Doctoral Dissertation

Proposal of Estimation Methods for Constituents and Their Monoisotopic Masses in Mass Spectrometry

質量分析における成分とそのモノアイソトピック質量の推定手法の提案

Taichi Tomono

February 2025

Graduate School of Engineering

Osaka University

Preface

This dissertation presents my research on physical parameter estimation using sparse spectrum learning methods in Mass Spectrometry (MS), conducted during my Ph.D. studies at the Department of Information and Communications Technology, Graduate School of Engineering, Osaka University.

Mass spectrometry is a powerful analytical technique employed across various applications, including drug development, quality assurance, food inspection, and monitoring environmental pollutants. Recently, the production of antibodies and nucleic acid pharmaceuticals has led to the formation of impurities with various modifications. These impurities can adversely affect drug stability, pharmacokinetics, and efficacy, making it essential to accurately distinguish and quantify them. This dissertation focuses on estimating the number of constituents and their monoisotopic masses in mass spectrometry, addressing these critical issues. Traditional methods have proven insufficient for meeting these requirements.

The dissertation is structured as follows:

Chapter 1 outlines the background, motivation, and purpose of this research. Mass spectrometry is a versatile analytical technique used in drug development, quality assurance, food inspection, and environmental pollutant monitoring. Recent advancements in antibody and nucleic acid pharmaceuticals have led to the production of impurities that affect drug stability, pharmacokinetics, and efficacy, underscoring the importance of this research for pharmaceutical quality control.

Chapter 2 delves into modeling mass spectrometry and Bayesian inference to estimate the number of constituents and their monoisotopic masses from an MS spectrum. By modeling mass spectrometry for various constituent counts using parameters like

monoisotopic mass and ion counts, and employing Markov chain Monte Carlo methods (MCMC) to explore those parameters, we determine the optimal parameters and maximum posterior probabilities. The chapter discusses how we compare these probabilities across models to select the optimal constituent counts and estimate their properties.

Chapter 3 addresses challenges related to the vanishing gradient problem in sparse spectra with a high-speed parameter search method. Standard optimization techniques struggle with MS spectra's sparse and predominantly flat nature, which can lead to vanishing gradients. To overcome this, we refine our approach by blurring comparative spectra and gradually reducing the blur, thus enabling more accurate estimation without the extensive time demands of previous MCMC methods.

Chapter 4 integrates a hybrid mass spectrometry (MS/MS) system into the physical model, enhancing the accuracy of estimation. By incorporating additional MS/MS spectra, the model leverages more information, which improves parameter estimation accuracy and reduces mass errors.

Chapter 5 concludes the dissertation, summarizing the findings and their implications for future research and practical applications.

Acknowledgment

I conducted this dissertation under the guidance of Prof. Takashi Washio. I am deeply grateful to him for sharing his knowledge on how to proceed with research, how to communicate my work to others, and various aspects of information processing. His guidance was indispensable in shaping this dissertation. I also owe a great deal of gratitude to Prof. Satoshi Hara from the University of Electro-Communications, who regularly provided advice, especially from a mathematical perspective, and gave me invaluable ideas for solving the vanishing gradient problem.

Thanks are also due to my co-authors, Yusuke Nakai, who discussed estimation methods with me, and Kazuma Takahara, who advised on algorithm acceleration.

I am deeply grateful to Prof. Shohei Shimizu, who served as the chief examiner, and to Prof. Kazunori Komatani and Assoc. Prof. Ryu Takeda, who served as co-examiners, for their invaluable guidance and constructive feedback throughout the dissertation review process. Their comments and suggestions improved the quality of this work.

I would also like to extend my sincere appreciation to other professors in the Graduate School of Engineering, Osaka University, including Prof. Tetsuya Takine, Assoc. Prof. Yoshiaki Inoue, Prof. Akihiro Maruta, Assoc. Prof. Ken Mishina, Prof. Yuichi Tanaka, Assoc. Prof. Hiroshi Higashi, Prof. Hideki Ochiai, Prof. Atsuko Miyaji, and Prof. Kyo Inoue. I acknowledge their guidance and support during my research.

I am grateful for the support of Dr. Matthew J. Holland, Ms. Hiroko Okada, and other lab members. I would also like to express my gratitude to the members of Shimadzu Corporation. I would also like to thank Dr. Junko Iida and Masanobu Shiga, who created the opportunity for me to join Osaka University. I express my deep appreciation to my superiors, Yoshihiro Ueno and Yusuke Tagawa, who arranged for me to conduct my

research at Osaka University. I am also thankful to Daisuke Hiramatsu and Dr. Hiroaki Nakanishi for supporting my research.

Most of all, I want to express my utmost gratitude to my family, who supported me throughout my research life.

Contents

Abbreviations

Abbreviation	Full Term
MS	Mass Spectrometry
MS/MS	Hybrid Mass Spectrometry
$\boldsymbol{m/z}$	Mass-to-Charge Ratio
ESI	Electrospray Ionization
MALDI	Matrix-Assisted Laser Desorption/Ionization
CI	Chemical Ionization
TOF	Time-of-Flight
Q-TOF	Quadrupole Time-of-Flight
FT-ICR	Fourier Transform Ion Cyclotron Resonance
MAP	Maximum A Posteriori
MCMC	Markov Chain Monte Carlo
HMC	Hamiltonian Monte Carlo
NUTS	No-U-Turn Sampler
SVI	Stochastic Variational Inference
BIC	Bayesian Information Criterion
SAI	Spectral Annealing Inference
ELBO	Evidence Lower Bound
BIC	Bayesian Information Criterion
KL Divergence	Kullback-Leibler Divergence

Abbreviation	Full Term
Adam	Adaptive Moment Estimation
SGD	Stochastic Gradient Descent
PSF	Point Spread Function
PDF	Probability Density Function
SD	Standard Deviation
cos	Cosine
NMR	Nuclear Magnetic Resonance
XPS	X-ray Photoelectron Spectroscopy
XRD	X-ray Diffraction
XRF (EDX)	X-ray Fluorescence (Energy Dispersive X-ray Spectroscopy)
XAS	X-ray Absorption Spectroscopy

Chapter 1.

Introduction

1.1. Background

In the pharmaceutical industry, the development of antibody-based and nucleic acid-based therapeutics has rapidly accelerated in recent years. Alongside these advancements, a range of modified molecular species—commonly referred to as impurities—has emerged during manufacturing and formulation processes. These impurities can have a significant impact on the safety and effectiveness of pharmaceutical products, potentially altering their stability, pharmacokinetics, or biological activity [1]–[4]. Therefore, it is essential to detect and characterize these impurity profiles as part of rigorous quality control and assurance practices.

A fundamental part of this analysis involves understanding the molecular mass of constituents, particularly the monoisotopic mass, which refers to the exact mass of a molecule using the most abundant isotopes. This parameter is crucial for identifying subtle differences in molecular structure that may lead to impurity formation. Moreover, quantifying the ion concentrations of these molecular species provides insight into their relative abundance and possible influence on the drug product.

Mass spectrometry (MS) has become a key analytical technique in this context. It allows for both qualitative and quantitative examination of mixtures, helping to detect, identify, and quantify impurities with high sensitivity and specificity. As such, MS is extensively employed in drug development and production.

1.2. What is Mass Spectrometry?

Mass spectrometry is an analytical method that identifies and quantifies chemical substances by converting them into ions and measuring their mass-to-charge (m/z) ratios. The process typically follows a structured workflow: sample introduction and preparation, ionization of the analytes, separation of ions based on m/z in the mass analyzer, and finally, ion detection. Several ionization techniques are commonly used depending on the nature of the analyte and analytical goals. These include Electrospray Ionization (ESI), Matrix-Assisted Laser Desorption/Ionization (MALDI), and Chemical Ionization (CI).

After ionization, separator such as Time-of-Flight (TOF), quadrupole filters, or ion traps are employed to segregate ions according to their m/z values. The resulting output, known as a mass spectrum, displays peaks that correspond to different ions and their relative intensities. Through interpretation of these spectra, analysts can deduce the chemical structure and identity of constituents in the sample, enabling tasks such as identification of unknown substances, structural elucidation of compounds, and assessment of sample purity. Figure 1-1 illustrates a typical mass spectrometry setup.

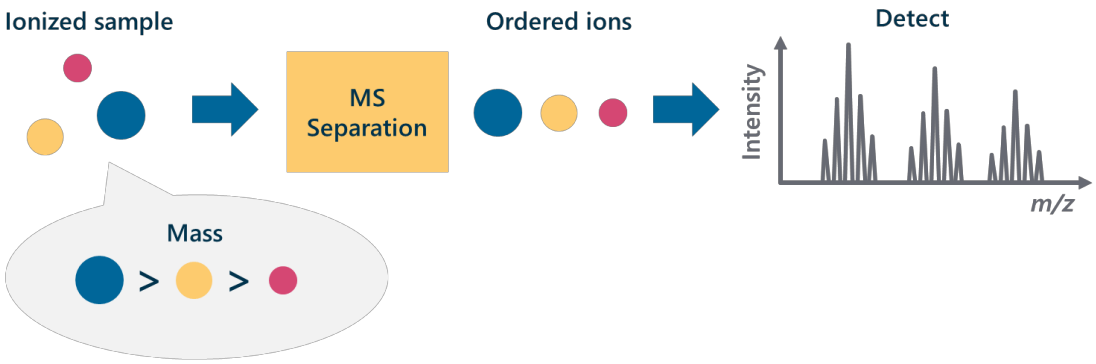

Figure 1-1. Schematic diagram of mass spectrometry.

Mass spectrometry is particularly valued for its sensitivity and specificity, making it indispensable in analytical laboratories. It can detect trace amounts of materials, which is essential for applications requiring the detection of very low concentrations of substances, such as in the detection of contaminants in food products or environmental samples. In pharmaceutical development, MS is used to confirm the identity of compounds, determine molecular structure, and assess the purity of the final product.

Advanced configurations, such as hybrid mass spectrometry (MS/MS), offer deeper insights by enabling structural analysis of ions. In this approach, precursor ions selected by the first analyzer (MS1) are fragmented within a collision cell, usually by interaction with an inert gas like argon. The resulting product ions are then analyzed in a second stage (MS2). This two-stage separation allows for detailed structural information that cannot be obtained through a single stage alone. Figure 1-2 depicts the layout of a typical MS/MS system.

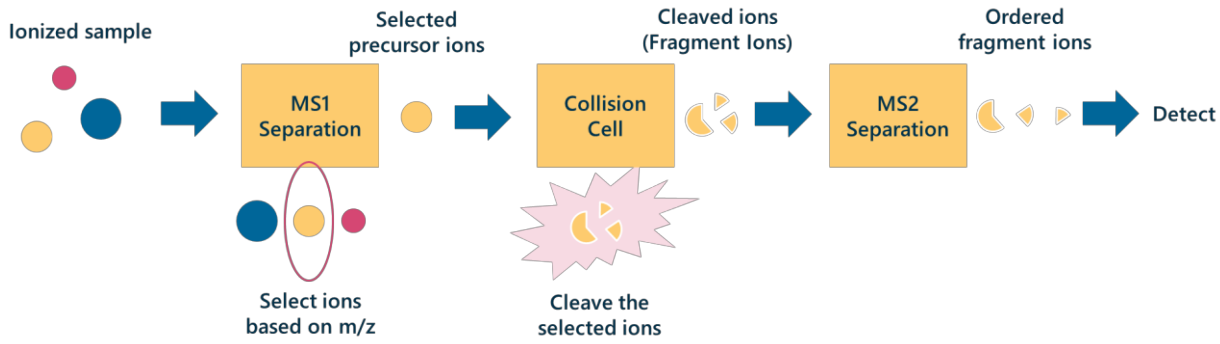

Figure 1-2. Schematic diagram of hybrid mass spectrometry (MS/MS).

The versatility of mass spectrometry makes it a powerful tool not only in scientific research but also in industries like biotechnology, environmental sciences, and forensic science. It plays a pivotal role in proteomics, metabolomics, and toxicology by providing precise molecular weight information and structural data. This allows researchers and

professionals to undertake a wide range of tasks from basic biological research to complex clinical diagnostics and therapeutic monitoring, offering invaluable insights into the molecular mechanisms of diseases and the effects of therapeutic interventions.

1.3. Issues in Mass Spectrometry

In current mass spectrometry practices, it remains a significant challenge to accurately detect and characterize impurities that occur in medium- to high-molecular-weight substances, especially when these impurities arise from subtle chemical modifications. These molecules—such as proteins or large nucleic acid chains—often undergo slight changes during synthesis or storage, resulting in forms that are chemically similar to the desired product but may still influence its behavior or efficacy. Conventional separation techniques, like chromatography, which aim to isolate individual constituents based on their chemical properties, often fall short in distinguishing these nearly identical impurities [5].

Furthermore, even when using mass spectrometry itself, it becomes increasingly difficult to resolve such impurities due to the complexity of the resulting spectra. One contributing factor is the presence of isotopic variants—molecules that differ only in the natural isotopes of their atoms—which produce overlapping signals. Another complicating factor is the formation of multivalent ions, especially common in techniques like electrospray ionization (ESI), where a single molecule carries multiple electric charges. These multicharged states generate numerous peaks for each species, further crowding the mass spectrum and making it hard to distinguish individual constituents.

High-resolution mass spectrometers, such as those utilizing Fourier Transform Ion Cyclotron Resonance (FT-ICR) [6]–[8], can resolve minute differences in m/z, but these instruments are typically expensive and bulky, restricting their use to specialized

laboratories. More commonly, laboratories rely on instruments like Triple Quadrupole MS and Quadrupole Time-of-Flight MS (Q-TOF-MS), which, while practical and accessible, may lack the resolution needed for distinguishing isomeric or closely related impurities. As a result, analytical software plays a crucial role in augmenting mass spectrometric data interpretation.

Various software solutions have been developed to extract detailed mass information from complex spectra. For instance, algorithms that perform wavelet-based spectral analysis [9] can generate peak lists from raw data. However, when analyzing spectra of medium to high-molecular-weight compounds ionized by methods like ESI [10]–[12], the interpretation becomes more difficult. ESI often produces ions with multiple charges, which broadens the isotopic distribution and complicates the identification of monoisotopic peaks. Simple peak-picking techniques often fall short in these scenarios [13].

1.4. Related Works

To address these challenges, researchers have developed various algorithms for deconvoluting charge states and deisotoping multivalent spectra. One such approach involves fitting Gaussian models to observed peaks using nonlinear least squares methods [14]. Charge deconvolution is the process of determining the neutral mass of an ion from its various charged forms, which is essential for accurately interpreting complex mass spectra.

The ReSpect algorithm, which employs a Maximum Entropy strategy [15], has seen widespread use [16]–[18]. It estimates m/z values by applying statistical constraints on the charge distribution to identify the most likely monoisotopic masses. However, this method does not explicitly estimate the number of unique molecular species (denoted as

k), nor does it evaluate discrete likelihoods, such as the probability of observing k versus $k + 1$ constituents. Additionally, as the complexity of a spectrum increases, the entropy term in the optimization function may cause the algorithm to overfit, resulting in overestimation of constituent numbers and inaccuracies in both monoisotopic mass and ion count predictions [19].

Recently, Bayesian approaches such as UniDec have been introduced to improve deconvolution performance [20], [21]. UniDec, inspired by the Richardson-Lucy deconvolution algorithm [22], [23], offers faster performance than ReSpect. Nonetheless, it too encounters limitations when it comes to evaluating the likelihood of a specific number of constituents within the spectrum.

1.5. Purpose and Direction of Our Research

The primary goal of our research is to evaluate the probability of constituent counts from spectral data analyzed using Mass Spectrometry (MS), and to determine optimal physical parameters such as monoisotopic masses. This is crucial for detecting and analyzing impurities in the manufacture and development of pharmaceuticals.

We use Bayesian inference to leverage prior knowledge. This enables probabilistic evaluation and accurate estimation of physical parameters. Additionally, by modeling for each possible number of constituents, we become able to evaluate discrete probabilities that means which number of constituents is optimal. Figure 1-3 shows the overview of our analysis method.

First, prior knowledge, such as parameter ranges and probability distributions, is incorporated to explore the parameter space using Markov Chain Monte Carlo (MCMC) or Stochastic Variational Inference (SVI). MCMC is a probabilistic sampling method that

generates samples from a posterior distribution by constructing a Markov chain. This allows us to approximate posterior probabilities even in high-dimensional parameter spaces. SVI, on the other hand, is a deterministic approach for approximating posterior distributions. It optimizes variational parameters by iteratively minimizing the Kullback-Leibler divergence between the true posterior and an approximating distribution.

Proposal parameters are input into the physical model of mass spectrometry, which then generates estimated spectra. By comparing these with observed spectra, the likelihood of the parameters is obtained. This process is repeated to derive subsequent parameters from this likelihood and prior knowledge. The log-likelihood is utilized in MCMC to calculate the acceptance probability for the next sampling and in SVI to compute the objective function to be minimized.

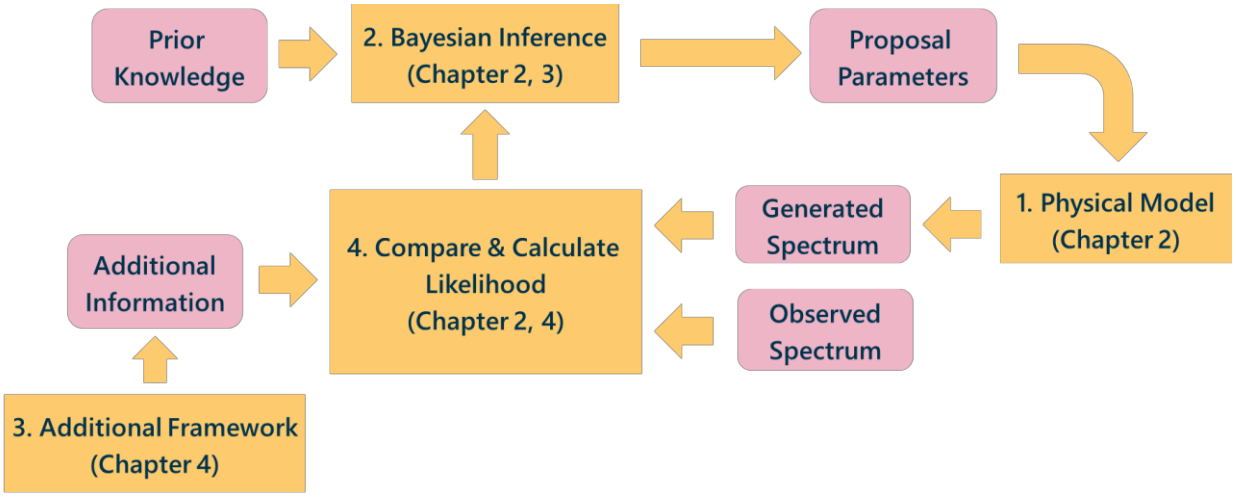

Figure 1-3. Overview of analysis method.

After performing this process for models corresponding to possible numbers of constituents, we compare their posterior probabilities to determine the most plausible number of constituents and their parameters.

1.6. Technical Issues Tackled in This Thesis

To realize this approach, we must address several technical challenges as follows:

1. Building Physical Models: it is necessary to model the mass spectrometry system using parameters such as the monoisotopic mass.
2. Exploring Sparse Posterior Probability Space: The posterior probability of the monoisotopic mass exhibits multiple steep peaks and is locally abrupt, presenting a significant challenge in how to explore this sparse parameter space. Such a parameter space can induce the vanishing gradient problem, making simple gradient-based methods of parameter exploration inappropriate.
3. Enhancement of Information Quantity: To improve accuracy, it is crucial to integrate information beyond the MS spectra, leveraging complementary data sources such as MS/MS spectra.
4. Establishing Appropriate Likelihood Estimation Methods: Developing accurate methods for estimating the likelihood of MS spectra is necessary to ensure reliable parameter estimation.

1.7. Summary of Contributions

This dissertation addresses the technical challenges outlined in Section 1.6 through the following contributions:

In Chapter 2, we addressed Technical Issue 1 (Building Physical Models) by constructing a mass spectrometry model based on Bayesian inference. This model incorporates parameters such as monoisotopic mass to estimate the number of constituents and their identities from MS spectra, leveraging prior knowledge. Additionally, we initiated parameter exploration in sparse posterior probability spaces using MCMC, partially addressing Technical Issue 2 (Exploring Sparse Posterior Probability Spaces). Chapter 2 is based on Tomono, Hara, Nakai, Takahara, Iida and Washio (2023a) [24].

In Chapter 3, to solve Technical Issue 2 (Exploring Sparse Posterior Probability Spaces), we developed a faster parameter exploration technique that mitigates the vanishing gradient problem. This was achieved by integrating gradient-based methods with techniques tailored to handle sparsity in the parameter space. Chapter 3 is based on Tomono, Hara, Iida and Washio (2024b) [25].

In Chapter 4, in response to Technical Issue 3 (Enhancement of Information Utilization), we incorporated additional data, such as MS/MS spectral information, to refine the accuracy of the analysis. Additionally, we tackled Technical Issue 4 (Establishing Appropriate Likelihood Estimation Methods) by improving the methods for likelihood estimation, enhancing the precision of determining the number of constituents and their monoisotopic masses. Chapter 4 is based on Tomono, Hara, Iida and Washio (2024c, 2024d) [26], [27].

Chapter 2.

Study on Estimating the Number of Constituents and Their Identities from MS Spectrum

2.1. Overview

Chapter 2 is based on Tomono, Hara, Nakai, Takahara, Iida and Washio (2023a) [24]. In this chapter, we newly propose a method to select the optimal number of constituents by comparing the probability of each constituent count, and to estimate the monoisotopic mass and ion counts under that condition. This can suggest the presence of impurities in pharmaceuticals, assist in the search for better synthesis conditions for middle to high molecular pharmaceuticals, and be useful for quality assurance in factories.

MS spectra are determined by the m/z (mass-to-charge) ratio and intensity axis. Essentially, MS spectra are defined by the ion quantities, monoisotopic mass, isotopic distribution, charge distribution of each constituent in a sample, and the detector's response. The detector's response is known, so by modeling the MS spectrum from parameters that dictate ion quantities, monoisotopic mass, isotopic distribution, and charge distribution, and fitting these models to the observed spectrum, we can accurately estimate the monoisotopic mass and ion quantities.

First, we model the mass spectrometry system based on parameters like the mass and charge of each constituent, assuming a certain number of constituents in a sample. Here, a constituent is defined as a substance with a specific monoisotopic mass. We then perform a MAP (Maximum A Posteriori) estimation of these parameters from the observed spectrum. By comparing the maximum posterior probability in models with

different numbers of constituents, we determine the model with the most appropriate number of constituents.

However, this model has a large dimensionality of the number of constituents multiplied by 6, where 6 represents the number of parameters per constituent in the physical model. Moreover, the posterior probability for one of the parameters, the monoisotopic mass, is flat over a large portion of the search space and has several sharp peaks locally. Hence, gradient-based methods are not suitable for this case due to anticipated gradient vanishing. Figure 2-1 is a schematic diagram of this issue. When the spectrum is sparse, changes in parameters do not affect the posterior probability of the spectrum. This leads to the problem of vanishing gradients.

