University of Warsaw

Doctoral School of Exact and Natural Sciences

Piotr Radziński

Student no. 345426

Computational methods for leading-edge mass spectrometry techniques

PhD's dissertation

in COMPUTER SCIENCE

Supervisors:

prof. Anna Gambin
Institute of Informatics,
University of Warsaw

Jakub Karasiński, PhD
Biological and Chemical Research Centre,
University of Warsaw

January 2025

Supervisor's statement

Hereby I confirm that the presented thesis was prepared under my supervision and that it fulfils the requirements for the degree of PhD of Computer Science.

Date Supervisor's signature

Author's statement

Hereby I declare that the presented thesis was prepared by me and none of its contents was obtained by means that are against the law.

The thesis has never before been a subject of any procedure of obtaining an academic degree.

Moreover, I declare that the present version of the thesis is identical to the attached electronic version.

Date Author's signature

Abstract

Computational methods for leading-edge mass spectrometry techniques

This dissertation explores three topics related to data modeling and analysis in mass spectrometry, each corresponding to a different measurement technique. The first part concerns the determination of monoisotopic mass, whose signals are often not directly visible in the mass spectra of high molecular mass molecules due to a low signal-to-noise ratio. However, knowing this mass is essential for molecular identification. Two novel algorithms are presented to address this issue, enabling the determination of monoisotopic mass for proteins and oligonucleotides. Next, the dissertation introduces an image compression algorithm using neural networks for mass spectrometry imaging (MSI) data. This method significantly enhances the computational efficiency of data analysis on the compressed images. Finally, the last part of the dissertation focuses on atomic mass spectrometry. After a brief overview of selected measurement techniques using a multicollector inductively coupled plasma mass spectrometer, a new algorithm is proposed for estimating uncertainty in the sample-standard bracketing calibration method for exact isotope ratio measurements.

Metody obliczeniowe dla wiodących technik spektrometrii mas

W rozprawie omówiono trzy zagadnienia związane z modelowaniem i analizą danych pochodzących z eksperymentów spektrometrii mas. Dotyczą one przetwarzania danych wywodzących się z trzech odrębnych obszarów. Pierwsze zagadnienie dotyczy wyznaczania masy monoizotopowej, której sygnał często nie jest widoczny na widmach mas cząsteczek o dużej masie cząsteczkowej. Z drugiej strony, znajomość tej masy jest niezbędna w procesie identyfikacji cząsteczek. Przedstawiono dwa nowe algorytmy dla tego problemu, które pozwalają na wyznaczenie masy izotopowej dla sygnałów białkowych oraz oligonukleotydowych. Następnie zaprezentowano algorytm kompresji obrazów spektrometrycznych (technologia MSI – ang. mass spectrometry imaging) przy użyciu sieci neuronowych, dzięki któremu algorytmy analizy danych MSI stają się efektywne obliczeniowo. Ostatnia część pracy poświęcona jest pierwiastkowej spektrometrii mas. Po krótkim omówieniu wybranych metod pomiarowych wykorzystujących technikę MC-ICP-MS (ang. multicollector inductively coupled plasma mass spectrometry), prezentujemy nowy algorytm szacowania niepewności w metodzie kalibrowania naprzemiennego w dokładnych pomiarach stosunków izotopowych.

Keywords

Mass Spectrometry, Data Analysis, Machine Learning, Monoisotopic Mass, Mass Spectrometry Imagining, Encoder-Decoder, Multicollector Inductively Coupled Plasma Mass Spectrometer, Isotopic Ratio, Sample-Standard Bracketing, Measurements' Uncertainty, Monte Carlo Simulations

Thesis domain (Socrates-Erasmus subject area codes)

11.3 Informatics, Computer Science

Subject classification

Applied computing
- Physical sciences and engineering
  - Chemistry
  - Mathematics and statistics

Computing methodologies
- Machine learning
- Modeling and simulation

Tytuł pracy w języku polskim

Metody obliczeniowe dla wiodących technik spektrometrii mas

Contents

List of Figures

List of Tables

1 Introduction

Mass spectrometry, a subfield of analytical chemistry, involves the use of *mass spectrometers* and the analysis of their output. Mass spectrometers can be roughly understood as a highly accurate weighing scales designed to measure the mass of molecules and even atoms. To fully appreciate the depth of this field, we must first explore the intricate structure of atoms, the fundamental building blocks of matter. Understanding atomic structure is essential to grasp the full capabilities and applications of mass spectrometry in various scientific domains.

Atoms are composed of protons, neutrons (excluding ^{1}H), and electrons. Since electrons have negligible mass compared to protons and neutrons, they are typically not considered in mass spectrometry. Protons and neutrons, collectively known as *nucleons*, have similar masses. Due to nuclear forces, the mass of a *nucleus* – center of an atom consisting of nucleons – is not simply the sum of its isolated nucleons' masses; adding a neutron results in a total mass slightly less than the combined mass of a nucleus and the additional neutron. To identify the chemical element an atom represents, we count its number of protons. For instance, hydrogen has a single proton, while sulfur has sixteen. However, atoms of the same element can have different numbers of neutrons. These variants, called *isotopes*, differ in mass but generally share the same chemical properties. Most elements have several stable isotopes that occur naturally with certain probabilities; see Table 1.1 for examples.

Let us consider the implications of such an atomic structure. Suppose we have a given number of molecules with an established chemical formula, differing only in the isotopes of

isotope	mass	abundance
^{12}C	12	0.9893
^{13}C	13.0034	0.0107
^{74}Se	73.9225	0.0089
^{76}Se	75.9192	0.0937
^{77}Se	76.9199	0.0763
^{78}Se	77.9173	0.2377
^{80}Se	79.9165	0.4961
^{82}Se	81.9167	0.0873

Table 1.1: Stable isotopes of hydrogen (H), carbon (C) and selenium (Se) with their masses [Da] and abundances in terrestrial matter (Laeter *et al.*, 2003).

their constituent atoms. The mass of these molecules will vary depending on the number of neutrons each atom contains. Moreover, even the molecules with the exact neutron count, such as ^{12}C^{17}O and ^{13}C^{16}O, will have subtly different masses (approx. 0.001 Da or less). The mentioned variants of the CO molecule are examples of *isotopologues* – molecules that share the same chemical formula but differ in their isotopic composition.

Let us consider the probabilities of specific isotopologues occurring using the example of a fullerene C_{70} isotopologue with the isotopic formula $^{12}C_{67}{}^{13}C_3$. To determine the probability that the fullerene contains sixty-seven ^{12}C atoms and three ^{13}C atoms, we compute $\frac{70!}{67!\cdot 3!}\cdot\mathbb{P}(^{12}\mathrm{C})^{67}\cdot\mathbb{P}(^{13}\mathrm{C})^{3}$, where the sequential terms represent the number of possible atomic arrangements, the probability of sixty-seven carbon atoms being ^{12}C, and the probability of three carbon atoms being ^{13}C, respectively. Note, that it follows multinomial distribution. In general the probability of an isotopologue with the form ${}^{i_1}\phi_{\upsilon_1}\ldots{}^{i_k}\phi_{\upsilon_k}$, where k different isotopes are present and the total number of atoms is $n=\sum_{i=1}^{k}\upsilon_i$, can be expressed as

$$\mathbb{P}({}^{i_1}\phi_{\upsilon_1}\ldots{}^{i_k}\phi_{\upsilon_k}) = \frac{n!}{\prod_{j=1}^{k}\upsilon_j}\times\prod_{j=1}^{k}\mathbb{P}({}^{i_j}\phi)^{\upsilon_j}. \tag{1.1}$$

Going one step further, the probability of an isotopologue consisting of different elements can be determined by multiplying the probabilities corresponding to the isotopologue's single-element components. If, for a given molecule, we construct a histogram of its isotopologues' masses versus their probabilities, we obtain an *isotopic envelope*, which represents a probability distribution over the molecule's isotopologues. This observation is substantial as it allows us to apply probabilistic tools to analyze isotopic envelopes.

Finally, by *mass spectrum*, we refer to a broader-sense histogram, typically displaying signals from a mixture of multiple molecules, with their isotopic envelopes combined. Note that combining preserves the interpretation of mass spectra as probability distributions, as is the case with isotopic envelopes. Unfortunately, this description assumes a mathemat-

ically perfect scenario; the mass spectra described under these conditions can be termed *theoretical* or *simulated*. In real-world conditions, the resulting patterns in *experimental* mass spectra often become much more complex.

For a molecule to be detected in mass spectrometry, it must be ionized, typically involving the addition of a hydrogen cation. However, the number of ions that attach to individual molecules may be difficult to control. The number of ions attached to the molecule, referred to as *charge*, also alters the molecule's mass. This distinction is crucial since the instrument records not the mass itself but the *mass-to-charge ratio* (m/z). An example of an experimental mass spectrum is presented in Figure 1.1. Notably, even though it represents a single chemical compound, signals appear in two distinct locations due to the two charge states generated during ionization.

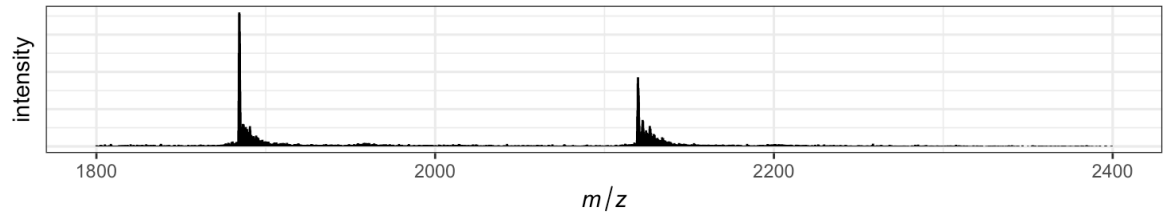

Figure 1.1: An example of experimental mass spectrum of equine apo-myoglobin, showcasing scan with two well-visible charge states. Average mass: 16951.5 Da.

Another reason why experimental spectra differ from theoretical ones is the finite resolution of mass analyzers. As a result, the unprocessed spectrum appears more "continuous", exhibiting a Gaussian-like distribution at each detected peak rather than discrete signals. This type of spectrum is referred to as a *profile* spectrum. However, it is more practical to use *centroided* spectrum, derived by aggregating the profile spectrum into discrete peaks, each representing the most frequent point of the detected signal. This aggregation process converts the continuous signal into distinct peaks, enhancing clarity and reducing computational complexity in subsequent analyses.

Finally, experimental spectra are plagued by various noise types, related to, a.o., low ion statistics (i.e., low-frequency noise), chemical contamination, or background disruptions (e.g., microshocks caused by nearby road transport). Managing these noises can cause a significant challenge in mass spectrometry. For instance, in proteomics, where the molecules under study are relatively heavy, even high-resolution spectra often reveal only the most intense peaks as clearly distinguishable from the surrounding noise. Consequently, there is a potential loss of crucial information. A monoisotopic mass peak can be an example here; it is vital for identifying the measured biomolecules yet often gets hidden under noise. Therefore, the first research we present in this thesis focuses on the determination of monoisotopic mass.

1.1. Determination of monoisotopic mass

Monoisotopic mass is the mass of a molecule that consists of only the most abundant naturally occurring stable isotopes. Unfortunately, for primary elements that constitute biomolecules – such as hydrogen, carbon, and oxygen – the most abundant stable isotope is also the lightest. Consequently, for biological compounds, the monoisotopic peak is typically the lightest as well. Considering probabilistic laws, it occurs that for biomolecules heavier than approx. 10 kDa, the lightest peak, which naturally resides at the tail of the multinomial distribution, is often lost in noise or goes undetected by mass spectrometers. Nevertheless, monoisotopic mass is considered more reliable than alternatives such as average mass and the mass of the most abundant peak, and therefore, it commonly plays an essential role in various applications. First, however, it has to be predicted.

Our research group has been collaborating with prof. Dirk Valkenborg's laboratory at Hasselt University, Belgium, on mass spectrometry issues. Among our group's most significant achievements are the development of algorithms for isotopic envelope simulation (Dittwald *et al.*, 2013; Łącki *et al.*, 2017); the application of Wasserstein distance in mass spectrometry (Majewski *et al.*, 2018; Ciach *et al.*, 2020); and finally, the `MIND` algorithm for determining monoisotopic mass, as detailed by Lermyte *et al.* (2019). In Chapter 2, we present a continuation of this interlaboratory work, developing two follow-up algorithms that extend `MIND`'s methodology to specific types of mass spectrometry data.

The first algorithm presented in this thesis, named `Envemind`, is derived from the term "envelope". Unlike the `MIND` algorithm, which relies on the mass of the most abundant peak for the prediction process, the `Envemind` algorithm utilizes the complete informational potential of an envelope, incorporating its average mass, variance, and shape into its computations. Consequently, while both algorithms are designed for protein spectra, `Envemind` requires spectra of much higher resolution to function effectively. Nevertheless, this requirement allows the newer algorithm to more effectively deal with specific challenges in monoisotopic mass determination, particularly by better addressing the problem of so-called *off-by-one dalton errors*, where predictions deviate from the true values by 1, 2, or more daltons.

The `Envemind` algorithm operates in two interrelated stages. Initially, it takes the experimental spectrum and searches a selectively specific part of the chemical formulas' space for a protein with the best-fitting theoretical spectrum, measured in terms of Wasserstein distance. For this process, we combined existing solutions with newly developed strategies for searching the space, enabling us to exclude numerous molecules from computations that clearly had no potential to fit. Once the fitting is complete, the second stage involves applying a predictive model – initially trained on a dataset of simulated spectra – to the fitted spectrum. This approach enables the extraction of spectral features, such as variance

and shape, which are typically nearly impossible to extract directly from experimental data. As a result, the algorithm achieves exceptionally high accuracy in monoisotopic mass prediction.

Extending the capabilities of the MIND algorithm to oligonucleotide analysis, MIND4OLIGOS is introduced as our second development in Chapter 2. Since molecular masses of oligonucleotides are generally lower than those of proteins, it occurred that MIND's methodology yields fewer candidates for the monoisotopic mass. Therefore, our work extended beyond training the model on a different dataset; we also developed a method to choose the best candidate for monoisotopic mass, but that was not so straightforwardly available in the case of proteins. Furthermore, our development process included exploring a wide range of additional predictors to enhance the selection of the best monoisotopic mass candidate. Although these predictors proved to be nearly faultless in terms of off-by-one dalton errors and achieved very high accuracy on simulated data, they also demonstrated a high sensitivity to noise. This highlighted a key limitation: while it's possible to further improve prediction accuracy with these tools, their practical application depends heavily on advancements in the resolution of mass spectrometers.

1.2. Compression of mass spectrometry images

Going one step further, we can expand beyond working with single mass spectra by incorporating a spatial dimension into our research through Mass Spectrometry Imaging (MSI). Unlike conventional photography, where each pixel represents a specific color, MSI introduces a more complex scenario. Simplifying, MSI involves using a laser to scan a properly preprocessed tissue cross-section. This laser ablate molecules from precise locations, which are carried with neutral gas into a mass spectrometer for measurement. Consequently, each pixel in an MS image contains a mass spectrum, providing a rich source of information about the tissue's biological and chemical composition.

MSI's ability to precisely map the spatial distribution of proteins, lipids, and metabolites offers detailed insights into the molecular composition across various tissues and their states. As such, this technique significantly contributes to biomarker discovery, enhances therapeutic methods, and deepens our understanding of disease mechanisms, underscoring its vital role in advancing precision medicine (Chughtai and Heeren, 2010; Schwamborn, 2012; Longuespée *et al.*, 2016; Vaysse *et al.*, 2017). However, to fully utilize the benefits of MS imaging, we must first address a fundamental challenge: the size of the images. The extensive memory requirements and computational load associated with these large data volumes can hinder the efficient and feasible application of essential machine learning techniques and neural networks for in-depth analysis (Alexandrov, 2020).

Thus far, researchers have primarily focused on optimizing the data processing efficiency of computational algorithms to adapt to the structure of MSI data within the constraints of memory and CPU resources. However, our approach tackles this challenge from a different perspective. Instead of merely speeding up analytical algorithms, we have developed a compression algorithm designed to preprocess MSI data, significantly reducing its size and making it more manageable for subsequent analysis, detailed in Chapter 3. For instance, we were able to compress an MS mouse bladder image from an initial size of 1.5 GB to just 8.5 MB.

The algorithm employs contrastive learning, a technique that teaches the model to differentiate between similar and dissimilar data points by comparing instances (Le-Khac *et al.*, 2020). The technique trains the model to generate consistent outputs for similar inputs and varied outputs for dissimilar ones without requiring explicit knowledge of the dataset. Through the application of contrastive learning, our algorithm efficiently learns high-level features from MSI data, effectively compressing the data while preserving the essential information required for subsequent analysis.

To validate our algorithm, we collaborated with prof. Benjamin Balluff from the Maastricht MultiModal Molecular Imaging Institute at Maastricht University in the Netherlands, an expert in the MSI field. The most fundamental analytical method applied to MSI data is segmentation. First, however, researchers typically apply the t-distributed stochastic neighbor embedding (t-SNE) algorithm to estimate the number of clusters expected in subsequent segmentation. Although even this initial step is often challenging due to the large data size, we ran the t-SNE algorithm smoothly on encoded images. Next, we found that the segmentation task was easily applicable to the encoded images as well, and it yielded even higher accuracy (according to established baseline models) compared to raw data. Moreover, in the end, it is important to highlight that the encoding transformation is reversible: encoded images can be decoded to recover the original input, ensuring that no critical information is lost during the compression process. This reversibility opens up an additional application of our encoder, allowing for more efficient storage of extensive MSI databases.

1.3. Analysis of measurements' uncertainty for atomic mass spectrometry

Till this point, the discussion has centered on what can be termed *molecular mass spectrometry*, as it relates exclusively to molecules. Now, we will move to a distinct field known as *elemental* or *atomic mass spectrometry*. Most commonly, atomic mass spectrometry is associated with the *inductively coupled plasma mass spectrometry* (ICP-MS) and the *atomic absorption spectrometry* (AAS), widely used to determine the level of elements in a given

sample. However, in this dissertation, we focus on a specific and far less common type of ICP mass spectrometry, referred to by the addition of the prefix *multicollector*; thus, MC-ICP-MS.

Initially, the term "multicollector" might seem misleading; the MC-ICP-MS can measure only a narrow m/z range compared to other mass spectrometers. However, what sets this instrument apart is its unique ability to measure all isotopes within that narrow range simultaneously. This capability effectively limits the impact of plasma fluctuations, as the multicollector monitors various isotopes that have been subjected to the exact same conditions. Therefore, in essence, the usage of MC-ICP-MS enables measurements of the variation in the isotopic composition of objects in the most accurate way currently available or, roughly speaking, allows us to study whether an isotopic composition of a given chemical element from a sample of unusual origin deviates from – in notable simplification – the composition of a standard.

The process, which begins with measuring the electric current generated by ions that strike a conductive surface of the detector and ends by yielding a specific isotopic ratio, is long and complex. To interpret the raw signals provided by MC-ICP-MS, the application of so-called *calibration methods* is mandatory. Unfortunately, each step in these methods affects the measurement's uncertainty in some way. In our research, we focused on analyzing this uncertainty, particularly for the sample-standard bracketing (SSB) correction method, which led to the development of a Monte Carlo-based algorithm to estimate the expanded uncertainty.

Given the rarity of the described technology, Chapter 4 begins with a brief introduction to how the spectrometer works, the exact nature of its measurements, and a review of the fields where knowledge of accurate isotopic composition proves beneficial. Following this overview, we describe three calibration methods – SSB and two additional ones that we will revisit in the context of future research in Chapter 6. Finally, at the heart of Chapter 4, we provide a detailed description of our uncertainty estimation algorithm and discuss its contribution to advancing the field.

1.4. Statistical methods for medical research

The interdisciplinary nature of our work naturally leads to a wealth of collaboration opportunities. In Chapter 5, we present brief summaries of our medical research projects that have been already published. These summaries capture the essence of our collaboration with various scientific and medical experts, illustrating how our research extends beyond mass spectrometry yet contributes to the broader scientific dialogue. Readers interested in detailed descriptions of data involved in the studies, methodologies employed, and the re-

sults obtained will find comprehensive information in the original publications.

At the beginning of the chapter, we discuss studies related to the COVID-19 pandemic, which coincided with the first year of our PhD studies. During the early stages of the pandemic, as global lockdowns were implemented, we, along with a large international group led by prof. Ali Jawaid, conducted a comprehensive survey on mental health. This survey included over 13300 participants worldwide and aimed to assess the effects of lockdowns on mental health. The research resulted in two publications: the first identified factors significantly impacting mental health during lockdowns, while the second specifically focused on individuals with pre-pandemic psychiatric conditions and how these conditions were affected during the lockdowns.

Our subsequent collaboration, arising after the delta wave, involved pediatricians from the Warsaw University of Medicine (WUM). We analyzed survey data from a children's hospital to characterize the profiles of individuals who initially introduced infection into households.

