A compositional data model to predict the isotope distribution for average peptides using a compositional spline model

AGTEN, Annelies; VILENNE, Frédérique; PROSTKO, Piotr & VALKENBORG, Dirk
(2024) A compositional data model to predict the isotope distribution for average peptides using a compositional spline model. In: Proteomics, 24 (8) (Art N° 2300154).

DOI: 10.1002/pmic.202300154
Handle: http://hdl.handle.net/1942/42029

A compositional data model to predict the isotope distribution for average peptides using a compositional spline model.

Agten Annelies[1*], Vilenne Frédérique[1, 2], Prostko Piotr[1], Valkenborg Dirk[1]

[1]Data Science Institute, Hasselt University, Diepenbeek, Belgium
[2] Health, Flemish Institute for Technological Research (VITO), Mol, Belgium

*Corresponding author:
annelies.agten@uhasselt.be
Universiteit Hasselt – Campus Diepenbeek, Agoralaan Gebouw D, B-3590 Diepenbeek

Abbreviations

ALR – Additive Log-Ratio
FDR – False Discovery Rate
LC – Liquid Chromatography
MSE – Mean Squared Error
ppm – part per million
PRIDE – Proteomics Identifications Database
UPS – Universal Proteomics Standard

Keywords

Average peptide
Compositional data
Isotope distribution
Spline regression
Sulphur prediction

Total number of words

Including Abstract, Statement of significance, and References: 7592

Abstract

We propose an updated approach for approximating the isotope distribution of average peptides given their monoisotopic mass. Our methodology involves in-silico cleavage of the entire UNIPROT database of Human reviewed proteins using Trypsin, generating a theoretical peptide dataset. The isotope distribution is computed using BRAIN. We apply a compositional data modelling strategy that utilizes an additive log-ratio transformation for the isotope probabilities followed by a penalized spline regression. Furthermore, due to the impact of the number of sulphur atoms on the course of the isotope distribution, we develop separate models for peptides containing zero up to five sulphur atoms. Additionally, we propose three methods to estimate the number of sulphur atoms based on an observed isotope distribution. The performance of the spline models and the sulphur prediction approaches is evaluated using a mean squared error and a modified Pearson's χ^2 goodness-of-fit measure on an experimental UPS2 data set. Our analysis reveals that the variability in spectral accuracy, i.e., the variability between MS1 scans, contributes more to the errors than the approximation of the theoretical isotope distribution by our proposed average peptide model. Moreover, we find that the accuracy of predicting the number of sulphur atoms based on the observed isotope distribution is limited by measurement accuracy.

Statement of Significance

The observed isotope distribution plays a crucial role in pre-processing mass spectrometry data for shotgun proteomics. In this manuscript, we introduce an updated approach for approximating the isotope distribution of average peptides based on their monoisotopic mass. With our method we aim to address some of the limitations in previous studies. In this regard, we utilize the latest protein database and elemental isotope definition instead of relying on an Averagine model or theoretical peptides. Our proposed methodology incorporates a data modelling approach that properly accounts the compositional nature of the isotope probabilities. Moreover, we employ penalized spline regression to model the isotope distribution as a function of the monoisotopic mass, which offers a precise and flexible means of modelling the data. Furthermore, recognizing the influence of sulphur on the isotope distribution, we introduce different models for peptides containing from zero up to five sulphur atoms, resulting in more accurate predictions. Finally, we propose different methods to predict the number of sulphur atoms in a peptide, providing valuable insights for mass spectra identification. By addressing the limitations of previous publications and utilizing advanced modelling techniques, our approach contributes to the advancement of isotope distribution approximation and spectral identification.

Introduction

The observed isotope distribution is a valuable tool for pre-processing mass spectrometry data in shotgun proteomics experiments. Using the observed isotope distribution, the data volume can be significantly reduced, allowing downstream procedures for peptide identification or quantification to focus solely on relevant peptides.[1,2] However, employing the isotope pattern for data clean-up presents a catch-22 situation. In high-throughput and peptide-centric proteomics, mass spectra often contain crowded mixtures of unknown peptide molecules. Therefore, in order to leverage the informative isotope patterns of these molecules, it is necessary to first determine their identity and atomic composition for accurate computation of the isotope distribution.[3,4] This creates a gridlock since the desired improvement in identification relies on the isotope distribution, which, in turn, necessitate knowledge on the molecule's identity.

Fortunately, applications utilizing the isotope pattern can work with an approximation of the isotope distribution, rather than an exact calculation. Thus, it suffices to provide these applications with an estimate of how an isotope distribution would appear for an average peptide. To this end, several estimation procedures have been developed in the past decades.[5-10] For example, the scaled Averagine method proposed by Senko et al. is based on the multinomial expansion to calculate the average isotopic distribution for peptides of any mass. However, the method is computationally intensive and not accurate enough for low mass peptides.[5,9] Another approach, introduced by Breen et al., approximates the multinomial expansion using a Poisson model.[6] Although the method is fast, it lacks accuracy when dealing with peptides containing sulphur. Moreover, Valkenborg et al. proposed a method that models the consecutive ratios of isotope peaks using a polynomial model, ignoring the compositional nature of the data.[9] A reviews paper that lists several applications and limitations of the isotope distribution has recently been published by Claesen et al.[11]

The current generation of (high-resolution) mass spectrometers have yielded increased sensitivity and improved detection limits. Moreover, time-of-flight instruments, which have regained popularity in recent years, have witnessed substantial improvements in spectral accuracy. Consequently, the isotope distribution of molecules can be measured with greater sensitivity, resolution and mass accuracy. This, along with more comprehensive protein databases, refined elemental isotope definitions, and enhancements in algorithmic and statistical methods, warrants an update to these models to predict the isotope distribution for an average peptide, given a specific mass value. Recently, a compositional model to predict the average isotope distribution of DNA and RNA oligonucleotides was proposed.[12] Although the method is efficient and accurate, it uses a high order polynomial model which is susceptible to boundary effects.

In this manuscript, we propose a similar method, with the modification of using a spline regression approach, to predict the average isotope distribution of peptides, taking into account the number of sulphur atoms contained in the peptide. In order to avoid confusion, before proceeding, a couple of concepts should be defined. Isotopologues are molecular entities sharing the same chemical formula and atomic bonding arrangement, yet differing by the number of neutrons in at least one atom. An example includes CH_4, CH_3D, and CH_2D_2. These variations in isotopic composition give rise to the fine isotope distribution, which characterizes the likelihood of each isotopic variant's occurrence within a molecule. Ignoring small mass deviations from integer values, these isotopic variants can be grouped into aggregated isotopic variants. The aggregated isotope distribution provides information about the number and probabilities of occurrence of these grouped variants. However, throughout the manuscript we will be working with the aggregated isotope distribution and we will refer to the aggregated isotope peaks as isotopic variants. For example, the monoisotopic peak is the first isotopic variant, the fifth isotopic peak aggregated all the M+4 isotopologues.

Our proposed model utilizes tryptic peptides from the UNIPROT database. The theoretical isotope distribution is computed by the BRAIN algorithm[3] incorporating the most recent NIST definition for the elemental isotopes[13]. The resulting isotope information that is compositional in nature, i.e. it sums to one, is modelled using a penalized spline regression approach using the monoisotopic mass as input. Since the number of sulphurs has a large influence on the isotope distribution of peptides, we model the average isotope distribution for each distinct sulphur-containing category found in the peptide database separately, for peptides containing zero up to five sulphurs.

In mass spectrometry data analysis, the molecule under investigation is often unknown. However, knowledge on the numbers of sulphur atoms in a peptide is required to apply the correct model since we propose different models for each sulphur category respectively. Recently, a machine learning approach was proposed to predict the number of sulphur atoms in peptides using the theoretical aggregated isotope distribution.[14] However, this method uses information on the masses and isotope probabilities of the observed isotope distribution. In this manuscript, we propose two methods to predict the number of sulphur atoms based on the observed monoisotopic mass. For the first method, we compute the average isotope distribution given an observed monoisotopic mass using each proposed sulphur model separately. The 'correct' sulphur model is then simply the model that produces the best fit, i.e. the smallest error compared to the observed isotope distribution. The second method is based on Bayes' theorem for conditional probabilities and calculates the probability that a certain observed isotope distribution is generated by a peptide containing zero up to five sulphur atoms.