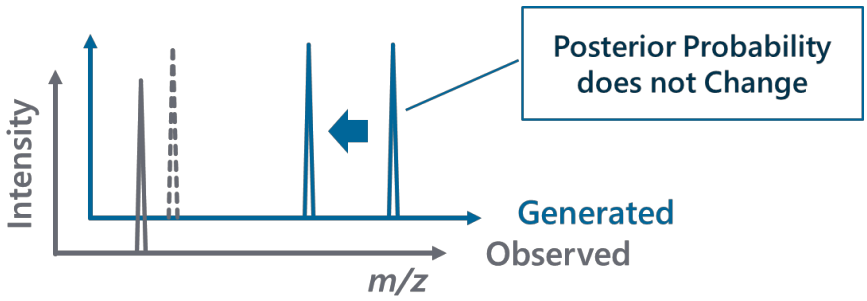

Figure 2-1. Schematic diagram of the vanishing gradient problem in sparse spectra.

Therefore, to estimate the parameters, we combine the No-U-Turn Sampler (NUTS [28]), a type of Markov Chain Monte Carlo (MCMC), with Simulated Annealing [29]. The purpose of using Simulated Annealing is to introduce a temperature parameter. By selecting a high-temperature exploration parameter distribution, we can actively explore parameters even in areas where the posterior probability is flat or has sharp peaks. This ensures a broader search across the parameter space, reducing the chance of overlooking

the global solution and getting trapped in local minima.

Furthermore, NUTS can explore parameters sparsely in areas with small gradients and can explore parameters in detail in areas with large gradients. Thus, introducing NUTS allows efficient exploration of the vast, high-dimensional parameter space.

On the other hand, while MCMC is good at searching for global solutions, it does not always reach the optimal solution within a certain number of search steps. Therefore, we use the parameters with the highest posterior probabilities obtained from NUTS and Simulated Annealing as initial values and apply stochastic variational inference. By doing this, we search for the optimal parameter where the posterior probability is maximized in the vicinity of that initial value, aiming to improve the accuracy of parameter estimation.

However, simultaneously searching for parameters for all possible numbers of constituents leads to a curse of dimensionality, where the search space explosively expands as the number of constituents increases, potentially reducing search efficiency and accuracy. To avoid this problem, we sequentially increase the number of constituents from $k = 1$ to the maximum conceivable number $k = k_{max}$. The value of k_{max} is determined based on prior knowledge, such as the expected complexity of the sample or physical constraints. For k constituents calculate the optimal parameters and their posterior probabilities. These posterior probabilities are then used to efficiently focus the parameter search areas for the $k + 1$ constituents.

To balance the complexity of the model (number of constituents) and its fit (loss against the data), in addition to the prior distribution of each parameter, we introduce a prior distribution for the number of constituents. We also incorporate a prior distribution on the differences between the monoisotopic masses of multiple constituents. For analytical purposes, we have defined a single constituent as a substance with a distinct

monoisotopic mass, thereby ensuring that their masses don't mutually take the same value. When seeking to separate isomers, it is essential to integrate other techniques such as fragmentation, ion mobility spectrometry, and chromatography, in addition to the proposed method. We first construct a model with $k = 1$ constituent, obtain the optimal parameters and the maximum posterior probability based on the above prior distributions and observed data.

Next, we construct a model with $k = 2$ constituents. For one of the two constituents, we use a prior distribution centered on the optimal parameters already estimated for $k = 1$, narrowing its range. This suppresses the significant increase in the parameter search space. Based on this new prior distribution, we estimate the optimal parameters and obtain the maximum posterior probability.

Subsequently, we seek the maximum posterior probability for each model with constituent numbers up to the upper limit k_{max} by efficiently exploring the optimal parameters in the same manner.

Finally, we compare the maximum posterior probabilities corresponding to each model with different numbers of constituents. We select the model with the highest probability and obtain the estimates for the monoisotopic masses and ion counts.

The analytical workflow is shown in Figure 2-2. The input is the MS Spectrum. The outputs of estimation are the number of constituents in the analyte, their monoisotopic masses, and ion quantities.

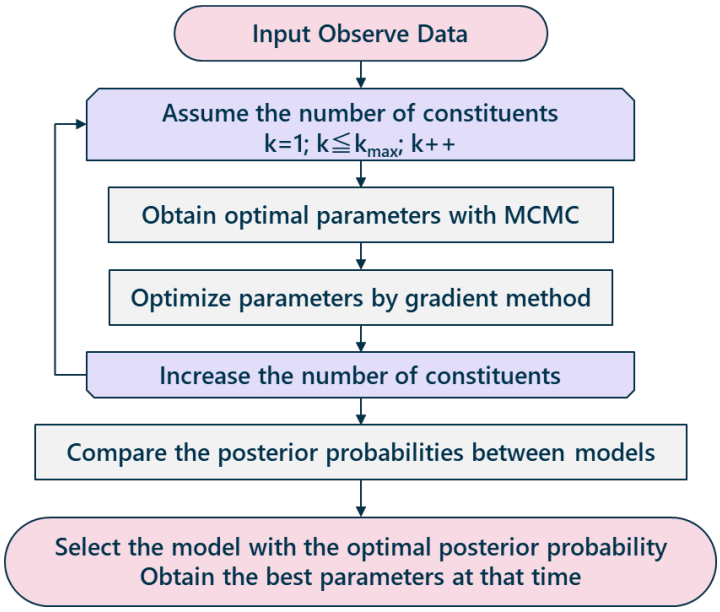

Figure 2-2. Schematic diagram of analytical workflow.

2.2. Proposed Method

2.2.1. Physically Modeling MS

The spectrum in mass spectrometry is composed of two primary axes: the mass-to-charge (m/z) axis and the intensity axis. The spectra are determined by the distribution of sample mass and charge. Specifically, the mass $p_j(m)$ and charge $q_j(z)$ distributions for each constituent are defined as follows:

$$p_j(m) = \sum_{\{M_j\}} \left[\delta\left(m - \sum_{j_\chi=1}^{n_j} m_{j_\chi} \right) \prod_{j_\chi=1}^{n_j} u_{j_\chi} \right], \tag{1}$$

$$q_j(z) = \sum_{\{Q_j\}} \left[\delta\left(z - \sum_{j_b=1}^{l_j} q_{j_b} \right) \prod_{j_b=1}^{l_j} v_{j_b} \right], \tag{2}$$

m: $a\ variable\ in\ the\ mass\ space\ where\ m \geq 0$,

z: $a\ variable\ representing\ the\ absolute\ value\ of\ charge,$

$where\ z \geq 1\ and\ z\ is\ an\ integer,$

j: $constituent\ IDs\ (j = 1,2,\cdots,k)$,

k: $number\ of\ constituents\ in\ the\ sample$,

n_j: $total\ number\ of\ atoms\ in\ constituent\ j$,

j_χ: $index\ of\ atoms\ in\ constituent\ j$,

m_{j_χ}: $mass\ of\ atom\ j_\chi$,

M_j: $vector\ of\ masses\ for\ atoms\ \left(m_{j_1}, m_{j_2}, \dots, m_{j_{n_j}}\right)$,

u_{j_χ}: $natural\ isotopic\ abundance\ of\ atom\ j_\chi$,

l_j: $total\ number\ of\ chargeable\ sites\ in\ constituent\ j$,

j_b: $index\ of\ chargeable\ sites\ in\ constituent\ j$,

q_{j_b}: $charge\ of\ chargable\ site\ j_b$,

Q_j: $vector\ of\ charges\ for\ chargable\ sites\ \left(q_{j_1}, q_{j_2}, \dots, q_{j_{l_j}}\right)$,

v_{j_b}: $probability\ that\ the\ chargeable\ site\ j_b\ attains\ its\ charge\ q_{j_b}$, and

δ: $Kronecker\ delta\ function.$

The number of parameters in this model, which are based on the count of elements and chargeable sites, makes practical computation and search unfeasible due to their high count. To manage this, we approximate isotope and charge distributions using a binomial distribution, which simplifies the complexity of the model and ensures that mass spectrometry analysis remains computationally feasible.

The spectrum in mass spectrometry can be approximated using the following model [24]. The probability distribution of mass of constituent j can be described by a binomial distribution $\tilde{p}_j(\omega_j)$. Here, $\omega_j = \text{round}\left(\frac{m-m_j'}{\varepsilon}\right)$ is the increase in neutron number from the monoisotopic ions of constituent j, where m_j' represents the monoisotopic mass of constituent j. m represents a variable in the mass space, and $m \geq 0$. ε represents the mass of neutron, 1.008664 Da. We postulate $\omega_j \geq 0$, because, in the biochemical domain, the most abundant isotope is usually also the lightest. In this model, we assume that n_j atoms within a molecule can be replaced by isotopes with a mass increase of ε Da at a probability of u_j. Additionally, for the charge distribution $\tilde{q}_j(z)$, we assume that l_j chargeable sites can acquire a charge of +1 (in the case the mass spectrometry system is in positive mode) at a charge rate of v_j. z denotes the variable representing the absolute value of charge, where $z \geq 1$ and z is an integer.

The mathematical expressions of the distributions generated by these binominal processes are:

$$\tilde{p}_j(\omega_j) = \begin{cases} \binom{n_j}{\omega_j} {u_j}^{\omega_j}\left(1-u_j\right)^{n_j-\omega_j} & \text{for } \omega_j \geq 0, \\ 0 & \text{otherwise, and} \end{cases} \tag{3}$$

$$\tilde{q}_j(z) = \binom{l_j}{z} {v_j}^{z}\left(1-v_j\right)^{l_j-z}. \tag{4}$$

Here,

u_j: $isotopic\ replacing\ rate\ of\ constituent\ j$,

v_j: $charge\ rate\ of\ chargeable\ sites\ of\ constituent\ j$, and

ε: $the\ mass\ of\ a\ neutron$.

Typically, the spectrum obtained from a mass spectrometer is represented along the mass-to-charge ratio m/z axis. Here, we define φ as the variable representing m/z. The total number of ions belonging to a set, *i.e.*, a constituent j, is denoted by I_j. Each ion in the set is indexed by i_j. The mass and charge of each individual ion i_j are denoted as $\omega_{i_j} \sim \tilde{p}_j$ and $z_{i_j} \sim \tilde{q}_j$. When an ion i_j is detected, its observed ideal spectrum would be $\delta\left(\varphi - \left(m_j' + \varepsilon\omega_{i_j}\right)/z_{i_j}\right)$ where δ is Kronecker delta function. Regardless of its charge state or mass, a single ion contributes to the observed spectrum as a single delta function. Therefore, the ideal spectrum formed by this set of ions (from $i_j = 1$ to I_j), $D_j(\varphi)$, can be represented as shown in equation (5). In this equation, φ is a variable representing the mass-to-charge ratio, and δ denotes the Kronecker delta function.

$$D_j(\varphi) = \sum_{i_j=1}^{I_j} \delta\left(\varphi - \left(m_j' + \varepsilon\omega_{i_j}\right)/z_{i_j}\right). \tag{5}$$

The theoretical probability distribution $U_j(\varphi)$ of the ions belonging to constituent j on the φ axis is determined solely by ω_j and z, which are mutually independent. Their independence comes from the facts that ω_j is a function of m, and a chemical property z is hardly affected by the isotope mass m. Accordingly, $U_j(\varphi)$ is obtained by summing the product of the probabilities of ω_j, the probabilities of z, and the Kronecker delta function $\delta\left(\varphi - \left(m_j' + \varepsilon\omega_j\right)/z\right)$ over all ω_j and z as follows.

$$U_j(\varphi) = \sum_{z=1}^{\infty} \sum_{\omega_j=1}^{\infty} \tilde{p}_j(\omega_j) \cdot \tilde{q}_j(z) \cdot \delta\left(\varphi - \left(m_j' + \varepsilon\omega_j\right)/z\right). \tag{6}$$

As previously stated, regardless of its charge state or mass, a single ion contributes as a single delta function. Therefore, the observed spectrum of ions is proportional to the

probability distribution of ions along the φ axis. According to the Glivenko-Cantelli Theorem [30], [31], the empirical spectrum $D_j(\varphi)$ converges uniformly to the theoretical distribution $U_j(\varphi)$ as sample size increases as far as our physical assumptions argued in the former explanation is valid. Therefore, the ideal spectrum of constituent j, $D_j(\varphi)$, can be approximated by $U_j(\varphi)$ as shown in equation (7).

$$D_j(\varphi) = \sum_{i_j=1}^{I_j} \delta\left(\varphi - \left(m_j' + \varepsilon\omega_{i_j}\right)/z_{i_j}\right) \approx I_j \cdot U_j(\varphi) \quad (I_j \gg 1). \tag{7}$$

Due to the point spread of the detector's response $R(\varphi)$, the observed spectrum becomes the convolution of approximated spectrum of constituent j, denoted as $I_j \cdot U_j(\varphi)$, with $R(\varphi)$, resulting in $I_j \cdot (U_j * R)(\varphi)$. Consequently, the summation of the spectra over all constituents contained in the sample yields the spectrum estimated to be observed, $\hat{S}_{ms}(\varphi)$ as shown in Equation (8). In this context, k represents the number of constituents in the sample.

$$\hat{S}_{ms}(\varphi) = \sum_{j=1}^{k} I_j \cdot (U_j * R)(\varphi). \tag{8}$$

2.2.2. Sensitivity Analysis of Parameters

Before exploring parameters, a sensitivity analysis was conducted within the exploration range of each parameter. Parameters that were manually fitted to the spectrum were taken as the true values. From these, only one parameter was varied within the exploration range to generate a spectrum, and the difference from the observed data was calculated.

As a result, as shown in Figure 2-3, it was found that the monoisotopic mass exhibits a steep sensitivity characteristic. This is due to the peak width of the detector response

being at most 0.05 Th at m/z: 1,972, which is extremely small (0.0025%) relative to the mass space analyzable by Q-TOF, ranging from 10 to 40,000 Th. It was also confirmed that the monoisotopic mass exhibits multi-modality within the exploration space.

The parameters of charge state influence the macro distribution shape, but do not affect the intervals between the comb-like peaks of mass-to-charge ratio, resulting in a broad sensitivity characteristic as shown in Figure 2-5 and Figure 2-6. Likewise, the isotopic parameters influence the micro peak width, but do not significantly change the mass, resulting in a broad sensitivity characteristic as seen in Figure 2-7 and Figure 2-8.

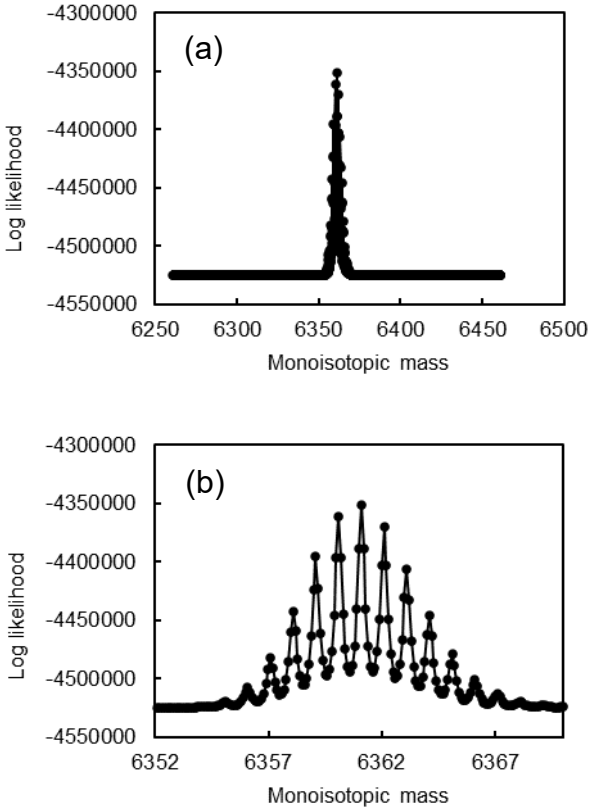

Figure 2-3. Sensitivity characteristic of monoisotopic mass.

(a) Overall view; and (b) Enlarged view.

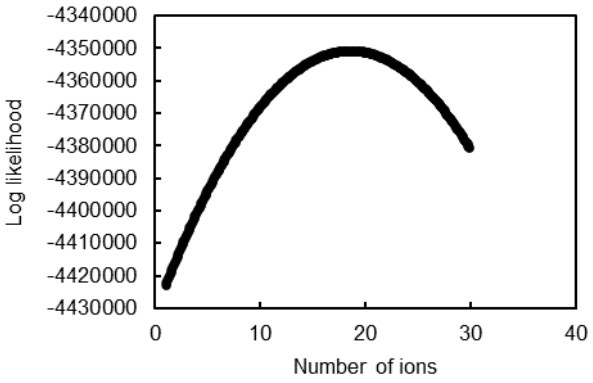

Figure 2-4. Sensitivity characteristic of ion counts.

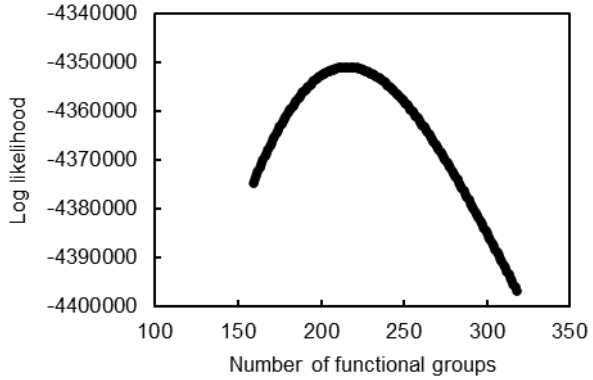

Figure 2-5. Sensitivity characteristic of representative number of functional groups.

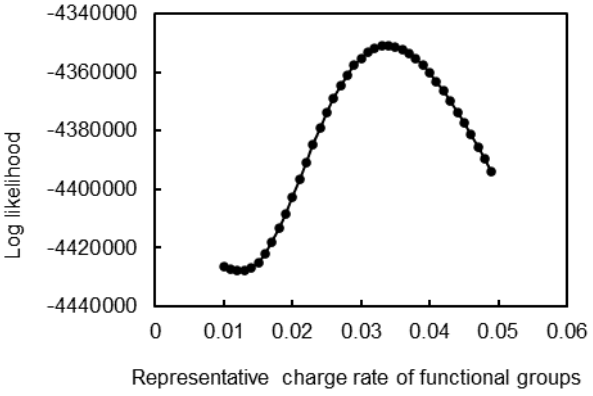

Figure 2-6. Sensitivity characteristic of representative charge rate.

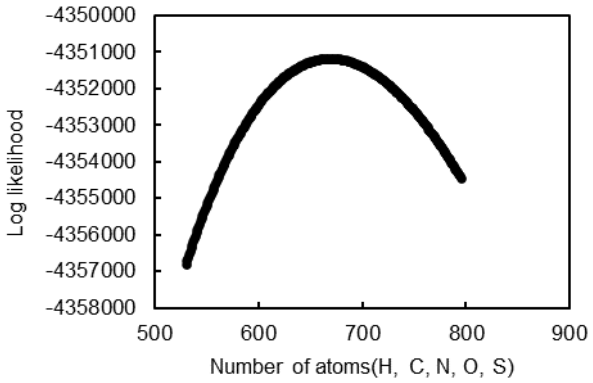

Figure 2-7. Sensitivity characteristic of representative number of atoms.

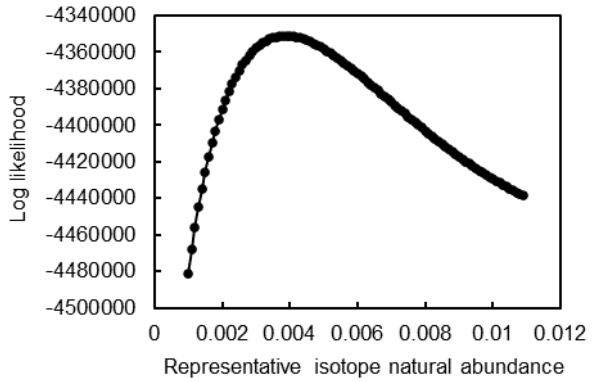

Figure 2-8. Sensitivity characteristic of representative isotopic abundance.

2.2.3. Bayesian Inference of Number of Constituents and Parameters

When the observation data from the mass spectrometer $S_{obs}(\varphi)$ is obtained, assuming the number of constituents as k, the posterior probability distribution $P_k(\theta_k|S_{obs})$ for parameters θ_k: $[(m'_1, I_1, n_1, u_1, l_1, v_1), \cdots, (m'_k, I_k, n_k, u_k, l_k, v_k)]$ is defined as per Bayes' theorem. Note that $P_k(\theta_k|S_{obs})$ represents the likelihood of parameters θ_k when $S_{obs}(t)$ is provided, and $P_k(\theta_k)$ denotes the prior distribution.

$$P_k(\theta_k|S_{obs}) \propto P_k(S_{obs}|\theta_k)P_k(\theta_k). \quad (9)$$

We determine the posterior probability and optimal parameters by maximizing logarithmic posterior probability LP_k, defined as Equation (10). Here, β represents the inverse temperature. We use Simulated Annealing to ensure active parameter exploration in flat areas or sharp peaks of posterior probability. This is achieved by multiplying the inverse temperature β (< 1) to the posterior probability. Initially starting from a low inverse temperature value (i.e., high temperature) and gradually increasing to a higher value (i.e., low temperature). At low inverse temperatures (high temperatures), the system explores a wide parameter space. Conversely, at high inverse temperatures (low temperatures), the system converges to the optimal solution. This time, we set the temperature change in three stages: $\beta = 0.2^5 \rightarrow 0.2^4 \rightarrow 0.2^3$, and

$$LP_k \coloneqq \beta\log\big(P_k(S_{obs}|\theta_k)\big) + \log\big(P_k(\theta_k)\big). \tag{10}$$

Here, in addition to the prior distribution of each parameter (uniform distribution), we incorporate a regularization term, $w_{bic}(k)$ to achieve a suitable balance between model complexity (number of constituents) and model fit (loss with respect to data). We also introduce a regularization term, $w_{ex}(k, m'_1 \dots m'_k)$, to prevent multiple constituents within the same model from assuming the same monoisotopic mass. Hence, we introduce the following logarithmic prior distribution:

$$\log\big(P_k(\theta_k)\big) \propto -\big(w_{bic}(k) + w_{ex}(k, m'_1 \dots m'_k)\big). \tag{11}$$

To determine the appropriate number of constituents k, we define the regularization term $w_{bic}(k)$ representing the complexity of the model with k based on the Bayesian Information Criterion (BIC) [32]. The BIC is a statistical measure that balances the fit to

the data and model complexity [32], [33]. Here, N represents the dimension of the observation data $S_{obs}(\varphi)$, which in this study is the number of data points in the mass-to-charge ratio (φ) direction. For example, if $S_{obs}(\varphi)$ represents the signal from a TOF-type MS, N corresponds to the value obtained by dividing the observation time by the time resolution of the detection system.

$$w_{bic}(k) = \lambda \cdot \frac{k}{2} \cdot \log N \text{, and} \tag{12}$$

$$\lambda: 300\ (hyperparameter).$$

Furthermore, we define a constituent by its unique monoisotopic mass. Therefore, if the estimated values of the monoisotopic mass parameters of multiple constituents are the same in the algorithm, the count of constituents won't be accurate. Here, we define the logarithmic prior distribution (regularization term) w_{ex} as shown in Equation (13), using a penalty that increases exponentially according to the difference in estimated monoisotopic mass values, as shown in Figure 2-9. The integral of the spectrum $\int_0^\infty S_{obs}(\varphi)\, d\varphi$ is also multiplied as a coefficient to ensure that the impact of the penalty does not change depending on the scale of the observed data. Here, m_i and m_j represent the monoisotopic masses of the ith and jth constituents, respectively. This $w_{ex}(k, m_1' \ldots m_k')$ increases as the monoisotopic masses of constituents become closer, preventing the algorithm from estimating the same constituent for both constituent i and constituent j.

$$w_{ex}(k, m_1', \ldots, m_k') = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{a \int_0^\infty S_{obs}(\varphi)\, d\varphi}{2b} \exp\left(-\frac{|m_i - m_j|}{b}\right), \tag{13}$$

$$a = 0.001, \text{and } b = 0.1\ (hyperparameter).$$

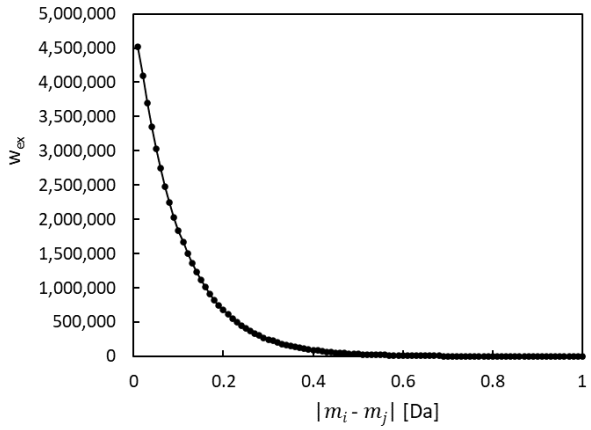

Figure 2-9. Characteristics of the regularization term w_{ex}.