Finally, in collaboration with pulmonologists from WUM, we analyzed a comprehensive database of patients with sarcoidosis. Our research identified a series of strong correlations between sTNF-α levels and various disease predictors, enabling us to propose a new sarcoidosis phenotype related to sTNF-α.

List of publications of major results from the thesis

Radziński, P., Skrajny, J., Moczulski, M., Ciach, M., Valkenborg, D., Balluff, B., Gambin, A., *Efficient compression of mass spectrometry images via contrastive learning-based encoding.* – Under review at Analytical Chemistry as of January 2025. –

Prostko, P., Radziński, P., Ciach, M., Liu, Y., Startek, M., Lermyte, F., De Vijlder, T., Gambin, A., Appeltans, S., Valkenborg, D. (2024). *MIND4OLIGOS: determining the monoisotopic mass of oligonucleotides observed in high-resolution mass spectrometry.* Analytical Chemistry, 96(23), 9343-9352.

Karasiński, J., Tetfejer, K., Radziński, P., Tupys, A., Gambin, A., Bulska, E., Halicz, L. (2024). *Coprecipitation as a One-Step Se Separation for Determination of Isotope Ratios Completed with Revised Uncertainty Evaluation.* Analytical Chemistry, 96(9), 3763–3771.

Radziński, P., Valkenborg, D., Startek, M. P., Gambin, A. (2022). *Envemind: Accurate Monoisotopic Mass Determination Based On Isotopic Envelope.* Journal of the American Society for Mass Spectrometry, 33(11), 2063-2069.

List of other publications

Goljan-Geremek, A., Radziński, P., Puścińska, E., Demkow, U. (2024). *Defining serum tumor necrosis factor α concentration-related endotype of sarcoidosis: a real-life, retrospective, observational Polish study.* Polish archives of internal medicine, 134(4), 16718.

Mańdziuk, J., Okarska-Napierała, M., Woźniak, W., Hryniewicka, A., Radziński, P., Gambin, A., Podsiadły, E., Demkow, U., Kuchar, E. (2023). *Monte Carlo Regression for Evaluating Children's Role in the Pandemic Spread on the Example of Delta COVID-19 Wave.* the Pediatric Infectious Disease Journal, 42(12), 1086-1092.

Płomecka, M., Gobbi, S., Neckels, R., Radziński, *et al.* (2021). *Factors associated with psychological disturbances during the COVID-19 pandemic: multicountry online study.* JMIR mental health, 8(8), e28736.

Gobbi, S., Płomecka, M. B., Ashraf, Z., Radziński, P., *et al.* (2020). *Worsening of preexisting psychiatric conditions during the COVID-19 pandemic.* Frontiers in psychiatry, 11, 581426.

Acknowledgements

First and foremost, I wish to express my deepest gratitude to prof. Anna Gambin, for her unwavering patience, support, and guidance – both in research and in life. Without her, I would not be on the path I am today.

I am deeply grateful to Jakub Karasiński and Michał Startek for their mentorship. Jakub's patience and willingness to explain complex concepts made learning chemistry manageable and amusing, while Michał's guidance helped me navigate the vast world of informatics.

Science is always a collaborative effort, and I thank all my university colleagues and coauthors who have been part of this experience. Your collaboration, insightful discussions, and shared challenges have made this journey both enriching and enjoyable. Mainly, I would like to mention Maria Bochenek, Michał Ciach, Barbara Domżał, Susanna Gobbi, Martyna Płomecka, Barbara Poszewiecka, Grzerzorz Preibisch, Grzegorz Skoraczyński, Jakub Skrajny and Dirk Valkenborg – the PhD journey would have been far more difficult and far less fulfilling without you.

A heartfelt thank you to my parents, Joanna and Paweł, for their almost perfectly balanced mix of understanding and gentle pressure – though, let's be honest, the latter often took the lead. Their support, in all its forms, has been truly indispensable.

Finally, undoubtedly, the financial support I received was invaluable during my research and preparation of this dissertation. Therefore, I would like to express my gratitude to the following institutions:

- Polish Ministry of Science and Higher Education, for its Excellence Initiative – Research University programme, granted to the University of Warsaw; contract no. 01/IDUB/2019/94. I worked on two projects founded by the programme:
  - *Combinatorial properties of isotopic envelopes as aid in chemical species annotation in mass spectrometry*,
  - *Elimination of the use of an isotope standard in the measurements of isotope ratios by means of multicollector mass spectrometry*.

  Moreover, the program financed my 3-months internship at Hasselt University and several international conference trips as well as granted me a *scholarship for the preparation of doctoral dissertations in accordance with Priority Research Areas issues*.
- Polish National Science Center for founding grants I was enrolled in:
  - OPUS 21 *Optimal-transport based algorithms for Mass Spectrometry and NMR*; no. 2021/41/B/ST6/03526,
  - OPUS 15 *Algorithmic challenges of mass spectrometry*; no. 2018/29/B/ST6/00681.

2

Determination of monoisotopic mass

Monoisotopic mass offers a definitive and unambiguous measure, unlike average and most abundant masses, which may be influenced by variations in isotopic abundances and spectra resolution. Unaffected by natural or technical fluctuations, monoisotopic mass remains consistent, making it a reliable metric. Consequently, it serves as an essential tool for predicting or identifying molecules across a wide range of applications. It is widely used to search databases to identify measured molecules (Hsieh *et al.*, 2010; Little *et al.*, 2011). Additionally, accurately determined monoisotopic mass is the most important tool used to predict the elemental composition of small molecules (Zubarev *et al.*, 1996) as well as their aggregated isotope distribution (Agten *et al.*, 2021), which can aid in identifying unknown compounds. Nevertheless, unfortunately, as previously noted, monoisotopic mass often remains undetected in heavier biomolecules, necessitating the usage of predictive techniques.

The first widespread method for monoisotopic mass determination was proposed by Senko *et al.* (1995). Their algorithm was dedicated to large biomolecules, with prediction based on an average mass of a given spectrum. The characteristic of prediction acquired by the method has relatively high trueness yet very low precision. Nevertheless, the approach is commonly used even nowadays due to its simplicity. Still, precise measurement of the average mass can be challenging due to diverse noises appearing in mass spectra, like chemical, background, stochastic and more (Claesen *et al.*, 2015). Hence, in recent years, scientists looked for more stable features on which prediction can be based. Chen *et al.*

(2013) proposed using a mass of the most abundant peak. The idea was later proceeded by Lermyte *et al.* (2019), who presented `MIND` algorithm that was devoted to proteins. The algorithm used two linear models – first, holding a general relationship between the most abundant peak and monoisotopic peak masses, and second, determining the exact distance between those peaks. As a result, the algorithm offered three candidates for monoisotopic mass with their corresponding probabilities of being the accurate one. The simplicity of this approach and usage of such a well-accessible feature as the most abundant mass causes it to be user-friendly. On the other hand, the most probable monoisotopic mass candidate was often – up to 33.5% of cases – incorrect.

In this chapter, we continue our exploration of monoisotopic mass prediction by detailing two developed algorithms: `Envemind`, which addresses prediction uncertainties for proteins, and `MIND4OLIGOS`, which adapts and enhances `MIND` methodology for application to oligonucleotides.

2.1. `Envemind`: an advanced prediction algorithm for high-resolution protein spectra

The procedure for determining monoisotopic mass is divided into two main parts. Initially, we identify a theoretical spectrum that best matches the given experimental one. Reminding, a theoretical spectrum refers to a simulated, noise-free spectrum, effectively acting as the probability mass function of a multinomial distribution with atomic masses as values and natural isotopic abundances as probabilities (Valkenborg *et al.*, 2012). By "best matching", we refer to a match obtained using a chosen spectra similarity measure. In our case, it is usually the Wasserstein distance, yet we will discuss this topic more in a further part of this chapter. Possessing a theoretical spectrum simplifies the prediction process, enabling the construction of a mathematical model that leverages precisely measured features – typically absent in experimental spectra, such as variance. This model, calibrated against theoretical spectra, yields highly accurate predictions.

For clarity, we first outline the prediction process using theoretical spectra before addressing the adaptation to experimental spectra. A control flow graph illustrating the main steps of the algorithm is presented in Figure 2.1.

2.1.1. Prediction model for theoretical spectra

All computations described in this section were performed using theoretical spectra generated exclusively by `IsoSpec` algorithm (Łącki *et al.*, 2017). These spectra were simulated based on chemical formulas of proteins randomly selected from the Uniprot database

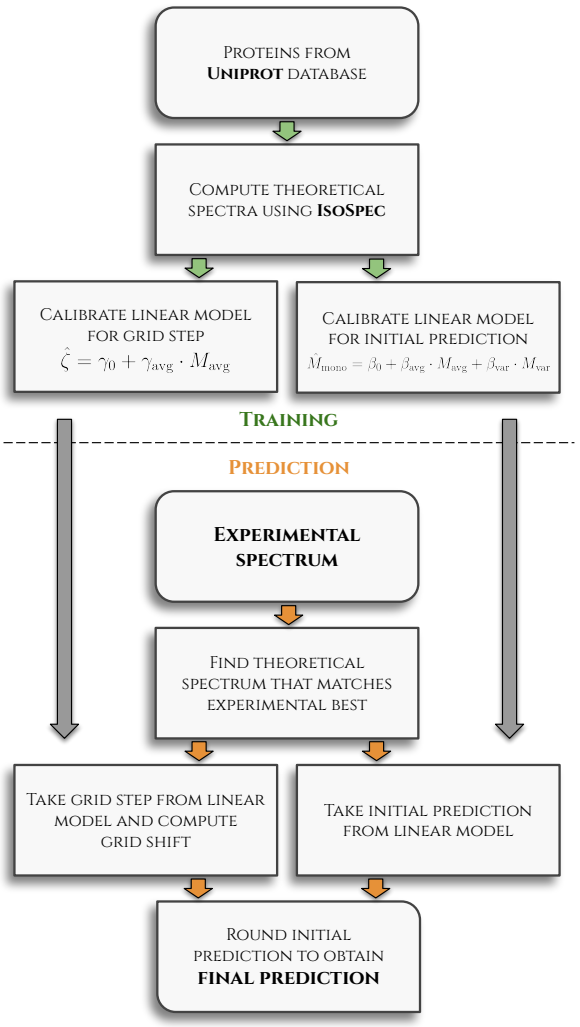

Figure 2.1: Control flow graph illustrating methodology of `Envemind` algorithm. It is divided into two parts. First, *training*, with calibration of linear models on theoretical spectra, and second, *prediction*, that deals with experimental spectrum by fitting procedure and linear models from the previous part.

(UniProt Consortium, 2015), with molecular masses ranging from 8 to 400 kDa. Theoretical spectra are presented in daltons, without incorporating any charge. Consequently, all parameters discussed herein are also expressed in daltons.

Let us denote the monoisotopic mass by M_{mono}. In the beginning, simple initial predictions are derived from a linear model that utilizes the average mass of a protein, M_{avg}, and the variance of its spectrum, M_{var}. This model has been trained and tested using 10-fold cross-validation. The functional form of the model is expressed as

$$\hat{M}_{\text{mono}} = \beta_0 + \beta_{\text{avg}} \cdot M_{\text{avg}} + \beta_{\text{var}} \cdot M_{\text{var}}, \tag{2.1}$$

with fitted coefficients $\beta_0 = -0.1456$, $\beta_{\text{avg}} = 0.9998$ and $\beta_{\text{var}} = -0.5982$, and mean absolute error (MAE) equal to 0.1383. While the linear predictor (2.1) may yield erroneously shifted outcomes for some proteins, the cross-validation results indicate an absolute prediction error below 0.5 Da for approximately 96.6% of proteins. Therefore, the next step aims

to reduce this bias by predicting a grid of potential locations for the monoisotopic mass. Recall that isotopologues in spectra aggregate into clusters, typically separated by approx. 1 Da. Thus, determining this grid allows us to round $\hat{M}_{\text{mono}}$ to the nearest grid point. We define the grid as

$$\mathcal{G}(\zeta, \Delta) = \{\zeta n + \Delta \,:\, n \in \mathbb{N}\}, \tag{2.2}$$

where ζ represents the distance between nodes on the grid, and Δ represents the grid's shift.

Estimation of the grid step ζ

The purpose of the ζ parameter is to control the spacing between peak clusters. To determine the distance between two consecutive grid nodes, envision a circle being rolled along the protein's spectrum, akin to a glue-coated roller that collects all peaks. The optimal grid step length, represented by the circle's circumference ζ, is achieved when all the peaks adhere within a small section of the circle, as illustrated in Figure 2.2. If all isotope peaks are gathered and confined to a compact region on this sticky roller, then the circle's circumference effectively matches the average distance between consecutive isotope clusters.

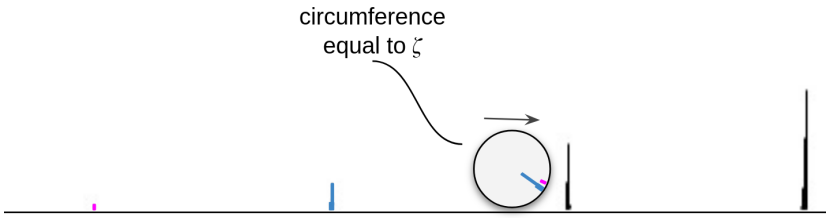

Figure 2.2: Circle that rolls across spectrum. For ζ that estimates well the distance between groups of peaks, peaks transformed into the circumference of a circle should overlap on a small fragment.

More formally, the sticky roller procedure described above transforms all peaks in the spectrum $\mathcal{S}$, i.e., pairs $p = (p^{\text{mass}}, p^{\text{prob}})$, to complex unit circle, rotates them to average zero (to avoid problems with logarithm specification on complex plane), and then transforms to the interval $[-\zeta/2, \zeta/2]$. The final transformation is expressed as

$$P_\zeta(z) = \frac{\zeta}{2\pi i} \log\left[\exp\left(\frac{2\pi i z}{\zeta} - i\text{Im}\Big[\log\Big(\sum_{p\in\mathcal{S}} p^{\text{prob}} \cdot \exp\left(2\pi i p^{\text{mass}}/\zeta\right)\Big)\Big]\right)\right]. \tag{2.3}$$

To quantify the concentration of peaks, we treat the spectrum $\mathcal{S}$ as a random variable and make use of the notion of variance. Thus, the optimal ζ, denoted as $\zeta^\star$, is the one that minimizes the variance of the transformed spectrum, i.e.,

$$\zeta^\star = \underset{\zeta\in\mathbb{R}}{\text{argmin}}\, \text{Var} P_\zeta(\mathcal{S}). \tag{2.4}$$

Unfortunately, the variance minimization procedure requires several minutes of computation for average-sized protein, and this time increases exponentially for larger molecules,

which may limit practical applications. Notably, we observed that the optimal grid step, $\zeta^\star$, shows a slight correlation with the protein's average mass (Pearson: 0.12, Kendall: 0.07, Spearman: 0.11). Based on this observation, we adjusted a linear model to approximate $\zeta^\star$ as a function of M_{avg} formulated as

$$\hat{\zeta} = \gamma_0 + \gamma_{\text{avg}} \cdot M_{\text{avg}}, \tag{2.5}$$

with coefficients $\gamma_0 = 1.002355$ and $\gamma_{\text{avg}} = 6.9584 \cdot 10^{-10}$, and MAE equal to $1.3102 \cdot 10^{-4}$ in 10-fold cross-validation. The fitted regression line is presented in Figure 2.3. For future reference, it is important to note that since γ_{avg} is very small, any measurement error in M_{avg} results in a negligible change to $\hat{\zeta}$. Consequently, $\hat{\zeta}$ can be reliably computed using the average mass of the experimental spectrum as well.

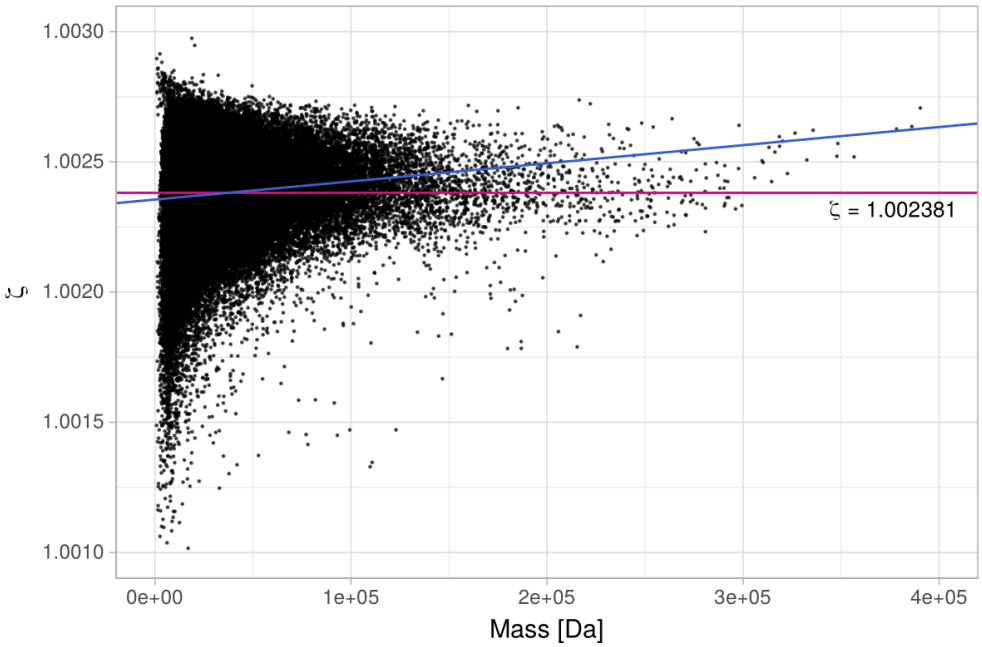

Figure 2.3: Average theoretical mass versus ζ that minimizes variance on a complex circle plot. The pink line represents the average ζ, equal to 1.002381, while the blue line is adjusted by linear regression as described in Equation (2.5).

Estimation of the grid shift Δ

Once we have chosen $\hat{\zeta}$, we can focus on the grid shift parameter Δ. In this setting, the Δ matches the spectrum best if it minimizes the distance between grid points and spectrum peaks, and can be mathematically expressed as

$$\hat{\Delta} = \underset{\Delta \in [0, \hat{\zeta}]}{\operatorname{argmin}} \sum_{p \in \mathcal{S}} p^{\text{prob}} \cdot \min_{g \in \mathcal{G}(\hat{\zeta}, \Delta)} |p^{\text{mass}} - g|. \tag{2.6}$$

However, this process can equivalently be understood as transforming the spectrum to a complex unit circle and finding the mean point in the complex space, i.e.,

$$\hat{\Delta} = \mathrm{Re}\left[\frac{\hat{\zeta}}{2\pi i}\log\left(\sum_{p\in\mathcal{S}} p^{\text{prob}}\cdot\exp\left(2\pi i p^{\text{mass}}/\hat{\zeta}\right)\right)\right]. \tag{2.7}$$

While both approaches yield the same $\hat{\Delta}$, implementing the second approach significantly enhances computational efficiency, making it feasible to calculate the grid shift for each protein independently.

Final prediction

With both parameters computed, we can proceed to the final prediction on the theoretical spectra. Recall that, according to the described procedure, the final step in predicting the monoisotopic mass M_{mono} involves rounding the initial prediction $\hat{M}_{\text{mono}}$ to the closest point on the grid $\mathcal{G}(\hat{\zeta},\hat{\Delta})$.

However, it turns out that the distances between clusters of aggregated peaks in a given spectrum are not perfectly uniform. When analyzing the distribution of inter-cluster distances centered on the most abundant peak, those in the left tail (for smaller masses) tend to increase slightly. This causes the distribution of errors after rounding to be marginally shifted from zero. As a remedy, we introduced the term λ, which adjusts the distribution of errors to have an expected value of zero, resulting in a subtly more accurate prediction.

To explain the computation of λ, consider that we have dependent variables Y_j and their corresponding predictions $\hat{Y}_j$. The mean error we aim to correct is expressed in parts per million (ppm), and it can be formulated as

$$c = \underset{j}{\mathrm{mean}}\left[\frac{Y_j-\hat{Y}_j}{Y_j}\cdot 10^6\right]. \tag{2.8}$$

Since, in our case, $c \neq 0$ – indicating that the distribution of ppm errors is not centred around zero – we can improve the prediction by adjusting the distribution to have a mean of zero. To achieve this, we seek a value z that satisfies the equation

$$\frac{\hat{Y}_j+z}{Y_j}\cdot 10^6 = \frac{\hat{Y}_j}{Y_j}\cdot 10^6 - c \quad\Longrightarrow\quad z = -\frac{cY_j}{10^6}. \tag{2.9}$$

Note that usually, during the prediction of $\hat{Y}_j$, we don't know the value of Y_j, and such considerations would typically be impractical. However, in our specific case, this limitation can be overcome by using the initial prediction $\hat{M}_{\text{mono}}$ to adjust the distribution of final predictions. For our data, $c \approx 0.12$, which resulted in $\lambda = -1.1982\cdot 10^{-7}$.

