 2012 SIAM International Conference on Data Mining*; Society for Industrial and Applied Mathematics, 2012; pp 851–862. https://doi.org/10.1137/1.9781611972825.73.

(23) Qian, Y.; Jia, S.; Zhou, J.; Robles-Kelly, A. Hyperspectral Unmixing via L1/2 Sparsity-Constrained Nonnegative Matrix Factorization. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49* (11), 4282–4297. https://doi.org/10.1109/TGRS.2011.2144605.

(24) Eng, J. K.; McCormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am.*

*Soc. Mass Spectrom.* **1994**, *5* (11), 976–989. https://doi.org/10.1016/1044-0305(94)80016-2.

(25) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and Validation of a Spectral Library Searching Method for Peptide Identification from MS/MS. *Proteomics* **2007**, *7* (5), 655–667. https://doi.org/10.1002/pmic.200600625.

(26) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M.-Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nat. Biotechnol.* **2012**, *30* (10), 918–920. https://doi.org/10.1038/nbt.2377.

(27) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *Electrophoresis* **1999**, *20* (18), 3551–3567. https://doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.

(28) Okuda, S.; Watanabe, Y.; Moriya, Y.; Kawano, S.; Yamamoto, T.; Matsumoto, M.; Takami, T.; Kobayashi, D.; Araki, N.; Yoshizawa, A. C.; Tabata, T.; Sugiyama, N.; Goto, S.; Ishihama, Y. JPOSTrepo: An International Standard Data Repository for Proteomes. *Nucleic Acids Res.* **2017**, *45* (D1), D1107–D1111. https://doi.org/10.1093/nar/gkw1080.

(29) Iwasaki, M.; Miwa, S.; Ikegami, T.; Tomita, M.; Tanaka, N.; Ishihama, Y. One-Dimensional Capillary Liquid Chromatographic Separation Coupled with Tandem Mass Spectrometry Unveils the Escherichia Coli Proteome on a Microarray Scale. *Anal. Chem.* **2010**, *82* (7), 2616–2620. https://doi.org/10.1021/ac100343q.

(30) Rappsilber, J.; Mann, M.; Ishihama, Y. Protocol for Micro-Purification, Enrichment, Pre-Fractionation and Storage of Peptides for Proteomics Using StageTips. *Nat. Protoc.* **2007**, *2* (8), 1896–1906. https://doi.org/10.1038/nprot.2007.261.

(31) Masuda, T.; Saito, N.; Tomita, M.; Ishihama, Y. Unbiased Quantitation of Escherichia Coli Membrane Proteome Using Phase Transfer Surfactants. *Mol. Cell. Proteomics MCP* **2009**, *8* (12), 2770–2777. https://doi.org/10.1074/mcp.M900240-MCP200.

(32) MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J. Skyline: An Open Source Document Editor for Creating and Analyzing Targeted Proteomics Experiments. *Bioinforma. Oxf. Engl.* **2010**, *26* (7), 966–968. https://doi.org/10.1093/bioinformatics/btq054.

(33) Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C. Using Annotated Peptide Mass Spectrum Libraries for Protein Identification. *J. Proteome Res.* **2006**, *5* (8), 1843–1849. https://doi.org/10.1021/pr0602085.

(34) Frewen, B. E.; Merrihew, G. E.; Wu, C. C.; Noble, W. S.; MacCoss, M. J. Analysis of Peptide MS/MS Spectra from Large-Scale Proteomics Experiments Using Spectrum Libraries. *Anal. Chem.* **2006**, *78* (16), 5678–5684. https://doi.org/10.1021/ac060279n.