Here, by substituting the parameter θ_k generated from MCMC into model (8), we obtain the spectrum as $\hat{S}(\varphi)$. We assume a normal distribution for the noise. The standard deviation of the noise denoted as σ is set to 2,000. We obtain likelihood from the normal distribution based on the differences between generated and observed spectra at each φ. Since, in reality, spectral data consists of a set of N discrete data points along the φ axis, the integral over φ can be approximated as a discrete sum over N measurement points.

$$\log\big(P_k(S_{obs}|\theta_k)\big) = \int \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left|\hat{S}_{ms}(\varphi) - S_{obs}(\varphi)\right|^2}{2\sigma^2}\right)\right)d\varphi$$
$$\approx -\frac{1}{2\sigma^2}\int \left|\hat{S}_{ms}(\varphi) - S_{obs}(\varphi)\right|^2 d\varphi - N\log(\sigma) - \frac{N}{2}\log(2\pi). \quad (14)$$

Consequently, the logarithm of the posterior probability distribution is as follows:

$$LP_k \coloneqq \beta\log\big(P_k(S_{obs}|\theta_k)\big) + \log\big(P_k(\theta_k)\big)$$
$$= \beta\left(-\frac{1}{2\sigma^2}\int \left|\hat{S}_{ms}(\varphi) - S_{obs}(\varphi)\right|^2 d\varphi - N\log(\sigma) - \frac{N}{2}\log(2\pi)\right)$$
$$-\big(w_{bic}(k) + w_{ex}(k, m'_1 \dots m'_k)\big). \quad (15)$$

To obtain the maximum posterior probability and parameters θ_k that maximize the

posterior probability (formula (15)), we conduct sampling from this posterior probability distribution using MCMC.

2.2.4. Parameter Exploration and Optimization

From the posterior probability distribution $P_k(\theta_k|S_{obs})$, we sample the parameter θ_k to select the one that maximizes the posterior probability. We employ the No-U-Turn Sampler (NUTS) for sampling, a recent and popular variant of the Markov Chain Monte Carlo (MCMC) method. NUTS is a type of MCMC, especially a derivative of the Hamiltonian Monte Carlo method (HMC) [34], [35].

After executing MCMC, the parameters of the maximum posterior probability obtained are inherited as initial values, and optimization of the parameters is performed using Stochastic Variational Inference (SVI [36]–[38]).

2.2.4.1. Parameter Exploration Using the Markov Chain Monte Carlo Method

HMC uses concepts from physics to efficiently sample from high-dimensional probability distributions. However, choosing an appropriate number of leapfrog [39] steps (the number of steps the parameter moves during simulation) in HMC can be challenging. If there are too few steps, the sampler cannot effectively move across the exploration space; and too many, it will U-turn back toward its starting point.

NUTS dynamically selects an appropriate number of steps to explore the Hamiltonian's energy surface based on the principle of stopping the step before the sampler begins a U-turn. This addresses the problem of adjusting the HMC leapfrog steps, enabling efficient sampling from high-dimensional probability distributions. The max tree depth, equivalent to the maximum number of search steps in a single iteration, was set to 10.

The Hamiltonian is defined as:

$$H(x,\zeta) = V(x) + K(\zeta). \tag{16}$$

where x is the variable (position) we want to sample from the target probability distribution, ζ is an auxiliary variable (momentum), $V(x)$ represents potential energy, and $K(\zeta)$ represents kinetic energy.

First, we randomly initialize the momentum ζ for the current position ζ. Using the leap-frog method, we compute a new (x,ζ) pair. In this process, the momentum ζ is initially updated by half a step $\frac{\eta}{2}$:

$$\zeta\left(t+\frac{\eta}{2}\right) = \zeta(t) - \frac{\eta}{2}\cdot\frac{\partial V(x)}{\partial x(t)}. \tag{17}$$

where η is the step size, and t represents the current time step. Next, the position x is updated for the one step:

$$x(t+\eta) = x(t) + \eta\cdot\frac{\partial K}{\partial\zeta\left(t+\frac{\eta}{2}\right)}. \tag{18}$$

Finally, the momentum ζ is updated by another half step:

$$\zeta(t+\eta) = \zeta\left(t+\frac{\eta}{2}\right) - \frac{\eta}{2}\cdot\frac{\partial V}{\partial x(t+\eta)}. \tag{19}$$

If $\zeta(t)\cdot\zeta(t+\eta)\leq 0$, it is determined that the sampler has made a U-turn, and the exploration is terminated. After all steps are completed, an acceptance/rejection step is performed using the Metropolis method [40], [41]. During this step, the difference in Hamiltonian energy is computed to get $\Delta H = H(x(t+\eta),\zeta(t+\eta)) - H(x(t),\zeta(t))$. Here, $(x(t),\zeta(t))$ is the current sample and $(x(t+\eta),\zeta(t+\eta))$ is the new sample. If the Hamiltonian energy of the new sample is lower or equal to the current one $(\Delta H\leq 0)$, the new sample is accepted. Conversely, if the Hamiltonian energy of the new sample is

higher ($\Delta H > 0$), it's accepted with probability $min(1, \exp(-\Delta H))$.

For this study, the step size η of the sampling algorithm was tuned to achieve an acceptance rate of 0.5. If the acceptance rate is too high, only steps near the current parameter value might be accepted, possibly preventing full exploration of the parameter space. If the rate is too low, many proposed steps will be rejected, increasing the time taken for sampling.

The number of samples in this study was set to 1,000. Although the initial state of MCMC is chosen randomly, this state often lies in a domain different from the target probability distribution. Reaching closer to the target distribution requires a certain number of steps (iterations). However, samples generated in this initial phase often do not reflect the posterior probability distribution correctly. Therefore, we discard samples from this initial phase. This process is called "Burn-in," and was set to 1,000 samples in this study. Before the 1,000 step sampling for the parameter search, this burn-in sampling was performed.

There's also autocorrelation between samples produced by MCMC, implying that consecutive samples depend on each other. This autocorrelation can impact statistical estimation. To reduce the correlation between acquired samples, we sampled every other step. Furthermore, MCMC sampling depends on its initial state, which increases the risk of getting trapped in local optima, especially in high-dimensional spaces. By sampling from multiple initial values, we can explore the parameter space more broadly and reduce this risk. In this study, we started from four different initial values.

The domain definitions for each parameter are as per Table 2-1. When the number of constituents is $k = k'(> 2)$, from among the k' constituents, the prior distribution of parameters for constituents 1 to $k' - 1$ is determined using a narrowed prior distribution

centered around the optimal parameters estimated in the model for $k = k' - 1$. Based on this new prior distribution, we estimate the optimal parameters and acquire the maximum posterior probability.

$m_j'|_{k=k'}$ represents the monoisotopic mass when the number of constituents is k'. When $j < k'$, the search range is limited to $\pm \Delta m$ from the value obtained at $k = k' - 1$. For $j = k'$, the entire pre-set search range is explored, as shown in Table 2-1. The same applies to $I_j|_{k=k'}$, $n_j|_{k=k'}$, $u_j|_{k=k'}$, $l_j|_{k=k'}$ and $v_j|_{k=k'}$. As the number of constituents increases, the area that a single constituent occupies in the observed data spectrum becomes smaller. Therefore, the lower limit of $I_j|_{k=k'}$ is divided by k'.

Table 2-1. The domain of the parameters.

Parameter	Range	Constant
$m_j'\|_{k=k'}$	$\begin{cases} [m_{min}, m_{max}] & for\ j = k', and \\ [(m_j'\|_{k=k'-1} - \Delta m), (m_j'\|_{k=k'-1} + \Delta m)] & for\ j < k'. \end{cases}$	$m_{min} = 6300.0$ $m_{max} = 6400.0$ $\Delta m = 4.0$
$I_j\|_{k=k'}$	$\begin{cases} [I_{min}, I_{max}] & for\ j = k', and \\ \left[\frac{I_j\|_{k=k'-1}}{3k'}, I_{max}\right] & for\ j < k'. \end{cases}$	$I_{min} = 10000$ $I_{max} = 300000$
$n_j\|_{k=k'}$	$\begin{cases} [n_{min}, n_{max}] & for\ j = k', and \\ \left[\left(n_j\|_{k=k'-1} * (1 - \Delta n)\right), \left(n_j\|_{k=k'-1} * (1 + \Delta n)\right)\right] & for\ j < k'. \end{cases}$	$n_{min} = \frac{m_j'\|_{k=k'}}{12.0}$ $n_{max} = \frac{m_j'\|_{k=k'}}{8.0}$ $\Delta n = 0.05$
$u_j\|_{k=k'}$	$\begin{cases} [u_{min}, u_{max}] & for\ j = k', and \\ [(u_j\|_{k=k'-1} - \Delta u), (u_j\|_{k=k'-1} + \Delta u)] & for\ j < k'. \end{cases}$	$u_{min} = 0.001$ $u_{max} = 0.011$ $\Delta u = 0.001$
$l_j\|_{k=k'}$	$\begin{cases} [l_{min}, l_{max}] & for\ j = k', and \\ l_j\|_{k=k'-1} & for\ j < k'. \end{cases}$	$l_{min} = \frac{m_j'\|_{k=k'}}{40.0}$ $l_{max} = \frac{m_j'\|_{k=k'}}{20.0}$
$v_j\|_{k=k'}$	$\begin{cases} [v_{min}, v_{max}] & for\ j = k', and \\ v_j\|_{k=k'-1} & for\ j < k'. \end{cases}$	$v_{min} = 0.01,$ $v_{max} = 0.05$

2.2.4.2. Parameter Optimization by Stochastic Variational Inference

After executing MCMC, the parameters corresponding to the maximum posterior probability are inherited as initial values, and optimization is performed using SVI. SVI replaces the complex posterior probability distribution with a more manageable approximate distribution (variational posterior $Q_k(\theta_k|\mu_k)$), minimizing the Kullback-Leibler (KL) divergence between the approximate and true posterior distributions. Since the KL divergence cannot be computed directly, we instead maximize the Evidence Lower Bound (ELBO) [42] as a surrogate objective function. The ELBO is defined as the expected log likelihood of the observed data under the variational distribution, adjusted by a regularization term that penalizes the divergence between the true posterior and the variational distribution. For this study, only the MAP values were needed, so $Q_k(\theta_k|\mu_k)$ is defined by a delta function $\delta(\theta_k - \mu_k)$ to approximate the posterior probability distribution of each number of constituents. μ_k is a point in the parameter space θ_k and serves as a candidate for the parameter set $\theta_{k_{map}}$ that maximizes the posterior probability. In the maximization of ELBO, since the variational distribution $Q_k(\theta_k|\mu_k)$ is defined as a delta function, the integral involving $\log Q_k(\theta_k|\mu_k)$ simplifies as its contribution becomes negligible except at μ_k. Thus, for practical purposes within this optimization framework, we can consider its impact to be zero, focusing solely on the log likelihood component evaluated at μ_k. Therefore, the desired $\theta_{k_{map}}$ is given by equation (20).

$$\begin{aligned}\theta_{k_{map}} &= \arg\max_{\mu_k}\big(\mathrm{ELBO}(\theta_k|\mu_k)\big) \\ &= \arg\max_{\mu_k}\Big(\mathbb{E}_{Q_k(\theta_k|\mu_k)}[\log P_k(S_{obs}|\theta_k) - \log Q_k(\theta_k|\mu_k)]\Big).\end{aligned} \tag{20}$$

Since $Q_k(\theta_k|\mu_k)$ is delta function $\delta(\theta_k - \mu_k)$,

$$\theta_{k_{map}} = \arg\max_{\mu_k}\left(\log P_k(S_{obs}|\mu_k) - \log Q_k(\theta_k|\mu_k)\right). \tag{21}$$

Given that $Q_k(\theta_k|\mu_k)$ is represented as a delta function, its contribution to the ELBO

becomes negligible except at μ_k, simplifying the calculation by effectively eliminating the $\log Q_k(\theta_k|\mu_k)$ term in the optimization.

$$\theta_{k_{map}} = \arg\max_{\mu_k} (\log P_k(S_{obs}|\mu_k)). \tag{22}$$

To maximize the ELBO, that is, to minimize the negative ELBO, Adam [43] (Adaptive Moment Estimation), a type of Stochastic Gradient Descent (SGD), is used. Adam is widely used in machine learning. By individually adjusting the learning rate α for each parameter, Adam allows parameters with steeper gradients to receive smaller updates, while parameters with gentler gradients receive larger updates, automatically scaling the problem. Additionally, Adam reduces the oscillations that were a challenge with SGD by considering both the first moment v_t and the second moment s_t of past gradients.

The parameter θ is updated in three steps:

1. Compute the gradient G_t of the loss function (in this case, the negative ELBO) at the current step t.
2. Update the first moment v_t and the second moment s_t as follows:

$$v_t = \beta_1 v_{t-1} + (1-\beta_1) G_t, \text{and} \tag{23}$$

$$s_t = \beta_2 s_{t-1} + (1-\beta_2) {G_t}^2. \tag{24}$$

3. Update the parameter θ using the adjusted moments:

$$\theta_{t+1} = \theta_t - \alpha \frac{\mathrm{v}_t}{\sqrt{s_t + \varepsilon}}. \tag{25}$$

Here, $\beta_1 = 0.9, \beta_2 = 0.999, \text{and } \varepsilon = 10^{-8}$. The initial value of the learning rate α was set to 0.0005, and the parameter update is performed 20,000 times.

2.2.5. Workflow for Estimating Constituents in a Sample

The overall picture of the workflow to determine the optimal parameters and posterior probability for each assumed number of constituents from the observational data of the mass spectrometer is as shown in Figure 2-10.

First, as described in 2.2.3, (i) input the observational data of the mass spectrometer with dimensions of flight time and ion counts. Then (ii) assume that the number of constituents, k, contained in the sample is 1. (iii) Set the inverse temperature to 0.2^5.

Next, as described in 2.2.4.1, (iv) sample $1{,}000 \times 4$ times from the posterior probability distribution, (v) set the MAP solution obtained by MCMC as the initial value for the next MCMC. Then (vi) divide the inverse temperature by 0.2. Repeat steps (iv) to (vi) three times. The number of iterations was determined experimentally.

As described in 2.2.4.2, (vii) set the parameter of the maximum posterior probability obtained by MCMC as the initial value for SVI. Then (viii) optimize the parameters with SVI, and (ix) set the parameter of the maximum posterior probability obtained by SVI as the initial value for the next MCMC.

Following 2.2.3, increase the number of constituents, k, by 1. Repeat steps (iii) to (vi). (x) Continue this until the maximum possible number of constituents, k_{max}. Finally, (xi) compare the maximum posterior probabilities of models from constituents $k = 1$ to k_{max}, and (xii) select the model with the largest posterior probability. Also, obtain the optimal parameters at that time.

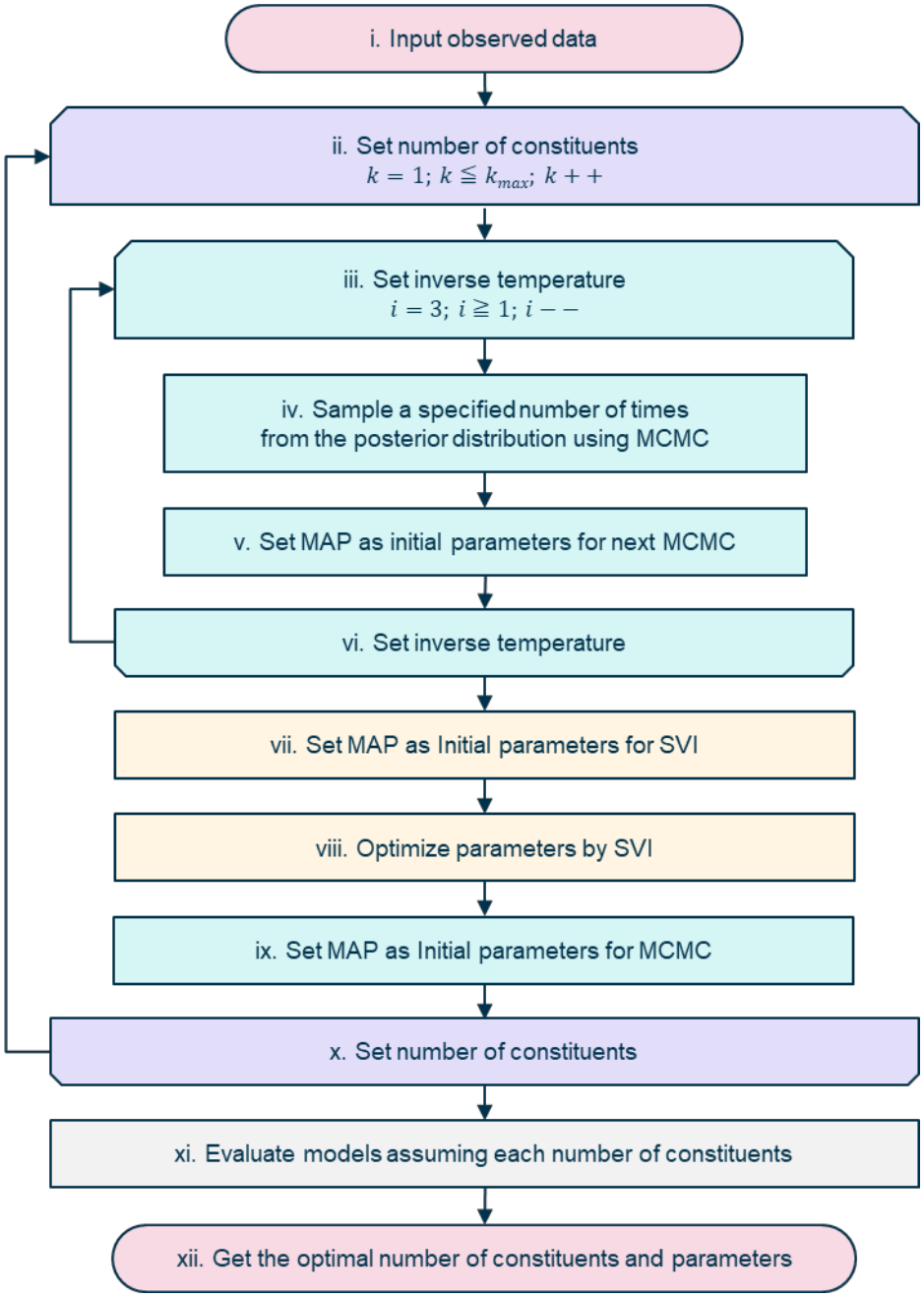

Figure 2-10. Estimation process overall workflow.

2.3. Results

2.3.1. Validation Environment

The specifications of the PC used for verifying the proposed method, as well as the software versions, are as follows. The proposed method handles data with 1 million dimensions along the time axis, requiring a large memory size. Additionally, to rapidly explore a wide 6-dimensional parameter space $\left(m_j', I_j, n_j, u_j, l_j, v_j\right)$ using MCMC, the high-speed probabilistic programming library, NumPyro, along with its compatible CUDA and GPU, were used.

Table 2-2. Validation environment.

CPU	Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz
GPU	Tesla V100-DGXS-16GB
RAM	256GB
OS	Ubuntu 20.04.6 LTS
Software	Python 3.8.10 Numpyro 0.11.0 jax 0.4.7 CUDA 11.8

2.3.2. Creation of Simulation Data for Validation

Based on the nucleic acid drug Fomivirsen [44] (ID: A), four impurity constituents with modified base sequences were added, and spectra for a total of five constituents were generated via simulation. Specific values are as per Table 2-3. This enables the replication of a system where the principal constituent's isotopic distribution and the impurity spectra are mixed.

Ion counts for each constituent were set at 20,000. To facilitate the interpretation of results and to ensure that the algorithm treats each constituent fairly, we will conduct evaluations using a 1:1 concentration ratio for each component in the proposed method. The number of atoms for each element in each constituent was obtained from the molecular formula of the respective constituent. Natural isotopic abundance ratios u_j followed the NIST Atomic Weights and Isotopic Compositions for All Elements [45]. The representative functional group number l_j and the representative charge rate v_j were set to 224 and 0.035, respectively, to ensure that the generated spectra resembled real data.

The procedure involved sampling from the multinomial distribution represented by Equations (1) and (2) 20,000 times (total incoming ion counts) for each constituent. Subsequently, spectra were formed following the procedures in Equations (6) and (8).
The mutation from C (Cytosine) to U (Uracil) is called deamination and is generated in the synthesis process due to solvent conditions and thermal stress [46], [47].

Table 2-3. Settings for constituent spectrum generation.

ID	Sequence	Molecular Formula	Monoisotopic Mass m_j' [Da]	Representative Functional Group Number l_j	Representative Charge Rate v_j	Ion Counts
A	gcgttt-gctcttcttctt-gcg	$C_{204}H_{263}N_{63}O_{134}P_{20}$	6361.088	224	0.035	200 000
B	gcgttt-gutcttcttctt-gcg	$C_{204}H_{262}N_{62}O_{135}P_{20}$	6362.072	224	0.035	200 000
C	gugttt-gutcttcttctt-gcg	$C_{204}H_{261}N_{61}O_{136}P_{20}$	6363.057	224	0.035	200 000
D	gugttt-gutcttcttctt-gug	$C_{204}H_{260}N_{60}O_{137}P_{20}$	6364.042	224	0.035	200 000
E	gugttt-gutcttuttctt-gug	$C_{204}H_{259}N_{59}O_{138}P_{20}$	6365.027	224	0.035	200 000

The spectra of the generated single constituents A to E were combined according to equation (8) in the 15 combinations listed in Table 2-4. This allows for a comprehensive combination of 2-3 constituents based on constituent A, as well as an evaluation of each individual constituent. We use these as test data.