Summarizing our model for theoretical protein data, the final prediction of the monoisotopic mass is accurately represented by the following equation

$$\hat{\hat{M}}_{\text{mono}} = \underset{g\in\mathcal{G}(\hat{\zeta},\hat{\Delta})}{\mathrm{argmin}}\left|g-\hat{M}_{\text{mono}}\right| + \lambda\cdot\hat{M}_{\text{mono}}. \tag{2.10}$$

Distributions of errors for both the initial and final predictions, which illustrate the impact of the rounding process on the proposed grid, are presented in Figure 2.4.

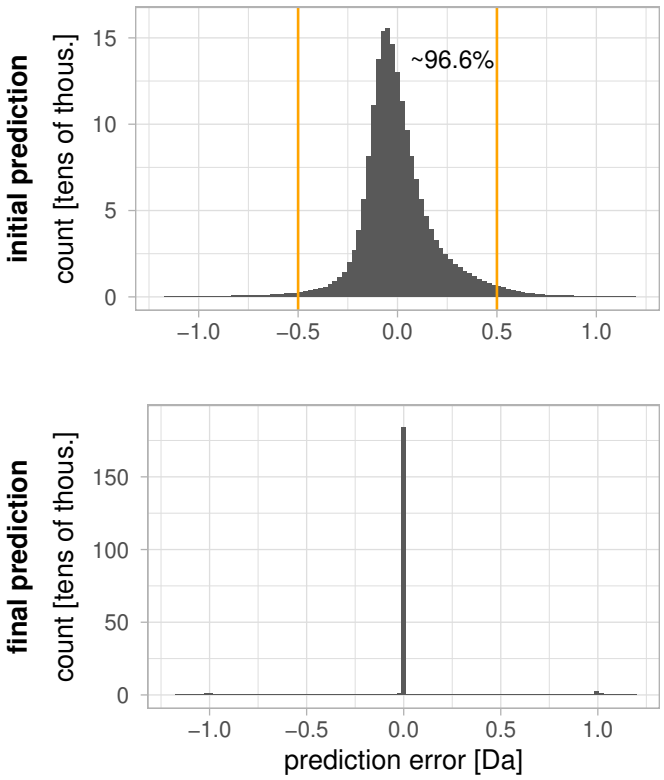

Figure 2.4: Comparison between the linear model's initial prediction and the final prediction made by adjusting the grid and rounding the initial prediction to the closest point on it. Orange lines designate an interval inside which predictions will round to the most appropriate point on the grid $\mathcal{G}(\hat{\zeta}, \hat{\Delta})$.

2.1.2. Matching of theoretical spectra to experimental ones

Once the model for theoretical spectra is established, there arises a necessity to adapt it for experimental data as well. A significant challenge we face is that variance, which is essential for the previously described model, is difficult to derive from experimental spectra. Therefore, for each given experimental spectrum – denoted by E – we aim to construct a theoretical spectrum that closely matches the experimental one, as quantified by a chosen spectra similarity measure.

Our construction of the simulated spectrum builds upon the concept of *averagine* as proposed by Senko *et al.* (1995), although with more complexity. First, we define averagine as a hypothetical molecule with the chemical formula

$$\mathrm{C}_{4.9245}\ \mathrm{H}_{7.7724}\ \mathrm{N}_{1.3555}\ \mathrm{O}_{1.4600}\ \mathrm{S}_{0.0356},$$

which has an average mass of 110.47 Da. Following Senko's methodology, we updated the formula using the larger Uniprot database, resulting in slight modifications to the original averagine composition.

As mentioned earlier, Senko proposed an algorithm to approximate the chemical formulas of proteins, facilitating the calculation of monoisotopic mass. His method involves scaling the averagine model to match the experimental spectrum's average mass, rounding the resulting formula to whole numbers, and adjusting it with additional hydrogen atoms to align closely with the experimental data once more. Today, with the aid of the `IsoSpec` algorithm (Łącki *et al.*, 2017), we can efficiently simulate the theoretical spectrum for any given chemical formula. This algorithm has been proven to be optimal and operates in linear time, proportional to the size of the molecules. Consequently, we explore the 5-dimensional space of protein chemical formulas, where each dimension corresponds to the count of a specific chemical element (C, H, N, O, and S) in a molecule. Given this vast space, our aim extends beyond merely considering the mass of the protein to also encompassing the variations in the shapes of its isotopic envelope.

At this point, we suggest keeping an eye on Figure 2.5 when reading the following description, as it should enhance understanding. Let us define spectrum A_ρ^k with two parameters. A_0^0 represents Senko's averagine shifted this way, that its first aggregated peak to the left of its average mass is in the place of the most abundant peak of E (red frame in Figure 2.5).

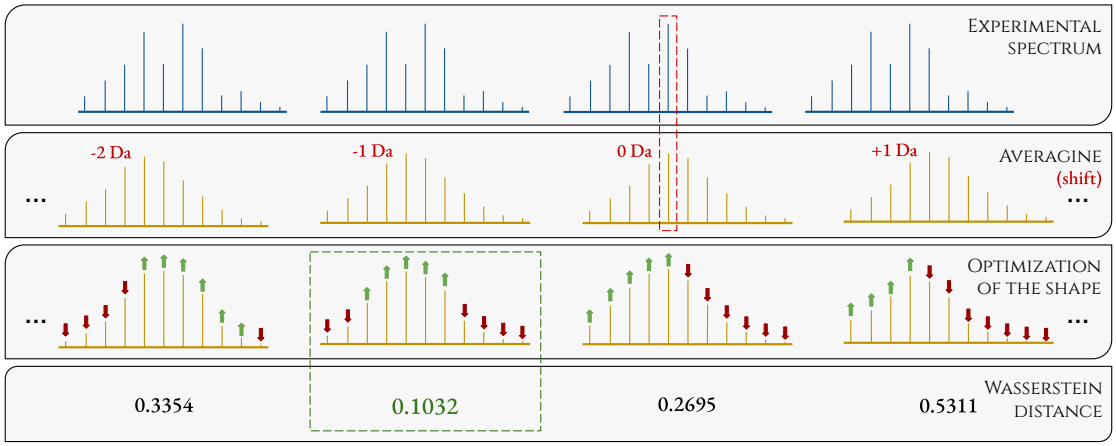

Figure 2.5: The simulated spectrum matching scheme for a given experimental one. In the first row, copies of the experimental spectrum are presented. The row below shows copies of averagine with different shifts. For example, a shift by 0 Da means that the most abundant aggregated peak of the simulated spectrum is in the place of the most abundant peak of the experimental spectrum (red frame). Then, we optimize variance for every averagine copy separately due to Wasserstein distance to the experimental spectrum. Arrows show if intensities grow or decrease due to the optimization. Finally, the spectrum with the lowest score will be used to further prediction by linear models (green frame).

The parameter k is responsible for shifting the spectrum, and it indicates that the spectrum is shifted by $k \cdot \hat{\zeta}$ Da. To ensure that peaks are accurately placed, we keep k as an integer. This method allows for the creation of multiple shifted copies of averagine, positioned accurately to ensure that the peaks of the simulated spectra align with those in the experimental one. Examples of such shifted copies of averagine can be seen in the second row of Figure 2.5.

The parameter ρ is responsible for altering the variance. We aim to modify the averagine model's chemical formula such that it affects the spectrum's variance without impacting its average mass. To achieve this optimally, we prove the following lemma.

Lemma 1. *Let $f(x) = \langle v, x\rangle$ and $\Lambda = \{x \in \mathbb{R}^n : \langle x, x\rangle = 1, \langle a, x\rangle = 0\}$, for given $a, v \in \mathbb{R}^n$. Then extrema of $f(x)$ on Λ are reached in*

$$x_1 = \frac{v - \text{proj}_a v}{||v - \text{proj}_a v||} \quad \text{and} \quad x_2 = -\frac{v - \text{proj}_a v}{||v - \text{proj}_a v||}.$$

Proof. Since Λ is a compact manifold, we can use the method of Lagrange multipliers to find extrema. We define the Lagrange function as

$$L(x, \lambda_1, \lambda_2) = \langle v, x\rangle + \lambda_1(\langle x, x\rangle - 1) + \lambda_2\langle a, x\rangle,$$

which – by taking derivatives over λ_1, λ_2 and x, respectively – gives us following system of equations

$$\begin{cases} \langle x, x\rangle - 1 = 0 \\ \langle a, x\rangle = 0 \\ v + 2\lambda_1 x + \lambda_2 a \ = 0 \end{cases}$$

the third equation provides us with $x = -\frac{v+\lambda_2 a}{2\lambda_1}$. Then, by inserting this result into the first equation, we obtain $\lambda_1 = \pm\frac{1}{2}||v + \lambda_2 a||$. Hence, we can write $x = \mp\frac{v+\lambda_2 a}{||v+\lambda_2 a||}$. Finally, by using the second equation and inserting the x into it, we obtain $\mp\langle a, v + \lambda_2 a\ \rangle = 0$, and therefore $\lambda_2 = -\frac{\langle a,v\rangle}{\langle a,a\ \rangle} = -\text{proj}_a v$, which provides us extrema $x = \mp\frac{v-\text{proj}_a v}{||v-\text{proj}_a v||}$. ∎

In our case, a represents a vector of average masses, and v a vector of variances, as outlined in Table 2.1. We aim to maximize the variance value, $\langle v, x\rangle$, while constraining the vector x to a unit ball, $\langle x, x\rangle = 1$, and ensuring it does not alter the average mass, i.e., $\langle a, x\rangle = 0$. The resulting variance gradient is also presented in Table 2.1.

element	C	H	N	O	S
average mass	12.0108	1.0079	14.0067	15.9994	32.0649
variance	0.0107	0.0001	0.0036	0.0086	0.1700
variance gradient	-0.3430	-0.0372	-0.4960	-0.5184	0.6050

Table 2.1: Average masses, variances and variance-growth gradient (in daltons) of basic protein-building elements that make up our protein space of chemical formulas. The variance-growth gradient indicates the direction in which the variances of the spectra increase the fastest, yet without altering the average masses of the proteins.

Specifically, ρ controls how much of the variance-growth gradient should be added to the initial averagine formula, yet before it is rounded to integers.

With both parameters explained, finally, they must be optimized to obtain the best-matching simulated spectrum. To compare spectra, we use the Wasserstein distance $\mathcal{W}$

(Rubner *et al.*, 2000; Villani, 2008). The Wasserstein distance between two probability distributions π and η over a metric space (Ω, d) is defined as the infimum of the expected value of the distance $d(X, Y)$, where X and Y are random variables with distributions π and η respectively, i.e.,

$$\mathcal{W}(\pi, \eta) = \inf_{\gamma \in \Gamma(\pi, \eta)} \int_{\Omega \times \Omega} d(x, y) \mathrm{d}\gamma, \tag{2.11}$$

where $\Gamma(\pi, \eta)$ denotes the set of all joint distributions γ on $\Omega \times \Omega$ with marginals π and η, called *transport plans*. However, in the case of one-dimensional distributions, the problem can be simplified as the optimal transport plan can be expressed by

$$\mathcal{W}(\pi, \eta) = \int_{\mathbb{R}} \left| F_\pi(x) - F_\eta(x) \right| \mathrm{d}x, \tag{2.12}$$

with F denoting the CDF function of the lower-indexed distribution. As this approach provides efficiency (computations are done in linear time) and yields relatively accurate comparison scores, we found the Wasserstein distance to be the optimal choice.

Going back to the optimization of our parameters, note that for every fixed shift k of the initial averagine, a different value of ρ will minimize the Wasserstein distance. Therefore, for a given k, the optimal quantity of ρ can be formulated as

$$\rho_k = \operatorname*{argmin}_{\rho \in \mathbb{R}} \mathcal{W}(E, A^k_\rho), \tag{2.13}$$

and subsequently, by optimizing the ρ_k values with respect to k, i.e.,

$$\hat{k} = \operatorname*{argmin}_{k \in \mathbb{Z}} \mathcal{W}(E, A^k_{\rho_k}), \tag{2.14}$$

we obtain the final best-matched simulated spectrum in the form of $A^{\hat{k}}_{\rho_{\hat{k}}}$. In simpler words, we optimize the variance separately for different copies of averagine, each shifted by $k \cdot \hat{\zeta}$ Da, and then select the spectrum that has the lowest Wasserstein distance to the experimental spectrum E. We use this spectrum to run the previously described prediction for theoretical spectra to determine the monoisotopic peak mass. Examples of the constructed spectra, alongside their theoretical and experimental counterparts, can be found in Figure 2.6.

Data

We selected proteins from the Uniprot database (UniProt Consortium, 2015) within an 8–400 kDa mass range to train and test linear models on theoretical spectra. The spectra were generated from chemical formulas using `IsoSpec` isotopic envelope calculator, ensuring coverage of at least 99% of intensity for each protein. The linear models (2.1) and (2.5) were cross-validated with datasets comprising approximately $1.9 \cdot 10^6$ and $8 \cdot 10^4$ randomly selected proteins, respectively.

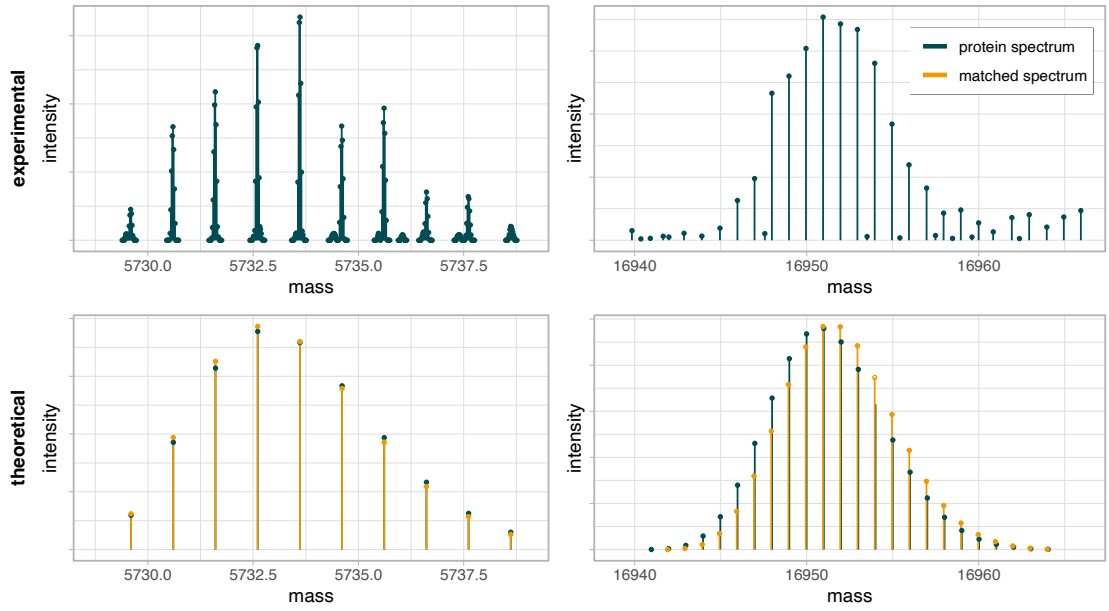

Figure 2.6: Two examples of matched simulated spectra. In each column, the experimental and theoretical spectra of insulin and apo-myoglobin are presented, respectively (blue). In the second row, simulated spectra that have been matched to the experimental ones are overlaid (orange). Note that in the right column, a poorly matched spectrum is presented, where an off-by-one-dalton error occurred.

We employed the same experimental spectra as used in `MIND` study for our proof-of-concept experiments; details of their acquisition are described in the Lermyte *et al.* (2019) manuscript. Our dataset includes:

- Bovine insulin: Merck's sigma catalogue no. I5500, featuring 50 scans with three well-visible charges; monoisotopic mass 5729.60 Da, average mass 5733.58 Da.
- Equine apo-myoglobin: Merck's sigma catalogue no. M0630, with 200 scans showing two well-visible charges and an additional 100 scans with three well-visible charges; monoisotopic mass 16940.97 Da, average mass 16951.50 Da.
- Equine cytochrome c: Merck's sigma catalogue no. C2506, comprising 400 scans with three well-visible charges; monoisotopic mass 12352.23 Da, average mass 12360.21 Da.

In total, we utilized 550 high-resolution experimental spectra segments (150 from bovine insulin and 400 from equine apo-myoglobin) along with 1500 spectra segments with low signal-to-noise ratio (300 from equine apo-myoglobin and 1200 from equine cytochrome c), each segment representing a selected interval with a single charge.

2.1.3. Performance and limitations

We compare `Envemind`'s performance with `MIND` algorithm, as it is well-known and widely used in pharmaceutical companies. The tests are divided into three categories based on

spectra quality. First, we used simulated spectra computed by `IsoSpec` algorithm (Łącki *et al.*, 2017), derived from the chemical formulas of actual proteins from the Uniprot database (UniProt Consortium, 2015). Then, we conducted tests on two groups of experimental spectra: one with well-visible isotopic envelopes (detectable "by-eye") and another with somewhat noisy spectra.

`Envemind` algorithm achieved a mean absolute error (MAE) of 0.51 ppm (0.0358 Da) in monoisotopic mass prediction on simulated spectra ranging from 8 to 400 kDa. For 96.6% of the proteins, no off-by-one dalton errors occurred, resulting in an MAE of 0.0526 ppm (0.0020 Da) for this subset. Since `MIND` algorithm was trained on proteins in the 8-60 kDa mass range, we divided and limited the comparison of the algorithms on simulated spectra into three mass range groups: 8–20, 20–40, and 40–60 kDa. Detailed results for cases without off-by-one dalton errors are provided in Table 2.2, while distributions of off-by-one dalton errors are shown in Figure 2.7.

kDa	`Envemind`	`MIND`
8–20	0.0693 (0.0009)	0.0670 (0.0009)
20–40	0.0474 (0.0014)	0.0549 (0.0016)
40–60	0.0393 (0.0019)	0.0478 (0.0023)

Table 2.2: Comparison of the accuracies of `Envemind` and `MIND` algorithms on simulated spectra, for cases where no off-by-one dalton errors occurred, across three mass range groups. Errors are provided in ppm (Da).

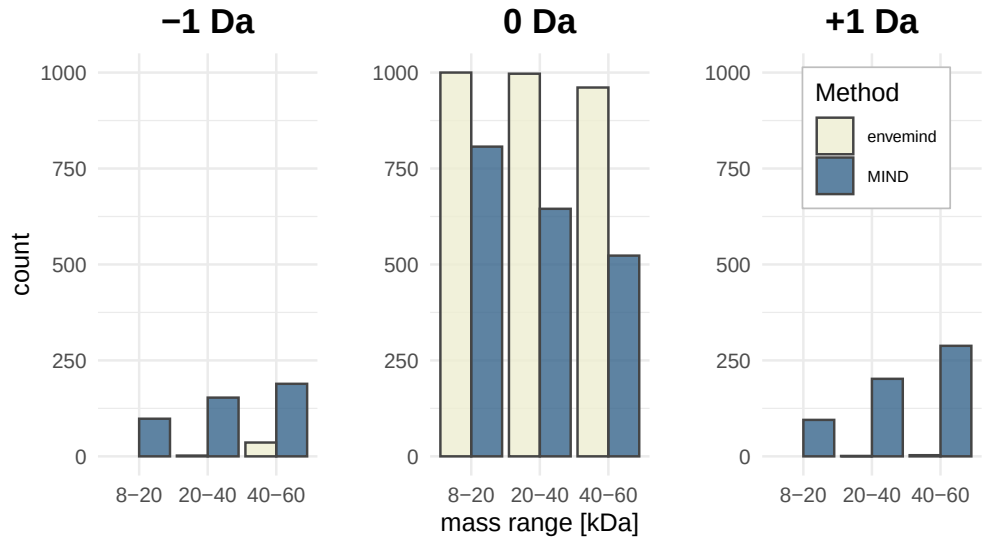

Figure 2.7: Comparison of the frequency of off-by-one dalton errors for `Envemind` and `MIND` algorithms. Tests were conducted on theoretical spectra of 1000 randomly chosen proteins, grouped into three different mass ranges. In rare cases where the absolute prediction error is approx. 2 Da or more, the error is included in the corresponding ± 1 Da bar.

For tests on spectra with visible isotopic envelopes, we used 550 spectra segments. It is worth noting that although the average mass of insulin falls slightly below the mass

range of the training set, it still yields accurate results. The MAE was 2.57 ppm (0.0338 Da). In 547 cases (99.5%), no off-by-one dalton errors occurred. When considering only these cases, the MAE drops to 2.06 ppm (0.0286 Da). Of the 550 spectra segments, 400 fit within the mass range of `MIND` algorithm. In these cases, `MIND` had no off-by-one dalton errors, while `Envemind` had 2 such errors out of 400. The MAE for cases without off-by-one dalton errors was 2.06 ppm (0.0349 Da) for `Envemind` and 1.64 ppm (0.0278 Da) for `MIND`.