(35) Deutsch, E. W.; Perez-Riverol, Y.; Chalkley, R. J.; Wilhelm, M.; Tate, S.; Sachsenberg, T.; Walzer, M.; Käll, L.; Delanghe, B.; Böcker, S.; Schymanski, E. L.; Wilmes, P.; Dorfer, V.; Kuster, B.; Volders, P.-J.; Jehmlich, N.; Vissers, J. P. C.; Wolan, D. W.; Wang, A. Y.; Mendoza, L.; Shofstahl, J.; Dowsey, A. W.; Griss, J.; Salek, R. M.; Neumann, S.; Binz, P.-A.; Lam, H.; Vizcaíno, J. A.; Bandeira, N.; Röst, H. Expanding the Use of Spectral Libraries in Proteomics. *J. Proteome Res.* **2018**, *17* (12), 4051–4060. https://doi.org/10.1021/acs.jproteome.8b00485.

(36) Oliphant, T. E. Python for Scientific Computing. *Comput. Sci. Eng.* **2007**, *9* (3), 10–20. https://doi.org/10.1109/MCSE.2007.58.

(37) Zhang, Z. Prediction of Low-Energy Collision-Induced Dissociation Spectra of Peptides with Three or More Charges. *Anal. Chem.* **2005**, *77* (19), 6364–6373. https://doi.org/10.1021/ac050857k.

(38) Frank, A. M. Predicting Intensity Ranks of Peptide Fragment Ions. *J. Proteome Res.* **2009**, *8* (5), 2226–2240. https://doi.org/10.1021/pr800677f.

(39) Degroeve, S.; Maddelein, D.; Martens, L. MS2PIP Prediction Server: Compute and Visualize MS2 Peak Intensity Predictions for CID and HCD Fragmentation. *Nucleic Acids Res.* **2015**, *43* (W1), W326-330. https://doi.org/10.1093/nar/gkv542.

(40) Renard, B. Y.; Kirchner, M.; Monigatti, F.; Ivanov, A. R.; Rappsilber, J.; Winter, D.; Steen, J. A. J.; Hamprecht, F. A.; Steen, H. When Less Can Yield More - Computational Preprocessing of MS/MS Spectra for Peptide Identification. *Proteomics* **2009**, *9* (21), 4978–4984. https://doi.org/10.1002/pmic.200900326.

(41) Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nat. Methods* **2007**, *4* (3), 207–214. https://doi.org/10.1038/nmeth1019.

(42) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *J. Proteome Res.* **2008**, *7* (1), 29–34. https://doi.org/10.1021/pr700600n.

(43) Masuda, T.; Sugiyama, N.; Tomita, M.; Ohtsuki, S.; Ishihama, Y. Mass Spectrometry-Compatible Subcellular Fractionation for Proteomics. *J. Proteome Res.* **2020**, *19* (1), 75–84. https://doi.org/10.1021/acs.jproteome.9b00347.

(44) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.-C.; Yates, J. R. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chem. Rev.* **2013**, *113* (4), 2343–2394. https://doi.org/10.1021/cr3003533.

(45) Listgarten, J.; Emili, A. Statistical and Computational Methods for Comparative Proteomic Profiling Using Liquid Chromatography-Tandem Mass Spectrometry. *Mol. Cell. Proteomics MCP* **2005**, *4* (4), 419–434. https://doi.org/10.1074/mcp.R500005-MCP200.

(46) Tsai, T.-H.; Wang, M.; Ressom, H. W. Preprocessing and Analysis of LC-MS-Based Proteomic Data. *Methods Mol. Biol. Clifton NJ* **2016**, *1362*, 63–76. https://doi.org/10.1007/978-1-4939-3106-4_3.

(47) Du, P.; Kibbe, W. A.; Lin, S. M. Improved Peak Detection in Mass Spectrum by Incorporating Continuous Wavelet Transform-Based Pattern Matching. *Bioinformatics* **2006**, *22* (17), 2059–2065. https://doi.org/10.1093/bioinformatics/btl355.

(48) Bauer, C.; Cramer, R.; Schuchhardt, J. Evaluation of Peak-Picking Algorithms for Protein Mass Spectrometry. *Methods Mol. Biol. Clifton NJ* **2011**, *696*, 341–352. https://doi.org/10.1007/978-1-60761-987-1_22.