Table 2-4. Combinations of constituents when generating spectra.

Mixture No.	Constituents
1	A,B,C
2	A,B,D
3	A,B,E
4	A,C,D
5	A,C,E
6	A,D,E
7	A,B
8	A,C
9	A,D
10	A,E
11	A
12	B
13	C
14	D
15	E

2.3.3. Evaluation of Constituent Count Estimation Accuracy

The results of estimating the number of constituents in the spectra of the test data (Mixture No.1~15) using our proposed method are as shown in Table 2-5. The values within the table represent the negative logarithm of the maximum posterior probability in the model of constituent count k. Therefore, the smallest value should be selected.

By choosing the most suitable number of constituents based on this criterion, the success rate for estimating the true number of constituents was 80% (12/15). Additionally, the presence or absence of impurities (distinguishing between $k = 1$ and $k \geqq 2$) could be determined with 100% accuracy. We believe this is sufficient as a standard for recognizing the presence and number of impurities in pharmaceuticals and taking appropriate measures.

The computation time required for the estimation was approximately 10 hours per constituent, resulting in a total of 50 hours under the condition of $k_{max} = 5$ set in this study.

Table 2-5. Negative logarithm of the maximum posterior probability assuming each constituent count.

(Orange background indicates the true number of constituents, blue text indicates the minimum value across models.)

Mixture No.	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
1	4,373,750	3,756,984	3,752,997	3,758,705	3,763,612
2	4,278,475	3,765,457	3,753,649	3,756,753	3,762,496
3	4,194,715	3,771,155	3,759,534	3,765,155	3,763,712
4	4,219,672	3,748,972	3,754,868	3,761,667	3,763,951
5	4,319,573	3,773,246	3,757,747	3,758,338	3,763,773
6	3,824,787	3,750,045	3,752,258	3,757,899	3,763,075
7	3,798,373	3,746,176	3,747,193	3,752,544	3,758,004
8	3,795,441	3,744,561	3,748,947	3,756,355	3,759,390
9	3,824,787	3,750,045	3,752,258	3,757,899	3,763,075
10	3,825,138	3,769,114	3,758,565	3,769,515	3,771,414
11	3,733,728	3,738,454	3,743,334	3,748,732	3,754,347
12	3,736,354	3,739,259	3,744,921	3,750,142	3,755,513
13	3,734,851	3,738,732	3,743,821	3,751,714	3,754,223
14	3,735,192	3,740,629	3,745,981	3,751,377	3,755,752
15	3,734,867	3,738,788	3,744,300	3,749,556	3,755,907

2.3.4. Accuracy of Parameter Estimation with Maximum Posterior

The optimal monoisotopic masses and ion counts estimated in the model where the posterior probability is maximum for each test data are shown in Table 2-6.

The monoisotopic mass had an average error of 1.348 Da and a maximum error of 4.931 Da. This is insufficient to determine how many mutations have occurred, making it unsuitable for examining the cause of impurity generation with a difference of 1 Da. Regarding the ion counts, there was an average error of 4% and a maximum error of 82%.

For instance, the standards for total desamido impurity and total impurities in injectable glucagon are 14% or less and 31% or less, respectively [48]. Therefore, the accuracy of the ion count estimation in the proposed method is insufficient to estimate the impact of impurities.

Table 2-6 (Part 1). Optimal monoisotopic masses and ion counts of the model with the maximum posterior probability.

Mixture No.	Constituents	Mass[Da] (Estimated)	Mass[Da] (True)	Absolute Error[Da]	Ion counts [ions] (Estimated)	Ion counts [ions] (True)	Relative Error[%]
1	A,B,C	6358.073	6361.088	-3.015	138 290	200 000	-31%
		6361.088	6362.072	-0.984	299 930	200 000	50%
		6363.047	6363.057	-0.010	172 510	200 000	-14%
2	A,B,D	6360.088	6361.088	-1.000	207 760	200 000	4%
		6361.043	6362.072	-1.029	270 470	200 000	35%
		6361.081	6364.042	-2.961	132 170	200 000	-34%
3	A,B,E	6359.047	6361.088	-2.041	299 970	200 000	50%
		6360.103	6362.072	-1.969	239 990	200 000	20%
		6366.008	6365.027	0.981	74 160	200 000	-63%
4	A,C,D	6360.088	6361.088	-1.000	298 940	200 000	49%
		6363.043	6363.057	-0.014	299 980	200 000	50%
		-	6364.042	-	-	200 000	-
5	A,C,E	6360.024	6361.088	-1.064	238 440	200 000	19%
		6361.07	6363.057	-1.987	296 510	200 000	48%
		6361.116	6365.027	-3.911	80 810	200 000	-60%
6	A,D,E	6360.079	6361.088	-1.009	297 500	200 000	49%
		6362.027	6364.042	-2.015	299 940	200 000	50%
		-	6365.027	-	-	200 000	-

Table 2-6 (Part 2). Optimal monoisotopic masses and ion counts of the model with the maximum posterior probability.

Mixture No.	Constituents	Mass[Da] (Estimated)	Mass[Da] (True)	Absolute Error[Da]	Ion counts [ions] (Estimated)	Ion counts [ions] (True)	Relative Error[%]
7	A,B	6357.088	6361.088	-4.000	191 670	200 000	-4%
		6362.073	6362.072	0.001	220 850	200 000	10%
8	A,C	6361.043	6361.088	-0.045	113 870	200 000	-43%
		6361.080	6363.057	-1.977	283 890	200 000	42%
9	A,D	6359.044	6361.088	-2.044	280 530	200 000	40%
		6359.111	6364.042	-4.931	138 700	200 000	-31%
10	A,E	6357.088	-	-	227 540	-	-
		6361.029	6361.088	-0.059	35 400	200 000	-82%
		6364.010	6365.027	-1.017	157 560	200 000	-21%
11	A	6361.088	6361.088	0.000	191 840	200 000	-4%
12	B	6361.072	6362.072	-1.000	207 340	200 000	4%
13	C	6363.058	6363.057	0.001	189 640	200 000	-5%
14	D	6363.042	6364.042	-1.000	205 290	200 000	3%
15	E	6365.027	6365.027	0.000	190 240	200 000	-5%

For reference, a comparison between the spectra reconstructed from the estimated parameters and the original signal is shown in Figure 2-11. The overall view in (a) represents the charge distribution, and the enlarged view in (b) represents the isotopic distribution. From these results, it is clear that the spectrum we generated closely matches the observed data. Despite the spectra matching, errors in parameter estimation occurred because of the high degree of freedom in isotopic parameters that trade-off with monoisotopic mass. Even if the monoisotopic mass was lower than the true value, by increasing the representative atomic number n_j or the representative isotopic natural abundance u_j, it's possible to make it fit the observed data to some extent.

Also, the estimated ion counts of each constituent showed errors of up to 82% from the true values. This is presumed to be due to the trade-off relationship between the ion counts of each constituent, with a decrease in the ion count of one constituent being compensated by an increase in another. This is further supported by the fact that the average error in ion counts settles at 4%.

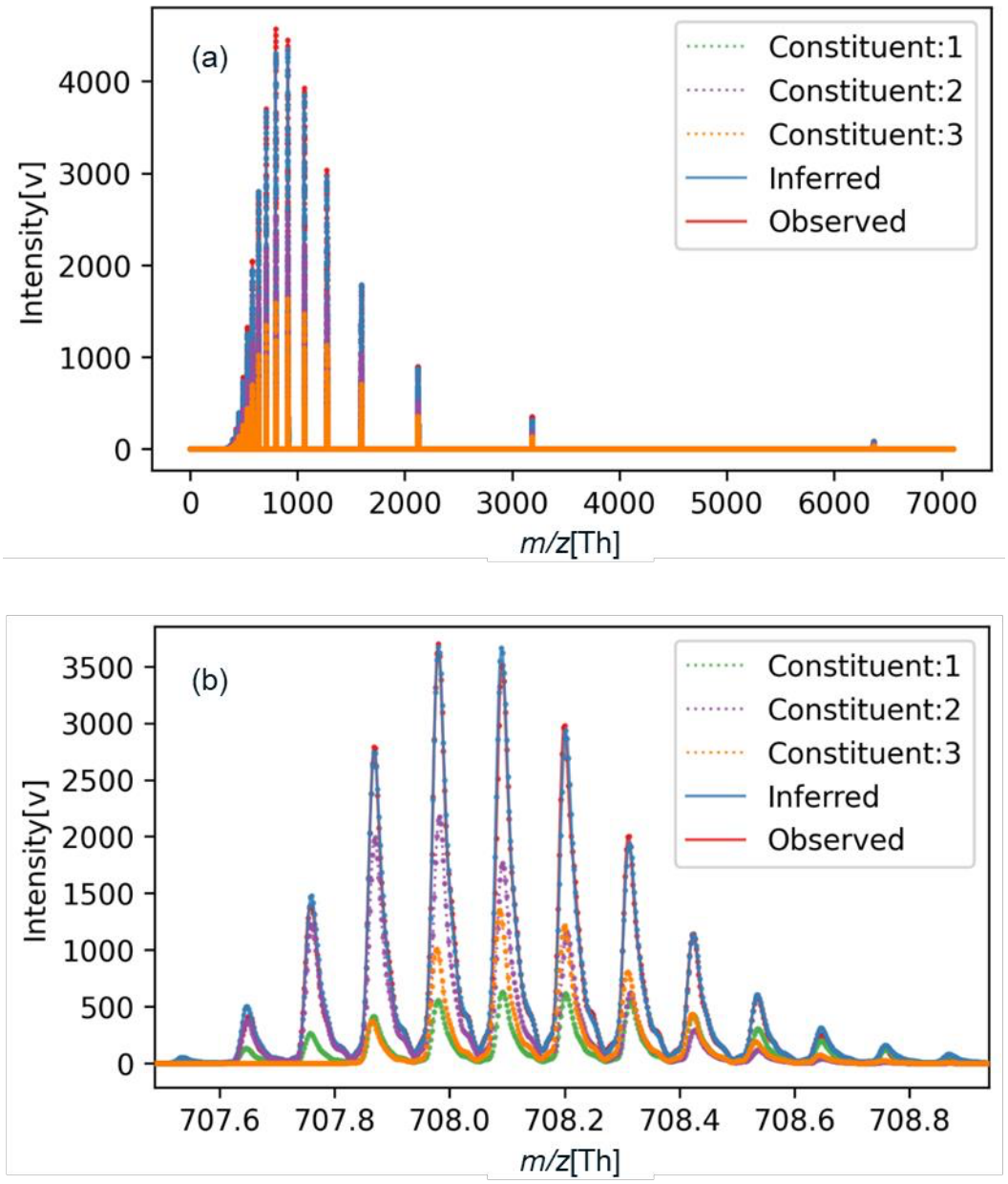

Figure 2-11. Comparison of observed and estimated spectra for Mixture No. 1.

(a) Overall view; and (b) Enlarged view.

2.3.5. Comparison with UniDec

Deconvolution of the test data was performed using the existing method, UniDec as well. Here, deconvolution refers to the process of extracting monoisotopic masses and ion counts from complex observed spectra. The results of deconvolution for each observed spectrum by UniDec are shown in Table 2-7. According to these results, the accuracy for the correct number of constituents was 13% (2/15). This is presumed to be because the UniDec algorithm, which obtains the number of constituents after multiple iterations of deconvolution, does not necessarily guarantee the number of constituents. Please note that this use of UniDec to determine the number of constituents is not its intended application. Under these conditions, UniDec completed the deconvolution process within a few seconds.

Table 2-7 (Part 1). Deconvolution results for each observed spectrum by UniDec.

Mixture No.	Constituents	Mass[Da] (Estimated)	Mass[Da] (True)	Absolute Error[Da]	Intensity [a.u.] (Estimated)	Intensity [a.u.] (True)	Relative Error[%]
1	A,B,C	6359.900	6361.088	-1.188	100.000	100.000	100%
		6360.900	6362.072	-1.172	54.614	100.000	55%
		-	6363.057	-	-	100.000	-
2	A,B,D	6359.900	6361.088	-1.188	100.000	100.000	100%
		6360.900	6362.072	-1.172	68.122	100.000	68%
		6361.800	6364.042	-2.242	23.490	100.000	23%
3	A,B,E	6359.900	-	-	100.000	-	-
		6360.900	-	-	47.326	-	-
		6361.800	6361.088	0.712	22.533	100.000	23%
		6362.800	6362.072	0.728	13.473	100.000	13%
		6363.800	6365.027	-1.227	13.496	100.000	13%
4	A,C,D	6359.900	-	-	100.000	-	-
		6360.900	6361.088	-0.188	94.673	100.000	95%
		6361.800	6363.057	-1.257	64.641	100.000	65%
		6362.800	6364.042	-1.242	19.369	100.000	19%

Table 2-7 (Part 2). Deconvolution results for each observed spectrum by UniDec.

Mixture No.	Constituents	Mass[Da] (Estimated)	Mass[Da] (True)	Absolute Error[Da]	Intensity [a.u.] (Estimated)	Intensity [a.u.](True)	Relative Error[%]
5	A,C,E	6359.900	-	-	100.000	-	-
		6360.900	-	-	63.684	-	-
		6361.800	6361.088	0.712	56.992	100.000	57%
		6362.800	6363.057	-0.257	33.851	100.000	34%
		6363.800	6365.027	-1.227	19.330	100.000	19%
6	A,D,E	6359.900	-	-	100.000	-	-
		6360.900	-	-	53.209	-	-
		6361.800	6361.088	0.712	61.898	100.000	62%
		6362.800	6364.042	-1.242	70.845	100.000	71%
		6363.800	6365.027	-	39.057	100.000	39%
7	A,B	6359.900	6361.088	-1.188	100.000	100.000	100%
		6361.000	6362.072	-1.072	11.538	100.000	12%
8	A,C	6359.900	-	-	100.000	-	-
		6361.000	6361.088	-0.088	40.696	100.000	41%
		6361.800	6363.057	-1.257	10.199	100.000	10%
9	A,D	6359.900	-	-	100.000	-	-
		6361.000	6361.088	-0.088	41.897	100.000	42%
		6361.800	-	-	26.937	-	-
		6362.800	6364.042	-1.242	16.351	100.000	16%

Table 2-7 (Part 3). Deconvolution results for each observed spectrum by UniDec.

Mixture No.	Constituents	Mass[Da] (Estimated)	Mass[Da] (True)	Absolute Error[Da]	Intensity [a.u.] (Estimated)	Intensity [a.u.](True)	Relative Error[%]
10	A,E	6359.000	-	-	19.045	-	-
		6359.900	-	-	100.000	-	-
		6360.900	-	-	27.472	-	-
		6361.800	6361.088	0.712	19.107	100.000	19%
		6362.800	-	-	27.075	-	-
		6363.900	-	-	33.325	-	-
		6364.800	6365.027	-0.227	13.659	100.000	14%
11	A	6358.900	-	-	48.595	-	-
		6359.800	6361.088	-1.288	100.000	100.000	100%
12	B	6359.900	-	-	40.161	-	-
		6360.800	6362.072	-1.272	100.000	100.000	100%
13	C	6360.900	-	-	41.609	-	-
		6361.800	6363.057	-1.257	100.000	100.000	100%
14	D	6361.800	-	-	52.753	-	-
		6362.800	6364.042	-1.242	100.000	100.000	100%
15	E	6362.800	-	-	54.440	-	-
		6363.900	6365.027	-1.127	100.000	100.000	100%

For the verification above, we used UniDec (Version 6.0.2). The particular set parameters during this verification are shown in Table 2-8. The Mass Range was set to the same range as the proposed method, and Sample Mass Every (Da) was set to 0.1 to sufficiently detect impurities with a difference of 1 Da. For parameters not mentioned, default values were used.

Table 2-8. UniDec setting parameters.

Parameter		Setting value
UniDec Parameters	Charge Range	1 to 50
	Mass Range	6300 to 6400 Da
	Sample Mass Every (Da)	0.1
Additional Deconvolution Parameters	Isotopes	Mono
Peak Selection and Plotting	Peak Detection Range (Da)	0.1
	Peak Detection Threshold	0.01

*The other settings are using default values.

2.4. Discussion

Using NUTS, Simulated Annealing, and stochastic variational inference, we estimated parameters such as monoisotopic masses from observed data, achieving an accuracy of 80% in selecting the correct number of constituents, which is significantly higher than the 13% accuracy of existing methods. This is thought to be due to the fact that we created models for each number of constituents, allowing for the comparative evaluation and selection of models for each number of constituents. This made it possible to suggest the presence of impurities in pharmaceuticals, which is useful for searching for better synthesis conditions for middle to high molecular weight pharmaceuticals, and for quality

assurance in factories.

On the other hand, as shown in Table 2-6, the estimated monoisotopic mass for constituent j had a maximum error of 4.931Da from the true value. This is thought to be due to the trade-off relationship between the monoisotopic mass m_j' and the parameters n_j and u_j that determine the isotopic distribution of constituent j. Additionally, there was a relative error of several tens of percent from the true value in the ion counts of each estimated constituent. This is speculated to be because the ion counts of each constituent trade off with each other, with a decrease in one ion being compensated for by an increase in another ion. A potential solution to these problems is to represent monoisotopic masses and ion counts as probability distributions. By considering the uncertainty in monoisotopic masses and ion counts of constituents in the sample, improvements in estimation satisfaction can be expected.

Furthermore, it took about 50 hours for deconvolution assuming 5 constituents per data. This is long compared to the few seconds to a few minutes processing time of UniDec. Also, this processing time is expected to increase almost linearly with the assumed number of constituents. Therefore, it is expected to take a long time when analyzing samples with many constituents, such as serum or environmental samples.

2.5. Conclusion of This Chapter

In this chapter, we aimed to model mass spectrometry, probabilistically estimate the number of constituents in a sample, and accurately determine their monoisotopic masses and ion quantities when identifying the optimal number of constituents. To achieve these goals, we assumed various numbers of constituents within the sample and developed a mass spectrometry model based on parameters such as monoisotopic masses and ion counts. We then applied methods like the No-U-Turn Sampler (NUTS), Simulated Annealing, and stochastic variational inference to find the maximum posterior probability for each modeled number of constituents compared against observed data. These efforts enabled us to accurately estimate the number of constituents, as well as to simultaneously determine parameters such as monoisotopic masses and ion counts. However, challenges remain due to the inaccuracy in estimating monoisotopic masses and ion counts, and the substantial computational time required.

Chapter 3. Study on Accelerating Estimations Using Simulated Annealing and Stochastic Variational Inference

3.1. Overview

Chapter 3 is based on Tomono, Hara, Iida and Washio (2024b) [25]. In the previous chapter, we estimated the number of constituents based on their monoisotopic masses and ion counts. We used various assumed constituent counts to model these parameters and then derived the maximum posterior probability and optimal model parameters for each constituent count using the No-U-Turn Sampler (NUTS), Simulated Annealing, and Stochastic Variational Inference (SVI). This process required extensive computing time, rendering the method impractical for routine use.

Therefore, we decided to perform all parameter estimations using the faster SVI method, entirely replacing the time-consuming Markov Chain Monte Carlo (MCMC) approach. However, as described in the previous chapter, using Stochastic Variational Inference alone is insufficient for exploring parameters extensively due to the Vanishing Gradient Problem. This issue arises because the posterior probability of the monoisotopic mass is mostly flat with several sharp peaks localized in certain areas. Consequently, changes in the generated spectrum do not lead to significant changes in the model's posterior probability, which prevents effective gradient calculation. Applying simple optimization methods to such data often leads to vanishing gradients, making it difficult to effectively explore parameters. If a modified version of SVI capable of addressing this issue could be devised, it would allow for faster and more efficient parameter estimation using

only the improved SVI method.

To address this challenge, we have developed a method that involves gradually convolving Gaussians along the m/z axis between observed and generated spectra, ensuring that gradients always occur during comparison. We have named this method Spectral Annealing Inference (SAI). SAI combines SVI and spectral annealing by Point Spread Function (PSF) to explore optimal parameters while avoiding vanishing gradients and local optima. Figure 3-1 is a schematic diagram that illustrates the mechanism of SAI. It involves convolving the PSF with the spectrum to create gradients for parameter optimization, and this process is repeated while narrowing the variance of the PSF. Ultimately, the PSF becomes a Kronecker delta function, allowing for the determination of parameters and posterior probabilities based directly on the observed spectrum.

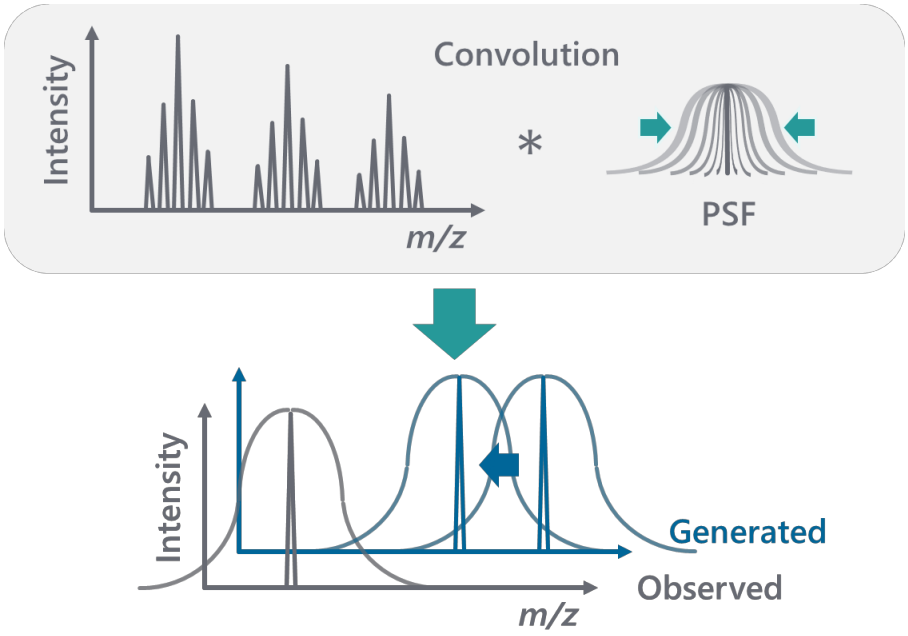

Figure 3-1. Schematic diagram of the gradient generation process in spectral annealing inference (SAI).

After calculating optimal parameters and posterior probabilities in all models by SAI, we select the most probable number of constituents, as well as their monoisotopic masses and ion counts.

3.2. Proposed Method

We use the same physical model as described in the previous chapter. To explore and optimize the parameters, we employ Stochastic Variational Inference (SVI) to estimate the Maximum A Posteriori (MAP) values of each parameter and to determine the model's highest posterior probability.