Subsequently, we tested `Envemind` on 1500 spectral segments that exhibited lower quality than suitable for optimal performance. Off-by-one dalton errors occurred in 1258 of these cases. The MAE was 2.53 ppm (0.0367 Da) for segments unaffected by these errors. For `MIND`, there were 799 off-by-one dalton errors, with the MAE of 2.13 ppm (0.0280 Da) in the remaining cases.

At the conclusion of this section, let us remind ourselves of two essential assumptions of `Envemind` algorithm. The first relates to data quality; notably, its performance significantly improves as the quality of the data increases. Thus, `Envemind` is particularly recommended for use with the highest quality spectra, where the majority of the isotopic envelope is clearly visible, especially when the user aims to minimize the probability of off-by-one dalton errors. Additionally, its capability to handle a wide mass range provides another advantage over competing algorithms. On the other hand, the `MIND` algorithm requires only the mass of the most abundant peak for its predictions, making it an effective tool for analyzing noisy spectra where only a few peaks are visible. The second assumption is that `Envemind` operates solely on spectra from a single molecule. Therefore, the molecule under study must be isolated during the measurement process, or its signal must be mathematically resolved using *deconvolution*. In mass spectrometry, deconvolution is the process of separating overlapping signals that arise from multiple molecules or their fragments. As a result, we aim to determine the isotopic envelopes of each molecule in the mass spectrum and estimate their relative abundances. Suitable algorithms, such as `masserstein` (Ciach *et al.*, 2020), can be used for this purpose.

2.1.4. Exploring the space of chemical formulas

Let us discuss the matching of theoretical spectra to experimental ones more deeply. We have previously described a construction method that compromises good predictions in a short computational time by use of the Wasserstein distance. However, numerous other measures for comparing spectra exist and offer valuable insights into spectra similarity, but they generally require significantly more computational time than the Wasserstein distance, like `masserstein` algorithm (Ciach *et al.*, 2020). Fortunately, the `Envemind` algorithm is very flexible, allowing easy integration of new measures as they are developed. This could improve predictions and extend `Envemind`'s utility to spectra where the isotopic envelope

is not clearly visible, though at the cost of longer computational time.

Moreover, the entire process of obtaining a simulated spectrum can be replaced. We propose an alternative method that we anticipate may be more accurate. Currently, the method is limited by the lengthy computational times required by modern measures, rendering it inefficient. However, as advancements in computational technologies and methods continue, this approach can become more feasible and efficient in the future.

The fundamental concept is straightforward: we aim to take various chemical formulas, compute their theoretical spectra, and determine which one aligns best with the experimental spectrum. Unfortunately, the 5-dimensional space of proteins' chemical formulas is densely populated, and our ability to perform comparisons is constrained by computational capabilities, limiting the analysis to a very restricted subspace. Thus, in this section, we describe two methods for selecting an appropriate subspace.

Baveragine: extending the averagine model

The first approach involves constructing a vector that, along with averagine, spans the subspace for minimizing distance. To this end, we simulated spectra for all chemical formulas listed in the Uniprot database. Subsequently, we conducted Principal Component Analysis (PCA) to characterize variations within the data. Through this analysis, the first principal component corresponded to averagine, while the second, termed *baveragine*, offered additional insights into the protein space that would not be possible with the use of averagine alone.

Next, we perform several transformations to enhance the interpretability of the method by avoiding negative coefficients in the baveragine vector and preventing vector subtraction when spanning the chemical formulas space. For this purpose, we added a sufficient number of averagine units to baveragine to ensure all coefficients are non-negative. Then, we subtracted the transformed baveragine from averagine to make one of its coefficients equal to zero. A scheme of these transformations can be found in Figure 2.8; note that these transformations do not affect the space that is spanned.

To utilize the constructed vectors for identifying a set of potential points for minimization, we first compute both vectors: A, which is the transformed averagine scaled to match the mass of the protein for which we aim to predict the monoisotopic peak, and similarly point B for the transformed baveragine. Finally, our goal is to select all points that approximately lie between points A and B. Let us denote the space of all possible chemical formulas consisting of C, H, N, O, and S elements by $\mathcal{V} = \mathbb{N}^5$. The set of points can then be defined as

$$\mathcal{L}(A, B) = \{\vartheta \in \mathcal{V} : \exists_{\theta \in \vec{AB}} \exists_{j \in \{\mathrm{c,\,h,\,n,\,o,\,s}\}} |\theta_j - \vartheta_j| \leqslant 1\}, \tag{2.15}$$

where $\vec{AB}$ is an interval between points A and B. A scheme that illustrates the concept of

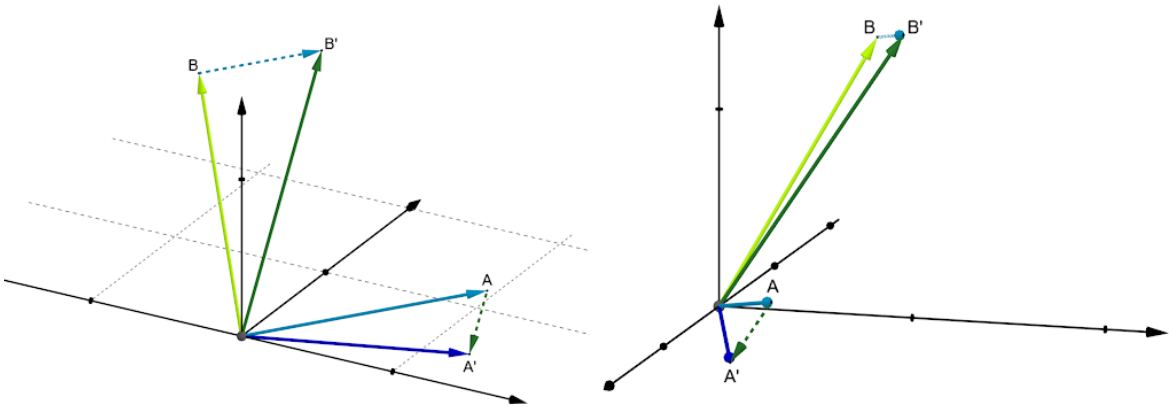

Figure 2.8: Three-dimensional scheme of the transformation of averagine and baveragine for improved interpretability. The goal is to avoid negative coefficients in the baveragine vector and prevent vector subtraction when spanning the chemical formulas space. First, a sufficient number of averagine units are added to the baveragine vector to ensure all coefficients are non-negative ($B \rightarrow B'$). Then, the transformed baveragine is subtracted in just the right amount from the averagine vector to make one of its coefficients equal to zero ($A \rightarrow A'$).

such a set in a 2-dimensional space is presented in Figure 2.9.

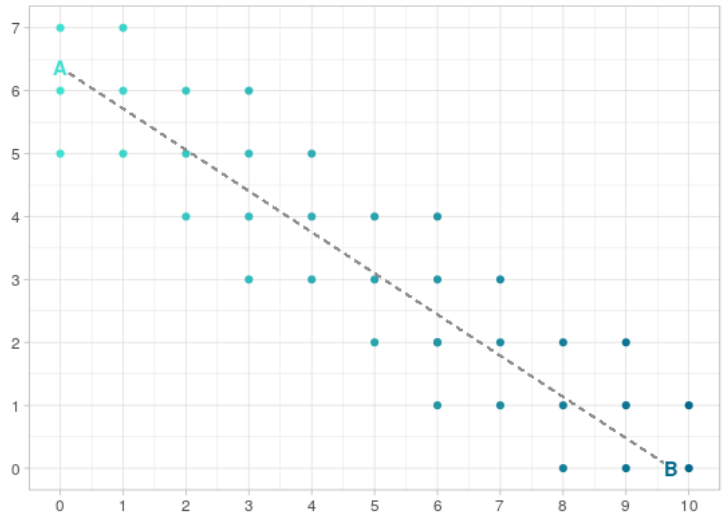

Figure 2.9: Two-dimensional scheme of point selection lying between transformed averagine and baveragine. If an interval connecting transformed averagine and baveragine passes through a unit cube aligned with natural numbers, then each vertex of the cube is selected.

Selecting chemical formulas with optimized average mass and variance

The most straightforward idea here can be to take all points with similar masses as the experimental spectrum and with chemical formulas that are not very distant from an average protein. However, since in a five-dimensional space, such a defined set grows exponentially as the distance parameter changes, we propose a more sophisticated approach. We start by identifying the highest peak in the experimental spectrum. We then calculate an averagine protein of the given mass using the averagine model, where the coordinates may be real numbers, not necessarily integers. Next, we generate an initial set of potential chemical for-

mulas. These formulas aim to have a closely matching average mass but do not necessarily have a similar isotopic distribution, i.e.,

$$\mathcal{H}_1 = \{\vartheta \in \mathcal{V} : \vartheta = \text{rdh}(A(M_{\text{MostAb}}) + n \cdot A^{\text{norm}} + m \cdot \nu),\ n, m \in \{-k_1, \ldots, k_1\}\}, \quad (2.16)$$

where

- $\text{rdh}(\vartheta)$ – function that first rounds ϑ to the nearest integer point in $\mathcal{V}$, and then adds as many hydrogen atoms as needed to match the mass as closely as possible to the initial ϑ;
- $A(M_{\text{MostAb}})$ – averagine vector with mass equal to the mass of the highest peak in the experimental spectrum;
- A^{norm} – averagine vector normalized to have length 1;
- ν – variance-growth gradient described earlier, which can be found in Table 2.1;
- k_1 – parameter responsible for set size. A larger k_1 provides a better fit to the spectrum but highly increases computation time.

Note, that stochastic noise may cause the highest experimental peak, that we used to construct $\mathcal{H}_1$, to differ from the theoretical highest peak of the molecule of interest. Therefore, we expand the $\mathcal{H}_1$ set by adding or subtracting up to 2 hydrogen atoms to each of its points, i.e.,

$$\mathcal{H}_2 = \{\vartheta \in \mathcal{V} : \vartheta = \vartheta^\star + n \cdot \boldsymbol{e}_{\text{h}},\ \vartheta^\star \in \mathcal{H}_1,\ n \in \{-2, \ldots, 2\}\}, \quad (2.17)$$

where $\boldsymbol{e}_{\text{h}}$ stands for a vector that is equal to 1 at the position responsible for the hydrogen dimension and 0 otherwise.

Then, we use a chosen similarity measure to determine which formula from the $\mathcal{H}_2$ set best fits the experimental spectrum. Note, however, that since the construction process of sets $\mathcal{H}_1$ and $\mathcal{H}_2$ primarily focused on average mass, we expect that the resulting formula will provide an appropriate average mass only. Now, we are also able to consider the shape in a similar manner. From the chosen chemical formula, we take the average mass and variance and use them to construct a new starting point. We create an averagine with the given mass, and then we adjust the variance by adding the ν variance-growth gradient. We denote the new starting point by $s_0 = \text{argmin}_{\vartheta \in \mathcal{H}_2} \mathcal{W}(E, \vartheta)$, where $\mathcal{W}$ represents similarity measure, yet not limited to the Wasserstein distance. This time, the potential points are intended to fit the shape of the spectrum as well

$$\mathcal{H}_3 = \{\vartheta \in \mathcal{V} : \vartheta = \text{rdh}(s_0 + \textstyle\sum_{i=1}^{3} n_i \cdot s_i),\ n_i \in \{-k_3, \ldots, k_3\}\}, \quad (2.18)$$

where k_3 is the parameter responsible for controlling set size, and s_i are the three orthonormal solutions of the following system of equations

$$\begin{cases} 0.2979\vartheta_{\mathrm{c}} + 0.0250\vartheta_{\mathrm{h}} + 0.3474\vartheta_{\mathrm{n}} + 0.3968\vartheta_{\mathrm{o}} + 0.7953\vartheta_{\mathrm{s}} = 0 & \text{Average Mass} \\ 0.0630\vartheta_{\mathrm{c}} + 0.0007\vartheta_{\mathrm{h}} + 0.0211\vartheta_{\mathrm{n}} + 0.0504\vartheta_{\mathrm{o}} + 0.9965\vartheta_{\mathrm{s}} = 0 & \text{Variance} \end{cases} \tag{2.19}$$

Note that the coefficients of the system are normalized values from Table 2.1. As a result, the solutions s_i do not alter the average mass and variance when added to any chemical formula.

Spectra similarity evaluation within the established subspace

Finally, within a subset of the chemical formula space – such as $\mathcal{L}(A, B)$, $\mathcal{H}_3$, or any other – a similarity measure must be applied. For instance, the `masserstein` algorithm (Ciach *et al.*, 2020) accounts for the occurrence of noise in mass spectra, thus providing highly reliable results when used to measure spectra similarity. Unfortunately, for example, it takes approximately 10 minutes to complete the matching process using `masserstein` on the $\mathcal{H}_3$ set with $k_1 = 2$ and $k_3 = 10$ with the use of a 2.3 GHz Intel Core i7 processor. Therefore, while this approach may yield more accurate results, it remains impractical in the current landscape.

In the end, we argue that monoisotopic peak mass prediction can be based on such an artificial, yet well-fitted, spectrum. Notice that the algorithm adapts to the inherently noisy experimental spectra, where only certain peaks are clearly visible. Consequently, the obtained spectrum may differ from the actual one, but it should keep the same average mass and variance, which is crucial. On the other hand, one might question whether the monoisotopic mass can be directly determined from the matched spectrum. Our tests indicate that both approaches yielded almost identical results; however, the linear model provides a safer solution – i.e., more robust against errors resulting from atypical inputs – since it was trained on actual proteins. In particular, it fits spectra composed of basic protein's chemical elements in their typical proportions, which may significantly affect the tails of the multinomial distribution.

2.2. `MIND4OLIGOS`: an adaptation of `MIND` methodology for oligonucleotides

Oligonucleotides, short chains of nucleotides, are the building blocks of DNA and RNA. Since they are primary genetic materials in all organisms, insightful knowledge of oligonucleotides opens up a wide range of applications. First, oligonucleotide-based drugs are

increasingly important in therapeutics; their ability to target previously inaccessible tissues facilitates the use of precision genetic medicine for diseases that are currently untreatable (Roberts *et al.*, 2020; Egli and Manoharan, 2023). Moreover, targeting abilities of oligonucleotides occur to be relevant in diagnostics (Kumar Kulabhusan *et al.*, 2020). Finally, oligonucleotides can be used in basic research and biotechnology applications, like in the study of molecular mechanics and genome engineering (Ma and Salaita, 2019; Glazier *et al.*, 2020).

Therefore, since mass spectrometry is the primary technique used in oligonucleotide research, adapting a well-tested algorithm like `MIND` to oligonucleotides was advantageous for enhancing this field. Moreover, due to the limited variability in smaller molecules, which have fewer possible combinations of atoms, we were not only able to adapt the algorithm but also to determine whether the predictions are certain – assuming the accuracy of the most abundant peak mass provided to the algorithm.

2.2.1. Preprocessing: the true most abundant peak's mass determination

Since `MIND` model is based on the most abundant peak's mass, ensuring that we know its true (theoretical) mass is essential. Often, however, the theoretically most abundant peak does not have to have the highest intensity in experimental spectra due to poor ion statistics. Therefore, initially, we must determine the theoretical peak. Fortunately, Lermyte *et al.* (2019) offers a simple heuristic that addresses this issue.

Lermyte *et al.* observed the difference between the average mass of a given protein and its theoretical most abundant peak mass holds between 0.1 and 1.2 Da. Thus, in most cases, if the difference value is above 1 Da, a peak to the left of the true most abundant peak incorrectly possesses the highest intensity and the next to the right of it should be chosen as the true one. Adequately, a value below 0 Da indicates that a peak to the left should be chosen. More formally, to obtain the true most abundant peak mass, we use the formula

$$M_{\text{MostAb}} = M^{\star}_{\text{MostAb}} + \lfloor M_{\text{avg}} - M^{\star}_{\text{MostAb}} \rfloor,$$

where M_{MostAb} and $M^{\star}_{\text{MostAb}}$ denote true and observed most abundant peaks, respectively, M_{avg} stands for average mass of the spectrum, and $\lfloor \, . \, \rfloor$ denotes a floor function. We tested on 10^5 randomly chosen oligonucleotides that the relation between average mass and true most abundant peak holds, with all differences between 0.06 and 1.17 Da. Hence, the heuristic can be applied in our case. More details on the heuristic can be found in the mentioned manuscript.

2.2.2. Adaptation of MIND methodology

With the corrected most abundant peak mass M_{MostAb} ready, we can now introduce the MIND model and explore the differences that arise when applying it to oligonucleotides. The primary aim of the algorithm is to determine the most probable distance between the monoisotopic mass and the most abundant peak mass, referred to as an *offset*, which must be (in approximation) an integer. The algorithm initiates with the training of a linear model $M_{\text{mono}} \sim M_{\text{MostAb}}$. Although the linear nature of this relationship is intuitive, it leads to notable insights when analyzing the residuals of the model $M_{\text{mono}} - \hat{M}_{\text{mono}}$, which highlight the significance of the offset.

Figure 2.10 displays the residuals for the model trained on all possible theoretical oligonucleotides with lengths between 5 and 92 nucleobases. Each "line" in Figure 2.10 corresponds to a different offset; the first line corresponds to an offset equal to 0 Da, the second to 1 Da, and so forth, up to 13 Da. As can be observed, for oligonucleotides, there are never more than two possible offset candidates. Generally, as the size of a biomolecule increases, the number of candidates also grows due to the increasing variety of possible combinations of isotopes. Thus, in the case of proteins, where 3 or more candidates usually appear, the MIND model addresses this challenge using a second linear model.

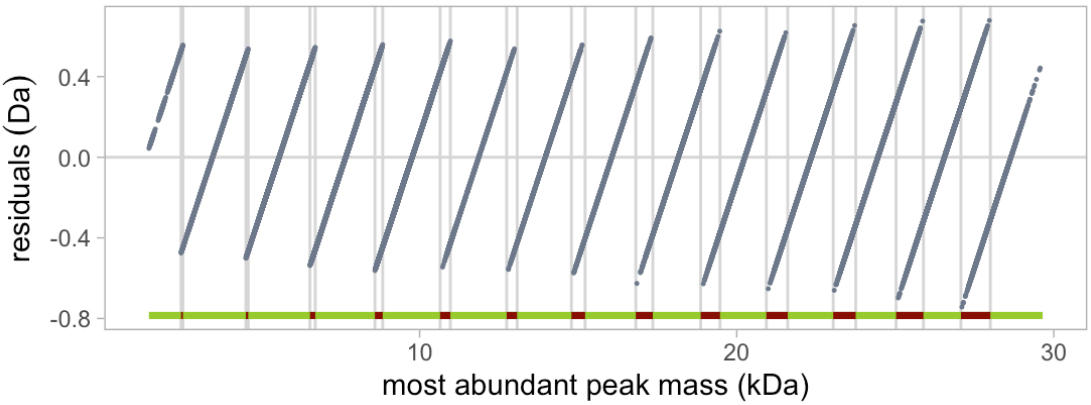

Figure 2.10: Residuals from the linear model $M_{\text{mono}} \sim M_{\text{MostAb}}$ are displayed in 14 linear-like patterns as a function of the most abundant mass. Each "line" corresponds to the distance between monoisotopic and most abundant masses – the offset; the first line on the left corresponds to a distance of 0 Da, the second to 1 Da, and so forth, up to 13 Da. For some masses of the most abundant peak, lines may overlap. In such cases, two possible offset candidates are identified. Intervals in which predictions are uncertain, constituting 21.2% of the training mass range, are marked in red.

In the case of MIND4OLIGOS, we implemented an even more straightforward solution. Initially, it is important to note that for the majority of the most abundant peak masses, no intervention is required at all, as there is only one candidate for the offset. Only for the remaining 21.2% of the masses, which are highlighted in red in Figure 2.10, a simple adjustment is necessary. For each mass interval containing two offset candidates, we quantified the proportion of theoretical oligonucleotides favoring each of the two distances by utilizing a 50 Da-width sliding-window to aggregate votes across the interval. At each

position within this interval, one candidate for the offset typically predominates. In fact, this method identifies a point where the support for each candidate reaches equilibrium. The process in illustrated in Figure 2.11. Accordingly, to the left of this equilibrium point, we use the offset associated with the upper line, whereas to the right, we use the offset associated with the lower line.

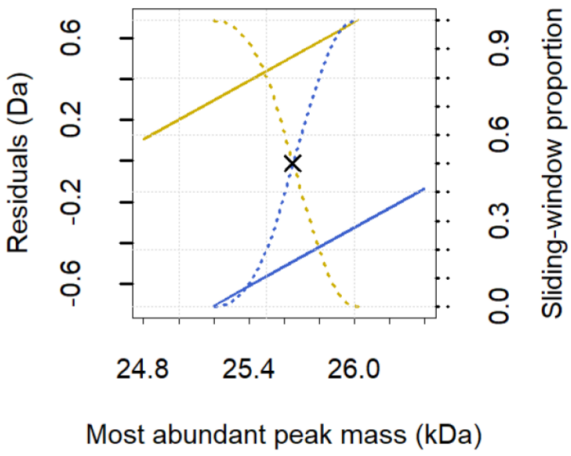

Figure 2.11: Evolution of sliding-window proportions within the overlapping mass region. This plot visualizes proportions computed using a 50 Da-wide sliding window, which counts the number of oligonucleotides along the overlapping gold and blue residual lines. The black ✗ marks the 50% proportion point; its x-axis coordinate indicates the optimal location for switching the offset candidate.