The methods we propose in this manuscript are all computationally simple and accurate in predicting the expected isotopic distribution for an average peptide and the number of Sulphur atoms in the peptide given a certain monoisotopic mass. The prediction model is available as an online tool at https://dsi-uhasselt.shinyapps.io/pointless4peptides/. The online tool provides the predicted isotope distribution for an average peptide of a particular mass. The web interface presents the isotope information as a table and graphical representation.

Materials and Methods

Experimental Data

In this manuscript, we used a publicly available data set obtained from an LC-MS/MS experiment to validate our findings. The experiment involved measuring a UPS2-kit that consists of a mixture 48 precisely quantified proteins spanning five orders of magnitude ranging from 50 pmol to 500 amol. To obtain peptide samples, the UPS2 proteins in the experiment were enzymatically cleaved using trypsin. The resulting mixture of peptides was then subjected to liquid chromatography (LC) separation for 120 minutes on an LTQ Orbitrap Velos instrument. The data set is accessible through the PRIDE repository under the identifier PXD000331. For a more comprehensive understanding of the measurement process, we recommend consulting the accompanying publication.[15]

The data was measured in duplicates, but for the purpose of this manuscript we focused on a single file (A11-12042.raw file). The Thermo Fisher .raw file was converted into the mzML format using MSConvert (version 3.0.20351) deploying the vendor peak picking option. The experimental data was subjected to database search analysis using MS-GF+ (SearchGUI version 4.1.3) with a reverse target-decoy approach, ensuring a false discovery rate (FDR) of 1%. The precursor mass tolerance was set to 5 ppm, and the fragment mass tolerance was set to 0.6 Da. Carbamidomethyl (C) was set as fixed

modification, oxidation (M) as variable modification. Trypsin is specified as the digestion enzyme, allowing for up to two missed cleavages. The database search results were imported into PeptideShaker, and only spectra identified with 100% confidence were retained. The resulting feature table was then exported for further analysis.

A distinct advantage of using the UPS2 data is that we have precise knowledge about the proteins present in the data, enabling us to anticipate the peptides and their expected masses within the observed spectra. Our objective was to curate an experimental dataset limited to fifty peptides. To select which peptides to include in our validation dataset, we focused on identifying the most intense peptides. Using the extracted ion chromatogram plotting the intensity of the observed signal in a 0.2 Da mass interval around the identified peptide mass and a 30s interval around the retention time we recorded the total ion count. The choice for these parameters is in alignment with the experimental setting described by Ahrne et al. The recorded total ion counts were used to rank and order the spectra from highest to lowest TIC. Additionally, we only retained spectra where the first five peaks of the isotope distribution were observed. Subsequently, from this pool of spectra containing at least five observed isotopic peaks, we selected the top 50 most intense peptides to comprise our validation data set. For each of these chosen peptides, we extracted both the masses and intensities of the first five isotope peaks, using a 5 ppm mass tolerance. All subsequent analyses were performed using R (version 4.2.3).

Theoretical Data

The models presented in this manuscript focus on human data. Although the average peptides are generally very similar across different species, separate models can be developed for each organism. For our theoretical data set, we obtained the entire UNIPROT database, comprising all Human reviewed proteins (20,350 protein entries on 24/4/2020). These proteins were digested in-silico using Trypsin, which is the most commonly used protease in peptide-centric proteomics[16], but specialized models could be conceived for each protease separately. After removing peptides with a redundant amino acid sequence and non-unique elemental compositions, and restricting the data set to peptides containing up to five sulphur atoms, we identified 284,947 peptides suitable for analysis. These peptides cover a mass range from 350.126 Da to 4999.689 Da. It should be noted that the choice to only include peptides up to five sulphur atoms is arbitrary. However, by doing so, we cover 99.5% of the unique atomic compositions. Figure 1 illustrates the distribution of peptides across different sulphur categories as a function of the monoisotopic mass. It is evident that the number of peptides varies significantly depending on the number of sulphur atoms present in the peptide. The distribution is highly unbalanced, with our theoretical dataset consisting of 118,256 peptides with zero sulphur atoms, 92,359 peptides with one sulphur atom, 48,271 peptides with two sulphur atoms, 18,051 peptides with three sulphur atoms, 5,919 peptides with four sulphur atoms, and 2,091 peptides with five sulphur atoms. Moreover, a significant portion of the peptides in our theoretical data set have a monoisotopic mass between 1500 Da and 2500 Da, indicating that this is the most prevalent mass range.

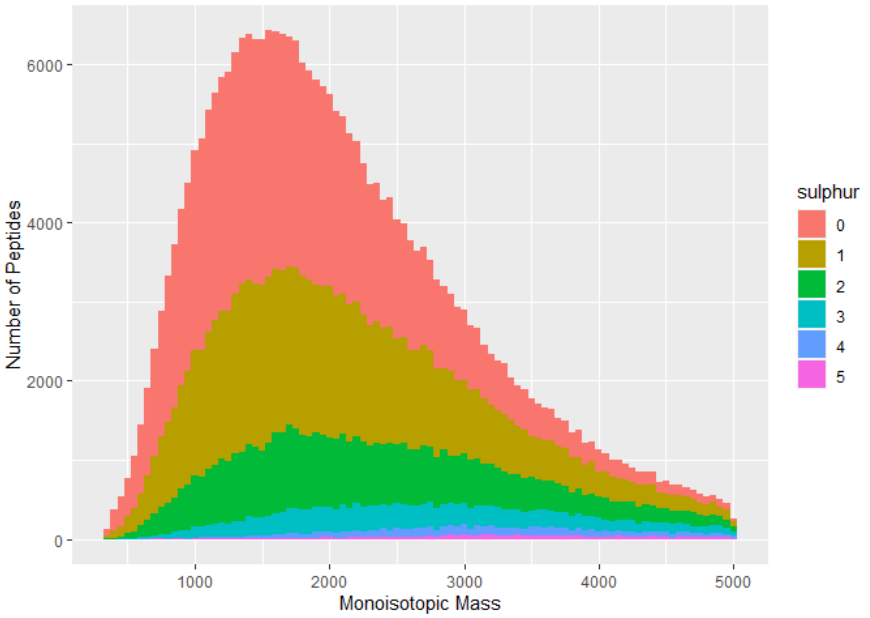

Figure 1: Distribution of the peptides in our theoretical data set as a function of the monoisotopic mass. The colours correspond to peptides containing zero up to five sulphur atoms. The distribution is highly unbalanced, with our theoretical dataset consisting of 118,256 peptides with zero sulphur atoms, 92,359 peptides with one sulphur atom, 48,271 peptides with two sulphur atoms, 18,051 peptides with three sulphur atoms, 5,919 peptides with four sulphur atoms, and 2,091 peptides with five sulphur atoms.

To compute the aggregated isotope distribution for all included peptides, we employ the BRAIN algorithm, using the most recent NIST[13] definition for the elemental isotopes. We restrict our analysis to the first six aggregated isotope peaks. Calculating the aggregated isotope distribution, that denotes the probability to observe a certain isotopic variant of a molecule, has previously been described as sampling ions from a multinomial distribution.[17-21] This implies that, theoretically, the isotope distribution is compositional in nature, where the probabilities of the different isotopes sum to one. The objective of this study is to predict the expected isotope distribution of an average peptide based on its monoisotopic mass using a regression approach. To properly account for the compositional nature of the data, we adopt the modelling approach introduced by Aitchison.[22-24] In essence, compositional data, such as the isotope distribution, can be represented on a high-dimensional simplex. Through Aitchison geometry, this representation can be transformed into real space using a well-characterized isomorphism.[24] However, the isotope distribution does not conform to a classical multinomial distribution, as the number of isotopes and the distribution of their probabilities vary depending on the peptide's mass. For low mass peptides in our theoretical data set (mass < 2500 Da), the first six aggregated isotope peaks cover 100% of the isotope probabilities. However, for the highest mass peptides, the first six aggregated isotope peaks cover only 84% of the entire isotope distribution, thereby breaking the compositional data structure. As our goal is to develop a model that is applicable across the entire mass range of Human peptides, we introduce the concept of a pseudo-isotope or closure term. This pseudo-isotope acts as a sinkhole, collecting the remaining isotope probabilities in one additional isotope variant. By introducing this pseudo-isotope, we ensure that 100% of the isotope information is captured by the first 6 isotopic components and the additional closure component.

Average isotope distribution estimation

Compositional data transformation

We have described our approach in detail for DNA and RNA molecules previously.[12] The data transformation approach for peptides is completely the same, so we restrict ourselves to a short summary. For more extended information, the reader is referred to the aforementioned publication.