The optimization problem under this setup can be solved using conventional numerical optimization techniques. Recall that k represents the assumed number of constituents in the sample. In this case, we used Adam [43], a type of stochastic gradient descent widely used in machine learning, to find the value of μ_k that maximizes the likelihood function. The resulting $\theta_{k_{map}}$ is the MAP estimation we sought.

However, the MS spectra to be compared are mostly flat with several localized sharp peaks. Simply applying SVI to such data can result in vanishing gradients, making it difficult to effectively explore parameters. Therefore, to create appropriate gradients of the likelihood function, we convolve a Gaussian distribution $g(\varphi)$ with both the observed spectra S_{obs} and the estimated spectra $\hat{S}_{ms}(\varphi)$ along the mass-to-charge ratio (φ) axis. We define the mean of $g(\varphi)$ as zero and the variance as T_s, and $g(\varphi)$ is represented as shown in Equation (26).

$$g(\varphi) = \frac{1}{\sqrt{2\pi {T_s}^2}} \exp\left(-\frac{1}{2{T_s}^2}(\varphi)^2\right). \tag{26}$$

Then, we performed SVI and iteratively narrowing the variance of $g(\varphi)$, T_s, to effectively search for θ_k. This process, resembling annealing, is termed Spectral Annealing Inference (SAI) in this paper. Let s denote the step of this iteration, and s_{max} denote the total number of iterations. We define T_s as shown in Equation (27). Narrowing the PSF step-by-step according to the iteration count s, this process repeatedly refines the MAP estimation.

$$T_s = \lambda \left(\frac{s_{max} - s}{s_{max}} \right)^4 \ (s = 0,1,2, \dots, s_{max}). \tag{27}$$

For this study, s_{max} is experimentally set to 46, and the coefficient λ is set to 8750. When $s = s_{max}$, the spectrum after convolution becomes identical to the spectrum before convolution.

The blurred spectra at each step are represented as shown in Equations (28) and (29).

$$S'_{obs}(\varphi) = (S_{obs} * g)(\varphi), \text{and} \tag{28}$$

$$\hat{S}'_{ms}(\varphi) = \left(\hat{S}_{ms} * g\right)(\varphi). \tag{29}$$

Using these blurred spectra, we derive the modified log-likelihood $L'_{mse_{ms}}$, and the logarithm of the posterior probability $\log\left(P_k(S'_{obs}|\theta_k)\right)$ is represented as shown in Equation (30). N represents the number of data points of the observation data $S'_{obs}(\varphi)$ and $\hat{S}'_{ms}(\varphi)$.

$$\log\left(P_k(S'_{obs}|\theta_k)\right) \approx -\frac{1}{2\sigma^2} \int \left|\hat{S}'_{ms}(\varphi) - S'_{obs}(\varphi)\right|^2 d\varphi + N \log(\sigma) + \frac{N}{2} \log(2\pi). \tag{30}$$

Here, we define the modified logarithmic likelihood LP'_k as follows:

$$\begin{aligned} LP'_k &\coloneqq \log\left(P_k(S'_{obs}|\theta_k)\right) + \log\left(P_k(\theta_k)\right) \\ &\approx -\frac{1}{2\sigma^2}\int \left|\hat{S}'_{ms}(\varphi) - S'_{obs}(\varphi)\right|^2 d\varphi + N\log(\sigma) + \frac{N}{2}\log(2\pi) \\ &\quad -w_{bic}(k) - w_{ex}(k, m'_1 \dots m'_k). \end{aligned} \tag{31}$$

At each iteration step s ($s = 0,1,2, \dots, s_{max}$), we maximize LP'_k to iteratively refine and determine the parameters θ_k and the posterior probability assuming a number of constituents k. θ_k from each iteration are carried forward to the next step.

By repeating this process from $k = 1$ to k_{max}, we obtain the posterior probabilities of each k. We then compare the posterior probabilities across all k and select the number of constituents with the highest posterior probability and its corresponding parameter set as the optimal choice.

3.3. Results

For the validation of our algorithm, we employed simulated MS data shown in Table 2-4. By using the same data as in Chapter 2, we can compare the estimation speed of the algorithm developed in the previous chapter. This simulation was based on the nucleic acid drug Fomivirsen and its four impurities, which exhibit mass differences ranging from 1 to 4 Daltons.

The overview of the results is presented in Table 3-1. For comparison, the estimation results from the previous chapter using MCMC are also included in the table. Compared to the methods in the previous chapter, the computation time has been significantly reduced while maintaining the accuracy of the number of constituents. However, the accuracy of the monoisotopic mass and ion quantities remained unchanged.

Table 3-1. Results of the estimated performance verification.

Metrics	Estimated by SAI	Estimated by MCMC
Accuracy of constituent numbers	80%（12/15）	80%（12/15）
Monoisotopic mass error	Avg 1.788Da Max 3.983Da	Avg 1.348Da, Max 4.931Da
Ion counts error	Avg 8% Max 89%	Avg 4%, Max 82%
Calculation time	15 minutes	50 hours

The posterior probabilities for the optimal parameters of each model are as shown in Table 3-2. Based on this, we selected the number of constituents and were able to maintain an accuracy rate of 80%.

Table 3-2. Negative logarithm of the maximum posterior probability assuming each constituent count.
(Orange background indicates the true number of constituents, blue text indicates the minimum value across models.)

Mixture No.	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
1	783,837,800	974,980,700	819,785,600	993,572,600	2,523,023,000
2	974,699,600	811,826,400	773,190,500	838,065,400	1,410,565,000
3	1,205,362,000	1,033,501,000	743,789,100	771,244,400	2,321,454,000
4	862,455,600	834,689,600	801,946,100	927,742,400	1,288,006,000
5	1,119,931,000	627,375,700	666,500,600	688,775,800	1,571,504,000
6	1,379,719,000	1,174,463,000	1,011,143,000	1,157,092,000	1,679,262,000
7	421,459,200	500,081,400	573,603,100	703,970,000	2,058,994,000
8	409,007,500	390,423,700	430,594,200	439,182,200	1,503,371,000
9	512,957,400	486,925,400	514,162,100	537,291,100	1,080,766,000
10	1,091,095,000	633,916,700	648,982,200	699,068,600	1,186,193,000
11	178,197,400	217,834,500	217,108,800	337,039,700	454,868,900
12	216,980,000	259,813,500	254,528,500	388,611,700	513,887,100
13	161,197,500	193,327,700	209,736,600	310,515,200	434,137,700
14	204,778,900	237,009,400	253,565,400	363,079,600	465,365,200
15	172,407,200	208,542,900	214,153,100	329,143,100	414,737,900

The monoisotopic masses and ion quantities at the time of maximum posterior probability for each model are shown in Table 3-3. There were no significant differences in the errors of the monoisotopic masses and ion quantities compared to those obtained through exploration using MCMC.

Table 3-3 (Part 1). Optimal monoisotopic masses and ion counts of the model with the maximum posterior probability.

Mixture No.	Constituents	Mass[Da] (Estimated)	Mass[Da] (True)	Absolute Error[Da]	Ion counts [ions] (Estimated)	Ion counts [ions] (True)	Relative Error[%]
1	A,B,C	6361.102	6361.088	0.014	568,570	200,000	184%
		-	6362.072	-	-	200,000	-
		-	6363.057	-	-	200,000	-
2	A,B,D	6360.102	6361.088	-0.986	273,322	200,000	37%
		6362.107	6362.072	0.035	152,619	200,000	-24%
		6364.074	6364.042	0.032	156,850	200,000	-22%
3	A,B,E	6360.101	6361.088	-0.987	349,775	200,000	75%
		6361.115	6362.072	-0.957	95,267	200,000	-52%
		6364.053	6365.027	-0.974	144,265	200,000	-28%
4	A,C,D	6358.094	6361.088	-2.994	156,851	200,000	-22%
		6360.113	6363.057	-2.944	103,492	200,000	-48%
		6362.082	6364.042	-1.960	336,362	200,000	68%
5	A,C,E	6360.102	6361.088	-0.986	359,303	200,000	80%
		6364.059	6363.057	1.002	219,789	200,000	10%
		-	6365.027	-	-	200,000	-
6	A,D,E	6357.106	6361.088	-3.983	178,322	200,000	-11%
		6360.091	6364.042	-3.951	127,659	200,000	-36%
		6362.062	6365.027	-2.965	293,079	200,000	47%

Table 3-3 (Part 2). Optimal monoisotopic masses and ion counts of the model with the maximum posterior probability.

Mixture No.	Constituents	Mass[Da] (Estimated)	Mass[Da] (True)	Absolute Error[Da]	Ion counts [ions] (Estimated)	Ion counts [ions] (True)	Relative Error[%]
7	A,B	6361.109	6361.088	0.021	377,724	200,000	89%
		-	6362.072	-	-	200,000	-
8	A,C	6358.094	6361.088	-2.994	99,031	200,000	-50%
		6361.103	6363.057	-1.954	293,385	200,000	47%
9	A,D	6357.091	6361.088	-3.997	121,317	200,000	-39%
		6360.097	6364.042	-3.945	271,892	200,000	36%
10	A,E	6357.116	6361.088	-3.972	211,758	200,000	-82%
		6361.055	6365.027	-3.972	192,770	200,000	-21%
11	A	6360.118	6361.088	-0.970	198,806	200,000	-4%
12	B	6361.102	6362.072	-0.970	199,874	200,000	4%
13	C	6362.087	6363.057	-0.970	197,888	200,000	-5%
14	D	6363.072	6364.042	-0.970	198,503	200,000	3%
15	E	6364.059	6365.027	-0.968	198,412	200,000	-5%

For Mixture No. 2, the comparison between the spectrum estimated using the optimal parameters and the observed spectrum is as shown in Figure 3-2. From this, it can be seen that the generated spectrum matches the observed spectrum, confirming that there are no issues with the estimation.

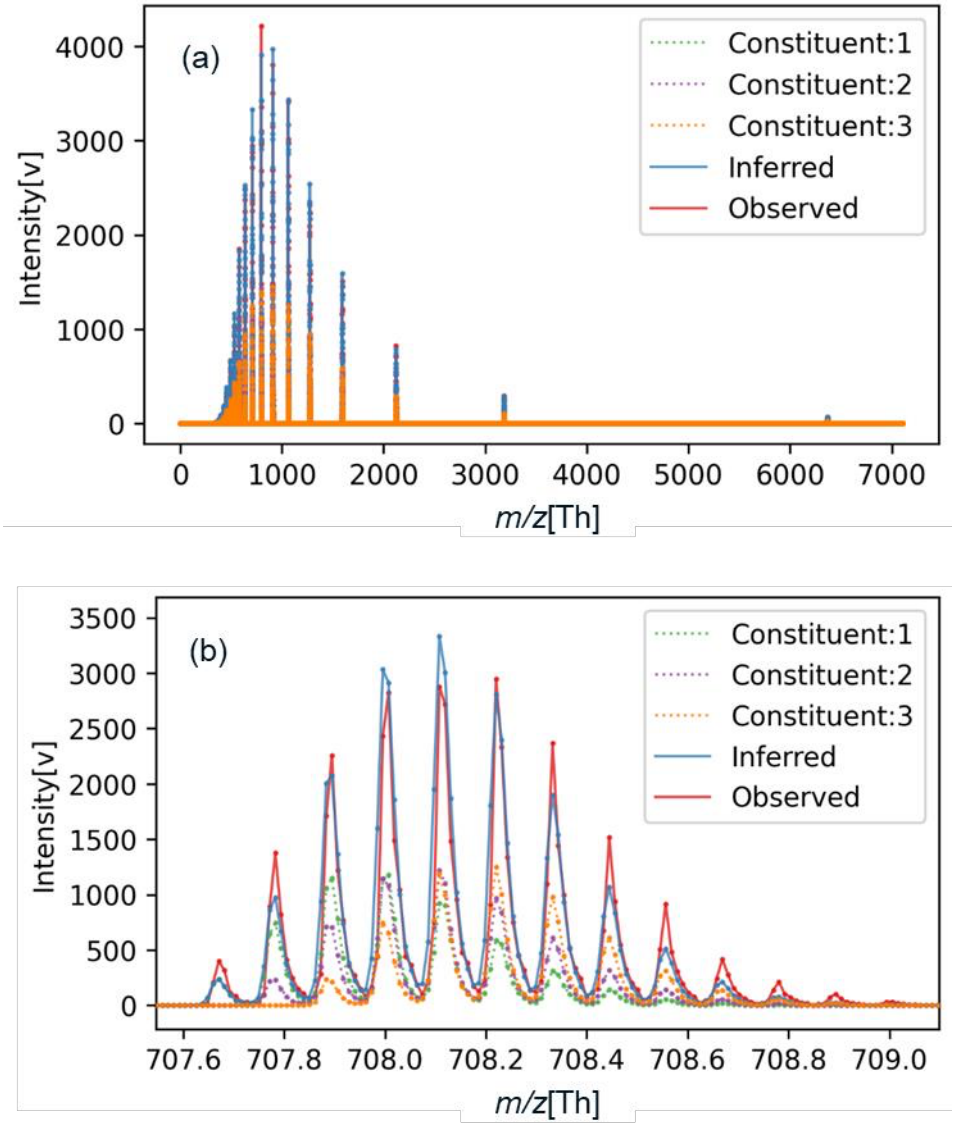

Figure 3-2. Comparison of an observed spectrum and an estimated spectrum.

(a) Overall view; and (b) Enlarged view.

3.4. Discussion

We successfully reduced the computation time from 50 hours to 15 minutes while maintaining an 80% accuracy in estimating the number of constituents. This indicates that the SAI method, which performs parameter exploration through gradient-based methods by convolving the Point Spread Function (PSF), was functioning effectively.

The maximum error observed in the estimated monoisotopic mass was 3.997 Da below the target. This is considered to be due to remaining challenges in the trade-offs among parameters.

Similarly, the relative errors in ion counts were several tens of percent below the target. This situation has not changed from the previous chapter, and we hypothesize that it results from trade-offs between different constituents. It implies that a decrease in the concentration of one constituent appears to be offset by an increase in another.

3.5. Conclusion of This Chapter

In this chapter, we aimed to accelerate the algorithm while maintaining the accuracy of the number of constituents. To this end, we developed the SAI method, which enables rapid parameter exploration without the issue of vanishing gradients, even for sparse spectra, by convolving the Point Spread Function (PSF). This allowed us to significantly reduce the computation time from 50 hours to 15 minutes. However, the low estimation accuracy for monoisotopic masses and ion quantities remains a challenge. To solve this issue, it is necessary to increase the usable information and impose new constraints on the model.

Chapter 4.

Study on Improving Estimation Accuracy by Incorporating a Physical Model into MS/MS Spectra

4.1. Overview

Chapter 4 is based on Tomono, Hara, Iida and Washio (2024c, 2024d) [26], [27]. In the prior chapter, we rapidly estimated the number of constituents and their monoisotopic masses and ion counts using Spectral Annealing Inference (SAI), which allows for estimation while avoiding the vanishing gradient problem by convolving with progressively narrowing PSFs. In this try, the speed of computation was drastically improved, but the accuracy of our results was insufficient.

To address the issue, this study introduces an improved methodology to accurately estimate the optimal number of constituents and their monoisotopic masses and ion counts using hybrid mass spectrometry (MS/MS) spectra. MS/MS is a technique that combines multiple mass spectrometry stages to obtain structural information about precursor ions. It involves isolating specific ions based on their mass-to-charge ratio in the first stage (MS1), fragmenting these ions in a collision cell, and analyzing the resulting fragment ions in the second stage (MS2). This allows for more detailed characterization of complex molecules that cannot be achieved with single-stage MS.

Our method initially models the physical MS and MS/MS system with all possible numbers of constituents. For each model with a different number of constituents, we estimate the optimal monoisotopic masses and ion counts and derived the posterior probabilities. This estimation is achieved by using SAI.

If the MS model and the MS/MS model are not properly linked, simply increasing

the number of estimated constituents will not impose any meaningful constraints, resulting in no improvement in performance. To overcome this, we mathematically combined the MS and MS/MS models, enabling us to utilize the MS/MS data to enhance the estimation of constituent information contained in the MS spectra. An overview diagram of the model extension is shown in Figure 4-1. First, we generate MS spectra using the same method as described in Chapter 2. For the ions contained in these spectra, we then apply a newly developed fragmentation model to obtain MS/MS spectra.

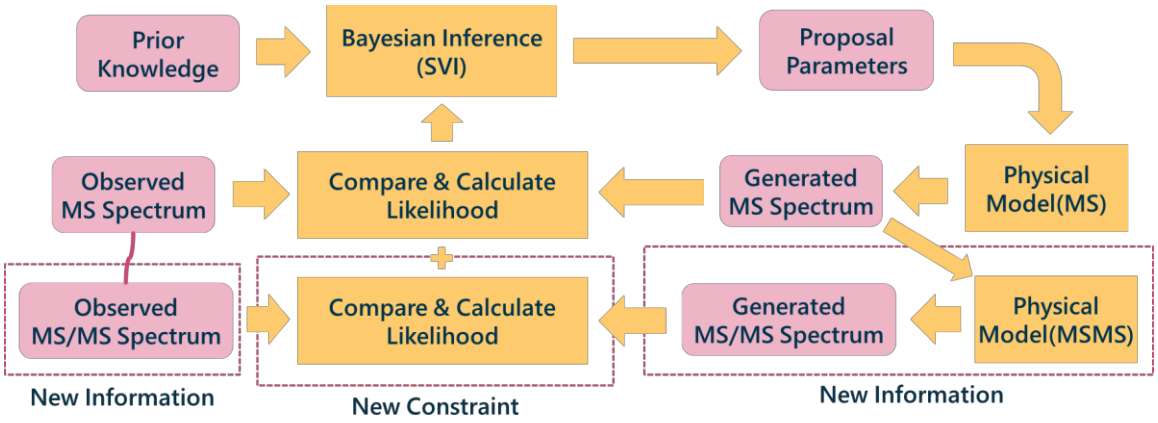

Figure 4-1. Overview of extended analysis method.

4.2. Proposed Method

4.2.1. Physical Model of Mass Spectrometers

We use the model constructed in Chapter 2 to generate MS spectra of intact ions and develop a new model to generate MS/MS spectra for fragment ions.

In this study, we consider a scenario where ions contained within a specific region of the MS spectrum, denoted as $peak_d$ $(d = 1\ to\ d_{max})$, are selected and forwarded to the subsequent stage for MS/MS spectral measurement. Neutral molecules formed during this collision-induced dissociation are not detected.

For the intact constituent j before fragmentation in the collision cell, we define a set of ions sharing the monoisotopic mass m'_f produced in the collision cell as constituent f ($f = 1\ to\ f_{max}$). We assume that totally f_{max} fragment constituents are produced. As with intact constituent j, we assume a binomial distribution as the isotopic distribution of fragment constituent f. Here we define the increase in neutron number as $\omega_f = \text{round}\left(\frac{m-m'_f}{\varepsilon}\right)$, where m represents a variable in the mass space, and ε represents the mass of a neutron as before. The distribution is denoted by $\tilde{p}_f(\omega_f)$, within the range of $\omega_f \geq 0$. In biomolecules such as nucleic acids and proteins, which consist of repeating structural units, it is reasonable to regard that elements are uniformly distributed across the ion of a precursor constituent. Therefore, we assume the number of atoms in an ion of a fragment constituent is roughly proportional to its monoisotopic mass. Accordingly, the number of atoms in constituent f, n_f, is evaluated as $n_j \cdot \frac{m'_f}{m'_j}$, where n_j denotes the number of atoms of constituent j, as defined in Section 2.2.1. Moreover, by similar argument on the uniformity of the chemical composition across the molecule of a precursor constituent, its fragments share the same chemical composition with the precursor constituent. Therefore, we assume the rate of isotopes in a fragment, u_f, is equal to the isotopic replacing rate of the precursor constituent j, u_j, which is also defined in Section 2.2.1. Consequently, the isotopic distribution $\tilde{p}_f(\omega_f)$ is represented as shown in Equation (32).

$$\tilde{p}_f(\omega_f) = \begin{cases} \binom{n_f}{\omega_f} {u_f}^{\omega_f} (1-u_f)^{n_f-\omega_f} & \text{for } \omega_f \geq 0, \text{and} \\ 0 & \text{otherwise.} \end{cases} \tag{32}$$

Additionally, we approximate the charge distribution of constituent f, $\tilde{q}_f(z)$, using a

binomial distribution, where z denotes a variable representing the absolute value of charge, as defined earlier. In a manner similar to the discussion on isotopes, it is reasonable to approximate that chargeable sites, such as phosphate groups in nucleic acids and side chains in proteins, are uniformly distributed across the entire precursor ion. Therefore, we assume that the number of chargeable sites that can acquire a charge is also roughly proportional to the monoisotopic mass of a fragment. Accordingly, the number of chargeable sites of constituent f, l_f, is calculated as $l_j \cdot \frac{m'_f}{m'_j}$, where l_j is defined in Section 2.2.1 as the total number of chargeable sites in constituent j.

$\tilde{q}_j(z)$ is represented as shown in Equation (4) as described in Section 2.2.1. For reference, it is restated below:

$$\tilde{q}_j(z) = \binom{l_j}{z} {v_j}^z \left(1 - v_j\right)^{l_j - z}. \tag{4}$$

Since the distribution of chargeable sites in the fragments are regarded as the same as those in the precursor constituent j, we also assume that the probability of the chargeable sites acquiring a charge, v_f, is equal to v_j. Thus, $\tilde{q}_f(z)$ can be expressed as shown in Equation (33).

$$\tilde{q}_f(z) = \binom{l_f}{z} {v_f}^z \left(1 - v_f\right)^{l_f - z}. \tag{33}$$

When the total number of ions of constituent j within $peak_d$ is given by I_{d_j} and the probability that a precursor constituent j dissociates into a fragment constituent f is denoted by $\rho_{j\to f}$ (where $\rho_{j\to f} < 1$), the expected number of ions of constituent f produced from constituent j within $peak_d$, $I_{d_j\to f}$, is calculated as $I_{d_j\to f} = \text{round}\left(I_{d_j} \cdot \rho_{j\to f}\right)$. Each ion in the $I_{d_j\to f}$ ions is indexed by $i_{d_j\to f}$. The mass and charge of each individual ion $i_{d_j\to f}$ are denoted as $\omega_{i_{d_j\to f}} \sim \tilde{p}_f$ and $z_{i_{d_j\to f}} \sim \tilde{q}_f$, respectively.