At the end, we utilize the offset estimation methodology to determine the monoisotopic mass. Thus, in conclusion, the formula for our prediction model can be expressed as

$$\hat{M}_{\text{mono}} = M_{\text{MostAb}} - \mathcal{I}(M_{\text{MostAb}}), \tag{2.20}$$

where $\mathcal{I}$ represents a function that calculates the offset based on the most abundant mass. Using the heuristic to estimate the equilibrium point, every offset value is associated with an interval where it is certain, or the most likely candidate. Consequently, the function $\mathcal{I}$ is defined such that $\mathcal{I}(M_{\text{MostAb}}) = j$ if $M_{\text{MostAb}} \in (m_j, m_{j+1}]$, with m_j representing the estimated equilibrium points.

2.2.3. Performance and limitations

First, we would like to clarify that although it may seem like the heuristic for selecting equilibrium points leads to a high number of incorrect predictions, this is not the case. In fact, oligonucleotides tend to cluster more densely near the centers of "lines". For example, let us consider the 25.15–26.05 kDa interval, one of the widest intervals with uncertain prediction outcomes. In this interval, the −1 and +1 off-by-one dalton errors occurred at rates of 4.34% and 4.14%, respectively, while 91.52% of cases were correctly classified across all theoretical oligonucleotides.

Moving our evaluation to testing MIND4OLIGOS on a validation set of 10^4 simulated DNA oligonucleotides, we introduced additional noise drawn from a uniform distribution – set to 10% of the isotope height – into the theoretical isotope distributions. This required the use of the most-abundant peak mass heuristic, ensuring that noise did not directly interfere with the mass values, which is crucial for an accurate monoisotopic mass prediction. The results showcased a classification accuracy of 92.87%, with error rates for −1 and +1 off-by-one dalton errors at 4.33% and 2.80%, respectively.

Next, we measured the isotopic envelopes of three full-length products, which are the intended results of oligonucleotide synthesis, in separate scans under the same chromatographic peaks and detected them in multiple charge states. These oligonucleotides are represented in 75, 45, and 270 isotopic envelopes, with off-by-one dalton errors occurring 2, 11, and 30 times, leading to error rates of 2.67%, 24.44%, and 11.11%, respectively. Note, however, that each of these errors was due to incorrectly identifying the most abundant mass. The results of our prediction method are depicted in Figure 2.12. Errors are predominantly concentrated around the 0 ppm histogram bin; nevertheless, several isotope envelopes exhibited deviations of a significantly larger magnitude, specifically in the range of tens of ppm or more, all corresponding to off-by-one dalton errors.

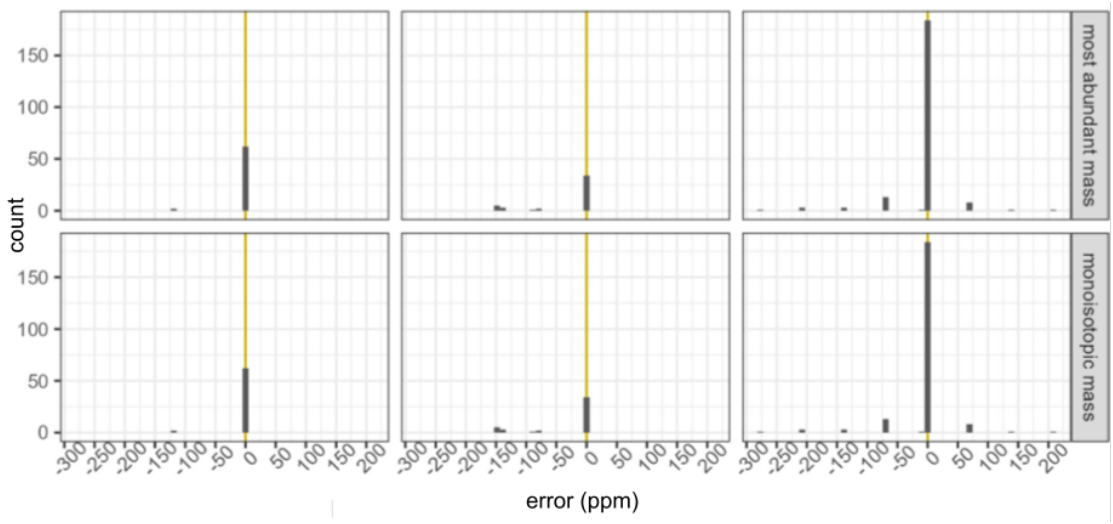

Figure 2.12: Results of the most abundant peak picking heuristic (first row) and the MIND4OLIGOS algorithm on experimental spectra (second row). Three different oligonucleotides are displayed across three columns, with 75, 45, and 270 isotopic envelopes, respectively. Deviations of magnitudes in the range of tens of ppm or more indicate off-by-one dalton errors, occurring in 2 (2.67%), 11 (24.44%), and 30 (11.11%) cases. Note that all the monoisotopic mass prediction off-by-one dalton errors result from incorrect identification of the most abundant masses.

In summary, the main advantage of the MIND4OLIGOS algorithm is its simplicity in application. Using only the most abundant peak mass not only limits noise influence on mass spectra but also facilitates the algorithm's use when working with mass spectra of mixtures, which is the most common scenario. Consequently, the algorithm can easily be included in pipelines, as it requires only very basic data preprocessing.

2.2.4. Usage of intensity ratios as an additional predictor

At this point, it is important to note that in the presented approach, despite generally high prediction accuracy, there are oligonucleotides for which the off-by-one dalton error will occur – regardless of the resolution of the input spectra. Therefore, a question arises: Is it possible to improve the method on intervals with uncertain predictions to avoid such errors?

To address this question, we explored several potential additional predictors alongside the most abundant peak mass and identified the most promising one: intensity ratios. These ratios, denoted by $\boldsymbol{r}$, are calculated on experimental spectra as the intensity of a peak with a lower mass divided by the intensity of a peak with a higher mass, computed from left to right. It occurs that oligonucleotides, in a space spanned by the most abundant mass and a given intensity ratio, group into point clouds that share the same offset value, see Figure 2.13 for an example. Therefore, we will base our alternative prediction approach on this observation.

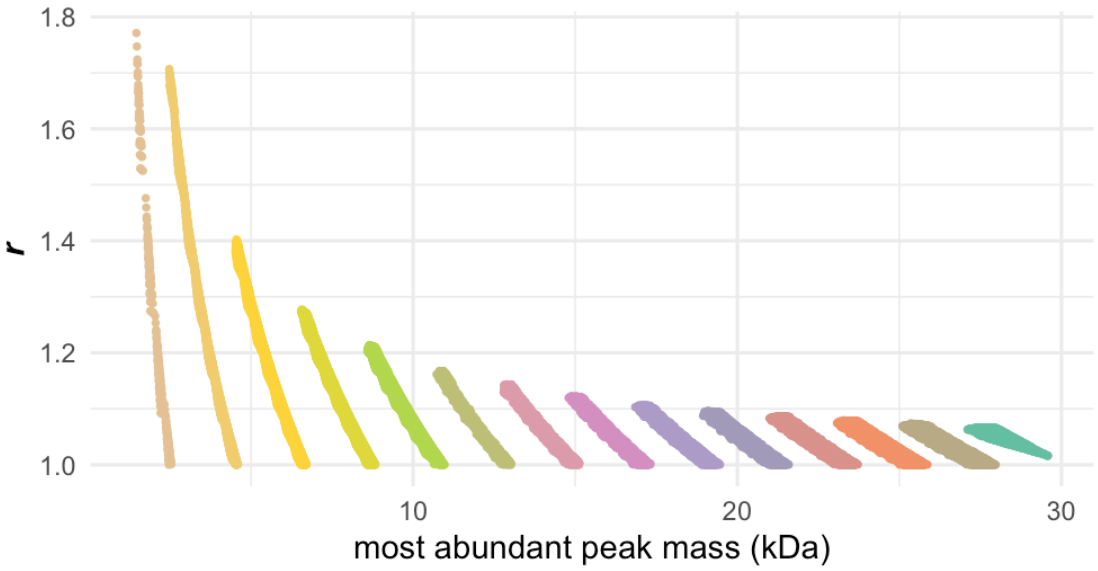

Figure 2.13: Dependency between the most abundant mass M_{MostAb} and the corresponding intensity ratio **r** (in this case, between the most abundant peak and the peak first to its right) for all theoretical oligonucleotide mass spectra. Different colors denote the various offsets $M_{\text{MostAb}} - M_{\text{mono}}$. Importantly, the monochromatic nature of each point cloud visually signifies grouping by offset value, with each sequential cloud representing an increasing offset – starting from zero for the first cloud, one for the second, and continuing in this pattern.

Since the linear part of Equation 2.20 still plays a crucial role, the addition of intensity ratios is intended to improve the accuracy of offset estimation. Consequently, the equation undergoes minimal changes and can be expressed as

$$\hat{M}_{\text{mono}} = M_{\text{MostAb}} - \mathcal{C}(M_{\text{MostAb}}, \boldsymbol{r}), \tag{2.21}$$

yet, in this case, we allow for any machine learning classifier, denoted as $\mathcal{C}$. To simplify our considerations, from now on, we will focus solely on the ratio between the most abundant

peak and the peak first to its right. This ratio offers two significant advantages: it always exists, and it should be more resistant to noise compared to ratios computed using lower-intensity peaks. Moreover, we will specify the $\mathcal{C}$ classifier as a k-Nearest Neighbors (kNN) model with $k = 1$.

Even in this basic model, we achieved near-perfect prediction accuracy for simulated oligonucleotide spectra, which were subjected to various, yet consistently low, noise levels relative to their theoretical intensity. Recognizing that the most abundant mass and the selected intensity ratio hold differing significances for kNN distance calculations, we calibrated weights to optimize prediction performance. Then, we conducted 10-fold cross-validation, yielding exceptionally high accuracy – specifically, up to less than one off-by-one dalton error per million predictions, assuming that the most abundant peak picking heuristic determined the mass correctly.

Nevertheless, noises in experimental spectra are characterized by much greater complexity and diversity. Unfortunately, as we increased the noise levels, the structured patterns presented in Figure 2.13 were lost – demonstrating that the freedom of oligonucleotide movement in the vertical axis increases with noise. Consequently, method validation on both experimental and high-noise simulated spectra failed to exceed the accuracy of the initial `MIND4OLIGOS` algorithm, even with the addition of other intensity ratios and the use of more advanced classifiers.

Essentially, we have demonstrated that as spectra of higher resolution become available in the future, the determination of monoisotopic mass will achieve perfect accuracy for any biomolecule with a mass below approximately 30 kDa. Moreover, the proposed methodology relies on easily accessible features, allowing for effortless integration into existing pipelines. For the present, however, we recommend `MIND4OLIGOS` as the optimal solution for the current technological landscape.

3

Encoding of mass spectrometry images

Biochemical composition of tissues, including concentrations and types of molecules, exhibits significant variations (Barr, 2018). The differences occur not only between distinct tissues but also between healthy and diseased sections. While these differences can offer valuable insights into disease progression, they also emphasize the complex and unique nature of each tissue type. Moreover, understanding these biochemical variations is foundational in advancing diagnostic precision and developing targeted treatments (Chughtai and Heeren, 2010; Schwamborn, 2012; Han *et al.*, 2019). As already mentioned in the Introduction, mass spectrometry imaging (MSI) is a crucial technique for acquiring tissue data. However, after data acquisition, a significant amount of analytical work is still required to understand and interpret the content of the MSI dataset.

The current landscape of MSI data analysis increasingly relies on machine learning and neural networks (Alexandrov, 2020). Sarkari *et al.* (2014) explored the use of k-means and fuzzy k-means clustering on MSI data, assessing how preprocessing steps and parameter adjustments impact the discovery of biologically significant patterns. Additionally, Hu *et al.* (2022) has proven a self-supervised clustering technique that utilizes contrastive learning and deep convolutional neural networks to effectively classify molecular colocalizations, thereby enabling the autonomous identification of co-localized molecules in MSI data. Furthermore, Guo *et al.* (2023) employed convolutional neural networks to improve the feature extraction and interpretability of complex biological MSI data.

However, researchers face major challenges when using these or similar analytical tools due to the high memory and computational demands of MS images, which often exceed tens of gigabytes. To overcome these limitations, researchers have worked on optimizing computational algorithms for MSI data processing. For example, Alexandrov and Kobarg (2011) proposed segmentation methods that either uniformly select neighbors or adaptively account for spectral similarities, achieving linear complexity and low memory requirements. Similarly, Dexter *et al.* (2017) introduced a graph-based algorithm incorporating a two-phase sampling method, improving segmentation efficiency for anatomical compounds and tissue types while reducing traditional methods' high CPU and memory usage.

3.1. Encoder-decoder architecture

To develop our encoding algorithm, we involved a classical deep learning autoencoder, a neural network designed to learn an efficient compressed data representation. An autoencoder consists of two parts: an encoder that compresses the input data into a lower-dimension space and a decoder that reconstructs the original input from this compressed representation. The algorithm aims to reduce data size while ensuring that the encoded spectra exhibit a regular probability distribution, making them well-suited for processing by neural networks and other analytical algorithms. We refer to these encoded spectra as *embeddings*. The size of the embeddings is controlled by a parameter; in our work, we typically set it to 64 or 128. Importantly, the encoding operation is reversible, allowing the original mass spectrum to be reconstructed from its embedding with only minor quality loss.

We consider each pixel of MS images as a one-dimensional intensity vector by aggregating mass spectra to m/z values rounded to the first decimal. Based on our experience, neural networks perform better with profile spectra, as their continuous structure helps to accentuate specific patterns in the input images, enhancing the network's ability to recognize spectral similarities effectively. Consequently, if an MS image contains centroided spectra, we recommend simulating a continuous profile by convolving each peak with a Gaussian distribution centered at zero, with its width defined by a specified standard deviation parameter.

3.1.1. Contrastive learning and loss functions

The process of contrastive learning begins with generating a noisy copy of each input spectrum, called *augmentation*. In our case, this noisy copy is obtained by applying small perturbations to the original spectrum, altering intensity values by up to 10% while preserving the overall spectral structure. A given original spectrum and its noisy copy are consid-

ered similar, while any other pair is considered dissimilar (*positive* or *negative pair*, respectively). To train a neural network to distinguish between similar and dissimilar spectra, we employed the so-called contrastive loss function. The function ensures that the model correctly identifies positive and negative pairs, or, in other words, rewards scenarios where the model's outputs for positive pairs are closely aligned, while the outputs for negative pairs are distinctly separated in the representation space. Formally, let us assume we have a batch of spectra $(z_1, z_2, \ldots, z_{2N})$, where each of the N original spectra has been augmented, resulting in a total of $2N$ spectra in the batch. The contrastive loss function can be expressed as

$$\text{CONTRASTIVE LOSS} = \alpha \sum_{i,j} \ell_{i,j} \cdot \mathbb{1}_{(z_i, z_j) \text{ positive pair}}, \tag{3.1}$$

where

$$\ell_{i,j} = -\log \frac{\exp\{(\text{sim}(z_i, z_j)/\gamma)\}}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp\{(\text{sim}(z_i, z_k)/\gamma)\}}, \tag{3.2}$$

α and γ are constants, sim stands for cosine similarity (i.e., $\text{sim}(x, y) = {}^{xy}/_{||x|| \cdot ||y||}$), and $\mathbb{1}_q$ is an indicator function equal to 1 when a given condition q is satisfied and 0 otherwise.

Simultaneously, we ensure that the embeddings follow a probability distribution with a predefined expected value or variance. Let us denote by w_i an embedding corresponding to spectrum z_i. First, we use the *mean loss* function given by

$$\text{MEAN LOSS} = \frac{1}{2N} \sum_{i=1}^{2N} \mu^2(w_i), \tag{3.3}$$

where $\mu(w_i)$ is the mean of the i^{th} embedding in the batch. The purpose of applying this loss function is to keep embeddings centered around zero. Moreover, we compute the standard loss function, defined as

$$\text{STANDARD LOSS} = \frac{1}{F_\text{e}} \sum_{i=1}^{F_\text{e}} \big(\sigma_i(w) - 1\big)^2, \tag{3.4}$$

where $\sigma_i(w)$ stands for the standard deviation of the i^{th} feature in the batch of embeddings and F_e is the number of features. By *features*, we understand individual positions in embedding vectors. Using standard loss in our neural network training process ensures that the distributions of features do not deviate significantly from the average value.

So far, each applied loss function has ensured that the encoder produces embeddings capable of differentiating spectra effectively and supporting the subsequent analysis. However, ensuring that the embeddings can be decoded back into their original form is crucial as well. For this purpose, we introduce one more loss function to compare the input spectrum z_i with its encoded and subsequently decoded substitute $\tilde{z}_i$. For the comparison, we applied the *mean squared error loss* function expressed as

$$\text{MSE LOSS} = \frac{1}{F_\text{s}} \sum_{i=1}^{F_\text{s}} (z_i - \tilde{z}_i)^2, \tag{3.5}$$

with F_s denoting the sample vector's length. A schematic representation of the encoder-decoder architecture, along with the applied loss functions, is shown in Figure 3.1.

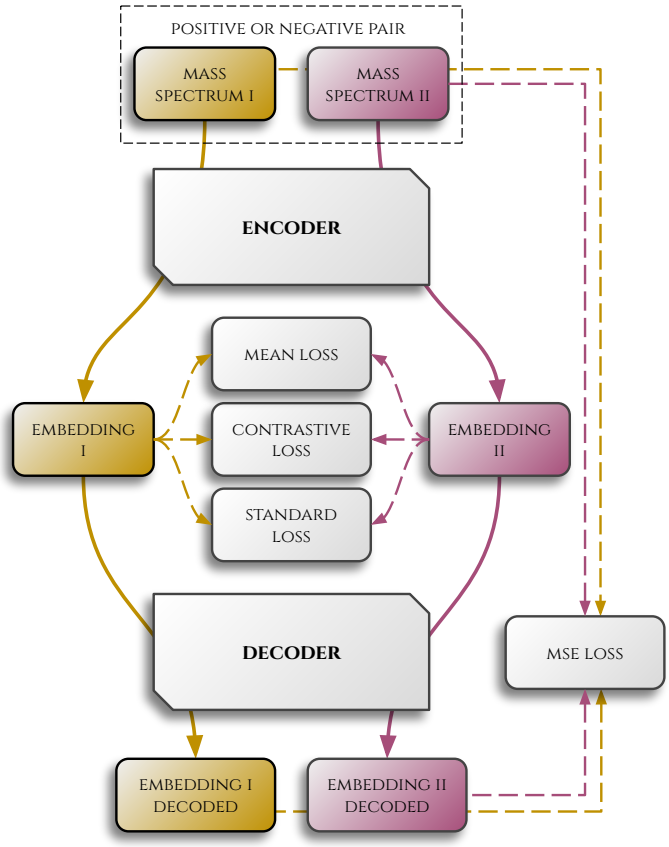

Figure 3.1: A graph illustrating the encoder-decoder learning process. First, a contrastive loss is computed to determine whether embeddings are close or distinct for positive and negative pairs, respectively. Simultaneously, mean and standard losses are calculated to ensure that the distribution of embeddings is well-structured. Then, embeddings are decoded, and an MSE loss is computed to evaluate whether the decoded and input spectra are consistent.

3.1.2. Layers and parameters

Beyond the carefully selected loss functions, our neural network architecture incorporates several specialized layers, each serving a distinct purpose in the encoding and decoding processes. These layers are designed to optimize data transformation, enhance feature extraction, and ensure effective representation learning. The key components include:

- linear layer – responsible for performing matrix multiplication followed by bias addition;
- normalization layer – ensures that input features are adjusted to have a mean of 0 and a standard deviation of 1, improving training stability and convergence speed;
- convolutional layer – facilitates the extraction of spatial hierarchies of features, allowing the network to identify patterns, textures, edges, and other local characteristics within the data;

- transposed convolutional layer – performs operations analogous to the convolutional layer but in reverse, effectively reconstructing embeddings during the decoding phase;

with a rectified linear unit (ReLU) serving as the activation function across these layers.

We optimized each dataset using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 10^{-3} and a weight decay of 10^{-5}. We set the batch size to 64, and training stopped if no improvement in the weighted sum of losses was observed after 30 epochs. While the loss functions require weighting to balance their contributions to the total loss, these weights are further fine-tuned individually for each dataset to address specific data characteristics and ensure optimal performance. All computations were performed on Google Colaboratory, utilizing a Tesla T4 GPU.