Let $\boldsymbol{x} = (x_1, x_2, \ldots, \mathrm{x}_7)$ denote the vector representing the first six aggregated isotope probabilities $(x_1, x_2, \ldots, \mathrm{x}_6)$ and the pseudo-isotope $(x_7 = 1 - \sum_{i=1}^{6} x_i)$. Then this compositional sample space can be accurately represented by the *6*-dimensional unit simplex

$$S^6 = \{(x_1, \ldots, x_7)\, \epsilon\, \mathbb{R}^7 \,|x_i \geq 0\, for\, (i = 1, \ldots, 7)\ \ and \sum_{i=1}^{7} x_i = 1\} \tag{1}$$

The elements of this simplex are nonnegative and constrained to sum to one by the creation of the closure term. In order to analyse the data using a regression approach, we transform the compositional data to the real space using an additive log-ratio transformation (ALR)

$$S^6 \rightarrow \mathbb{R}^6 \colon (x_1, x_2, \ldots, x_7) \rightarrow \left(ln\left(\frac{x_2}{x_1}\right), ln\left(\frac{x_3}{x_1}\right), \ldots, ln\left(\frac{x_7}{x_1}\right)\right) = (z_1, z_2, \ldots, z_6) \tag{2}$$

Note that the monoisotopic peak (x_1) is chosen as the reference component. We chose the monoisotopic peak as the reference because in general, for peptides, it is always observed in the mass spectrum. Moreover, the choice of the reference component does not influence the results or the interpretation of our modelling approach, since standard multivariate statistical models are invariant to the choice of reference category.[23,25]

Modelling approach

In our previous study, where we modelled the isotope distribution of an average DNA or RNA molecule based on its mass, we employed a univariate polynomial regression approach to model the isotopic envelope, transformed into ALR space.[12] However, we observed that the model was susceptible to boundary bias, with the polynomial estimators consistently over- or underestimating the true regression function at the edges of the data's support. Additionally, we found that a high-order polynomial of degree ten was necessary to adequately capture the data trend. To address these challenges, we decided to revert to spline regression. While polynomial regression can only capture the general trend of the data, as it uses a single polynomial to models the entire data set, spline regression offers greater flexibility and accuracy by fitting local trends more precisely. This is accomplished through the use of piecewise continuous functions composed of numerous basis functions that collectively model the dataset. Moreover, spline regression yields more stable estimates and reduces the boundary effects of the estimator.

However, one difficulty in spline regression lies in determining the appropriate number and placement of knots, which define the breakpoints for the piecewise continuous functions. To overcome this issue, we opted for penalized spline regression, where we used the monoisotopic mass as the independent variable to predict the ALRs. In this approach, we included a large amount of knots placed equidistantly, and let a penalization term take care of overfitting.[26] The basis functions for our penalized splines are B-splines, and the order of the spline and the estimation of the smoothing parameters are automatically determined using generalized cross validation.[26,27]

Let the monoisotopic mass of a peptide be denoted by m, then the resulting spline models are given by

$$ALR_j = \sum_{i=1}^{n} \beta_i B_{i,k}(m) \tag{3}$$

For each ALR ratio j $(j = 1, \dots, 6)$ univariately, where $B_{i,k}$ are the B-spline basis functions of order k at knots $m_1, \dots, m_n$ (where n is the total number of knots) and β_i are the spline coefficients.

In our previous research, we discussed the major impact of a sulphur atom on the pattern of the isotope distribution[28]. This diversity in the isotope distribution introduced by sulphur is also apparent in Figure 2A, which illustrates the isotope distributions of the peptides in our theoretical data set. Therefore, we propose six distinct models for average peptides, for each sulphur category separately. This means we create separate models for peptides containing zero up to five sulphur atoms. Since we model each ALR univariately, this results in a total of 36 models, specifically six ALRs multiplied by six different sulphur categories.

In addition to a predictive model for estimating the probabilities, a method is provided to predict the centroid masses of the corresponding isotopic variants. This approach essentially calculates the average mass differences between the aggregated isotope variants and the monoisotopic variant for all the peptides in the theoretical database, for each number of sulphur atoms separately. The outcome of this computation yields a simple vector of mass differences to which the monoisotopic mass of the molecule is added.

Back-transformation

The ALR ratios are back-transformed to the original simplex space using a modified softmax transformation to obtain the isotopic probabilities of the first six aggregated isotope peaks and the closure term. By adding an extra column of ones before employing the softmax transformation, we control the probabilities predicted by the unconstrained penalized spline regression approach to produce estimates that sum back to one.

$$\mathbb{R}^6 \to S^6 : (z_1, z_2, \dots, z_6) \to (x_1, x_2, \dots, x_7) = \left(\frac{1}{g(z)}, \frac{e^{z_1}}{g(z)}, \dots, \frac{e^{z_6}}{g(z)}\right) \tag{4}$$

$$with\ g(z) = 1 + e^{z_1} + \cdots + e^{z_6}$$

The Goodness-Of-Fit Statistic

The performance of our theoretical model is evaluated using the benchmark data set described in the experimental data section. The model predicts the first six isotopes as ALR transformed ratios and their probabilities back-transformed via the modified softmax function. We employ the same goodness-of-fit statistics as previously described.[12] The main difference compared to our previous publication lies in the assumption that the monoisotopic peak is always observed for peptides, whereas this might not be the case for DNA/RNA molecules. However, it is worth noting that we do not always observe all six peaks in the experimental mass spectra. As a result, the calculation of the goodness-of-fit metrics is limited to the number of observed peaks. Therefore, we define $(o_1, \dots, o_l)$ as the observed isotope pattern, where o_1 represents the monoisotopic peak and o_k represents the last observed isotope variant.

The first goodness-of-fit statistic is inspired by the mean squared error, which is commonly used to evaluate univariate linear regression models. In our case, we adopt this metric to assess the error in ALR space. To calculate the ALR goodness-of-fit metric, we first transform the observed spectrum to ALR space by applying the log-ratio transformation using the monoisotopic peak as reference:

$$(t_1, \dots, t_{l-1}) = \left(\ln(\frac{o_2}{o_1}), \dots, \ln(\frac{o_l}{o_1})\right) \tag{5}$$

The goodness-of-fit in ALR space can then be evaluated by calculating the mean squared error

$$MSE_{ALR} = \frac{1}{l-1}\sum_{i=1}^{l-1}(\mathrm{t_i} - z_i)^2 \tag{6}$$

Where $(z_1, \dots, z_{l-1})$ are the predicted ALR ratios using the spline model described in the previous section.

The second goodness-of-fit statistic we adopted is a metric in the simplex space (i.e. probabilities) that is inspired by the multinomial test introduced by Pearson[29]:

$$\chi^2_{simplex} = \frac{1}{l}\sum_{i=1}^{l}\frac{(O_i - E_i)^2}{E_i} \tag{7}$$

Where $E_i = Nx_i$ is the expected peak intensity, with N being the total peak intensity and x_i the predicted probabilities using the theoretical models, back-transformed using the modified softmax transformation. To fit the theoretical model, we introduced a closure term to make sure the aggregated isotope distribution sums to one. When not all isotopes are detected in a spectrum, the total intensity N is unknown. Therefore, N is calculated as the ratio of the sum of the observed intensities and the sum of the probabilities of the expected peaks.

$$N = \frac{\sum_{i=1}^{l} O_i}{\sum_{i=1}^{l} x_i} \tag{8}$$

Sulphur model selection

As mentioned in the previous section, we emphasized the strong dependence of the isotope distribution on the number of sulphur atoms present in a peptide. Consequently, we develop separate models for the isotope distribution for peptides containing zero up to five sulphur atoms. In our theoretical data set, all peptides are identified, enabling us to precisely determine the number of sulphur atoms and select the appropriate model to predict the isotope distribution. However, in practice, this level of information is not always available. Therefore, we aim to predict the number of sulphurs based on the observed isotope distribution of a peptide. To achieve this, we propose two different strategies.