When an ion $i_{d_j \to f}$ is detected, its observed ideal spectrum would be $\delta\left(\varphi - \left(m'_f + \varepsilon\omega_{i_{d_j \to f}}\right)/z_{i_{d_j \to f}}\right)$. Regardless of its charge state or mass, a single ion contributes to the observed spectrum as a single delta function as well as Equation (5). Therefore, the ideal spectrum formed by this set of ions (from $i_{d_j \to f} = 1$ to $I_{d_j \to f}$), $D_{d_j \to f}(\varphi)$, is represented as shown in Equation (34).

$$D_{d_j \to f}(\varphi) = \sum_{i_{d_j \to f}=1}^{I_{d_j \to f}} \delta\left(\varphi - \left(m'_f + \varepsilon\omega_{i_{d_j \to f}}\right)/z_{i_{d_j \to f}}\right). \tag{34}$$

The probability distribution $U_{d_j \to f}(\varphi)$ of constituent f, which is produced by the dissociation of constituent j included in $peak_d$, can be calculated using the same approach as for constituent j. However, when the increase in neutron number from the monoisotopic mass and the charge of the precursor ion of constituent j in the $peak_d$ is denoted as ω_{d_j} and z_{d_j}, the increase in neutron number and charge of the precursor ion of fragment f produced from constituent j in the $peak_d$, ω_f and z do not exceed ω_{d_j} and z_{d_j}. Therefore, the domain of the fragment spectrum is limited to $\omega_f < \omega_{d_j}$ and $z < z_{d_j}$. Consequently, the probability distribution of fragment f produced from the ions belonging to constituent j in $peak_d$ along the mass-to-charge ratio, φ, axis, $U_{d_j \to f}(\varphi)$ is described by Equation (35).

$$U_{d_j \to f}(\varphi) = \sum_{z=1}^{z_{d_j}} \sum_{\omega_f=1}^{\omega_{d_j}} \tilde{p}_f(\omega_f) \cdot \tilde{q}_f(z) \cdot \delta\left(\varphi - \left(m'_f + \varepsilon\omega_f\right)/z\right). \tag{35}$$

In a manner similar to the MS spectrum, the observed spectrum of ions is proportional to the probability distribution of ions along the φ axis. Then, the empirical

spectrum $D_{d_j \to f}(\varphi)$ converges uniformly to the theoretical distribution $U_{d_j \to f}(\varphi)$ as sample size increases. Consequently, the spectrum of fragment constituent f produced from constituent j in the $peak_d$, $D_{d_j \to f}(\varphi)$, is approximated by $U_{d_j \to f}(\varphi)$ as shown in Equation (36).

$$D_{d_j \to f}(\varphi) = \sum_{i_{d_j \to f}=1}^{I_{d_j \to f}} \delta\left(\varphi - \left(m'_f + \varepsilon\omega_{i_f}\right)/z_{i_f}\right) \approx I_{d_j \to f} \cdot U_{d_j \to f}(\varphi) \quad \left(I_{d_j \to f} \gg 1\right). \tag{36}$$

Therefore, the MS/MS spectrum for $peak_d$, $\hat{S}_{msms_d}(\varphi)$, is obtained by summing $I_{d_j \to f} \cdot U_{d_j \to f}(\varphi)$ over all j and f, as shown in Equation (36). Here, $R(\varphi)$ represents the point spread of the detector's response, as introduced in Section 2.2.1.

$$\hat{S}_{msms_d}(\varphi) = \sum_{j=1}^{k} \sum_{f=1}^{f_{max}} I_{d_j \to f} \cdot \left(U_{d_j \to f} * R\right)(\varphi). \tag{37}$$

Here, we set f_{max} to an appropriate number of potential fragment constituents. In actual estimation, the fitting progresses from the most prominent fragment constituents identified by the magnitude of the spectrum. To estimate the number of precursor constituents and their parameters, it is not necessary to identify all the fragment constituents, and it suffices to cover some key fragments. Consequently, f_{max} may be set to a number less than the actual number of fragment constituents produced. The value of f_{max} is determined based on prior knowledge.

4.2.2. Bayesian Inference of Number of Constituents and Parameters

As described in Section 2.2.1, the physical parameters of precursor ions, such as monoisotopic mass, chargeable sites, number of atoms, isotopic replacement rate, and charge rate, have already been defined.

Assuming the number of constituents as k, the extended set of parameters for estimation, denoted as ${\theta'}_k$, is derived from the original parameter set θ_k. This extended parameter set is represented as:

$$
\begin{aligned}
{\theta'}_k = \Big\{ & m'_j, I_j, n_j, u_j, l_j, v_j, m'_f, I_{d_j}, \rho_{j\to f}, n_f, u_f, l_f, v_f \\
& \mid j = 1,2,\dots,k, d = 1,2,\dots,d_{max}, f = 1,2,\dots,f_{max} \Big\}.
\end{aligned}
$$

Here, ${\theta'}_k$ is defined for each combination of a precursor constituent j, a fragment constituent f and a peak d.

We specifically calculate $m'_j, I_j, n_j, u_j, l_j, v_j, m'_f, I_{d_j}\ and\ \rho_{j\to f}$ using the iterative optimization algorithm, Adam, from the range specified in Table 4-1. Here, the range for the newly introduced dissociation rate, $\rho_{j\to f}$, is also defined. The initial values are randomly determined within the defined domain. Parameters $n_f, u_f, l_f\ and\ v_f$ are automatically determined as described in Section 4.2.1. The value of I_{d_j} is set to the number of ions contained within the peak interval of the MS spectrum generated from the precursor ion parameters. The m/z range of the peak interval is determined based on the settings used during actual analysis on the instrument.

Table 4-1. The domain of the parameters.

Parameter	Range	Constant
$m'_j\|_{k=k'}$	$\begin{cases}[m_{min}, m_{max}] & for\ j = k', and \\ [(m'_j\|_{k=k'-1} - \Delta m), (m'_j\|_{k=k'-1} + \Delta m)] & for\ j < k'.\end{cases}$	$m_{min} = 100.0$ $m_{max} = 10000.0$ $\Delta m = 4.0$
$I_j\|_{k=k'}$	$\begin{cases}[I_{min}, I_{max}] & for\ j = k', and \\ \left[\frac{I_j\|_{k=k'-1}}{3k'}, I_{max}\right] & for\ j < k'.\end{cases}$	$I_{min} = 100$ $I_{max} = 100000$
$n_j\|_{k=k'}$	$\begin{cases}[n_{min}, n_{max}] & for\ j = k', and \\ \left[\left(n_j\|_{k=k'-1} * (1 - \Delta n)\right), \left(n_j\|_{k=k'-1} * (1 + \Delta n)\right)\right] & for\ j < k'.\end{cases}$	$n_{min} = \frac{m'_j\|_{k=k'}}{16.0}$ $n_{max} = \frac{m'_j\|_{k=k'}}{6.0}$ $\Delta n = 0.05$
$u_j\|_{k=k'}$	$\begin{cases}[u_{min}, u_{max}] & for\ j = k', and \\ [(u_j\|_{k=k'-1} - \Delta u), (u_j\|_{k=k'-1} + \Delta u)] & for\ j < k'.\end{cases}$	$u_{min} = 0.0001$ $u_{max} = 0.01$ $\Delta u = 0.001$
$l_j\|_{k=k'}$	$\begin{cases}[l_{min}, l_{max}] & for\ j = k', and \\ l_j\|_{k=k'-1} & for\ j < k'.\end{cases}$	$l_{min} = 1.0$ $l_{max} = \frac{m'_j\|_{k=k'}}{20.0}$ $\Delta l = 1.0$
$v_j\|_{k=k'}$	$\begin{cases}[v_{min}, v_{max}] & for\ j = k', and \\ v_j\|_{k=k'-1} & for\ j < k'.\end{cases}$	$v_{min} = 0.01,$ $v_{max} = 1.0$
$m'_f\|_{k=k'}$	$\begin{cases}[m_{min}, m_{max}] & for\ j = k', and \\ [(m'_f\|_{k=k'-1} - \Delta m), (m'_f\|_{k=k'-1} + \Delta m)] & for\ j < k'.\end{cases}$	$m_{min} = 100.0$ $m_{max} = m'_j$ $\Delta m = 4.0$
$\rho_{j\to f}$	$\left[\rho_{j\to f\ min}, \rho_{j\to f\ max}\right]$	$\rho_{j\to f\ min} = 0.1$ $\rho_{j\to f\ max} = 1.0$

Substituting the number of atoms of constituent j, n_j, the isotopic replacing rate of constituent j, u_j into Equation (3) and the number of chargeable sites of constituent j, l_j, and the charge rate of chargeable sites of constituent j, v_j into Equation (4), and the

monoisotopic mass of constituent j, m_j' and the number of ions of constituent j, I_j, into Equation (5) yields the MS spectrum $\hat{S}_{ms}(\varphi)$ as derived from Equation (8). Further, substituting n_f, u_f into Equation (32), l_f, v_f into Equation (33), and $m_f', I_{d_j}, \rho_{j\to f}$ into Equation (34) leads to the derivation of the MS/MS spectra $\hat{S}_{msms_d}(\varphi)$ from Equation (37).

We consider a scenario in which we obtain a set of observed spectra $\boldsymbol{S}_{obs}$, consisting of MS spectrum $S_{obs_{ms}}$ and MS/MS spectra $S_{obs_{msms_d}}(d = 1,2,\dots,d_{max})$. In the extended parameter set ${\theta'}_k$, the posterior probability distribution ${P'}_k({\theta'}_k|S_{obs})$, where the combined MS and MS/MS spectra $\boldsymbol{S}_{obs}$ are observed, is defined according to Bayes' theorem as the following formula. Here ${P'}_k(\boldsymbol{S}_{obs}|{\theta'}_k)$ represents a likelihood of parameters ${\theta'}_k$ given under $\boldsymbol{S}_{obs}$. ${P'}_k({\theta'}_k)$ denotes a prior distribution.

$$ {P'}_k({\theta'}_k|\boldsymbol{S}_{obs}) \propto {P'}_k(\boldsymbol{S}_{obs}|{\theta'}_k){P'}_k({\theta'}_k). \tag{38} $$

We determine the posterior probability and optimal parameters by maximizing logarithmic posterior probability $LP"_k$, defined as:

$$ LP"_k \coloneqq \log\left({P'}_k(\boldsymbol{S}_{obs}|{\theta'}_k)\right) + \log\left({P'}_k({\theta'}_k)\right). \tag{39} $$

Here, we introduce two likelihoods derived from observation error models. The observed spectrum typically includes thermal noise from detection circuitry, which is assumed to follow a normal distribution. Therefore, we base the observational error, representing a deviation between observed data and true values, on this distribution. For estimation, we employ square error-based likelihood derived from the normal distribution. However, because low-intensity regions within the spectrum have less contribution to the overall error evaluation if we use a square error-based likelihood, relying solely on this likelihood reduces accuracy of parameter estimation where the errors in the low-intensity

spectral regions must be reflected. To overcome this difficulty, we additionally introduce a likelihood function sensitive to errors in the low-intensity parts of the spectrum. To evaluate the discrepancies between the observed and estimated spectra regardless of spectral intensity, we use the correlation coefficient along the φ axis as the additional likelihood. This coefficient, calculated by normalizing the inner product of both spectra against their intensities, excludes the influence of each spectrum's intensity, thus providing a measure that assesses the similarity of their shapes over the entire spectrum domain including the low-intensity region.

Let $L_{mse_{ms}}$ denote a logarithmic likelihood based on the normal error distribution of the MS spectrum and $L_{mse_{msms_d}}$ denote that of the MS/MS spectrum at peak d, respectively. The standard deviation of the normal distribution, σ, is set to 0.5 based on actual measurements. $L_{mse_{ms}}$ and $L_{mse_{msms_d}}$ are calculated by summing the logarithms of the probability densities of the error between the observed spectrum and estimated spectrum over φ. Here, N specifically denotes the number of data points on the φ axis within a single spectrum. $L_{mse_{ms}}$, $L_{mse_{msms_d}}$ are expressed as follows:

$$L_{mse_{ms}} = \int \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left|\hat{S}_{ms}(\varphi) - S_{obs_{ms}}(\varphi)\right|^2}{2\sigma^2}\right)\right)d\varphi$$
$$\approx -\frac{1}{2\sigma^2}\int \left|\hat{S}_{ms}(\varphi) - S_{obs_{ms}}(\varphi)\right|^2 d\varphi + N\log(\sigma) + \frac{N}{2}\log(2\pi), \text{and} \quad (40)$$

$$L_{mse_{msms_d}} = \int \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left|\hat{S}_{msms_d}(\varphi) - S_{obs_{msms_d}}(\varphi)\right|^2}{2\sigma^2}\right)\right)d\varphi$$
$$\approx -\frac{1}{2\sigma^2}\int \left|\hat{S}_{msms_d}(\varphi) - S_{obs_{msms_d}}(\varphi)\right|^2 d\varphi + N\log(\sigma) + \frac{N}{2}\log(2\pi). \quad (41)$$

To introduce the additional correlation-based likelihood, we employ the von Mises distribution as an error model, which is defined by the correlation coefficient between

two vectors representing the observed and estimated spectra. The logarithmic likelihoods based on the von Mises distribution are denoted as $L_{cos_{ms}}$ and $L_{cos_{msms_d}}$, respectively. The probability density function of the von Mises distribution is given by $f(\widehat{\boldsymbol{S}}) = \frac{1}{2\pi I_0(\gamma)} \exp\left\{\gamma \frac{\langle \widehat{\boldsymbol{S}}, \boldsymbol{S} \rangle}{|\widehat{\boldsymbol{S}}||\boldsymbol{S}|}\right\}$ [49]. Here, $\widehat{\boldsymbol{S}}$ and $\boldsymbol{S}$ represent estimated and observed spectra, respectively, viewed as vectors. $\langle \widehat{\boldsymbol{S}}, \boldsymbol{S} \rangle$ represents their inner product. The parameter γ represents concentration of the probability distribution. I_0 is a modified Bessel function of the first kind of order zero, and $2\pi I_0(\gamma)$ serves as normalization factor. γ is experimentally determined to be the aforementioned number of data points N. Consequently, the log-likelihoods, $L_{cos_{ms}}$ and $L_{cos_{msms_d}}$, are calculated as shown in Equations (17) and (18).

$$\begin{aligned} L_{cos_{ms}} &= \log\left(\frac{1}{2\pi I_0(N)} \exp\left(N \frac{\left\langle \hat{S}_{ms}(\varphi), S_{obs_{ms}}(\varphi) \right\rangle}{\left|\hat{S}_{ms}(\varphi)\right| \left|S_{obs_{ms}}(\varphi)\right|}\right)\right) \\ &= N \frac{\left\langle \hat{S}_{ms}(\varphi), S_{obs_{ms}}(\varphi) \right\rangle}{\left|\hat{S}_{ms}(\varphi)\right| \left|S_{obs_{ms}}(\varphi)\right|} - \log\left(2\pi I_0(N)\right), \text{and} \end{aligned} \tag{42}$$

$$\begin{aligned} L_{cos_{msms_d}} &= \log\left(\frac{1}{2\pi I_0(N)} \exp\left(N \frac{\left\langle \hat{S}_{msms_d}(\varphi), S_{obs_{msms_d}}(\varphi) \right\rangle}{\left|\hat{S}_{msms_d}(\varphi)\right| \left|S_{obs_{msms_d}}(\varphi)\right|}\right)\right) \\ &= N \frac{\left\langle \hat{S}_{msms_d}(\varphi), S_{obs_{msms_d}}(\varphi) \right\rangle}{\left|\hat{S}_{msms_d}(\varphi)\right| \left|S_{obs_{msms_d}}(\varphi)\right|} - \log\left(2\pi I_0(N)\right). \end{aligned} \tag{43}$$

The total log-likelihood of the estimated spectrum set ($\hat{S}_{ms}(\varphi)$, $\hat{S}_{msms_d}(\varphi)(d = 1,2, \dots, d_{max})$) under the observed spectrum set S_{obs} is expressed as shown in Equation (44).

$$\log\left(P'_k(S_{obs}|\theta'_k)\right) = L_{mse_{ms}} + \frac{1}{d_{max}} \sum_{d=1}^{d_{max}} L_{mse_{msms_d}} + L_{cos_{ms}} + \frac{1}{d_{max}} \sum_{d=1}^{d_{max}} L_{cos_{msms_d}}. \tag{44}$$

In determining the appropriate number of constituents k in Bayesian framework, we need to prevent the selection of overfitted complex models of its logarithmic posterior probability $LP"_k$. For doing so, we incorporate a modified penalty term $w'_{bic}(k)$ based on prior knowledge. $w'_{bic}(k)$ is defined using the Bayesian Information Criterion (BIC), a statistical measure that evaluates the trade-off between model fit and complexity [32], [33]. Incorporating $w'_{bic}(k)$ into the prior probability allows us to determine the appropriate number of constituents k. By applying $\lambda = 6.0 \times 10^7$ (a hyperparameter) and using the number of data points N in the spectrum, as defined earlier, $w'_{bic}(k)$ is represented as shown in Equation (45).

$$w'_{bic}(k) = \lambda \cdot \frac{k}{2} \cdot \log N\,. \tag{45}$$

Additionally, to ensure that the monoisotopic masses of the constituents do not overlap, we introduce a modified penalty function $w'_{ex}(k, m'_1 \ldots m'_k)$, inspired by the Laplace distribution. The reason why we use such a penalty is because we define a constituent by its unique monoisotopic mass. Here, we experimentally set the gain coefficient $a = 10 \times N$. If m'_i and m'_j differ by more than the mass of neutron, ε, they are certainly different constituents. Consequently, we also experimentally determine the appropriate value below ε as the threshold coefficient $b = 0.8$. We then define $w'_{ex}(k, m'_1 \ldots m'_k)$, represented by the assumed number of constituents k and the monoisotopic masses of each constituent, $m'_1 \ldots m'_k$, as shown in Equation (46).

$$w'_{ex}(k, m'_1 \ldots m'_k) = a \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \max\left(1 - \frac{\left|m'_i - m'_j\right|}{b}, 0\right). \tag{46}$$

This penalty function reaches its maximum value when the monoisotopic masses of

different constituents completely coincide.

By assuming a uniform prior distribution of each parameter, the logarithmic prior probability is defined as:

$$\log\left(P'_k(\theta'_k)\right) = -w'_{bic}(k) - w'_{ex}(k, m'_1 \dots m'_k). \tag{47}$$

Here, by substituting Equations (19) and (22) into Equation (14), we obtain the logarithmic posterior probability $LP"_k$ to be maximized as:

$$\begin{aligned} LP"_k &\coloneqq \log\left(P'_k(\boldsymbol{S}_{obs}|\theta'_k)\right) + \log\left(P'_k(\theta'_k)\right) \\ &= L_{mse_{ms}} + \frac{1}{d_{max}} \sum_{d=1}^{d_{max}} L_{mse_{msms_d}} + L_{cos_{ms}} + \frac{1}{d_{max}} \sum_{d=1}^{d_{max}} L_{cos_{msms_d}} \\ &\quad -w'_{bic}(k) - w'_{ex}(k, m'_1 \dots m'_k). \end{aligned} \tag{48}$$

4.2.3. Parameter Exploration and Optimization

The parameter exploration method is the same as in Chapter 3. To create appropriate gradients of the likelihood function, we convolve a Gaussian distribution $g(\varphi)$ with both the observed spectra $S_{obs_{ms}}$, $S_{obs_{msms}}$ and the estimated spectra $\hat{S}_{ms}(\varphi)$, $\hat{S}_{msms_d}(\varphi)$ (where $d = 1,2, \dots, d_{max}$). We define the mean of $g(\varphi)$ as zero and the variance as T_s, and $g(\varphi)$ is represented as shown in Equation (26).

We performed SVI and iteratively narrowing the variance of $g(\varphi)$, T_s, to effectively search for θ'_k, which is termed Spectral Annealing Inference (SAI) in this paper. Let s denote the step of this iteration, and s_{max} denote the total number of iterations. We define T_s as shown in Equation (27), which is reiterated here for clarity.

$$T_s = \lambda \left(\frac{s_{max} - s}{s_{max}} \right)^4 \ (s = 0,1,2, \dots, s_{max}). \tag{27}$$

For this study, s_{max} is set to 46, and the coefficient λ is set to 8750, the same value as

successfully used in Chapter 3 to find global solutions.

The blurred spectra at each step are represented as shown in Equations (49), (50), (51) and (52).

$$S'_{obs_{ms}}(\varphi) = \left(S_{obs_{ms}} * g\right)(\varphi), \tag{49}$$

$$S'_{obs_{msms}}(\varphi) = \left(S_{obs_{msms}} * g\right)(\varphi), \tag{50}$$

$$\hat{S}'_{ms}(\varphi) = \left(\hat{S}_{ms} * g\right)(\varphi), \text{and} \tag{51}$$

$$\hat{S}'_{msms}(\varphi) = \left(\hat{S}_{msms} * g\right)(\varphi). \tag{52}$$

Using these blurred spectra, we derive the modified log-likelihood $L'_{mse_{ms}}$, $L'_{mse_{msms_d}}$, $L'_{cos_{ms}}$ and $L'_{cos_{msms_d}}$, as defined in Equations (53), (54), (55), and (56), respectively.

$$L'_{mse_{ms}} = -\frac{1}{2\sigma^2}\int \left|\hat{S}'_{ms}(\varphi) - S'_{obs_{ms}}(\varphi)\right|^2 d\varphi + N\log(\sigma) + \frac{N}{2}\log(2\pi), \tag{53}$$

$$L'_{mse_{msms_d}} = -\frac{1}{2\sigma^2}\int \left|\hat{S}'_{msms_d}(\varphi) - S'_{obs_{msms_d}}(\varphi)\right|^2 d\varphi + N\log(\sigma) + \frac{N}{2}\log(2\pi), \tag{54}$$

$$L'_{cos_{ms}} = N\frac{\left\langle \hat{S}'_{ms}(\varphi), S'_{obs_{ms}}(\varphi)\right\rangle}{\left|\hat{S}'_{ms}(\varphi)\right|\left|S'_{obs_{ms}}(\varphi)\right|} - \log\left(2\pi I_0(N)\right), \text{and} \tag{55}$$

$$L'_{cos_{msms_d}} = N\frac{\left\langle \hat{S}'_{msms_d}(\varphi), S'_{obs_{msms_d}}(\varphi)\right\rangle}{\left|\hat{S}'_{msms_d}(\varphi)\right|\left|S'_{obs_{msms_d}}(\varphi)\right|} - \log\left(2\pi I_0(N)\right). \tag{56}$$

Let $\boldsymbol{S}'_{obs}$ denote the set of observed spectra $\boldsymbol{S}_{obs}$ blurred by the PSF. The logarithm of the likelihood $\log\left(P"_k(\boldsymbol{S}'_{obs}|\theta'_k)\right)$ is represented as follows:

$$\log\left(P"_k(\boldsymbol{S}'_{obs}|\theta'_k)\right) = L'_{mse_{ms}} + \frac{1}{d_{max}}\sum_{d=1}^{d_{max}} L'_{mse_{msms_d}} + L'_{cos_{ms}} + \frac{1}{d_{max}}\sum_{d=1}^{d_{max}} L'_{cos_{msms_d}}. \tag{57}$$

We introduce a prior distribution same as Chapter 3. By substituting Equation (57) in place of Equation (19) into Equation (14), the modified logarithmic likelihood ${LP'''}_k$ is obtained as follows:

$$\begin{aligned} {LP'''}_k &\coloneqq \log\left({P"}_k({S'}_{obs}|{\theta'}_k)\right) + \log\left({P'}_k({\theta'}_k)\right) \\ &\approx {L'}_{mse_{ms}} + \frac{1}{d_{max}} \sum_{d=1}^{d_{max}} {L'}_{mse_{msms_d}} + {L'}_{cos_{ms}} + \frac{1}{d_{max}} \sum_{d=1}^{d_{max}} {L'}_{cos_{msms_d}} \\ &\quad - {w'}_{bic}(k) - {w'}_{ex}(k, m'_1 \dots m'_k). \end{aligned} \tag{58}$$

Same as in Chapter 3, at each iteration step s ($s = 0,1,2, \dots, s_{max}$), we maximize ${LP'''}_k$ to iteratively refine and determine the parameters ${\theta'}_k$ and the posterior probability assuming a number of constituents k. ${\theta'}_k$ from each iteration are carried forward to the next step.