3.2. Segmentation of MS images

To validate whether the compression process successfully preserves essential features, we conducted segmentation, one of the most critical tasks in MSI data analysis, and a key area where our encoding algorithm demonstrates significant utility. For this purpose, we selected the *k*-means algorithm for two primary reasons: first, it is widely adopted in the field, and second, it benefits significantly from the regularity in data distribution ensured by our encoding approach. Additionally, we propose a modification to the traditional *k*-means algorithm, named *iterative k-means*, which offers a valuable alternative in specific scenarios.

The iterative *k*-means algorithm enhances the traditional *k*-means clustering process by integrating it with Principal Component Analysis (PCA) to dynamically determine the optimal number of clusters based on the silhouette score (Rousseeuw, 1987) for each principal component. This approach iteratively adjusts the number of clusters until it meets or exceeds a predefined target value. The process is outlined as follows:

Input: Data points $\mathcal{D}$, maximal number of clusters K.

Initialize: Apply PCA to $\mathcal{D}$ to obtain principal components. Set the initial cluster count $k = 1$ and the currently considered principal component $i = 1$.

Step 1: For the i^{th} principal component, execute k-means with varying cluster counts, identifying the optimal number c, based on silhouette score. If $c = 1$, go to **Step 4**.

Step 2: Set $k = k \cdot c$ and $i = i + 1$.

Step 3: Repeat Steps 1 and 2 until $k \geqslant K$.

Step 4: Aggregate the data points into final clusters by consolidating the saved clustering results.

Note that the purpose of iterative k-means is to estimate the number of classes accurately. Therefore, applying it to images where only two or three classes are expected is unnecessary, as the results will be equivalent to those obtained with the standard k-means algorithm. We recommend using the iterative version for images where the number of classes is uncertain but suspected to be at least four.

Following the completion of segmentation, we refine the results by application of so-called *convolutional smoothing*. This technique involves an iteration over each pixel and executes a majority voting procedure. The class of the target pixel and its neighboring pixels is assessed, and the most frequent class among them is assigned as the new class for the target pixel.

3.2.1. Matching and accuracy computation

To assess the accuracy of segmentation results against the actual data (ground truth or baseline model), it is necessary to align the classes identified by the segmentation algorithm with those in the baseline model. To achieve this, we use a voting procedure known as *matching*. For each class identified by the segmentation algorithm, we analyze the corresponding pixels' classes in the baseline model image. The segmentation class is then matched to the baseline model class that appears most frequently among those pixels. A graphical representation of this process is provided in Figure 3.2.

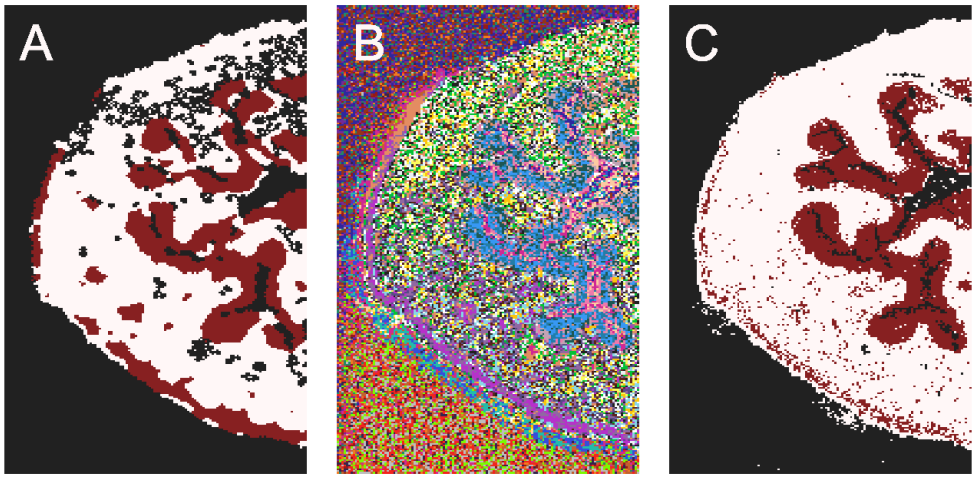

Figure 3.2: Graphical representation of the matching procedure. Panel (A) shows the baseline model image, illustrating the actual class distribution. Panel (B) displays the initial classes identified by the k-means segmentation algorithm. Panel (C) presents the results after applying the matching procedure, where the algorithm's classes are aligned with those of the baseline model to enable accuracy assessment. This alignment is achieved through a majority voting procedure, where each class identified by the algorithm is matched to the most frequently corresponding class in the baseline model.

Lastly, an image with matched classes enables verification of class alignment with the baseline model and facilitates the calculation of segmentation accuracy. However, it is important to note that in the MSI field, baseline models are themselves acquired using

specific histopatological, segmentation or other methods. Therefore, since accuracy is computed relative to these baseline models, it should be interpreted more as a correlation with the results of another method rather than an absolute measure.

A comprehensive visual summary outlining the key steps of the process described so far, arranged in a structured workflow to clarify their sequential structure, is provided in Figure 3.3.

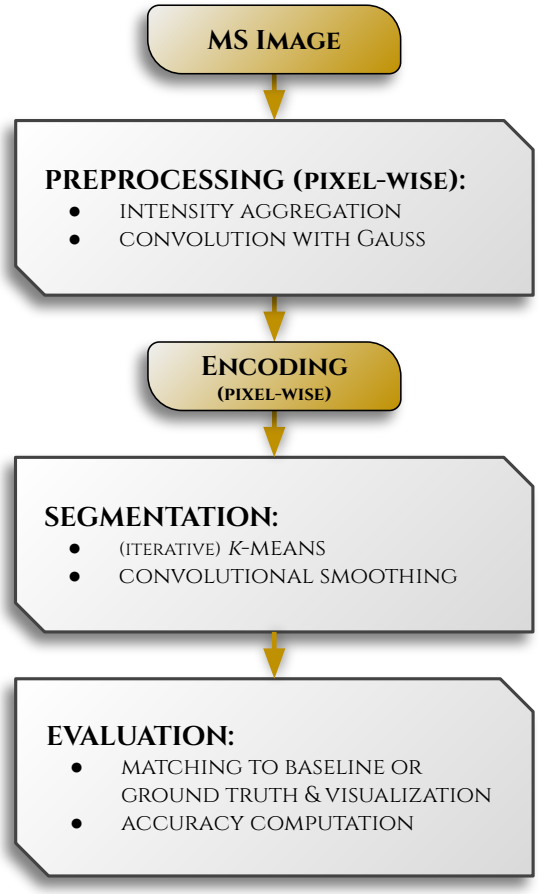

Figure 3.3: A workflow illustrating the processing of MS images, with a particular emphasis on the encoding step as the core focus of our work. The Segmentation and Evaluation steps are included primarily to demonstrate the effectiveness of the encoding process.

3.2.2. Images involved

To validate the performance and reliability of our encoding algorithm, we conducted experiments using two distinct datasets: a mouse bladder cross-section image and a set of biopsy images from patients with Barrett's esophagus. These datasets were selected because they provide reliable baseline models, which are essential for assessing the accuracy and effectiveness of our method.

The mouse bladder image was downloaded from the PRIDE database (ID: PXD001283, Perez-Riverol *et al.* (2022)). The image measures 260 × 134 pixels (34840 pixels in total),

with a spatial resolution of 10 μm. Further details on sample preparation, data acquisition, and processing can be found in Römpp *et al.* (2010) manuscript. Histological staining of the tissue section used to generate the image revealed eight distinct morphological regions. To obtain a ground truth segmentation of the MS image, we generated ion images of *m/z* 422.93 Da, 824.55 Da, and 851.64 Da. These ion images were overlaid to highlight regions with different chemical compositions. Using the overlaid ion images, and guided by the histological staining, we manually delineated different morphological regions.

Barrett's esophagus is a condition where the normal lining of the esophagus changes to a different type of tissue, often due to long-term acid reflux, increasing the risk of developing esophageal cancer. The dataset of esophagus biopsies are also available in the PRIDE database (ID: PXD028949) and comprise 19 MS images in profile mode, with sizes ranging from 1370 to 7137 pixels. Annotations by a trained pathologist distinguish between epithelial tissue and stroma, with epithelial tissue further classified into dysplasia levels: high-grade, low-grade, and non-dysplastic (healthy tissue). These annotations enable us to aggregate patient data and validate whether the segmentation performed on the images encoded by our algorithm accurately differentiates epithelial tissue from stroma. Further details on the dataset can be found in the Beuque *et al.* (2021) manuscript.

3.3. Encoding algorithm's performance

Finally, in this section, we evaluate compression efficiency and validate our algorithm through segmentation on the previously described data to ensure that essential biochemical information is preserved after encoding.

3.3.1. Mouse urinary bladder image

In the first dataset, which consists of the mouse bladder image, our objective was to compare segmentation performance between raw and encoded data. The image originally had a size of approximately 1.5 GB. The encoder training process, lasting between 1.5 and 2.5 hours depending on the parameters, successfully compressed each pixel's spectrum into a 64-element NumPy array. This compression reduced the file size to 8.5 MB, achieving a 99.4% reduction in memory usage.

Next, we applied the t-SNE algorithm to the image, as this algorithm is commonly used for the preliminary analysis of MS images to gain a better understanding of the measured data, particularly for estimating the number of biochemical classes present in the analyzed images. However, the high computational demands of t-SNE often make it challenging to apply to raw MS images. In contrast, its application to our encoded image proved successful. Notably, the t-SNE analysis on the image was completed in approximately 50 minutes,

suggesting that at least three biochemical categories can be expected, as shown in Figure 3.4.

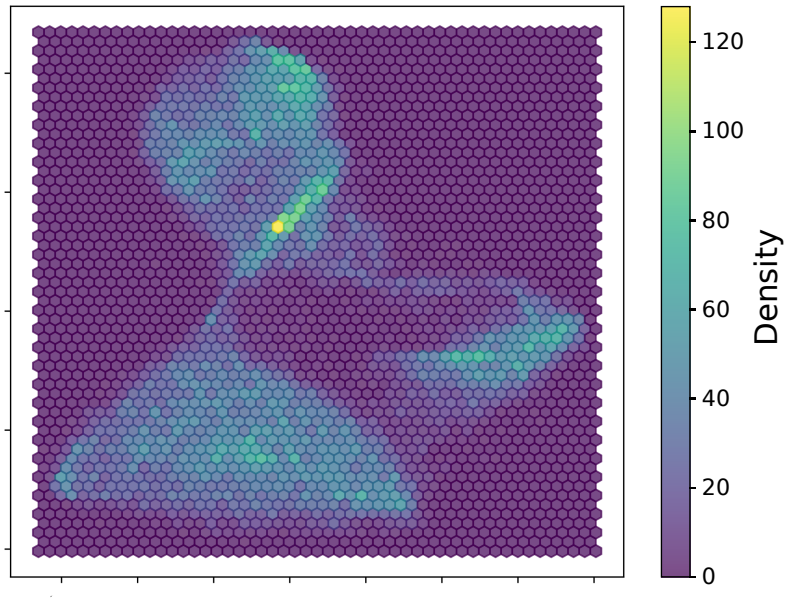

Figure 3.4: Example output of the t-SNE algorithm applied to the encoded image of the mouse bladder. A single selected dimension is presented, illustrating the minimal number of distinct biochemical classes expected from segmentation.

Finally, we conducted a segmentation task to verify that the encoding process did not result in the loss of any essential information. The k-means algorithm was applied to both raw and encoded images, while the iterative k-means algorithm was used exclusively on encoded images due to the high computational demands of applying it directly to raw data. Additionally, we evaluated clustering algorithms on the highest spectral peaks by calculating the mean intensity for each spectral index across the image and selecting the 128 indices with the highest means. Segmentation accuracies are summarized in Table 3.1, with visual representations provided in Figure 3.5.

method	data	accuracy (%)
k-means	raw image	42.67
k-means	top-128	41.28
k-means	encoded image	59.01
iter. k-means	encoded image	64.44

Table 3.1: Accuracies of segmentation tasks into seven classes for the mouse bladder image using the k-means algorithm. Segmentation was performed on raw and encoded images, as well as on the highest 128 peaks (denoted as top-128) from each pixel in the raw image. Additionally, the iterative k-means algorithm was applied exclusively to the encoded image. Note that accuracy computation was feasible due to the availability of a baseline model, which is often not the case in real-life scenarios.

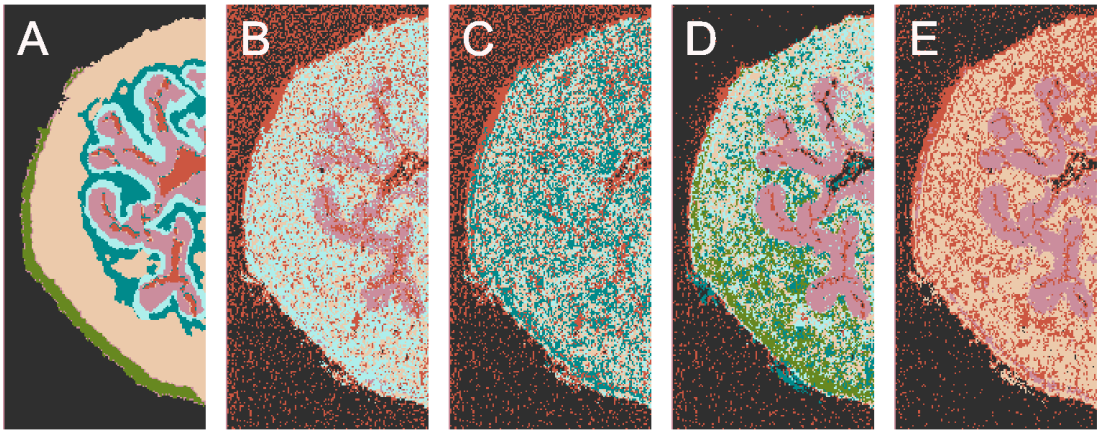

Figure 3.5: Segmentation results for the mouse urinary bladder MS image. Panel (A) shows the baseline model, as described in the Images involved section. Subsequent panels present segmentation results: (B) using k-means on the raw image, (C) using k-means on the 128 highest peaks of the raw image, (D) using k-means on the encoded image, and (E) using iterative k-means on the encoded image. Segmentation accuracies are summarized in Table 3.1. A matching procedure was applied to each segmentation result to enhance visualization clarity.

3.3.2. Barrett's esophagus biopsy images

In contrast to the mouse bladder image, Barrett's esophagus biopsy images are significantly larger – a size more commonly encountered in mass spectrometry imaging. This allowed us to demonstrate the encoder algorithm's performance on datasets substantially larger both in terms of individual image size and patient count. Consequently, we faced a key challenge our work aims to address: computing segmentation directly on raw images proved infeasible due to excessive CPU requirements. Once again employed the commonly used naive approach of applying a segmentation algorithm only to the highest peaks in the pixels' spectra, with the number of peaks fixed at 128.

The complete dataset totals 54.21 GB in size. Training the encoder to compress the data into 128-length arrays across the entire dataset took approximately 2 hours, resulting in a compressed size of 81.26 MB – representing a 99.85% reduction in memory usage. Each patient's images were encoded individually, allowing the encoder to focus on patient-specific tissue variations rather than inter-patient differences. Segmentations were performed to distinguish between tissue types: epithelial tissue and stroma. Table 3.2 summarizes the segmentation results both for individual cross-section images and aggregated by patients' four groups of overall dysplasia classifications: non-dysplastic (healthy), low-grade dysplasia non-progressive, low-grade dysplasia progressive, and high-grade dysplasia.

	images' storage size		computation	tissue type seg. acc. (%)	
	raw (GB)	encoded (MB)	time (min)	encoded image	top-128
A1	0.59	0.89	1.34	63.39	59.76
A2	0.93	1.40	2.10	72.33	63.02
B2	0.43	0.64	0.96	78.01	60.68
C3	1.13	1.70	2.55	73.75	74.50
E1	0.61	0.91	1.37	63.55	57.37
F2	0.55	0.82	1.23	69.94	62.04
I1	0.62	0.93	1.40	67.27	63.49
J2	1.07	1.60	2.40	74.48	72.25
K3	1.07	1.60	2.40	63.57	62.48
M1	0.55	0.83	1.25	77.49	65.68
N2	0.54	0.81	1.22	83.69	79.53
O3	1.74	2.61	3.91	72.79	71.07
Q1	0.87	1.30	1.95	74.94	74.10
R2	0.55	0.82	1.23	63.34	59.78
S3	0.34	0.51	0.77	71.44	70.16
ND	$\Sigma = 11.59$	$\Sigma = 17.37$	$\Sigma = 26.08$	avg. = 71.33	avg. = 66.39
A3	0.63	0.94	1.41	73.65	64.67
C1	1.20	1.80	2.70	74.91	73.66
D2	1.13	1.70	2.55	71.95	68.08
E3	0.31	0.47	0.71	78.53	77.78
H2	0.97	1.45	2.18	81.78	74.07
I3	0.93	1.40	2.10	68.20	54.19
K1	0.80	1.20	1.80	57.32	56.34
L2	2.34	3.51	5.26	80.01	75.78
M3	1.93	2.90	4.35	63.44	67.19
O1	1.81	2.71	4.07	67.48	61.43
P2	2.12	3.18	4.77	88.88	64.93
Q3	0.62	0.93	1.40	61.08	63.16
S1	0.47	0.71	1.07	62.24	54.92
T2	0.77	1.15	1.73	56.39	57.20
LGs	$\Sigma = 16.03$	$\Sigma = 24.05$	$\Sigma = 36.10$	avg. = 70.42	avg. = 65.24
B1	0.49	0.74	1.11	78.91	77.18
C2	0.58	0.87	1.31	71.96	66.98
D3	0.57	0.85	1.28	67.82	62.27
F1	0.31	0.46	0.69	61.78	65.89
H3	0.45	0.67	1.01	58.97	60.05
J1	1.00	1.50	2.25	74.24	74.01
K2	0.87	1.30	1.95	60.16	60.00
L3	1.65	2.47	3.71	77.09	59.51
N1	2.27	3.40	5.10	71.16	60.57
O2	1.58	2.37	3.55	68.17	70.24
P3	2.32	3.48	5.22	66.43	70.54
R1	1.20	1.80	2.70	67.06	64.45
S2	1.00	1.50	2.25	67.05	64.75
T3	1.40	2.10	3.15	61.51	56.01
LGp	$\Sigma = 15.69$	$\Sigma = 23.51$	$\Sigma = 35.28$	avg. = 68.02	avg. = 65.18
D1	0.30	0.45	0.68	59.44	61.80
E2	1.07	1.60	2.40	63.88	62.08
F3	0.80	1.20	1.80	89.86	75.98
H1	1.76	2.64	3.96	72.47	62.72
I2	0.63	0.95	1.43	64.74	58.27
J3	0.27	0.41	0.62	63.69	60.15
L1	0.93	1.39	2.09	61.99	54.28
M2	1.27	1.90	2.85	60.65	63.63
N3	0.44	0.66	0.99	62.24	52.84
P1	0.72	1.08	1.62	80.68	71.90
Q2	0.87	1.30	1.95	57.52	58.57
R3	1.07	1.60	2.40	71.17	65.06
T1	0.77	1.15	1.73	60.94	60.29
HG	$\Sigma = 10.90$	$\Sigma = 16.33$	$\Sigma = 24.52$	avg. = 66.87	avg. = 62.12

Table 3.2: Summary of the compression process and tissue segmentation results. Images (denoted by A1, ..., T3) were grouped due to patients' overall classification as follow: (ND) non-dysplastic, (LGs) low-grade dysplasia non-progressive, (LGp) low-grade dysplasia progressive, and (HG) high-grade dysplasia. The left side of the table details the compression process, including data size before and after encoding and encoder training times. The right side presents segmentation outcomes, distinguishing tissue types – epithelial tissue and stroma. Segmentation was performed using the k-means algorithm, with the 128 highest peaks ("top-128") used as a baseline due to the high computational demands of applying k-means directly to raw images.

3.4. Expanding labels on partially annotated images

Finally, we would like to discuss how segmentation can be further enhanced through the use of a so-called *classification head*.

Since the encoding algorithm relies on unsupervised learning, it autonomously identifies the most relevant features of the mass spectra. While this generally works effectively, it can sometimes lead to excessive classes in following segmentation, requiring class merging "by hand" afterwards. For segmentation tasks with predefined specifications, this process can be adjusted more effectively to the specific case. Additionally, again, although we expect segmentation on encoded images to yield more accurate results than on raw ones, it is important to determine which algorithm – k-means or others – will perform better.

In many cases, only a portion of an image contains histopathological labelling, or a single labelled image is available, while others from different cross-sections remain unlabeled. Due to MSI data's complexity and high dimensionality, full labelling is often impractical, leading researchers to annotate only key regions of interest. However, we can easily extend any labelling across entire datasets by combining our encoding algorithm with the classification head (Huynh and Elhamifar, 2020). The head is an additional neural network trained on the labelled section of an image to learn patterns critical to specific classes. Therefore, it can evaluate the similarity of unlabeled pixels to these patterns and extend labelling across MS images derived from the same tissue and experiment. Also, it eliminates the need to choose a segmentation algorithm, as the head performs the segmentation autonomously.