The first strategy relies on the previously described goodness-of-fit statistics in ALR and simplex space. The spline models are fitted using only the monoisotopic mass obtained from the observed mass spectrum. Based on this monoisotopic mass, the average isotope distribution can be predicted using the separate models corresponding to 0, 1, …, 5 sulphurs. By comparing these predictions with the observed distribution in an experimental spectrum, we can calculate the MSE_{ALR} and the $\chi^2_{simplex}$ for each sulphur model separately. The idea is that the 'correct model', i.e. the model that is fitted based on the correct number of sulphur atoms present in the observed peptide, produces the smallest error. Therefore, the number of sulphur atoms can be predicted by:

$$\hat{S} = \arg\min_{s}\{MSE_{ALR,S0}, \dots, MSE_{ALR,S5}\} \tag{9}$$

Or

$$\hat{S} = \arg\min_{s}\{\chi^2_{simplex,S0}, \dots, \chi^2_{simplex,S5}\} \tag{10}$$

For the second strategy we adopt Bayes' theorem for conditional probabilities. Let A and B be two events, then Bayes' theorem states:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{11}$$

We are interested in estimating the probability that a certain peptide has n sulphur atoms, with n varying between zero and five, given the observed isotope distribution. Hence, we are interested in calculating the conditional probability of the sulphur content being equal to n, given the observed isotope distribution:

$$P(S = n\,|(t_1, \dots, t_{l-1})\,) = \frac{P\big((t_1, \dots, t_{l-1})|S = n\big)P(S = n)}{P(t_1, \dots, t_{l-1})} \tag{12}$$

Where we call $P\big((t_1, \dots, t_{l-1})|S = n\big)$ the 'likelihood', i.e. the probability that we observe this isotope distribution given that we assume the peptide contains n sulphur atoms. The probability $P(S = n)$ is the 'prior', which is the probability of observing a peptide that contains n sulphur atoms and $P(t_1, \dots, t_{l-1})$ is a normalizing factor which, based on the law of total probability, can be calculated by

$$P(t_1, \dots, t_{l-1}) = \sum_{n=0}^{5} P\big((t_1, \dots, t_{l-1})|S = n\big)P(S = n) \tag{13}$$

The probability of having n sulphur atoms given the observed isotope distribution $(o_1, \dots, o_l)$ is referred to as the 'posterior probability' $P(S = n\,|(t_1, \dots, t_{l-1})\,)$. It is important to note that this posterior probability is calculated in ALR space. This means that we first transform the observed isotope distribution using the additive log-ratio transformation. This transformation is necessary

because we may not always observe the first six isotope peaks, resulting in partial isotope distributions. By working in ALR space, we can employ models that can appropriately handle such partial isotope distributions.

To compute the likelihood $P\big((t_1, \dots, t_{l-1})|S = n\big)$, we start by predicting the ALR ratios using the monoisotopic mass extracted from the observed mass spectrum. This prediction is done separately for each of the six sulphur models. For each sulphur model, we calculate the normal probability density function using the mean and variance estimated by our spline models. This probability density function represents the likelihood that the observed ALRs are generated from a normal distribution with the estimated mean and variance.

The prior probability $P(S = n)$ corresponds to the probability of observing a peptide with n sulphur atoms, which depends on the peptide's mass. To calculate this probability, we determine the mean number of sulphurs for peptides of a given mass. We employ a sliding window approach on the monoisotopic mass, using a window size of 50 Da and a 10 Da overlap on our theoretical data. In each mass bin, we calculate the total number of peptides in our data set with zero up to five sulphur atoms. From this information, we obtain the mean number of sulphurs per mass bin, denoted as lambda in our analysis. To model the relationship between peptide length and the number of sulphur atoms, we use a count process, where lambda represents the mean count per mass bin. We consider several distribution options for fitting, including the Poisson, negative binomial, zero-inflated Poisson and Conway-Maxwell Poisson distributions.

Using the likelihood of the data $P\big((t_1, \dots, t_{l-1})|S = n\big)$ and the prior probability $P(S = n)$, we calculate the posterior probabilities $P(S = n\,|(t_1, \dots, t_{l-1})\,)$ for each of the sulphur models, given the monoisotopic mass from the observed spectrum. The posterior probability indicates the probability that an observed isotope distribution is generated by a peptide that contains n sulphur atoms. Therefore, the model that produces the highest posterior probability is considered the most likely prediction for the number of sulphur atoms (n) in the peptide.

Results

Average Isotope Distribution Modelling

Figure 2A showcases scatterplots depicting the first six isotopes of all peptides present in our theoretical data set, encompassing a total of 284,947 peptides. These peptides have monoisotopic masses ranging from 350.126 Da to 4999.689 Da. The x-axis represents the monoisotopic mass of the peptides, while the y-axis represents their respective probabilities. Each data point is color-coded to indicate the number of sulphur atoms in the peptide, ranging from zero to five. The plots clearly demonstrate the substantial influence of the sulphur atom count on the isotope distribution of peptides. Figure 2B specifically focuses on the scatterplot of the closure term, which sums up the probabilities of the first six aggregated isotope probabilities to achieve a total of one. For low mass peptides, the closure term is equal to or very close to zero. However, as the mass of the peptides increases, the closure term gradually increases as well, reaching a value of up to 0.15 for high mass peptides. Figure 2C shows the average theoretical aggregated isotope distributions of molecules around 2000 Da and 4000 Da containing 0 to 5 sulphur atoms. This represents a vertical cross section of the black lines in the scatterplots in Figure 2A and Figure 2B at a mass of 2000 Da and 4000 Da. Note that the masses on the x-axis in Figure 2C were offset by 0.1 Da for each peptide containing an increasing number of sulphur atoms to visualize the probabilities more clearly.

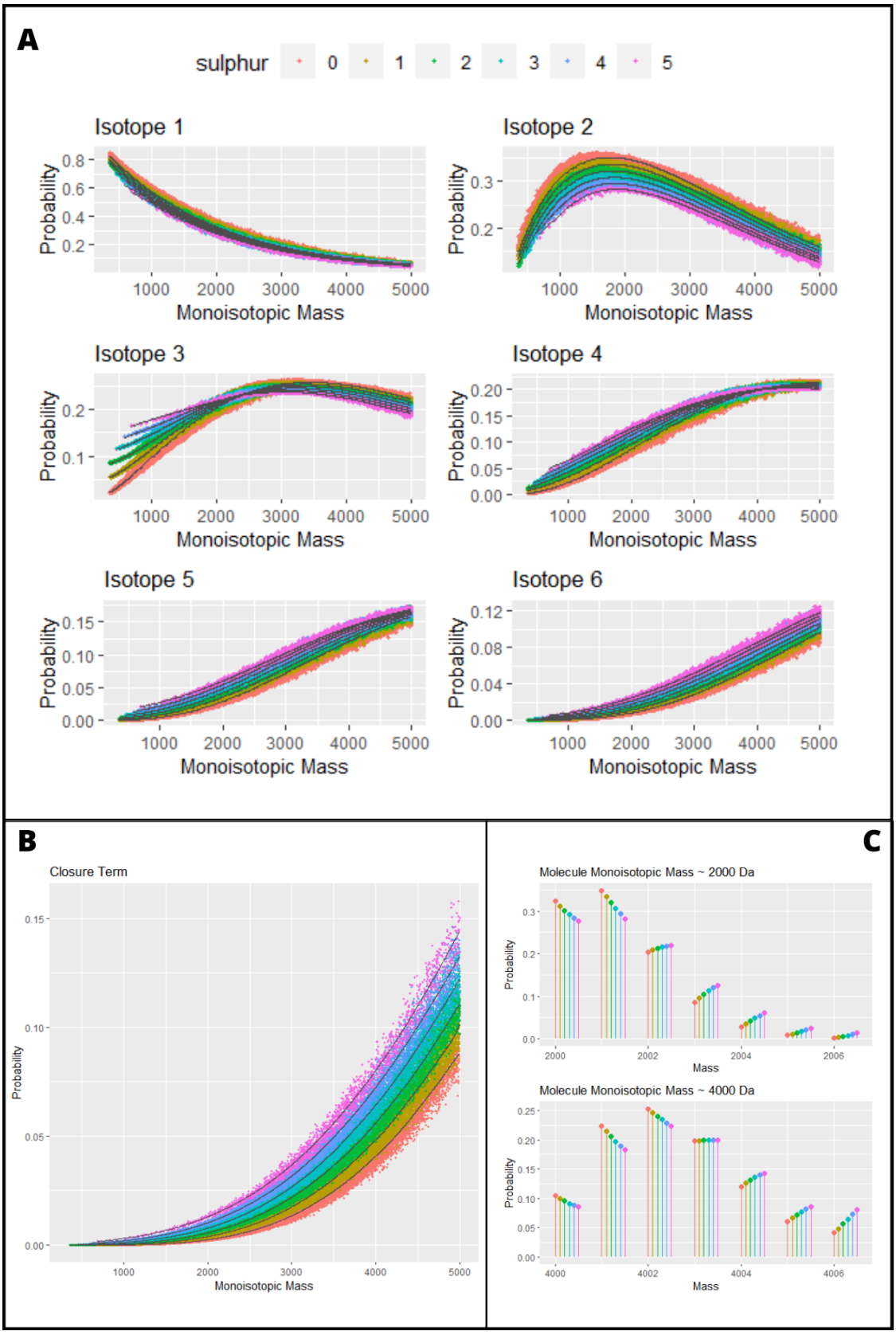

Figure 2: (A) Scatterplots showcasing the first six isotope probabilities (y-axis) of all peptides in our theoretical data set in relation to the monoisotopic mass (x-axis). (B) Scatterplot showcasing the closure term (y-axis), which sums the probabilities of the first six isotopes to one in relation to the monoisotopic mass (x-axis). The data points in all scatterplots are color-coded to indicate peptides containing zero to five sulphur atoms. The black lines present the predicted isotope probabilities obtained after applying the modified softmax transformation on the ALRs predicted using our penalized spline regression models.