As in the previous chapter, by repeating this process from $k = 1$ to k_{max}, we obtain the posterior probabilities of each k. We then compare the posterior probabilities across all k and select the number of constituents with the highest posterior probability and its corresponding parameter set as the optimal choice.

4.3. Results

In this section, we detail the outcomes of our experiments to validate the estimation accuracy of constituent counts, monoisotopic mass, and ion quantities in our proposed method. All the experiments were conducted exclusively using numerical simulations. These simulations generated data to mimic real-world mass spectrometry analyses. We specifically focused on simulating the mass spectra of nucleic acid drugs and their impurities, such as Fomivirsen and its altered sequences. This is because current analytical

methodologies have challenges in accurately identifying these substances, due to the complexities arising from their isotopic and charge distributions. We compared the performance of our proposed method against established baseline method, UniDec. The performance was evaluated based on accuracy of constituent count estimation, deviations in monoisotopic mass, and discrepancies in ion quantities.

4.3.1. Validation Environment

The specifications of a computer used to verify the proposed method, as well as the software versions, are summarized in Table 4-2. The proposed method handled data with high dimensions along the time axis, requiring a large memory size. Additionally, to rapidly explore the parameter space using SVI, the high-speed probabilistic programming library, NumPyro, along with its compatible CUDA and GPU, were used.

Table 4-2. Computational environment used for validation.

CPU	Intel(R) Xeon(R) Platinum 8280 CPU @ 2.70GHz
GPU	NVIDIA A100
RAM	1,024 GB
OS	Ubuntu 20.04.6 LTS
Software	Python 3.10.12
	Numpyro 0.14.0
	jax 0.4.14
	CUDA 12.1

4.3.2. Creation of Simulation Data for Validation

Based on the nucleic acid drug Fomivirsen [44] (ID: A), two impurities with modified base sequences were added, and MS spectra for a total of three constituents were generated using simulation methods presented in the prior research [24]. Specific details were

provided in Table 4-3. This setup replicated a system where the principal constituent's isotopic distribution was mixed with the spectra of the impurities. The mutation from C (Cytosine) to U (Uracil), known as deamination, can occur during the synthesis process due to solvent conditions and thermal stress [46], [47].

Table 4-3. Settings for constituent spectrum generation.

ID	Sequence	Molecular Formula	Monoisotopic Mass m_j' [Da]
A	gcgtttgctcttcttcttgcg	$C_{204}H_{263}N_{63}O_{134}P_{20}$	6361.088
B	gcgtttgutcttcttcttgcg	$C_{204}H_{262}N_{62}O_{135}P_{20}$	6362.072
C	gugtttgutcttcttcttgcg	$C_{204}H_{261}N_{61}O_{136}P_{20}$	6363.057

The single constituents A to C were combined according to the 10 combinations listed in Table 4-4. To verify the accuracy of ion count estimation, the ion counts of constituents A, B, and C were mixed at a ratio of 20,000:2,000. This was because we wanted to validate if our proposed algorithm tends to provide moderate ratios of multiple constituents even when their actual ratios were highly imbalanced. When the ratio of ion counts between constituents was 10:1, the algorithm should not excessively provide less imbalanced ratios. This setup enabled the analysis of complex mixtures consisting of a few constituents.

Table 4-4. Combinations of constituents when generating spectra.

Mixture No.	Ion Counts		
	Constituents A	Constituents B	Constituents C
1	20 000	20 000	20 000
2	20 000	20 000	2 000
3	20 000	2 000	20 000
4	2 000	20 000	20 000
5	20 000	2 000	2 000
6	2 000	20 000	2 000
7	2 000	2 000	20 000
8	20 000	20 000	-
9	20 000	2 000	-
10	2 000	20 000	-

We set the number of chargeable sites of constituent j, l_j, to 224 and the charge rate of constituent j, v_j, to 0.035. This was done to ensure that the generated spectra closely resembled real data. Then, we generated the test spectra listed in Table 4-4.

Next, we generated the MS/MS spectra of these mixtures. The sequences, molecular formulas, monoisotopic masses, and conversion rates of the fragments generated from the dissociation of constituents A, B, and C are defined in Table 4-5. The MS/MS spectra were generated using these parameters. This time, we selected five peaks in ascending order of m/z from the most prominent isotopic distribution, based on practical memory usage constraints. Additionally, and we assumed two fragment constituents, informed by prior knowledge of dissociation behavior. Thus, d_{max} was 5, and f_{max} was 2.

Table 4-5. Settings for constituent spectrum generation.

Precursor	Fragment ID	Sequence	Molecular Formula	Monoisotopic Mass m_j' [Da]	Conversion Rate ρ
A	F1	gcgtt	$C_{49}H_{63}N_{17}O_{30}P_4$	1494.077	0.3
	F2	tgctcttct	$C_{87}H_{114}N_{24}O_{57}P_8$	2655.810	0.3
	F3	tcttgcg	$C_{68}H_{88}N_{22}O_{43}P_6$	2087.450	0.3
B	F1	gcgtt	$C_{49}H_{63}N_{17}O_{30}P_4$	1494.077	0.3
	F4	tgutcttct	$C_{87}H_{113}N_{23}O_{58}P_8$	2656.795	0.3
	F3	tcttgcg	$C_{68}H_{88}N_{22}O_{43}P_6$	2087.450	0.3
C	F5	gugtt	$C_{49}H_{63}N_{17}O_{30}P_4$	1495.061	0.3
	F4	tgutcttct	$C_{87}H_{113}N_{23}O_{58}P_8$	2656.795	0.3
	F3	tcttgcg	$C_{68}H_{87}N_{21}O_{44}P_6$	2088.435	0.3

4.3.3. Evaluation of Accuracy in Estimated Constituent Counts

We estimated the optimal parameters for assumed constituent count models. Table 4-6 presents the logarithm of the maximum posterior probabilities of each model. By selecting the constituent count that maximizes the logarithm of the posterior probability in each mixture, we estimated the number of constituents present in each mixture. Our method successfully estimated the true number of constituents in 80% of cases (8 out of 10 mixture data). In the two cases where estimation failed, it is possible that the algorithm converged to a different local minimum. We believe this result is a sufficient benchmark for identifying the presence and number of impurities in pharmaceuticals and implementing appropriate corrective measures.

Table 4-6. Negative logarithmic the maximum posterior probability assuming each constituent count.

Mixture No.	True k	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
1	3	47,105,180,000	47,393,730,000	**47,483,530,000**	47,346,420,000	47,254,300,000
2	3	47,146,890,000	47,366,930,000	**47,449,330,000**	47,313,370,000	47,175,010,000
3	3	47,014,840,000	47,244,320,000	**47,373,050,000**	47,250,990,000	47,086,080,000
4	3	47,131,240,000	47,395,780,000	**47,471,380,000**	47,379,240,000	47,237,030,000
5	3	47,064,820,000	47,151,820,000	**47,280,770,000**	47,011,780,000	47,037,130,000
6	3	46,905,570,000	47,412,680,000	**47,418,450,000**	47,312,960,000	47,170,700,000
7	3	45,634,240,000	**46,312,830,000**	46,127,830,000	46,126,200,000	45,988,160,000
8	2	47,152,970,000	**47,406,670,000**	47,272,770,000	47,361,900,000	47,229,700,000
9	2	47,063,080,000	**47,126,520,000**	47,088,670,000	46,960,900,000	46,818,110,000
10	2	47,119,250,000	47,376,490,000	**47,405,410,000**	47,277,430,000	47,172,650,000

4.3.4. Accuracy of Parameter Estimation

To compare the estimation results, we performed deconvolution on the same mixture data using UniDec, a popular deconvolution software. For this verification, we used UniDec (Version 7.0.1). The specific parameter settings used during this verification are shown in Table 4-7. The Mass Range was aligned to the same range as the proposed method, and Sample Mass Every (Da) was set to 0.1 to ensure sufficient detection of impurities with a difference of 1 Da, as described in Chapter 2. Default values were used for parameters not mentioned.

Table 4-8 shows the optimal monoisotopic mass of the models of the selected number of constituents for each mixture, as described in Table 4-4, estimated by our algorithm. The median error was −0.005 Da, the average error in monoisotopic mass was −0.282 Da, and the maximum error was −1.840 Da, as shown in Table 4-9. The standard deviation

was 0.552 Da. The distribution of these errors is shown in Figure 4-2. As observed in the box plot in Figure 4-2, the errors in the monoisotopic masses estimated by the proposed method are discretely distributed approximately 1 Da apart, corresponding to the mass differences between isotopes. The extreme case of No. 6, which produced the maximum error of −1.840 Da, can also be explained by this discrete distribution. This large error is likely caused by the posterior probabilities of the monoisotopic masses being distributed in a comb-like pattern [24], increasing the chances of the algorithm converging to a local minimum 1-2 steps away. However, no clear trend was observed between the ion count ratios of the constituents and the error magnitudes. Using the mean as the representative value and all data from No. 1 to No. 10, the 95% confidence interval calculated using the t-distribution [50] ranges from −0.721 Da to +0.157 Da. This indicates the method could potentially be used to investigate the causes of impurities that occur with a difference of 1 Da [51], [52].

However, the estimated ion counts for each constituent showed errors with a median of 1.1 times the true values, averaging up to twice the true values, with some errors reaching up to twelve times the true values, as shown in Table 4-10. This discrepancy was thought to be due to the trade-off relationship between the ion counts of different constituents; that was, a decrease in the ion count of one constituent was compensated by an increase in another. This was further supported by the fact that the average error across the total ion counts of all constituents stabilized at 8% of the true value. For instance, the standard for total desamido impurities and total impurities in injectable glucagon were, respectively, below 14% and 31%. Therefore, the accuracy of ion count estimation in our proposed method was insufficient to assess the impact of impurities.

The accuracy of estimating the number of constituents was 40% (4 out of 10). This

was obtained by comparing the number of estimated monoisotopic masses output by UniDec with the true number of constituents, as shown in Table 4-8. This was thought to be because the UniDec algorithm, which iterated through multiple deconvolutions to arrive at the number of constituents, did not necessarily guarantee the accuracy of the constituent count. Note that using UniDec to determine the number of constituents was not its intended application. The median error of the monoisotopic mass estimated using UniDec was −0.008 Da, which is slightly worse than that of the proposed method. On the other hand, the average error was 0.091 Da, and the maximum error was 0.993 Da, both slightly better than those of the proposed method. However, in principle, accurate estimation on the monoisotopic mass required precise identification of the number of constituents. As shown in Table 4-11, the error in estimating the number of ions was, on average, 3.2 times the true value and up to 17 times at maximum. This result was not better than that of the proposed method.

Table 4-7. UniDec setting parameters.

Parameter		Setting value
UniDec Parameters	Charge Range	1 - 20
	Mass Range	6000 - 6800
	Sample Mass Every (Da)	0.1
Additional Deconvolution Parameters	Isotopes	Mono
Peak Selection and Plotting	Peak Detection Range (Da)	0.1
	Peak Detection Threshold	0.1

*The other parameters were set at their default values.

Table 4-8 (Part 1). Optimal monoisotopic masses of the model with the maximum posterior probability.

Mixture No.	TRUE	SAI		UniDec	
	Mass [Da]	Mass [Da]	Error [Da]	Mass [Da]	Error [Da]
1	6361.088	6361.086	-0.002	6361.070	-0.018
	6362.072	6362.273	0.201	6362.070	-0.002
	6363.057	6363.055	-0.002	-	-
2	6361.088	6361.084	-0.004	6361.080	-0.008
	6362.072	6362.277	0.205	6362.070	-0.002
	6363.057	6363.056	-0.001	-	-
3	6361.088	6361.099	0.011	6361.080	-0.008
	6362.072	6362.059	-0.013	6362.070	-0.002
	6363.057	6363.265	0.208	-	-
4	6361.088	6360.231	-0.857	6361.070	-0.018
	6362.072	6362.066	-0.006	6362.060	-0.012
	6363.057	6363.054	-0.003	6364.050	0.993
5	6361.088	6360.146	-0.942	6361.080	-0.008
	6362.072	6361.073	-0.999	-	-
	6363.057	6363.282	0.225	-	-

Table 4-8 (Part 2). Optimal monoisotopic masses of the model with the maximum posterior probability.

Mixture No.	TRUE	SAI		UniDec	
	Mass [Da]	Mass [Da]	Error [Da]	Mass [Da]	Error [Da]
6	6361.088	6359.248	-1.840	6361.070	-0.018
	6362.072	6361.069	-1.003	6362.070	-0.002
	6363.057	6362.068	-0.989	-	-
7	6361.088	-	-	6361.060	-0.028
	6362.072	6362.064	-0.008	6362.060	-0.012
	6363.057	6363.264	0.207	6364.050	0.993
8	6361.088	6360.224	-0.864	6361.080	-0.008
	6362.072	6361.072	-1.000	6362.080	0.008
9	6361.088	6361.102	0.014	6361.090	0.002
	6362.072	6362.041	-0.031	-	-
10	-	6360.075	-	-	-
	6361.088	6361.071	-0.017	6361.070	-0.018
	6362.072	6362.257	0.185	6362.070	-0.002

Table 4-9. Statistical summary of monoisotopic mass estimation errors for SAI and UniDec.

	SAI Error [Da]	UniDec Error [Da]
Max.	0.225	0.993
Min.	-1.840	-0.028
Average	-0.282	0.091
Median	-0.005	-0.008
SD	0.552	0.301

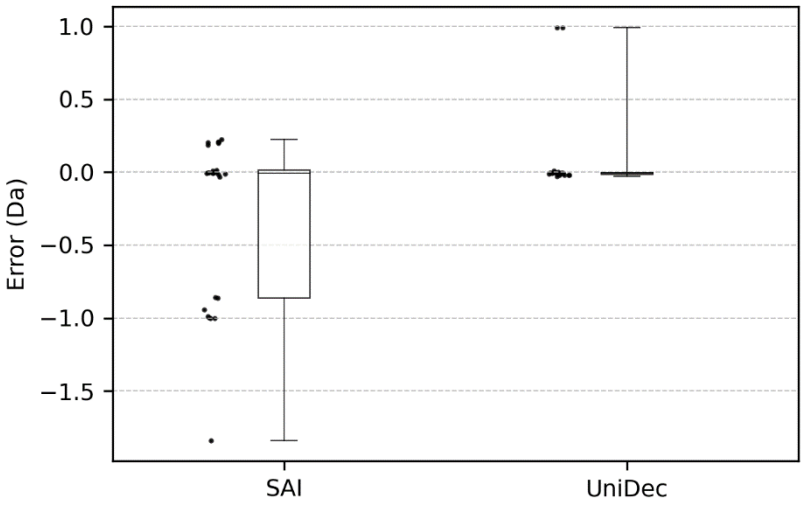

Figure 4-2. Distribution of errors in the estimated monoisotopic masses. (Excluding points that the algorithm could not estimate.)

Table 4-10 (Part 1). Optimal ion counts and relative quantities of the model with the maximum posterior probability.

Mixture No.	TRUE	SAI		UniDec	
	Count	Count	Error [%]	Relative Quantity	Error [%]
1	20000	33690	68.4	100.0	200.0
	20000	8179	-59.1	41.1	23.2
	20000	22228	11.1	-	-
2	20000	31058	55.3	100.0	110.0
	20000	5900	-70.5	18.0	-62.3
	2000	8098	304.9	-	-
3	20000	13215	-33.9	100.0	110.0
	2000	26190	1209.5	34.1	615.7
	20000	5580	-72.1	-	-
4	2000	10643	432.1	85.8	1700.8
	20000	19684	-1.6	100.0	110.0
	20000	17031	-14.8	14.6	-69.4
5	20000	6889	-65.6	100.0	20.0
	2000	16208	710.4	-	-
	2000	2328	16.4	-	-

Table 4-10 (Part 2). Optimal ion counts and relative quantities of the model with the maximum posterior probability.

Mixture No.	TRUE	SAI		UniDec	
	Count	Count	Error [%]	Relative Quantity	Error [%]
6	2000	5143	157.2	100.0	1100.0
	20000	3697	-81.5	56.4	-32.3
	2000	17125	756.3	-	-
7	2000	-	-	57.0	583.5
	2000	22439	1022.0	100.0	1100.0
	20000	3287	-83.6	26.1	-68.6
8	20000	10689	-46.6	100.0	100.0
	20000	34062	70.3	15.0	-70.0
9	20000	18739	-6.3	100.0	10.0
	2000	3369	68.4	-	-
10	-	616	-	-	-
	2000	21303	965.1	100.0	1000.0
	20000	3867	-80.7	48.6	-46.6

Table 4-11. Statistical summary of ion counts estimation errors for SAI and UniDec.

	SAI Error[%]	UniDec Error[%]
Max.	1209.487	1700.772
Min.	-83.564	-69.964
Average	201.203	321.698
Median	13.772	105.000
SD	383.640	503.549

For reference, Figure 4-3 presents a comparison between the spectrum of Mixture No.1 and the spectrum reconstructed from its estimated parameters. Figure 4-3 (a) provides an overview of the charge distribution, while Figure 4-3 (b) offers a detailed view of the isotopic distribution. The gray vertical dashed lines in Figure 4-3 (a) and (b) indicate the m/z of the fragmented ions. Additionally, Figure 4-3 (c) and (d) display the MS/MS spectrum of the fragmented ion groups and its detailed view, respectively. The five graphs correspond to the five peaks in Figure 4-3 (b), each representing the MS/MS spectra of the ions contained in those peaks when they are fragmented. These results demonstrated that the generated spectrum closely matched with the observed data. Furthermore, the appearance of the MS/MS spectra was consistent with findings from prior research cited in references [53]–[55].

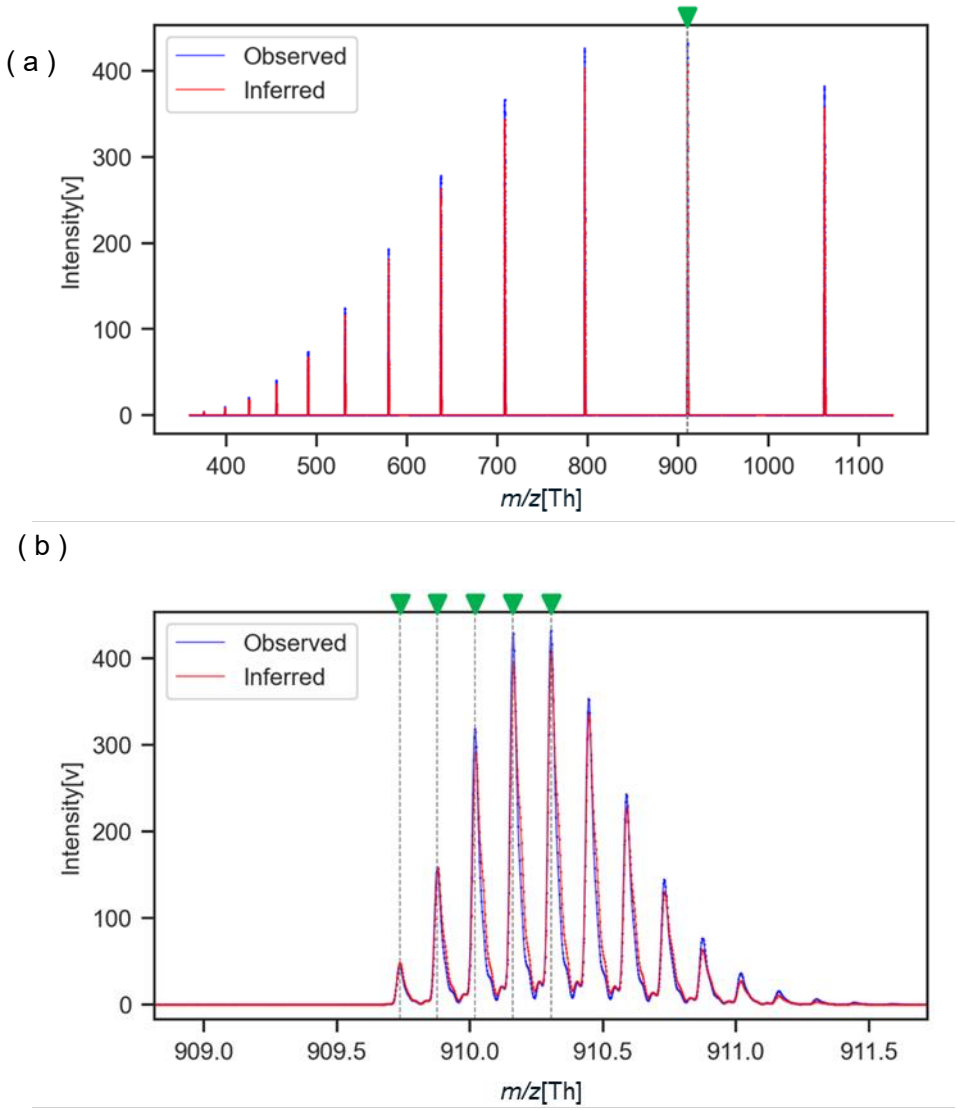

Figure 4-3 (Part 1). Comparison of observed and estimated spectra for Mixture No. 1.

(a) MS spectrum overall view; and (b) MS spectrum enlarged view.