(49) Keller, B. O.; Sui, J.; Young, A. B.; Whittal, R. M. Interferences and Contaminants Encountered in Modern Mass Spectrometry. *Anal. Chim. Acta* **2008**, *627* (1), 71–81. https://doi.org/10.1016/j.aca.2008.04.043.

(50) Craig, R.; Beavis, R. C. TANDEM: Matching Proteins with Tandem Mass Spectra. *Bioinforma. Oxf. Engl.* **2004**, *20* (9), 1466–1467. https://doi.org/10.1093/bioinformatics/bth092.

(51) Kim, S.; Pevzner, P. A. MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. *Nat. Commun.* **2014**, *5* (1), 1–10. https://doi.org/10.1038/ncomms6277.

(52) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; Stein, S. E.; Aebersold, R. Building Consensus Spectral Libraries for Peptide Identification in Proteomics. *Nat. Methods* **2008**, *5* (10), 873–875. https://doi.org/10.1038/nmeth.1254.

(53) Yang, X.; Neta, P.; Stein, S. E. Quality Control for Building Libraries from Electrospray Ionization Tandem Mass Spectra. *Anal. Chem.* **2014**, *86* (13), 6393–6400. https://doi.org/10.1021/ac500711m.

(54) Yang, X.; Neta, P.; Stein, S. E. Extending a Tandem Mass Spectral Library to Include MS2 Spectra of Fragment Ions Produced In-Source and MSn Spectra. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (11), 2280–2287. https://doi.org/10.1007/s13361-017-1748-2.

(55) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: Powerful Software for Peptide de Novo Sequencing by Tandem Mass Spectrometry. *Rapid Commun. Mass Spectrom. RCM* **2003**, *17* (20), 2337–2342. https://doi.org/10.1002/rcm.1196.

(56) Frank, A.; Pevzner, P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Anal. Chem.* **2005**, *77* (4), 964–973. https://doi.org/10.1021/ac048788h.

(57) Fischer, B.; Roth, V.; Roos, F.; Grossmann, J.; Baginsky, S.; Widmayer, P.; Gruissem, W.; Buhmann, J. M. NovoHMM: A Hidden Markov Model for de Novo Peptide Sequencing. *Anal. Chem.* **2005**, *77* (22), 7265–7273. https://doi.org/10.1021/ac0508853.

(58) Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M. Accurate Proteome-Wide Label-Free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol. Cell. Proteomics MCP* **2014**, *13* (9), 2513–2526. https://doi.org/10.1074/mcp.M113.031591.

(59) Zhang, B.; Käll, L.; Zubarev, R. A. DeMix-Q: Quantification-Centered Data Processing Workflow. *Mol. Cell. Proteomics MCP* **2016**, *15* (4), 1467–1478. https://doi.org/10.1074/mcp.O115.055475.

(60) Serang, O.; Noble, W. A Review of Statistical Methods for Protein Identification Using Tandem Mass Spectrometry. *Stat. Interface* **2012**, *5* (1), 3–20.

(61) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **2011**, *10* (4), 1794–1805. https://doi.org/10.1021/pr101065j.

(62) Krokhin, O. V.; Ying, S.; Cortens, J. P.; Ghosh, D.; Spicer, V.; Ens, W.; Standing, K. G.; Beavis, R. C.; Wilkins, J. A. Use of Peptide Retention Time Prediction for Protein Identification by Off-Line Reversed-Phase HPLC-MALDI MS/MS. *Anal. Chem.* **2006**, *78* (17), 6265–6269. https://doi.org/10.1021/ac060251b.

(63) Pfeifer, N.; Leinenbach, A.; Huber, C. G.; Kohlbacher, O. Statistical Learning of Peptide Retention Behavior in Chromatographic Separations: A New Kernel-Based Approach for Computational Proteomics. *BMC Bioinformatics* **2007**, *8* (1), 468. https://doi.org/10.1186/1471-2105-8-468.