(C) The average aggregated isotope distributions for molecules with a monoisotopic mass of 2000 Da and 4000 Da. There is a horizontal offset in the mass of 0.1 to show the peaks for molecules containing a different number of sulphur atoms more clearly. From this Figure it is clear that the isotope distributions are highly dependent on the number of sulphur atoms in the peptide.

Figure 3 illustrates a scatterplot that depicts the relationship between the ALR-transformed isotopes, with the monoisotopic variant serving as the reference, and the monoisotopic mass. The ALR CLOSURE, derived from the ratio of the closure term and the monoisotopic variant, is also included in the plot. Each ALR-transformed isotope is independently modelled using a penalized spline regression model specific to each sulphur category. Consequently, a total of 36 models are fitted. The number of knots and the order of the penalized spline prediction model are automatically determined using generalized cross validation. The black lines depicted in Figure 3 represent the predicted values of the ALR-transformed isotopes obtained from the spline model. Subsequently, after applying the inverse transformation via the modified softmax function, the predicted isotope probabilities are derived. These predicted probabilities for the first 6 aggregated isotopes of our theoretical peptides, as well as the closure term, are visualized in Figure 2A and Figure 2B, respectively. In both figures, it is evident that the predicted values align with the centre of the data cloud. The accuracy of the fit is further supported by the residual plot presented in Supplementary Figure S1. For the majority of the ALRs, the residuals fall within an error interval ranging from -0.3 to 0.3. These results indicate that approximating the theoretical isotope distribution with an average model is justified, as it introduces only a small degree of deviation.

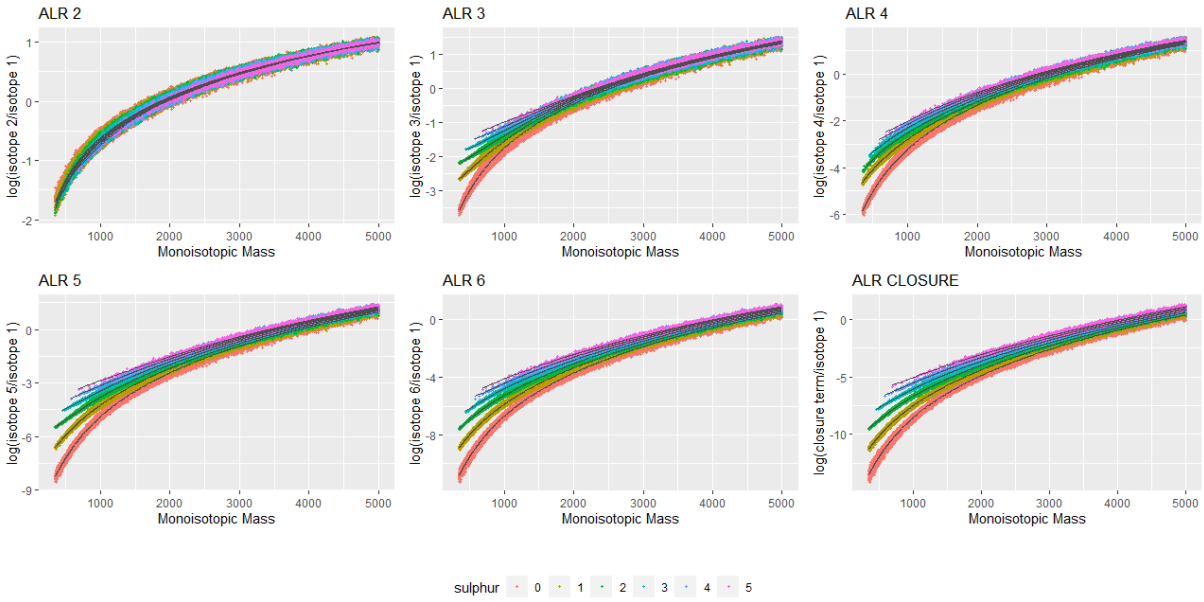

Figure 3: Scatterplots showcasing the ALR transformed isotopes (y-axis) in relation to the monoisotopic mass (x-axis) based on the first six isotope probabilities and the closure term. The monoisotopic mass is taken as the reference isotope for this transformation. ALR CLOSURE is the additive log-ratio transformation of the pseudo-isotope. The data points are color-coded to indicate peptides containing zero up to five sulphur atoms. The black lines present the predicted ALRs using our penalized spline regression models.

To gain a deeper understanding of the error metrics proposed in this study, we present the mean squared errors between the ALR-transformed isotopes computed using BRAIN and the penalized spline

regression model predictions, as well as the mean Pearson chi-squared error between the theoretical isotope probability values computed using BRAIN and the softmax-transformed predicted probabilities. These error measurements are illustrated in Supplementary Figure S2 and Figure S3, respectively. Notably, the size of the errors varies according to the number of sulphur atoms present in the peptide, but remains consistent across the entire mass range of the peptides. In addition to estimating the probabilities, the model offers a predictive feature for determining the centroid mass of the aggregated isotopic variant. This prediction involves calculating the average mass differences between each isotope variant and the corresponding monoisotopic variant across the peptides in our theoretical database, for each number of sulphurs separately. We present the results of these average mass differences in in Table 1. Noteably, there is only a minor discrepancy between the average mass differences of peptides with a varying number of sulphur atoms. The values in Table 1 are used to compute the centroid masses of the predicted isotope probabilitites by simple addition of these values to the observed mass of the monoisotopic variant.

Table 1: Average mass differences (in Da) between the centroid masses of the aggregated isotope variants and the monoisotopic variant for all peptides in the theoretical database, for each number of sulphur atoms separately. Isotope 1 is the monoisotopic variant.

	Isotope 1	**Isotope 2**	**Isotope 3**	**Isotope 4**	**Isotope 5**	**Isotope 6**
0 S atoms	0	1.002891678	2.005596818	3.008229827	4.010806086	5.013344473
1 S atom	0	1.00286816	2.004629099	3.00639507	4.008163627	5.009999388
2 S atoms	0	1.002843655	2.004123027	3.00558085	4.006825931	5.00814229
3 S atoms	0	1.00283248	2.004039656	3.005395842	4.006442102	5.007544965
4 S atoms	0	1.002829018	2.004165477	3.005488558	4.006517479	5.007537239
5 S atoms	0	1.0028162	2.004183866	3.005451329	4.006430802	5.007365125

Sulphur Distribution Modelling

The results of our sliding window approach, displaying the mean number of sulphur atoms (lambda) as a function of the monoisotopic mass is presented by the scatterplot in Figure 4A. A similar penalized spline regression model as described in the previous section is used to model lambda as a function of the monoisotopic mass, indicated by the red line in Figure 4A. This model enables us to predict the mean number of sulphur atoms for an average peptide given a specific monoisotopic mass. Furthermore, our sliding window analysis allows us to determine the probabilities of observing a peptide with n sulphur atoms, given the monoisotopic mass $P(S = n)$. These probabilities are presented in the scatterplots in Figure 4B. To determine the best-fit distribution (Poisson, Negative Binomial, Conway-Maxwell Poisson, or zero-inflated Poisson) for these probabilities, we assess the χ^2-error and mean square error approach as described in the Materials and Methods section. The results are presented in Supplementary Materials Table S1. The negative binomial distribution exhibits the best fit, as indicated by the smallest χ^2-error and mean square error. Consequently, we assume that the prior probabilities $P(S = n)$, as described in the Materials and Methods section, follow a negative binomial distribution. The mean of this distribution is equal to lambda, while the dispersion parameter is estimated based on the data. The fit of the binomial distribution is represented by the red lines in Figure 4B.