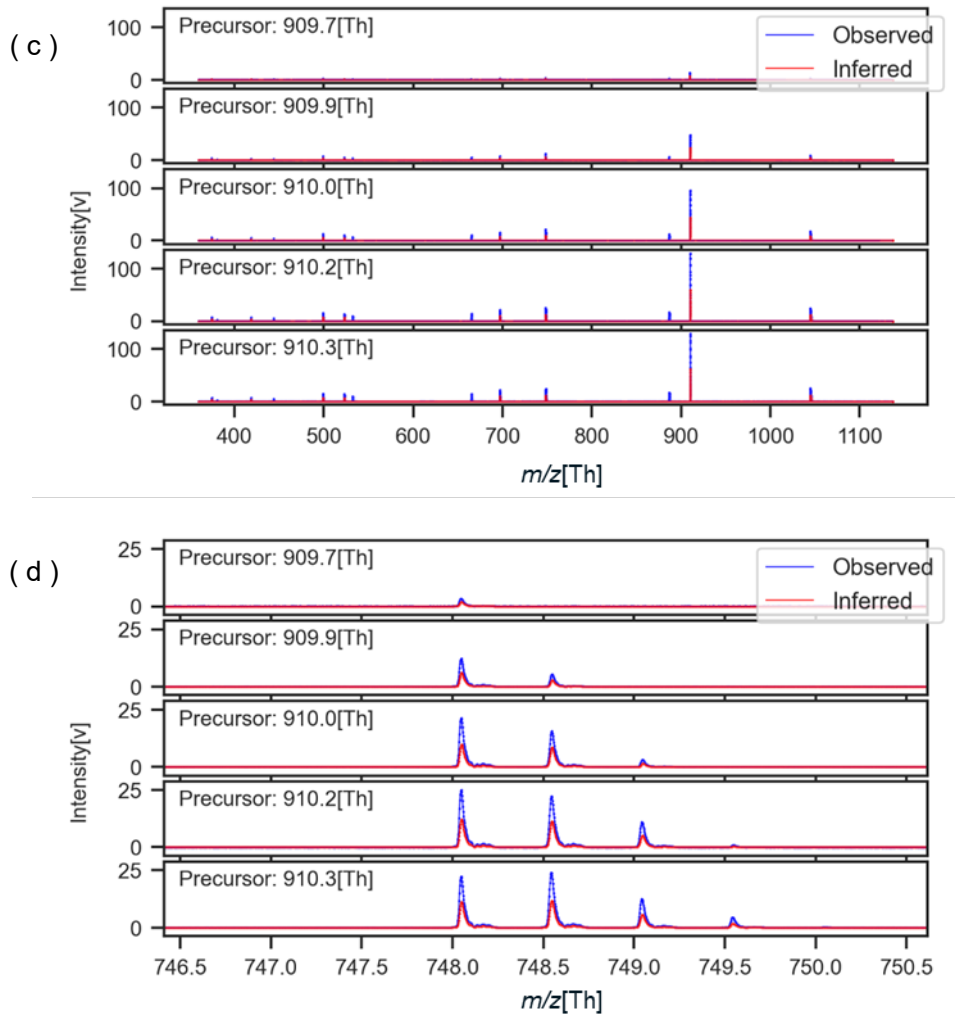

Figure 4-3 (Part 2). Comparison of observed and estimated spectra for Mixture No. 1. (c) MS/MS spectrum overall view; and (d) MS/MS spectrum enlarged view.

4.4. Discussion

We confirmed that our proposed method allowed for accurate estimation of parameters such as monoisotopic masses from simulated MS and MS/MS data of the nucleic acid drug Fomivirsen and its impurities, and it also successfully selected the correct number of constituents with an 80% accuracy, even in datasets containing more challenging ion count ratios of 10:1. These results were better compared to the 40% accuracy rate achieved with UniDec. This success was attributed to our approach of creating models for each constituent count, enabling comparative evaluation and selection of models for each constituent count. This capability suggests the presence of impurities in pharmaceuticals and could aid in the search for better synthesis conditions for medium to high molecular weight drugs, as well as in quality assurance in manufacturing facilities.

As shown in Table 4-8, we were able to estimate monoisotopic mass with higher accuracy than previous chapter's studies [24], with an average estimation error of 0.282 Da, which was an improvement over the 1.348 Da error reported in prior research. Although this accuracy was slightly inferior to UniDec's 0.091 Da, it was sufficient for distinguishing differences as small as 1 Da due to deamidation. We believe this improvement is due to the incorporation of the MS/MS spectra into the physical model, which increased the constraints on the model's degrees of freedom. Additionally, the use of the correlation-based likelihood contributed to more stringent constraints on the spectral shape.

As indicated in Table 4-10, the estimated ion quantities for each constituent showed an average relative error of twice the true value. Although a direct comparison with the prior studies, which used a 1:1 mixing ratio, was not straightforward due to our use of a 10:1 ratio, the results were favorable compared to UniDec, which had an average error of

3.2 times the true value. The errors observed in our proposed method might result from a trade-off among the ion quantities of each constituent, where a decrease in one was offset by an increase in another. Despite our expectations that incorporating MS/MS spectra would tighten estimation constraints and enhance both mass and ion quantity accuracy, the performance fell short of expectations, failing to reduce the relative error to below the 10% threshold required for impurity analysis in nucleic acid drugs. A possible solution to these issues would be to represent the ion quantities as probability distributions. By accounting for the uncertainty in the ion quantities of constituents in the sample, an improvement in estimation accuracy was expected.

Despite the sixfold increase in data volume—comprising one MS spectrum and five MS/MS spectra corresponding to five peaks—the analysis time per data point remained 13 hours. While this duration did not match the few seconds required by UniDec, it was less than half the time required by our previous method [24] described in Chapter 2 that use MCMC. Replacing the estimation mechanism with a neural network or similar approaches is one potential solution for achieving faster processing.

4.5. Conclusion of This Chapter

In this chapter, we assumed the numbers of constituents in a given sample and created models of MS and MS/MS mass spectrometry based on parameters such as monoisotopic mass and ion quantity. We then applied our proposed method from Chapter 3, Spectral Annealing Inference (SAI), which effectively seeks the maximum posterior probability by optimizing parameters for the observed data. After obtaining the maximum posterior probability for each constituent count model, we selected the model that had the highest maximum posterior probability across all models. As a result, we successfully estimated

the number of constituents and simultaneously estimated the monoisotopic mass with high accuracy. We think this achievement is attributed to the increased amount of constraint information provided by leveraging MS/MS spectra. While the accuracy of monoisotopic mass estimation was improved, future challenges include improving the accuracy of ion count estimation and achieving further computational speedup.

Chapter 5.

Conclusion and Future Challenges

Our objective was to accurately determine the number of constituents, monoisotopic masses, and ion counts from mass spectrometry (MS) data to contribute to impurity detection and analysis in pharmaceutical development and manufacturing.

In this study, we first constructed mass spectrometry models for each possible number of constituents and applied a Bayesian inference framework. This allowed us to develop a methodology for estimating the most probable number of constituents along with their corresponding parameters, such as monoisotopic masses and ion counts.

In Chapter 2, to handle systems with sparse posterior probability distributions, which are characteristic of mass spectrometry data, we initially employed MCMC (Markov Chain Monte Carlo) for parameter exploration. While this approach successfully determined the optimal number of constituents and associated parameters, such as monoisotopic masses, it required an extensive amount of computation time. Furthermore, the accuracy of monoisotopic mass and ion count estimation was insufficient for achieving the goal of detecting impurities with a mass difference of 1 Da.

In Chapter 3, we addressed this issue by developing a faster parameter exploration method to replace MCMC. We introduced a novel approach named Spectral Annealing Inference (SAI), which involves convolving spectra with a PSF (Point Spread Function) that progressively approaches a delta function, thereby enabling rapid and convergent estimation. As a result, estimation times were reduced from 50 hours to just 15 minutes.

However, challenges remained in improving the accuracy of estimated monoisotopic masses and ion counts.

To address these challenges, in Chapter 4, we incorporated MS/MS information and refined the likelihood function to enhance estimation accuracy. By mathematically combining the MS and MS/MS models, we utilized MS/MS spectra to improve the parameter estimation of MS spectra. As a result, monoisotopic mass estimation accuracy was improved to a level sufficient for distinguishing mass differences of 1 Da.

The results of this development will contribute to the detection of impurities, the evaluation of their impact, and the investigation of their causes in the manufacturing and development of biopharmaceuticals.

Nevertheless, challenges still remain in enhancing the accuracy of ion count estimation. Currently, there are no established guidelines for the quality control of nucleic acid-based pharmaceuticals [56], [57]. Therefore, the results of this study hold a certain significance for identifying the presence and quantity of impurities in pharmaceuticals and implementing appropriate corrective measures. That said, for future use in quality control, an estimation accuracy of less than 10% will likely be required.

Additionally, as the number of constituents increases, the computational time also grows, posing a limitation of the proposed method. Replacing the estimation framework with neural networks or similar advanced techniques to handle multi-constituent systems is a promising direction for future development. Such advancements will be crucial for expanding the application of this method to fields like metabolomics and environmental analysis.

Furthermore, the SAI method developed in this study may also be applicable to other spectroscopic techniques that produce sparse and complex signals, such as Nuclear

Magnetic Resonance (NMR) spectroscopy [58], Raman spectroscopy [59], and various X-ray-based methods. NMR is widely used in structural biology and organic chemistry to analyze molecular structures based on nuclear spin interactions. Raman spectroscopy provides information on vibrational modes, which is useful for material characterization.

Additionally, techniques like X-ray Photoelectron Spectroscopy (XPS) [60], X-ray Diffraction (XRD) [61], X-ray Fluorescence (XRF, also known as Energy-Dispersive X-ray Spectroscopy, EDX) [62], and X-ray Absorption Spectroscopy (XAS) [63] are commonly used in material science and chemistry. XPS analyzes surface composition by measuring the kinetic energy of emitted photoelectrons. XRD identifies crystal structures through diffraction patterns. XRF (EDX) determines elemental composition based on characteristic X-ray emissions, and XAS provides insight into local electronic structures and bonding environments.

By adapting SAI to these techniques, it may be possible to improve the extraction of physical parameters from spectral data. This approach could be useful in fields such as metabolomics, environmental analysis, and material characterization, where precise parameter estimation is important. For instance, SAI might help in analyzing NMR spectra for protein structure determination or in processing Raman and infrared spectroscopy [64] data for quality control. Its application to XPS, XRD, XRF (EDX), and XAS could also support more detailed structural and elemental analysis.

Further investigation is needed to assess the feasibility and effectiveness of applying SAI to these areas. However, the methodology presented in this study provides a potential foundation for refining spectral analysis across various analytical techniques.

List of Publications

Peer-reviewed Journal Article	Related Chapter
T. Tomono, S. Hara, Y. Nakai, K. Takahara, J. Iida, and T. Washio, "A Bayesian approach for constituent estimation in nucleic acid mixture models," *Front. Anal. Sci.*, vol. 3:1301602, Jan. 2024.	Chapter 2
T. Tomono, S. Hara, J. Iida, and T. Washio, "Enhancing constituent estimation in nucleic acid mixture models using spectral annealing inference and MS/MS information," *Front. Anal. Sci.*, vol. 5:1494615, Feb. 2025.	Chapter 4

Peer-reviewed International Conference Proceedings	Related Chapter
T. Tomono, S. Hara, J. Iida, and T. Washio, "Advanced stochastic variational inference for accurate constituent estimation in nucleic acid mixture models," in *Proceedings of the 72nd ASMS Conference on Mass Spectrometry and Allied Topics*, p. 127, Anaheim, CA, Jun. 2024.	Chapter 3
T. Tomono, S. Hara, J. Iida, and T. Washio, "Extending a Bayesian method for inferring constituent information using MS/MS data," in *SciX 2024 abstract book*, pp. 252–253, Raleigh, NC, Oct. 2024.	Chapter 4

References

[1] Y. S. Sanghvi, “A status update of modified oligonucleotides for chemotherapeutics applications,” *Curr. Protoc. Nucleic Acid Chem.*, vol. 4, no. 1, pp. 1–22, Sep. 2011.

[2] W. C. Weinberg *et al.*, “Development and regulation of monoclonal antibody products: Challenges and opportunities,” *Cancer Metastasis Rev.*, vol. 24, no. 4, pp. 569–584, Dec. 2005.

[3] S. Tamara, M. A. den Boer, and A. J. R. Heck, “High-resolution native mass spectrometry,” *Chem. Rev.*, vol. 122, no. 8, pp. 7269–7326, Apr. 2022.

[4] R. Pecori, S. Di Giorgio, J. Paulo Lorenzo, and F. Nina Papavasiliou, “Functions and consequences of AID/APOBEC-mediated DNA and RNA deamination,” *Nat. Rev. Genet.*, vol. 23, no. 8, pp. 505–518, Aug. 2022.

[5] D. Capaldi *et al.*, “Quality aspects of oligonucleotide drug development: Specifications for active pharmaceutical ingredients,” *Drug Inf. J.*, vol. 46, no. 5, pp. 611–626, Sep. 2012.

[6] A. G. Marshall, “Fourier transform ion cyclotron resonance mass spectrometry,” *Acc. Chem. Res.*, vol. 18, no. 10, pp. 316–322, Oct. 1985.

[7] A. G. Marshall, C. L. Hendrickson, and G. S. Jackson, “Fourier transform ion cyclotron resonance mass spectrometry: a primer,” *Mass Spectrom. Rev.*, vol. 17, no. 1, pp. 1–35, Jan. 1998.

[8] E. N. Nikolaev, Y. I. Kostyukevich, and G. N. Vladimirov, “Fourier transform ion cyclotron resonance (FT ICR) mass spectrometry: Theory and simulations: FT ICR MS,” *Mass Spectrom. Rev.*, vol. 35, no. 2, pp. 219–258, Mar. 2016.

[9] J. Zhang, E. Gonzalez, T. Hestilow, W. Haskins, and Y. Huang, “Review of peak detection algorithms in liquid-chromatography-mass spectrometry,” *Curr. Genomics*, vol. 10, no. 6, pp. 388–401, Sep. 2009.

[10] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse, “Electrospray ionization for mass spectrometry of large biomolecules,” *Science*, vol. 246, no. 4926, pp. 64–71, Oct. 1989.

[11] A. P. Bruins, “Mechanistic aspects of electrospray ionization,” *J. Chromatogr. A*, vol. 794, no. 1–2, pp. 345–357, Jan. 1998.

[12] M. Wilm, “Principles of electrospray ionization,” *Mol. Cell. Proteomics*, vol. 10, no. 7, p. M111.009407, Jul. 2011.

[13] M. W. Senko, S. C. Beu, and F. W. McLaffertycor, “Determination of monoisotopic

masses and ion populations for large biomolecules from resolved isotopic distributions," *J. Am. Soc. Mass Spectrom.*, vol. 6, no. 4, pp. 229–233, Apr. 1995.

[14] S. Dasari, P. A. Wilmarth, A. P. Reddy, L. J. G. Robertson, S. R. Nagalla, and L. L. David, "Quantification of isotopically overlapping deamidated and 18O-labeled peptides using isotopic envelope mixture modeling," *J. Proteome Res.*, vol. 8, no. 3, pp. 1263–1270, Mar. 2009.

[15] A. G. Ferrige, M. J. Seddon, B. N. Green, S. A. Jarvis, J. Skilling, and J. Staunton, "Disentangling electrospray spectra with maximum entropy," *Rapid Commun. Mass Spectrom.*, vol. 6, no. 11, pp. 707–711, Nov. 1992.

[16] R. L. Tranter, *Design and Analysis in Chemical Research*. Hoboken, NJ, USA: John Wiley & Sons, 2000.

[17] A. Ferrige, S. Ray, R. Alecio, S. Ye, and K. Waddell, "Electrospray-MS charge deconvolutions without compromise – an enhanced data reconstruction algorithm utilising variable peak modelling," in *Proceedings of the 51st ASMS Conference on Mass Spectrometry and Allied Topics*, Montreal, QC, Canada, Jun. 2024.

[18] Z. Zhang, S. Guan, and A. G. Marshall, "Enhancement of the effective resolution of mass spectra of high-mass biomolecules by maximum entropy-based deconvolution to eliminate the isotopic natural abundance distribution," *J. Am. Soc. Mass Spectrom.*, vol. 8, no. 6, pp. 659–670, Jun. 1997.

[19] A. D. Rolland and J. S. Prell, "Approaches to heterogeneity in native mass spectrometry," *Chem. Rev.*, vol. 122, no. 8, pp. 7909–7951, Apr. 2022.

[20] M. T. Marty, A. J. Baldwin, E. G. Marklund, G. K. A. Hochberg, J. L. P. Benesch, and C. V. Robinson, "Bayesian deconvolution of mass and ion mobility spectra: From binary interactions to polydisperse ensembles," *Anal. Chem.*, vol. 87, no. 8, pp. 4370–4376, Apr. 2015.

[21] M. T. Marty, "A universal score for deconvolution of intact protein and native electrospray mass spectra," *Anal. Chem.*, vol. 92, no. 6, pp. 4395–4401, Mar. 2020.

[22] W. H. Richardson, "Bayesian-based iterative method of image restoration," *J. Opt. Soc. Am.*, vol. 62, no. 1, p. 55, Jan. 1972.

[23] L. B. Lucy, "An iterative technique for the rectification of observed distributions," *Astron. J.*, vol. 79, p. 745, Jun. 1974.

[24] T. Tomono, S. Hara, Y. Nakai, K. Takahara, J. Iida, and T. Washio, "A Bayesian approach for constituent estimation in nucleic acid mixture models," *Front. Anal. Sci.*, vol. 3:1301602, Jan. 2024.

[25] T. Tomono, S. Hara, J. Iida, and T. Washio, "Advanced stochastic variational inference for accurate constituent estimation in nucleic acid mixture models," in

Proceedings of the 72nd ASMS Conference on Mass Spectrometry and Allied Topics, p. 127, Anaheim, CA, Jun. 2024.

[26] T. Tomono, S. Hara, J. Iida, and T. Washio, "Extending a Bayesian method for inferring constituent information using MS/MS data," in *SciX 2024 abstract book*, pp. 252–253, Raleigh, NC, Oct. 2024.

[27] T. Tomono, S. Hara, J. Iida, and T. Washio, "Enhancing constituent estimation in nucleic acid mixture models using spectral annealing inference and MS/MS information," *Front. Anal. Sci.*, vol. 5:1494615, Feb. 2025.

[28] M. D. Hoffman and A. Gelman, "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1593–1623, Apr. 2014.

[29] S. Kirkpatrick, C. D. Gelatt Jr, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, May 1983.

[30] V. Glivenko, "Sulla determinazione empirica delle leggi di probabilita," *Gion. Ist. Ital. Attauri*, vol. 4, pp. 92–99, 1933.

[31] F. P. Cantelli, "Considerazioni sulla legge uniforme dei grandi numeri e sulla generalizzazione di un fondamentale teorema del sig," *Paul Levy. Giorn. Ist. Ital. Attuari*, vol. 4, no. 3, pp. 327–350, 1933.

[32] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[33] A. A. Neath and J. E. Cavanaugh, "The Bayesian information criterion: background, derivation, and applications," *WIREs Computational Statistics*, vol. 4, no. 2, pp. 199–203, Mar. 2012.

[34] R. M. Neal, *Handbook of Markov Chain Monte Carlo, chapter 5: MCMC Using Hamiltonian Dynamics*. Boca Raton, FL, USA: CRC Press, 2011.

[35] M. Betancourt, "A conceptual introduction to hamiltonian monte carlo," *arXiv preprint*, Jan. 2017.

[36] D. Wingate and T. Weber, "Automated variational inference in probabilistic programming," *arXiv preprint*, Jan. 2013.

[37] R. Ranganath, S. Gerrish, and D. Blei, "Black box variational inference," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, Reykjavik, Iceland, Apr. 2014.

[38] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv preprint*, Dec. 2013.

[39] C. K. Birdsall and A. B. Langdon, *Plasma physics via computer simulation*. Boca Raton, FL, USA: CRC Press, 2004.

[40] R. L. Wasserstein, M. H. Kalos, and P. A. Whitlock, “Monte Carlo methods, volume 1: Basics,” *Technometrics*, vol. 31, no. 2, p. 269, May 1989.

[41] L. Tierney, “Markov Chains for exploring posterior distributions,” *Ann. Stat.*, vol. 22, no. 4, pp. 1701–1728, Dec. 1994.

[42] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: A review for statisticians,” *J. Am. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, Apr. 2017.

[43] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint*, Dec. 2014.

[44] C. M. Perry and J. A. Balfour, “Fomivirsen,” *Drugs*, vol. 57, no. 3, pp. 375–380, Mar. 1999.

[45] National Institute of Standards and Technology (NIST), “Atomic weights and isotopic compositions for all elements,” *National Institute of Standards and Technology (NIST)*. [Online]. Available: https://www.nist.gov/pml/atomic-weights-and-isotopic-compositions. [Accessed: 27-Jan-2025].

[46] J. Gao, H. Choudhry, and W. Cao, “Apolipoprotein B mRNA editing enzyme catalytic polypeptide-like family genes activation and regulation during tumorigenesis,” *Cancer Sci.*, vol. 109, no. 8, pp. 2375–2382, Aug. 2018.

[47] J. Stavnezer, “Complex regulation and function of activation-induced cytidine deaminase,” *Trends Immunol.*, vol. 32, no. 5, pp. 194–201, May 2011.

[48] Z. Bao, Y.-C. Cheng, M. Z. Luo, and J. Y. Zhang, “Comparison of the purity and impurity of glucagon-for-injection products under various stability conditions,” *Sci. Pharm.*, vol. 90, no. 2, p. 32, May 2022.

[49] K. V. Mardia and P. E. Jupp, *Directional Statistics*. Hoboken, NJ, USA: Wiley & Sons, Limited, John, 2008.

[50] Student, “The probable error of a mean,” *Biometrika*, vol. 6, no. 1, p. 1, Mar. 1908.

[51] C. Rentel, J. DaCosta, S. Roussis, J. Chan, D. Capaldi, and B. Mai, “Determination of oligonucleotide deamination by high resolution mass spectrometry,” *J. Pharm. Biomed. Anal.*, vol. 173, pp. 56–61, Sep. 2019.

[52] S. Pourshahian, “Therapeutic oligonucleotides, impurities, degradants, and their characterization by mass spectrometry,” *Mass Spectrom. Rev.*, vol. 40, no. 2, pp. 75–109, Mar. 2021.

[53] M. A. van Agthoven, Y. P. Y. Lam, P. B. O’Connor, C. Rolando, and M.-A. Delsuc, “Two-dimensional mass spectrometry: new perspectives for tandem mass spectrometry,” *Eur. Biophys. J.*, vol. 48, no. 3, pp. 213–229, Apr. 2019.

[54] L. J. Szalwinski, D. T. Holden, N. M. Morato, and R. G. Cooks, “2D MS/MS spectra recorded in the time domain using repetitive frequency sweeps in linear quadrupole

ion traps," *Anal. Chem.*, vol. 92, no. 14, pp. 10016–10023, Jul. 2020.
[55] L. E. Gonzalez, L. J. Szalwinski, T. C. Sams, E. T. Dziekonski, and R. G. Cooks, "Metabolomic and lipidomic profiling of Bacillus using two-dimensional tandem mass spectrometry," *Anal. Chem.*, vol. 94, no. 48, pp. 16838–16846, Dec. 2022.
[56] International Council for Harmonisation (ICH), *ICH Q2(R2): Validation of Analytical Procedures*. International Council for Harmonisation, 2023.
[57] World Health Organization (WHO), "Good practices for pharmaceutical quality control laboratories," *World Health Organ. Tech. Rep. Ser.*, no. 1052, p. 4, Apr. 2024.
[58] F. Bloch, "Nuclear induction," *Phys. Rev.*, vol. 70, no. 7–8, pp. 460–474, Oct. 1946.
[59] D. A. Long, *Raman spectroscopy*. New York, NY, USA: McGraw-Hill, 1977, p. 310.
[60] K. Siegbahn, *ESCA applied to free molecules*. Amsterdam, The Netherlands: North-Holland Publishing Company, 1969.
[61] B. D. Cullity and S. R. Stock, *Elements of X-ray diffraction: Pearson new international edition*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2001.
[62] R. Jenkins, *X-Ray fluorescence spectrometry*, 2nd ed. Hoboken, NJ, USA: John Wiley & Sons, 1999.
[63] D. C. Koningsberger and R. Prins, *X-ray absorption: principles, applications, techniques of EXAFS, SEXAFS and XANES*. New York, NY: John Wiley and Sons Inc., 1988.
[64] L. J. Bellamy, *The infra-red spectra of complex molecules*. New York, NY: Springer, 1975.