Moczulski (2024) demonstrates that such a head indeed improves accuracy; however, more extensive testing across diverse image datasets is required. If this approach proves successful, it could significantly assist in the labeling process, allowing a pathologist to histopathologically label only a portion of an image, with the remaining labeling completed autonomously.

4

Uncertainty estimation of measurements calibrated by sample-standard bracketing

As mentioned in the Introduction chapter, multicollector inductively coupled plasma mass spectrometers hold a unique role in the mass spectrometry landscape. To enhance the understanding of this uniqueness, let us begin with a description of how this instrument operates and explore several of its key advantages.

The measurement process begins with loading a sample into an ion source (see Figure 4.1), where plasma tears the sample's constituents into individual atoms and subsequently ionizes them. Here, two notable advantages appear. First, using the ICP source provides flexibility in sample preparation, allowing for the analysis of both liquid and solid samples. Second, ICP enables the analysis of a wide range of elements – including those with high ionization potentials that are challenging to measure with Thermal Ionization Mass Spectrometry (TIMS) – yet limited to "non-traditional" elements, i.e., other than e.g., carbon, hydrogen, nitrogen, oxygen or sulphur.

Next, the ion beam passes through the ion optics system, where it is focused using electrostatic lenses and deflectors. This ensures the beam is precisely aligned, and its trajectory is optimized for subsequent analysis. Once appropriately shaped, the ion beam enters the mass analyzer, which consists of electrostatic and magnetic components. The electrostatic analyzer filters ions based on their kinetic energy, allowing only those within the de-

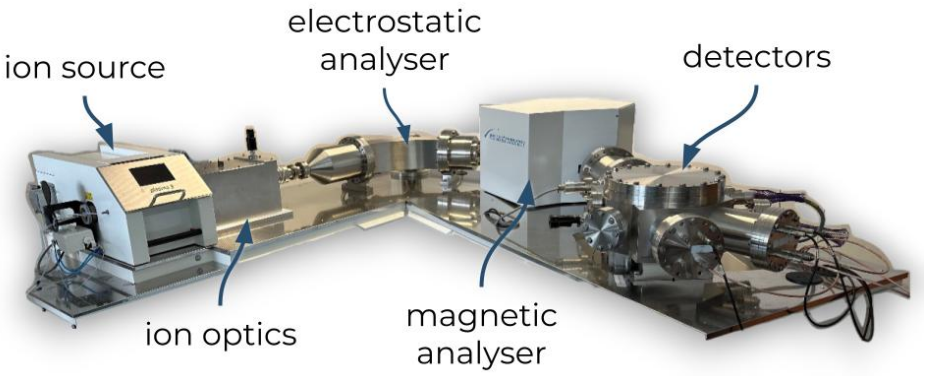

Figure 4.1: Image of the MC-ICP-MS (Plasma 3, Nu Instruments, UK) in our laboratory at the University of Warsaw Biological and Chemical Research Centre, with labeled components.

sired energy range to proceed. The magnetic analyzer then separates the ions based on their m/z ratio. By bending the ion trajectories in a magnetic field, ions of different masses are deflected to varying degrees, enabling the instrument to separate specific isotopes. This dual-stage separation process achieves monoisotopic ion beams while effectively removing artifacts from matrix constituents, i.e., undesired elements or isotopes in the sample.

The previous paragraph found its place in this dissertation to emphasize as much as possible: monoisotopic beams reaching the multicollector are shaped under precisely identical conditions. This is essential, as it ensures that neither ionisation nor transportation inside the instrument does not alter the isotope ratios. Consequently, finally, the multicollector can simultaneously capture each monoisotopic beam, enabling the most precise isotopic ratio measurements achievable with current technology.

4.1. Significance of accurate isotope ratio measurements

The applications of isotopic ratio analysis are diverse, yet the complexity and cost of measurements make large-scale use of MC-ICP-MS techniques challenging. Currently, these techniques are primarily applied to unique and scientifically significant samples. While the range of potential applications continues to expand, this section presents an overview of its current uses in bullet points, providing a clear illustration to help the reader envision the possibilities in this field. the main applications include:

- **Geochemistry and Cosmochemistry**: High-precision isotopic measurements, illuminating processes such as magma formation or fluid-rock interaction, are essential in geological sciences. For example, MC-ICP-MS is pivotal for examining rare earth elements' isotope ratio critical for dating and understanding the geological transformations from the planet's earliest formation stages providing insights into Earth's evolutionary history (Sindern, 2017). Another instance is in cosmochemistry, where

the analysis of magnesium (Mg) and titanium (Ti) in calcium-aluminium-rich (Ca–Al) inclusions from meteorites provide vital clues about the nucleosynthetic origins of the solar system and help map early solar system processes (Larsen *et al.*, 2018).

- **Archaeology**: In archaeological research, MC-ICP-MS provides insights into the origins and movements of ancient materials and works of art. For instance, the analysis of lead (Pb), strontium (Sr), or neodymium (Nd) isotopes has been used to identify the geographic sources of metals in ancient artefacts such as bronze tools, coins, or weapons. Such analyses allow researchers to trace the provenance of materials and offer a glimpse into ancient trade networks and cultural exchanges (Nord and Billström, 2018). Additionally, studies of isotopic composition aid in the authentication of artworks, like by analysis of lead white pigments in 16^{th} and 17^{th}-century paintings, which helped to confirm their origins and detect forgeries (Fortunato *et al.*, 2005).

- **Environmental Pollution**: Monitoring contamination via radioactive materials is highly effective with MC-ICP-MS due to these materials' significantly disturbed isotopic compositions, which allow for extremely sensitive detection. For instance, when detecting contamination with uranium, ICP-MS can only measure the total uranium content. However, natural uranium is abundant in the environment, making it difficult to detect contamination from sources like nuclear power plants or weapons factories. By measuring isotopic ratios, such as the proportion of ^{234}U, which is naturally 0.005% but may reach 0.5% in contaminated samples, MC-ICP-MS can readily identify contamination by uranium from the nuclear industry (Hou and Roos, 2008). Moreover, this technique can also effectively monitor other non-radioactive pollution sources, like releasing soluble, toxic selenium oxyanions into post-mining groundwater (Basu *et al.*, 2016). Here, let us reiterate that while other techniques allow for monitoring the level of contamination (of, e.g., total Cd or Pb), MC-ICP-MS enables us to track specific isotopes within the contamination and determine whether it originates from a suspected source.

- **Biomedical and Clinical Research**: MC-ICP-MS is utilized to analyze the isotopic composition of essential metals (e.g., Mg, Ca) and metaloids (e.g., Se) in biological tissues and fluids, aiding in the discovery of potential diagnostic markers (Reitsema, 2013; Costas-Rodríguez *et al.*, 2016).

4.2. Isotopic fractionation correction methods

One notable inconvenience is that the raw measurement signal from MC-ICP-MS is inherently meaningless on its own. To interpret the signal correctly, calibration is required using

a *standard* – an isotopic pair with a exactly known ratio. Although there are numerous calibration methods, they can generally be categorized into two groups: *internal* and *external* calibration. In internal calibration methods, the standard must consist of an isotopic pair where both isotopes have different mass numbers compared to the analyte's measured isotopic pair (and, therefore, are usually isotopes of a different element as well). Since the term "standard" can have varying interpretations within the ICP field, it is more precise to use the term *calibrator* when referring to the standard used in internal calibration methods. In contrast, external calibration methods require the standard to be exactly the same isotopic pair as the one present in the measured sample.

Here, we introduce three selected calibration methods. Beyond providing an outline of how calibration works, one of the methods serves as the main focus of this chapter, while the remaining two are presented to offer context for future research discussed in Chapter 6.

4.2.1. Internal standard

The Internal Standard (IS) correction method is widely adopted for its simplicity and effectiveness in enhancing measurement accuracy. The method relies on the assumption that the isotopic fractionation affecting the internal standard mirrors that of the isotopic pair to be measured, referred to as the *analyte*. This assumption enables the correction of the analyte's measured isotope ratios by applying the bias observed in the internal standard. The relationship between the analyte and internal standard is described by Russell's law , which can be expressed as

$$R_{i/j} = r_{i/j} \times \left(\frac{m_i}{m_j}\right)^f, \tag{4.1}$$

where f – the mass fractionation factor – can be computed as

$$f = \frac{\ln\left(\frac{R_{k/l}}{r_{k/l}}\right)}{\ln\left(\frac{m_k}{m_l}\right)}, \tag{4.2}$$

with R and r representing the actual and measured isotopic ratios, respectively, for the lower-indexed isotopic pair; m denoting the mass of the lower-indexed isotope; i and j referring to the analyte's isotopes, while k and l to the calibrator's isotopes.

While Russell's law provides a robust foundation for mass bias correction, its accuracy heavily depends on the assumption that the mass fractionation behaviour of the analyte and calibrator remains the same under identical measurement conditions. However, recent studies suggest this assumption may not always hold, particularly when the analyte and calibrator are different chemical elements (Suárez-Criado *et al.*, 2024). Therefore, careful selection of an appropriate internal standard remains crucial for minimizing uncertainty in isotope ratio measurements.

4.2.2. Optimized regression model

The next internal calibration method – optimized regression model (ORM) – is a relatively novel approach in plasma mass spectrometry, building on an accidental discovery made in France at the end of the 20th century by Maréchal *et al.* (1999). To understand this method, it is essential to introduce the concept of the *drift* effect, which occurs in every mass spectrometers and causes isotopic ratios of measured elements to change monotonically over time. This discovery identified a consistent linear relationship between two pairs of measured isotopic ratios as instrumental drift progresses. The relationship can be expressed as

$$\ln\left(r_{i/j}\right) = a + b \cdot \ln\left(r_{k/l}\right), \tag{4.3}$$

where a and b are constants, and the remaining parameters hold as defined previously. This relationship proves to be highly useful if we assume that the true isotopic ratio R can be substituted for the measured one r (or equivalently, from Russell's law, if we assume that $b = {}^{\ln \frac{m_i}{m_j}}/_{\ln \frac{m_k}{m_l}}$, which is expected when measurements are conducted carefully). Then, we can formulate

$$R_{i/j} = e^a \cdot R_{k/l}^b. \tag{4.4}$$

Therefore, the measurement methodology proceeds as follows: we aim to measure the analyte and calibrator at several (typically 4 or 5) different plasma power levels, which simulate the progression of instrumental drift. Then, we estimate the parameters a and b. Since the isotopic pair ${}^k/_l$ serves as the calibrator with a known isotopic ratio, substituting it into Equation 4.4 allows us to calculate the corrected isotopic ratio of the analyte $R_{i/j}$.

The key advantage of the ORM lies in the fact that its computations are free from instrumental fractionation f and, therefore, independent of the IS assumption that different isotope pairs undergo identical isotopic fractionation. This independence significantly broadens the applicability of MC-ICP-MS measurements to a wider range of isotopic pairs where the IS assumption fails. However, this comes at the cost of increased effort, as multiple measurement points must be recorded for a single obtained isotopic ratio.

4.2.3. Sample-standard bracketing

The final calibration method introduced – and central to this dissertation – is a representative of external calibration. As a reminder, external calibration requires using a calibrator of exactly the same isotopic pair as the analyte. In the context of sample-standard bracketing (SSB) calibration method, the term "standard" or "calibrator" is more appropriately referred to as the *bracketing standard*. This term summarizes the entire methodology, which involves three sequential measurements: first, of the bracketing standard; second, of the analyte sample; and third, of the bracketing standard again. For an example of SSB measurement,

refer to Figure 4.2. By assuming that the instrument's conditions during the analyte sample measurement represent an average of the conditions observed during the two bracketing standard measurements, we can state

$$\delta_{i/j}(‰) = \left[\frac{r_{i/j}}{\frac{1}{2}\left[r_{k/l}^{(1)} + r_{k/l}^{(2)}\right]} - 1 \right] \times 10^3. \tag{4.5}$$

Note that this definition yields the deviation of the analyte from the bracketing standard (in permil, denoted by δ) rather than the isotopic ratio itself. Depending on the field, this form of presenting results often predominates. Nevertheless, if the isotopic composition of the bracketing standard is known, the analyte's isotopic ratios can be easily computed if needed.

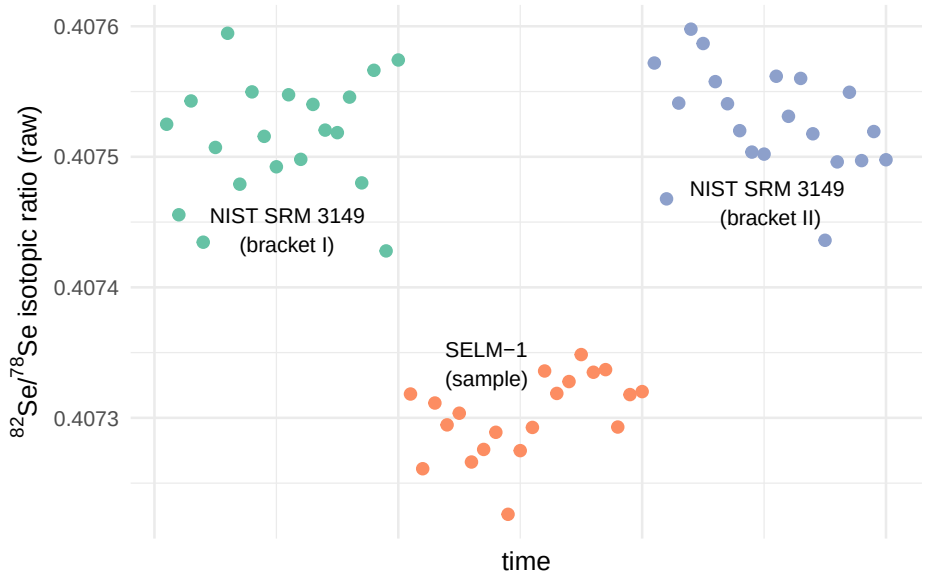

Figure 4.2: An example of an SSB measurement for the $^{82/78}$Se isotopic pair. The deviation of the analyte – named SELM-1 – is measured with respect to the bracketing standard, NIST SRM 3149. Each of the three point clouds is averaged to facilitate the application of Equation 4.5.

We can state that the SSB method offers the most straightforward approach to mass bias correction in MC-ICP-MS. Its primary advantage in addressing mass discrimination effects lies in the fact that each analyte's isotope is corrected individually using the exact same isotope as the calibrator. This ensures that the analyte and the calibrator are subjected to identical conditions, making SSB the most accurate form of correction available. However, a notable drawback of SSB is its inability to correct for matrix effects, to which it is highly sensitive. Consequently, if the sample is not perfectly purified, systematic errors are certain to occur.

Note, finally, that SSB's nature makes it able to be combined with internal calibration methods, resulting in hybrid approaches such as IS-SSB or ORM-SSB. These combined methods yield exceptionally improved accuracy but also introduce additional complexity

to the analytical procedure, requiring significantly more effort and higher costs to conduct measurements. Therefore, careful consideration of these factors is essential when selecting the appropriate calibration method for precise isotopic analyses.

4.3. Methodology of uncertainty estimation

To provide context, our research on uncertainty propagation was conducted alongside an investigation into a simplified method for selenium separation from various geological and biological samples. Traditionally, selenium separation relied on complex and time-consuming thiol resin methods, which are highly sensitive to reagent concentrations. In contrast, our proposed method utilizes coprecipitation of selenium with iron (III) hydroxide, followed by dissolution in hydrochloric acid. This approach significantly simplifies the sample preparation process, reduces the processing time to 3-4 hours after sample decomposition – compared to 2-3 days – and minimizes reliance on clean lab environments. Notably, the method maintains high accuracy in measuring Se isotope ratios, underscoring the potential for broader adoption of the coprecipitation technique in selenium isotope studies.

In this study, we utilized a variety of geological and biological standards and reference materials, including NIST SRM 3149, artificial seawater (ASW), open ocean seawater NASS-4, the United States Geological Survey reference materials SGR-1 (oil shale, Green River Formation), SCo-1 (cody shale), MAG-1 (marine mud), European reference material BC210a (selenized wheat flour), and the selenium-enriched yeast certified reference material SELM-1. Notably, these materials are relatively accessible and are either certified in terms of isotope composition or possess the composition published.

The secondary objective of the study was to calculate the anticipated level of uncertainty for selenium isotope ratio. In other words, we aimed to determine an δ value for which we can be 95% confident that it represents the upper limit of the possible difference between a δ value obtained from a new measurement and the expected δ value of the sample. By analyzing multiple Se samples collectively, we performed calculations that can be applied universally to any selenium isotopic results corrected using the SSB method.

Finally, with the above in mind – that we are working with multiple selenium samples and occasionally performing additional computations to handle them collectively – we can now proceed to describe our uncertainty computation methodology.

4.3.1. Baseline: ignoring error propagation

Typically, a sample of interest is measured several times, and the exact number can often be low due to limited availability. In many cases, the final uncertainty is computed solely

from the standard deviation of these repeated measurements, overlooking the fact that each individual measurement carries its own uncertainty. As a side note, this is not the worst scenario, as some researchers still do not "believe" in necessity of uncertainty consideration, treating each measurement as an absolute value. Nevertheless, we begin by calculating uncertainty based only on the standard deviation of the measurement series, establishing a baseline for comparison with our more comprehensive approach.

First, we examine the distribution of measured δ values relative to the true δ value. Let δ_j^i represent the δ value of i^{th} measurement of j^{th} sample, and δ_j^{true} the true reference value for the j^{th} sample. Using this notation, we can define the deviation of a measurement from the true value as

$$\Delta_j^i = \delta_j^i - \delta_j^{\text{true}}. \tag{4.6}$$

To ensure the correctness of our subsequent computations, we applied the Shapiro-Wilk test to verify whether Δ, considered as a random variable, is normally distributed (H_0: $\Delta \sim \mathcal{N}$, H_1: otherwise). Moreover, we used the one-sample Student's t-test to verify if the distribution is concentrated around zero (H_0: $\Delta^{\text{avg}} = 0$, H_1: $\Delta^{\text{avg}} \neq 0$). On the complete dataset, i.e., consisting of all samples we measured, we obtained p-values equal to 0.9566 and 0.3578, respectively; detailed p-values for individual samples can be found in Table 4.1. Verifying both hypotheses was essential since, from now on, we assume that Δ is normally distributed with an average value equal to zero, while keeping in mind which samples do not meet this assumption.

	Student's t-test p-value	Shapiro-Wilk test p-value
NIST	0.7558	0.5950
ASW	0.0371	0.1960
NASS-4	0.5207	0.4153
SELM-1	0.0701	0.9942
BC210a	1.0000	0.5675
SGR-1	0.0052	0.4166
SCo-1	0.3296	0.7336
MAG-1	0.6209	0.5174
jointly	0.3578	0.9566

Table 4.1: p-values of the one-sample Student's t-test and Shapiro-Wilk test for individual samples and the combined dataset. Low p-values from the t-test in the case of ASW and SGR-1 may indicate that either δ_j^{true} value we refer to, or the results of our measurements could be of poor trueness.

Next, we computed the standard deviation of Δ, denoted by σ_Δ. Once again, these computations were performed for each sample individually, as well as for the joint dataset. It

is worth noting that, in the case of individual samples, the standard deviation calculation is independent of the reference value δ_j^{true}, yet this dependency becomes essential in the joint analysis.

Finally, we considered the 0.975 quantile of a normal distribution with an expected value of zero (justified by the Student's t-test with p-value = 0.3578) and a standard deviation of $\sigma_\Delta = 0.0701$. It enabled us to determine 95% confidence bound for any newly measured δ value of a $^{82/78}$Se observation, i.e.,

$$\left|\delta - \delta^{\text{true}}\right| \leqslant \mathcal{N}_{0.975}(0, \sigma_\Delta) =: \mathrm{U}_\Delta, \tag{4.7}$$

where $\mathcal{N}_{0.975}(0, \sigma_\Delta) = 0.1374$. More precise quantiles for specific samples are provided in Table 4.2, presented later in this chapter.