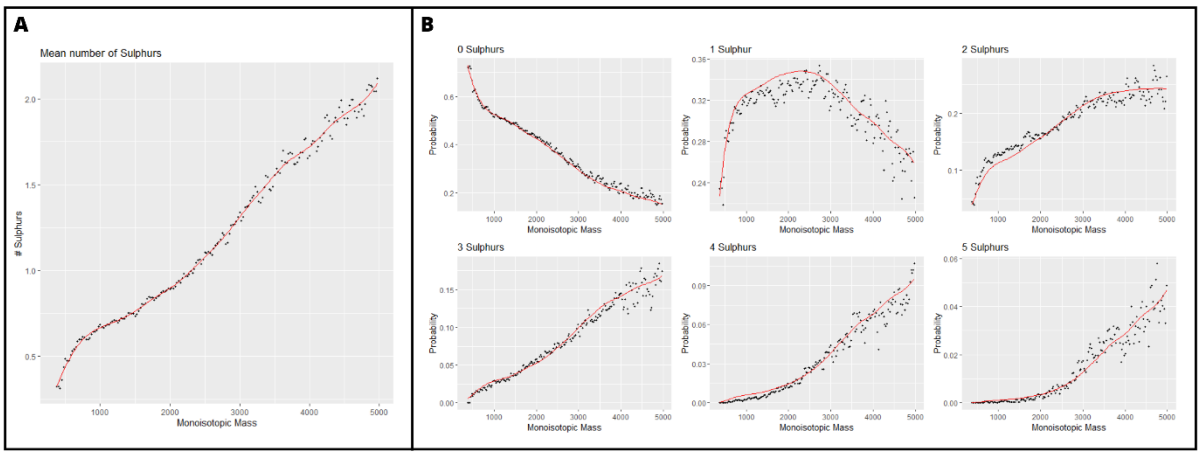

Figure 4: (A) Scatterplot displaying the results of our sliding window approach. The y-axis shows the mean number of sulphur atoms as a function of the monoisotopic mass (x-axis). The red line shows the result of a penalized spline regression model that uses the monoisotopic mass as input to predict lambda. (B) Scatterplot displaying the probability of observing n sulphur atoms (y-axis) in relation to the monoisotopic mass (x-axis). The red lines show the fit of the negative binomial distribution using lambda from (A) and a dispersion parameter estimated by the data.

Since we know the identity of the peptides in our theoretical database, we can use this to gain a deeper understanding of the achievable sulphur prediction accuracies. For this purpose, we use all available sulphur models to predict the ALRs and isotope probabilities for every peptide in our data set. Subsequently, we predict the number of sulphur atoms using the minimal MSE_{ALR} and minimal χ^2-error. We compare these predictions to the true sulphur count based on the peptides' elemental composition. Additionally, we predict the number of sulphur atoms using the maximal posterior probability approach, which incorporates both the data likelihood and prior information, and compare these predictions to the true values. The overall accuracy of sulphur prediction based on the minimal MSE_{ALR} approach is equal to 0.8076, while for the minimal χ^2-error approach it is equal to 0.8932. In case of the posterior probability approach, the total accuracy is 0.8023. These accuracies represent the highest achievable prediction accuracies in a theoretical and thus experimental setting. The overall prediction accuracies are quite similar, however, the minimal χ^2-error approach displays the best overall prediction outcomes. For more detailed information, including confusion matrices, please refer to Supplementary Figures S4-S6.

EXPERIMENTAL VALIDATION

The final penalized spline models are evaluated using experimental spectral data, as detailed in the Materials and Methods section. Our experimental data set comprises 50 peptides with monoisotopic mass values ranging from 893.4244 Da to 3577.7645 Da. Within this data set, we have 35 peptides without any sulphurs, 10 peptides with one sulphur atom, 3 peptides with two sulphur atoms, and 2 peptides with three sulphur atoms. We extracted the first five observed isotope peaks for all 50 peptides. By employing MS-GF+ for peptide identification, we know the elemental composition of the peptides, including the number of sulphur atoms. This knowledge allows us to determine the appropriate sulphur model in advance.

Firstly, to assess the potential error caused by instrument variability, we compared the experimental data with the theoretical isotope distribution computed by BRAIN. Note that when we refer to 'variability in spectral accuracy', we mean the variability between MS1 scans. In general, it is known that higher-order isotope peaks are measured less accurately in a mass spectrometer. Therefore, we computed the MSE_{ALR} and the mean Pearson χ^2-error between the observed isotope distribution and the expected isotope distribution using BRAIN, considering the first three, four and five observed

isotope peaks. Scatterplots displaying the results of this analysis are shown in Supplementary Figure S7. Generally, we observed that the MSE_{ALR} tends to increase as the number of considered isotope peaks increases. Particularly in the lower mass range of our experimental data set, the analysis demonstrates that the fifth isotope peak is measured less accurately. However, the observed errors (range shown in Table S2) align with those depicted in Figure S2, suggesting that the experimental variability in our validation data set is consistent with our expectations. On the other hand, the mean Pearson χ^2-error shows greater consistency across the different orders of isotope information used and is comparable to the expected errors shown in Figure S3 (range shown in Table S2). An important remark with respect to the scaling should be made here. As the peptide identity is known, we rescaled the extracted peak intensities of each of the 50 experimental peptide spectra such that they sum up to the summed probabilities of the first three, four or five peptide isotopes as computed using BRAIN. By implementing this scaling, we established direct comparability between the predicted probabilities from BRAIN and the observed probabilities in the experimental dataset. Consequently, we are able to assess instrument variability and compare it to the theoretically expected values.

Secondly, we compared the experimental data with the predictions generated by the penalized spline regression models to evaluate whether the error introduced by the average model is acceptable in comparison to the error originating from instrumental variability. Figure 5 displays scatterplots illustrating the results of the error analysis in relation to the monoisotopic mass, while Table 2 presents the range of the errors. It is important to note that the observed isotope probabilities were rescaled to enable a direct comparison with the errors expected due to instrument variability as depicted in Figure S7. We can clearly see that the errors computed between the observed isotope distribution and the predicted isotope distribution using the spline models are higher compared to the errors computed between the observed isotope distribution and the theoretical isotope distribution using BRAIN. However, the mean absolute difference between the theoretical MSE_{ALR} and the spline MSE_{ALR} is minimal, measuring to 0.005, 0.010, and 0.017 when considering three, four, or five isotope peaks, respectively. This finding suggests that the error introduced by using the average spline model is small compared to the expected error due to instrument variability. Thus, indicating that our penalized spline regression models accurately predict the isotope distribution of the peptides. Similarly, the mean absolute difference between the theoretical χ^2-error and the spline χ^2-error is consistently low, with a value of 0.0003 for the analysis involving three up to five isotope peaks. This further supports the conclusion that the spline models perform well in predicting the peptide's isotope distributions. A plot illustrating the best observed and the predicted isotope distribution for the best fitting peptide (i.e. lowest error) and the worst fitting peptide (i.e. highest error) is provided in Figure S8.

Table 2: Range of the MSE_{ALR} and the mean Pearson χ^2-error between the observed isotope distributions and the predicted isotope distributions computed using our penalized spline regression approach, considering the first three, four, and five observed isotope peaks.