(64) Moruz, L.; Tomazela, D.; Käll, L. Training, Selection, and Robust Calibration of Retention Time Models for Targeted Proteomics. *J. Proteome Res.* **2010**, *9* (10), 5209–5216. https://doi.org/10.1021/pr1005058.

(65) Petritis, K.; Kangas, L. J.; Ferguson, P. L.; Anderson, G. A.; Paša-Tolić, L.; Lipton, M. S.; Auberry, K. J.; Strittmatter, E. F.; Shen, Y.; Zhao, R.; Smith, R. D. Use of Artificial Neural Networks for the Accurate Prediction of Peptide Liquid Chromatography Elution Times in Proteome Analyses. *Anal. Chem.* **2003**, *75* (5), 1039–1048. https://doi.org/10.1021/ac0205154.

(66) Ma, C.; Ren, Y.; Yang, J.; Ren, Z.; Yang, H.; Liu, S. Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. *Anal. Chem.* **2018**, *90* (18), 10881–10888. https://doi.org/10.1021/acs.analchem.8b02386.

(67) Guan, S.; Moran, M. F.; Ma, B. Prediction of LC-MS/MS Properties of Peptides from Sequence by Deep Learning. *Mol. Cell. Proteomics MCP* **2019**, *18* (10), 2099–2107. https://doi.org/10.1074/mcp.TIR119.001412.

(68) Wehofsky, M.; Hoffmann, R. Automated Deconvolution and Deisotoping of Electrospray Mass Spectra. *J. Mass Spectrom.* **2002**, *37* (2), 223–229. https://doi.org/10.1002/jms.278.

(69) Gutierrez, M.; Handy, K.; Smith, R. Quantitative Evaluation of Algorithms for Isotopic Envelope Extraction via Extracted Ion Chromatogram Clustering. *J. Proteome Res.* **2018**, *17* (11), 3774–3779. https://doi.org/10.1021/acs.jproteome.8b00451.

(70) Tay, A. P.; Liang, A.; Hamey, J. J.; Hart-Smith, G.; Wilkins, M. R. MS2-Deisotoper: A Tool for Deisotoping High-Resolution MS/MS Spectra in Normal and Heavy Isotope-Labelled Samples. *PROTEOMICS* **2019**, *19* (17), 1800444. https://doi.org/10.1002/pmic.201800444.

(71) Gutierrez, M.; Handy, K.; Smith, R. XNet: A Bayesian Approach to Extracted Ion Chromatogram Clustering for Precursor Mass Spectrometry Data. *J. Proteome Res.* **2019**, *18* (7), 2771–2778. https://doi.org/10.1021/acs.jproteome.9b00068.

(72) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. Intensity-Based Protein Identification by Machine Learning from a Library of Tandem Mass Spectra. *Nat. Biotechnol.* **2004**, *22* (2), 214–219. https://doi.org/10.1038/nbt930.

(73) Degroeve, S.; Martens, L. MS2PIP: A Tool for MS/MS Peak Intensity Prediction. *Bioinformatics* **2013**, *29* (24), 3199–3203. https://doi.org/10.1093/bioinformatics/btt544.

(74) Zhou, X.-X.; Zeng, W.-F.; Chi, H.; Luo, C.; Liu, C.; Zhan, J.; He, S.-M.; Zhang, Z. PDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Anal. Chem.* **2017**, *89* (23), 12690–12697. https://doi.org/10.1021/acs.analchem.7b02566.

(75) Lin, Y.-M.; Chen, C.-T.; Chang, J.-M. MS2CNN: Predicting MS/MS Spectrum Based on Protein Sequence Using Deep Convolutional Neural Networks. *BMC Genomics* **2019**, *20* (9), 906. https://doi.org/10.1186/s12864-019-6297-6.

(76) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets. *Nat. Methods* **2007**, *4* (11), 923–925. https://doi.org/10.1038/nmeth1113.