		Lower	Upper
Using 3 isotope peaks	$\boldsymbol{MSE_{ALR}}$	3.690e-05	4.688e-02
	Mean $\boldsymbol{\chi^2}$-error	2.215e-05	3.180e-03
Using 4 isotope peaks	$\boldsymbol{MSE_{ALR}}$	1.396e-04	1.674e-01
	Mean $\boldsymbol{\chi^2}$-error	3.145e-05	4.994e-03
Using 5 isotope peaks	$\boldsymbol{MSE_{ALR}}$	8.702e-04	1.004e-00
	Mean $\boldsymbol{\chi^2}$-error	5.911e-05	4.566e-03

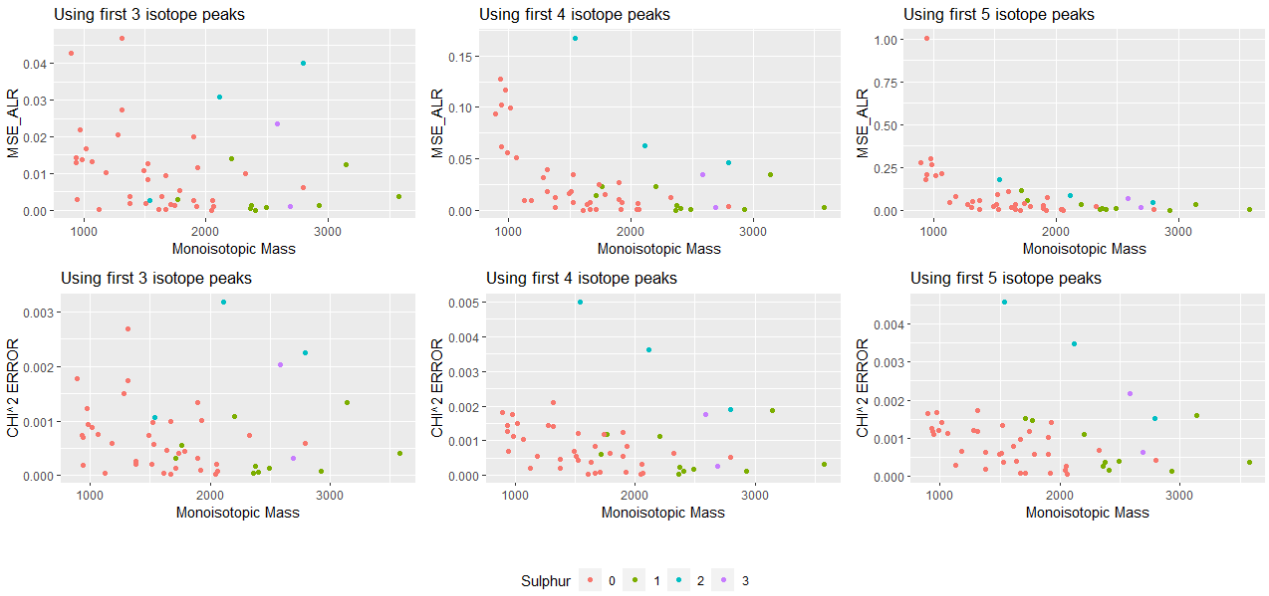

Figure 5: Scatterplots showing the MSE_{ALR} and the mean Pearson χ^2-error between the observed isotope distributions and the predicted isotope distributions computed using the penalized spline regression models, considering the first three, four, and five observed isotope peaks. The colours indicate the number of sulphur atoms in the peptide.

Lastly, we assess the accuracy of the prediction of the number of sulphurs atoms in our experimental peptides using the minimal MSE, minimal χ^2-error and maximal posterior probability approaches. We did this exercise using information from the first three, four and five isotope peaks in the observed spectra. The detailed results and confusion matrices for the nine different prediction exercises are presented in Supplementary Figures S9-S17. In general, the overall prediction accuracy remains consistent across the different approaches. sulphur prediction based on the first three or four isotope peaks show slightly better performance compared to predictions based on five isotope peaks. Moreover, the minimal χ^2-error and the maximal posterior probability approaches perform better than sulphur prediction based on the minimal mean squared error in ALR space. Overall, for peptides containing zero or one sulphur atom, the number of sulphurs can be predicted quite accurately, with sensitivities reaching up to 80% and 70% for these respective categories. This is achieved using the maximal posterior probability approach on the first three isotope peaks. However, peptides in the experimental data set containing two or three sulphur atoms are consistently misclassified as having zero or one sulphur atom(s) across all different methods. This discrepancy can be attributed to the higher measurement error observed in Figure S7. It is clear that the errors between the observed isotope distribution and the expected isotope distribution using BRAIN, especially from the peptides containing two sulphur atoms, are higher compared to the other peptides, indicating that these peptides were measured less accurately. This can explain the decreased accuracy in predicting the correct number of sulphur atoms. Furthermore, the experimental peptides with three sulphur atoms are predominantly found in higher mass range. In this mass range, there is more overlap in the theoretical isotope distributions, as depicted in Figure 2 and Figure 3. Hence, even slight deviations in the measured isotope distribution can affect the prediction of the number of sulphurs. Finally, it is worth noting that theoretically there are more peptides containing zero or one sulphur atoms than there are peptides containing more sulphur atoms. As a result, the prediction approaches, especially the posterior probability method, tend to be skewed towards predicting fewer sulphur atoms, which is consistent with the observed trend.

Software

The modelling approach described in this manuscript has been implemented in Shiny (a package in R programming language) and deployed at https://dsi-uhasselt.shinyapps.io/pointless4peptides/. The user interface is presented in the screenshot in Figure 6.

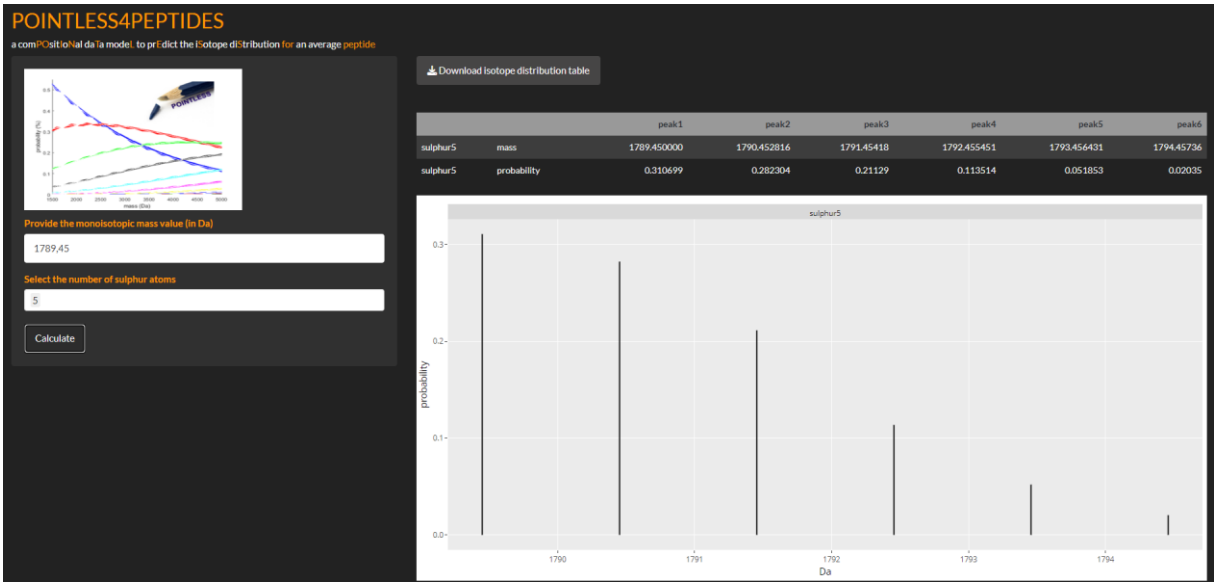

Figure 6: The modelling framework for the prediction of the isotope distribution for an average peptide based on the compositional spline model introduced in this manuscript has been made available to a wider public via an easy-to-use web application.

Firstly, the user has to provide the monoisotopic mass (in Da) of a molecule for which the aggregated isotope distribution needs to be calculated. Note that for any specified mass value that exceeds the mass range used in the model fitting (5000 Da), the isotope distribution will not be computed and warning messages will be shown. In the following step, the number of sulphur atoms needs to be specified. Currently, the app does not incorporate the automatic prediction of the number of sulphur atoms yet. Once the mass and sulphur number is specified, the "Calculate" button triggers the prediction of the isotopic envelope, and the predicted probabilities are directly displayed in the user interface as a downloadable table and graph.

Discussion

In this paper, we introduce an updated approach for approximating the isotope distribution of average peptides based on their given mass. Our method incorporates several advancements, including the utilization of the latest protein database and elemental isotope definition. Instead of relying on an Averagine model or a set of theoretical peptides, we employ the BRAIN algorithm to accurately compute the aggregated isotope distribution for all tryptic peptides in the database. By leveraging this extensive data, we can optimally model the isotope probabilities using a state-of-the art statistical algorithm that allows us to process the entire protein database to accurately predict the average isotope distribution for a peptide of a given mass.

It is important to note that our proposed method is not a continuation of the model presented in 2008.[9] The previous model tried to avoid the compositional nature of the data by giving estimates of the isotope ratios. Instead, our approach draws inspiration from the method of Gay et al., but optimized to deliver actual isotope probabilities rather than relying on more complex algorithms or statistical models.[7] We achieve this though the application of a compositional data model that utilizes

an additive log-transformation followed by a penalized spline regression model. Furthermore, our model is extended to estimate the first six isotopes, including the monoisotopic variants. The Poisson model proposed by Breen also provides an elegant solution to handle the compositional nature of the data.[6] However, the Poisson model is not flexible enough to handle sulphur in the data. To overcome this limitation, we model the isotope distributions for peptides containing up to five sulphur atoms separately. Moreover, by employing a statistical regression approach to model the peptide database, we can provide a measure of uncertainty on the predicted isotope probabilities.

Finally, we proposed different methods to predict the number of sulphur atoms in a peptide, given an observed mass spectrum. The proposed methods all rely on the accuracy with which the isotope distribution of the experimental peptides is measured. Our findings indicate that the error introduced by averaging the isotope distribution is small compared to the error introduced by instrument variability. However, it should be noted that the prediction accuracy for the number of sulphur atoms is limited by measurement accuracy in our experimental data set. Nevertheless, we utilize a posterior probability approach to predict the number of sulphur atoms, providing the probability that a certain observed isotope distribution was generated by a peptide containing n sulphur atoms (where n ranges from zero to five). To achieve this, we employ a prior representing the probability that a certain peptide has n sulphur atoms given a certain monoisotopic mass using the negative binomial distribution. These properties present opportunities for further exploration and exploitation in future research.

Conflicts of interest

The authors have declared no conflicts of interest.

Associated data

The study from Ahrné et al. publicly is available on the PRIDE repository with identifier PXD000331. The samples A11-12042.raw was selected for analysis.

References

[1] Horn, D. M., Zubarev, R. A., & McLafferty, F. W. (2000). Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J Am Soc Mass Spectrom, 11*(4), 320-332. doi: 10.1016/s1044-0305(99)00157-9

[2] Hsieh, E. J., Hoopmann, M. R., MacLean, B., & MacCoss, M. J. (2010). Comparison of database search strategies for high precursor mass accuracy MS/MS data. *J Proteome Res, 9*(2), 1138-1143. doi: 10.1021/pr900816a

[3] Dittwald, P., Claesen, J., Burzykowski, T., Valkenborg, D., & Gambin, A. (2013). BRAIN: A Universal Tool for High-Throughput Calculations of the Isotopic Distribution for Mass Spectrometry. *Analytical Chemistry, 85*(4), 1991-1994. doi: 10.1021/ac303439m

[4] Łącki, M. K., Startek, M., Valkenborg, D., & Gambin, A. (2017). IsoSpec: Hyperfast Fine Structure Calculator. *Analytical Chemistry, 89*(6), 3272-3277. doi: 10.1021/acs.analchem.6b01459

[5] Senko, M. W., Beu, S. C., & McLaffertycor, F. W. (1995). Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry, 6*(4), 229-233. doi: 10.1016/1044-0305(95)00017-8

[6] Breen, E. J., Hopwood, F. G., Williams, K. L., & Wilkins, M. R. (2000). Automatic Poisson peak harvesting for high throughput protein identification. *ELECTROPHORESIS, 21*(11), 2243-2251. doi: https://doi.org/10.1002/1522-2683(20000601)21:11<2243::AID-ELPS2243>3.0.CO;2-K
[7] Gay, S., Binz, P. A., Hochstrasser, D. F., & Appel, R. D. (1999). Modeling peptide mass fingerprinting data using the atomic composition of peptides. *ELECTROPHORESIS, 20*(18), 3527-3534. doi: 10.1002/(sici)1522-2683(19991201)20:18<3527::Aid-elps3527>3.0.Co;2-9
[8] Valkenborg, D., Assam, P., Thomas, G., Krols, L., Kas, K., & Burzykowski, T. (2007). Using a Poisson approximation to predict the isotopic distribution of sulphur-containing peptides in a peptide-centric proteomic approach. *Rapid Communications in Mass Spectrometry, 21*(20), 3387-3391. doi: https://doi.org/10.1002/rcm.3237
[9] Valkenborg, D., Jansen, I., & Burzykowski, T. (2008). A model-based method for the prediction of the isotopic distribution of peptides. *J Am Soc Mass Spectrom, 19*(5), 703-712. doi: 10.1016/j.jasms.2008.01.009
[10] Ghavidel, F. Z., Mertens, I., Baggerman, G., Laukens, K., Burzykowski, T., & Valkenborg, D. (2014). The use of the isotopic distribution as a complementary quality metric to assess tandem mass spectra results. *J Proteomics, 98*, 150-158. doi: 10.1016/j.jprot.2013.12.013
[11] Claesen, J., Rockwood, A., Gorshkov, M., & Valkenborg, D. The isotope distribution: A rose with thorns. *Mass Spectrometry Reviews, n/a*(n/a). doi: https://doi.org/10.1002/mas.21820
[12] Agten, A., Prostko, P., Geubbelmans, M., Liu, Y., De Vijlder, T., & Valkenborg, D. (2021). A Compositional Model to Predict the Aggregated Isotope Distribution for Average DNA and RNA Oligonucleotides. *Metabolites, 11*(6). doi: 10.3390/metabo11060400
[13] Coursey, J. S., Schwab, D. J., Tsai, J. J., & Dragoset, R. A. (2015). Atomic Weights and Isotopic Compositions (version 4.1) ([online]). Available from National Institute of Standards and Technology, Gaithersburg, MD http://physics.nist.gov/Comp
[14] Agten, A., Claesen, J., Burzykowski, T., & Valkenborg, D. (2023). Machine learning approach for the prediction of the number of sulphur atoms in peptides using the theoretical aggregated isotope distribution. *Rapid Communications in Mass Spectrometry, 37*(9), e9480. doi: https://doi.org/10.1002/rcm.9480
[15] Ahrné, E., Molzahn, L., Glatter, T., & Schmidt, A. (2013). Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *PROTEOMICS, 13*(17), 2567-2578. doi: https://doi.org/10.1002/pmic.201300135
[16] Vandermarliere, E., Mueller, M., & Martens, L. (2013). Getting intimate with trypsin, the leading protease in proteomics. *Mass Spectrom Rev, 32*(6), 453-465. doi: 10.1002/mas.21376
[17] MacCoss, M. J., Toth, M. J., & Matthews, D. E. (2001). Evaluation and optimization of ion-current ratio measurements by selected-ion-monitoring mass spectrometry. *Anal Chem, 73*(13), 2976-2984. doi: 10.1021/ac010041t
[18] Kaur, P., & O'Connor, P. B. (2004). Use of statistical methods for estimation of total number of charges in a mass spectrometry experiment. *Anal Chem, 76*(10), 2756-2762. doi: 10.1021/ac035334w
[19] Kaur, P., & O'Connor, P. B. (2007). Quantitative determination of isotope ratios from experimental isotopic distributions. *Anal Chem, 79*(3), 1198-1204. doi: 10.1021/ac061535z
[20] Yergey, J., Heller, D., Hansen, G., Cotter, R. J., & Fenselau, C. (1983). Isotopic distributions in mass spectra of large molecules. *Analytical Chemistry, 55*(2), 353-356. doi: 10.1021/ac00253a037
[21] Yergey, J. A. (2020). A general approach to calculating isotopic distributions for mass spectrometry. *Journal of Mass Spectrometry, 55*(8), e4498. doi: https://doi.org/10.1002/jms.4498
[22] Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological), 44*(2), 139-177.
[23] Aitchison, J. (1994). Principles of Compositional Data Analysis. *Lecture Notes-Monograph Series, 24*, 73-81.

[24] Aitchison, J., & Shen, S. M. (1980). Logistic-Normal Distributions: Some Properties and Uses. *Biometrika, 67*(2), 261-272. doi: 10.2307/2335470

[25] Aitchison, J., Vidal, C., Martín-Fernández, J., & Pawlowsky-Glahn, V. (2000). Logratio Analysis and Compositional Distance. *Mathematical Geology, 32*, 271-275. doi: 10.1023/A:1007529726302

[26] Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science, 11*(2), 89-121, 133.

[27] Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., & Schmid, M. (2019). A review of spline function procedures in R. *BMC Med Res Methodol, 19*(1), 46. doi: 10.1186/s12874-019-0666-3

[28] Valkenborg, D., Jansen, I., & Burzykowski, T. (2008). A model-based method for the prediction of the isotopic distribution of peptides. *Journal of the American Society for Mass Spectrometry, 19*(5), 703-712. doi: 10.1016/j.jasms.2008.01.009

[29] Pearson, K. (1992). On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics: Methodology and Distribution* (pp. 11-28). New York, NY: Springer New York.