4.3.2. Error propagation for the SSB method

The consideration of uncertainty propagation must begin at the earliest possible stage, which, in our case, involves the computation of δ values. Let us rewrite Equation 4.5 specifically for our measurement, i.e.,

$$\delta_{\mathrm{a}}(‰) = \left[\frac{\left(\frac{^{82}\mathrm{Se}}{^{78}\mathrm{Se}}\right)_{\mathrm{a}}}{\frac{1}{2}\left[\left(\frac{^{82}\mathrm{Se}}{^{78}\mathrm{Se}}\right)_{\mathrm{b}_1} + \left(\frac{^{82}\mathrm{Se}}{^{78}\mathrm{Se}}\right)_{\mathrm{b}_2}\right]} - 1 \right] \times 10^3, \tag{4.8}$$

where the index "a" represents the analyte measurement, while "b_1" and "b_2" correspond to the two measurements of the bracketing standard. Since, as illustrated in Figure 4.2, each measured sample consists of multiple signals from detectors, these signals are treated as average values with corresponding standard deviations. Let us denote the isotope ratios of the analyte and bracketing standards by μ, with appropriate indices, and similarly represent their standard errors by σ. To compute how the standard errors of these measurements propagate into the standard error of δ values – denoted by σ_δ – we can use either a Taylor series approximation (Lee and Forthofer, 2005) or, equivalently, differentiation of Eq. 4.8 (Sullivan *et al.*, 2020). Both approaches yield the same formulation for the propagated standard error of a given δ, expressed as

$$\sigma_\delta \approx 10^3 \cdot \left| \frac{2 \cdot \mu_{\mathrm{a}}}{\mu_{\mathrm{b}_1} + \mu_{\mathrm{b}_2}} \right| \sqrt{\left(\frac{\sigma_{\mathrm{a}}}{\mu_{\mathrm{a}}}\right)^2 + \left(\frac{\sqrt{\sigma_{\mathrm{b}_1}^2 + \sigma_{\mathrm{b}_2}^2}}{\mu_{\mathrm{b}_1} + \mu_{\mathrm{b}_2}}\right)^2}. \tag{4.9}$$

Figure 4.3 presents two exemplary series of measurements for the BC210a and SGR-1 samples, with NIST (SRM 3149) serving as the bracketing standard. The plot includes annotated propagated errors (σ_δ) for individual measurements, along with two reference values from the literature for the SGR-1 sample.

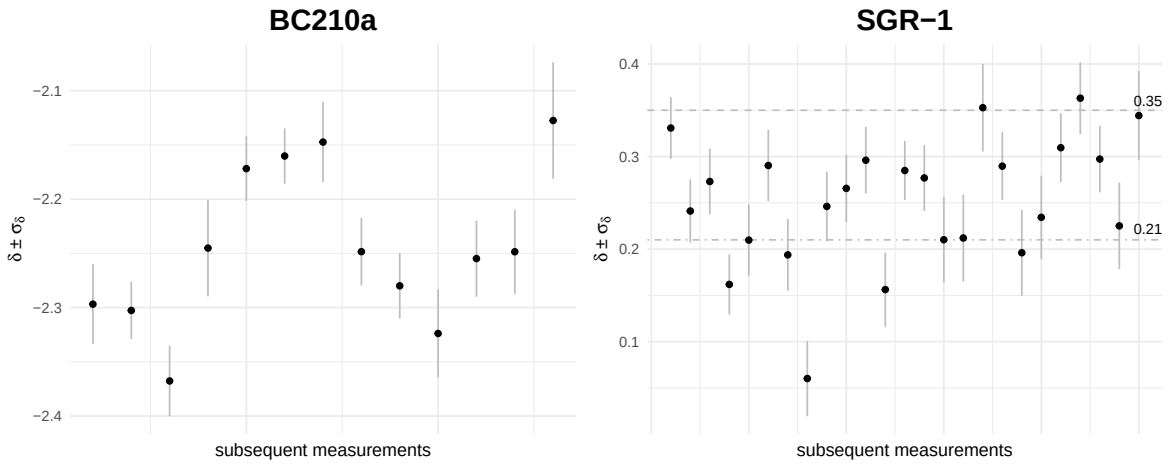

Figure 4.3: Exemplary sequences of measured δ values with corresponding propagated standard errors σ_δ for BC210a and SGR-1 samples. Two reference δ values for SGR-1 – 0.21 (Stüeken *et al.*, 2013) and 0.35 (Rouxel *et al.*, 2002) – are indicated by dashed lines.

4.3.3. Measurements' aggregation

So far, the computations described align with widely used methods. The final step, however, involves incorporating σ_δ into the U_Δ calculations, as neglecting σ_δ may result in underestimating the aggregated uncertainty. It is important to note that since each measured δ is associated with its own standard error, deriving an analytical formula to describe error propagation is a challenging task. The problem of aggregating results or measurements with individual uncertainties is commonly addressed using methods such as DerSimonian-Laird (DerSimonian and Laird, 1986) or Hartung-Knapp-Sidik-Jonkman (Sidik and Jonkman, 2007; Röver *et al.*, 2015). Thus, let us briefly review these two estimators and explain why we consider them insufficient, at least in the case of our data.

Both methods – abbreviated DSL and HKSJ – are based on the same model; therefore, we begin with a unified description. Let y_i represent the i^{th} measurement out of a total of k measurements, and let μ denote the unknown true value of the measured effect. Additionally, let υ_i capture the variation in the study, assumed to follow a normal distribution with a mean of 0 and a variance of τ^2, i.e., $\upsilon_i \sim \mathcal{N}(0, \tau^2)$. Similarly, ε_i represents a random error in the study, following a normal distribution with a mean of 0 and a variance of κ_i^2, i.e., $\varepsilon_i \sim \mathcal{N}(0, \kappa_i^2)$. Based on these assumptions, the model can be expressed as

$$y_i = \mu + \upsilon_i + \varepsilon_i, \tag{4.10}$$

resulting in each measurement y_i also following a normal distribution with parameters $\mathcal{N}(\mu, \tau^2 + \kappa_i^2)$.

Based on this model, the first aspect the methods investigate is heterogeneity in the study, which can be expressed using the following statistic

$$Q = \sum_{i=1}^{k} w_i (y_i - y_w)^2, \tag{4.11}$$

which should be understood as the distance from the weighted average effect, y_w, defined as $y_w = \sum_{i=1}^{k} w_i y_i / \sum_{i=1}^{k} w_i$, with both the effect and the distances weighted by the certainty of individual measurements, w_i, expressed as $w_i = {}^1/_{\sigma_i^2}$. Subsequently, using the Q statistic, DerSimonian and Laird estimated the variance in the study as

$$\hat{\tau}^2 = \max\left(0, \frac{Q-(k-1)}{\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}}\right), \tag{4.12}$$

by equating Q with its expected value and solving for τ^2. Note, however, that the estimator $\hat{\tau}^2$ is biased, as the calculations assume true κ_i^2 values, whereas, in practice, these are directly derived from the measurements. Nevertheless, possessing the $\hat{\tau}^2$ estimator allows us to update the weights and, consequently, the estimator for the size of the effect under study. By defining $w_i^\tau = {}^1/_{(\hat{\tau}^2 + \kappa_i^2)}$, we can state

$$\hat{\mu} = \frac{\sum_{i=1}^{k} w_i^\tau y_i}{\sum_{i=1}^{k} w_i^\tau}. \tag{4.13}$$

With all the above statistics at our disposal, we can now proceed to the uncertainty estimators, defined as follows

$$\mathrm{Var}(\hat{\mu})_{\mathrm{DSL}} = \frac{1}{\sum_{i=1}^{k} w_i^\tau} \quad \text{and} \quad \mathrm{Var}(\hat{\mu})_{\mathrm{HKSJ}} = \frac{\sum_{i=1}^{k} w_i^\tau (y_i - \hat{\mu})^2}{(k-1)\sum_{i=1}^{k} w_i^\tau}. \tag{4.14}$$

Note that both formulas use weights involving both $\hat{\tau}^2$ and κ_i^2 estimators. Unfortunately, based on our observations, such a definition can lead to issues as the distinction between random errors and in-study variance becomes blurred from the model's point of view. Since both estimators contribute to the aggregated uncertainty in the same manner, losing this distinction results in anomalous behavior of the aggregated uncertainty estimator, as demonstrated in Figure 4.4. Specifically, as input uncertainties (x-axis) increase, the uncertainty estimator should also increase, consistent with our primary assumption of $y_i \sim \mathcal{N}(\mu, \tau^2 + \kappa_i^2)$. However, in the interval 1.5-2.4, the uncertainty unexpectedly decreases. Since this example is derived from our data, the methods proved – at least for heterogeneity and effect sizes comparable to those in our study – to be inadequate.

Therefore, we decided to use Monte Carlo simulations as a safe estimation tool. For this purpose, we chose to disturb computations of Δ in Equation 4.6, by adding an individual random effect to each δ, proportional to its σ_δ. More formally, in each Monte Carlo run, we compute

$$\Delta_j^i = \delta_j^i + \xi_j^i - \delta_j^{\mathrm{true}}, \tag{4.15}$$

where the additional term ξ_j^i is sampled from $\mathcal{N}(0, \sigma_{\delta_j^i})$ distribution. Since each run introduces a new error ξ_j^i for every observation separately, we generate a uniquely perturbed

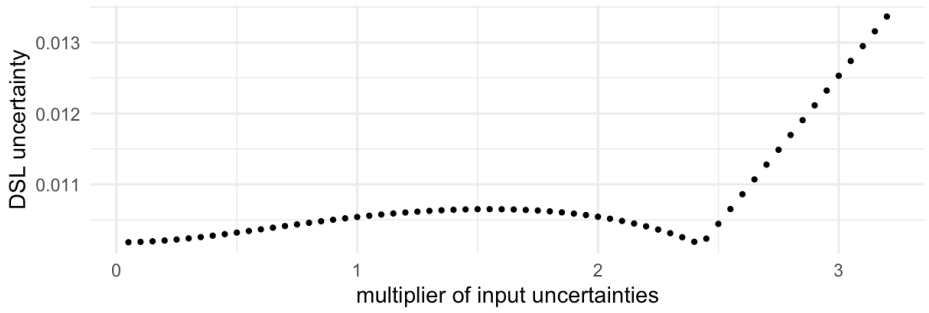

Figure 4.4: The DerSimonian-Laird uncertainty of aggregated measurements in relation to the size of input uncertainties. The data for this simulation were derived from the SELM-1 sample. Notably, in the interval between 1.5 and 2.4, the uncertainty unexpectedly decreases even though the input uncertainties of the individual measurements increase.

dataset in every simulation. Next, on each perturbed dataset, we recompute σ_Δ, and ultimately, we calculate the average standard deviation across all Monte Carlo runs, denoted as $\sigma_\Delta^{\mathrm{MC}}$.

The influence of the propagated standard error of individual samples on the computation of expected uncertainty proved significant. Using the Monte Carlo approach, we obtained $\sigma_\Delta^{\mathrm{MC}}$ = 0.0963 compared to σ_Δ = 0.0701 from the original, jointly considered data. Notably, while our goal was to construct a 95% confidence boundary for δ, additional simulations on random data suggest that the actual confidence level may exceed 95%. Consequently, the method can be regarded as conservative. In other words, the actual uncertainty is expected to lie between the baseline model estimation, representing a lower bound, and the Monte Carlo approach estimation, serving as an upper bound.

Detailed results for each sample are presented in Table 4.2, and we summarize its content here to facilitate the interpretation of the data. The $\delta^{\,\mathrm{true}}$ column contains the literature values (Rouxel *et al.*, 2002; Far *et al.*, 2010; Stüeken *et al.*, 2013; Karasiński *et al.*, 2020), with the one selected as a reference underlined. $\delta^{\,\mathrm{avg}}$ represents the average δ value from n measurements taken for a given sample. $\Delta^{\,\mathrm{avg}}$ is the difference between the average δ value ($\delta^{\,\mathrm{avg}}$) from our n measurements and the literature value ($\delta^{\,\mathrm{true}}$). Next, σ_Δ is the standard deviation calculated for all Δ values, i.e., for all n differences between individual measurements and the literature value. Note that σ_Δ does not account for the precision of the absolute isotope ratio measurements. Incorporating this precision via error propagation and estimating its impact using the Monte Carlo method yields the $\sigma_\Delta^{\mathrm{MC}}$ value. Finally, U_Δ and $\mathrm{U}_\Delta^{\mathrm{MC}}$ represent the 95% confidence expanded uncertainty values for σ_Δ and $\sigma_\Delta^{\mathrm{MC}}$, respectively.

Finally, we would like to emphasize an additional line of reasoning enabled by our approach. Suppose the σ_Δ and $\sigma_\Delta^{\mathrm{MC}}$ values are similar. In that case, it indicates that the absolute isotope ratios used to calculate individual δ values were measured with high precision, and the σ_Δ or U_Δ value primarily reflects significant variation between the individual δ

	δ^{true} (‰)	δ^{avg} (‰)	Δ^{avg} (‰)	σ_Δ	$\sigma_\Delta^{\text{MC}}$	U_Δ	$\text{U}_\Delta^{\text{MC}}$	n
NIST	0.00	-0.00	-0.00	0.07	0.09	0.13	0.17	54
ASW	0.00	0.04	0.04	0.07	0.08	0.13	0.17	15
NASS-4	0.00	0.02	0.02	0.09	0.10	0.17	0.20	7
SELM-1	-0.68, -0.66	-0.70	-0.02	0.07	0.08	0.14	0.16	53
BC210a	n.d.	-2.24	0.00	0.07	0.08	0.14	0.16	13
SGR-1	0.21, 0.35	0.25	0.04	0.07	0.08	0.14	0.16	25
SCo-1	0.175	0.19	0.01	0.04	0.15	0.07	0.29	9
MAG-1	0.21	0.22	0.01	0.04	0.16	0.08	0.32	9
jointly			0.00	0.07	0.10	0.14	0.19	185

Table 4.2: Statistics for $^{82/78}$Se isotope ratio measurements, computations and Monte Carlo simulations. $\text{U}_\Delta = \mathcal{N}_{0.975}(0, \sigma_\Delta)$ and $\text{U}_\Delta^{\text{MC}} = \mathcal{N}_{0.975}(0, \sigma_\Delta^{\text{MC}})$ are 0.975 quantiles of normal distributions with the given distribution parameters, they should be interpreted as 95% confidence bounds for a difference between measured δ and its true value, as shown in Equation 4.7. Underlined δ^{true} and δ^{avg} (if there were no δ^{true} value we could refer to) were used to compute Δ deviations. Note that for samples considered individually, σ_Δ is equivalent to the standard deviation of a given sample.

values (indicating poor repeatability or reproducibility). On the other hand, suppose there is a large difference between σ_Δ and $\sigma_\Delta^{\text{MC}}$ (or U_Δ and $\text{U}_\Delta^{\text{MC}}$). In that case, this suggests that the absolute isotope ratios used to calculate the δ values were measured with relatively low precision. Such differences can be observed, for instance, by comparing the results for the BC210a and SCo-1 samples.

5 Medical research

In this chapter, we outline our collaborative studies that have already been published. Given that these collaborations required expertise from various fields, it is vital to note that our responsibilities concentrated on data analysis and statistical modeling. In keeping with our intention to provide a brief overview of these studies, we often omit detailed methodologies and results to focus on illustrating our responsibilities, particularly in overcoming analytical challenges posed by unusual data. More comprehensive details can be found in the source publications.

5.1. COVID-19 pandemic

As the COVID-19 pandemic began, the world faced a health crisis of unprecedented scale. Identified in late 2019 and declared a global pandemic by the World Health Organization (WHO) in March 2020, the spread of SARS-CoV-2 prompted governments worldwide to implement widespread lockdowns and other public health strategies to control its transmission. These lockdowns and public health strategies, while effective in limiting the spread of the virus, introduced a series of unforeseen challenges, particularly in how they might affect mental health. As societies faced social isolation, economic uncertainties, and the ongoing threat of the virus, it quickly became apparent that these changes could significantly affect mental well-being. This necessitated immediate research efforts to understand better and address its wide-ranging psychological effects across various populations.

5.1.1. Factors influencing mental health during lockdown

❞ Płomecka, M., Gobbi, S., Neckels, R., Radziński, P., Skórko, B., Lazzeri, S., Almazidou, K., Dedić, A., Bakalović, A., Hrustić, Ashraf, Z., Es Haghi, S., Rodríguez-Pino, L., Waller, V., Jabeen, H., Alp, a B., Behnam, M., Shibli, D., Barańczuk-Turska, Z., Qureshi, S., Strutt, A. M., Jawaid, A. (2021). *Factors associated with psychological disturbances during the COVID-19 pandemic: multicountry online study.* JMIR mental health, 8(8), e28736.

In response to the lockdowns across the world, our international team, which included psychology, epidemiology, and data science experts, launched a comprehensive online survey in late March 2020. The survey was translated into 11 languages and spanned two weeks, capturing data from over 13300 individuals worldwide. Additionally, a follow-up survey was conducted with European participants a month after the initial lockdown measures were relaxed to acquire insights into the evolving impact of the pandemic on psychological well-being.

The survey inquired about various potential predictors, including demographic information, personality traits, exposure to COVID-19, compliance with preventative practices, and under-lockdown habits, like social contacts, social media usage and time spent exercising. Then, to quantitatively assess the psychological effects, the survey incorporated three validated psychological scales: the WHO's Self-Reporting Questionnaire-20 (SRQ) to evaluate general mental health; the Impact of Event Scale (IES) to measure the specific symptoms of post-traumatic stress disorder (PTSD); and the Beck Depression Inventory (BDI), which assesses the severity of depression.

The data analysis of the survey involved multiple linear and logistic regression models adjusted with Bonferroni correction, Poisson regression with log link function for modeling factors, and stepwise selection of the factors by the Akaike or Bayesian Information Criterion (AIC or BIC, respectively). Odds ratios (ORs) were calculated, and correlations were assessed using the Pearson test. Given that the statistical methods employed are standard, we move to the summary of our findings instead of discussing them in detail.

In the first assessment, female gender, preexisting psychiatric conditions, previous exposure to trauma, and working remotely were identified as key risk factors for heightened mental health issues, including general psychological disturbances, PTSD, and depression. For example, females had a significantly higher mean SRQ score (7.62) compared to males (5.29), and individuals with a worsening of preexisting psychiatric conditions exhibited sevenfold higher odds of depression (OR: 7.10). Moreover, introverted participants, those with less frequent interactions with friends and family, and individuals dissatisfied with state and employer responses were more likely to report elevated psychological symptoms. Conversely, optimism, the ability to share concerns with family and friends like usual, and

daily physical exercise emerged as protective factors. Optimistic individuals reported significantly lower mean BDI scores (8.86) compared to pessimists (18.41), while participants engaging in physical activity for at least 15 minutes daily had reduced SRQ and BDI scores. Satisfaction with employer and state responses was also associated with reduced symptoms, underscoring the role of institutional trust in mitigating mental health impacts. These findings underscore the critical interplay between demographic, lifestyle, and psychological resilience factors in shaping mental health outcomes during global crises.

The second assessment included 1077 European participants, all of whom had also participated in the first survey. Female gender, preexisting psychiatric conditions, and prior trauma remained significant predictors of elevated general psychological disturbances, as reflected in higher SRQ scores for women (mean coefficient: 0.27) and those with worsening psychiatric conditions (mean coefficient: 0.41). Additional risk factors included increased social media usage (mean coefficient: 0.19), home isolation alone (mean coefficient: 0.22), and the death of close contact due to COVID-19 (mean coefficient: 0.17). Similar to the first assessment, protective factors included optimism, physical exercise, and satisfaction with employer and state responses. Participants with an optimistic attitude about the pandemic's resolution demonstrated lower SRQ scores (mean coefficient: -0.26), and those engaging in daily physical activity for over an hour showed greater resilience against psychological disturbances. Interestingly, significant interaction effects emerged, indicating that urban residents and individuals working from home experienced different trajectories of mental health changes between assessments.

The analysis of a wide range of predictors in this study generated extensive results, only a few of which could be addressed in this chapter. While not all scientific standards may be met through an online survey, the urgency necessitated rapid mental health research. We can proudly state that our study was among the first worldwide to provide valuable insights during the early stages of the pandemic.

5.1.2. Worsening of preexisting psychiatric conditions

❞ Gobbi, S., Płomecka, M. B., Ashraf, Z., Radziński, P., Neckels, R., Lazzeri, S., Dedić, A., Bakalović, A., Hrustić, L., Skórko, B., Es Haghi, S., Almazidou, K., Rodríguez-Pino, L., Alp, a B., Jabeen, H., Waller, V., Shibli, D., Behnam, M., Arshad, A. H., Barańczuk-Turska, Z., Haq, Z., Qureshi, S., Jawaid, A. (2020). *Worsening of preexisting psychiatric conditions during the COVID-19 pandemic.* Frontiers in psychiatry, 11, 581426.

This follow-up study builds on our earlier work by explicitly investigating the worsening of preexisting psychiatric conditions during the COVID-19 pandemic. The study combined data from the survey of 2734 individuals who reported preexisting psychiatric conditions and a clinical cohort of 318 psychiatric patients from Texas. The Texas cohort

provided clinical verification of the findings, offering a deeper understanding of how psychiatric conditions evolved under pandemic-related stressors. This combined approach enabled a comprehensive analysis of both self-reported and clinically observed changes.

Globally, more than half of the participants reported a worsening of their psychiatric conditions. Demographic analysis revealed notable trends: the United States and Canada exhibited some of the highest worsening rates, while Turkey reported the lowest (Figure 5.1). The distribution of SRQ, IES, and BDI scores further emphasized the severity of the mental health crisis, as shown in Figure 5.2. These findings highlight the widespread and heterogeneous impact of the pandemic on psychiatric populations across diverse geographical regions.

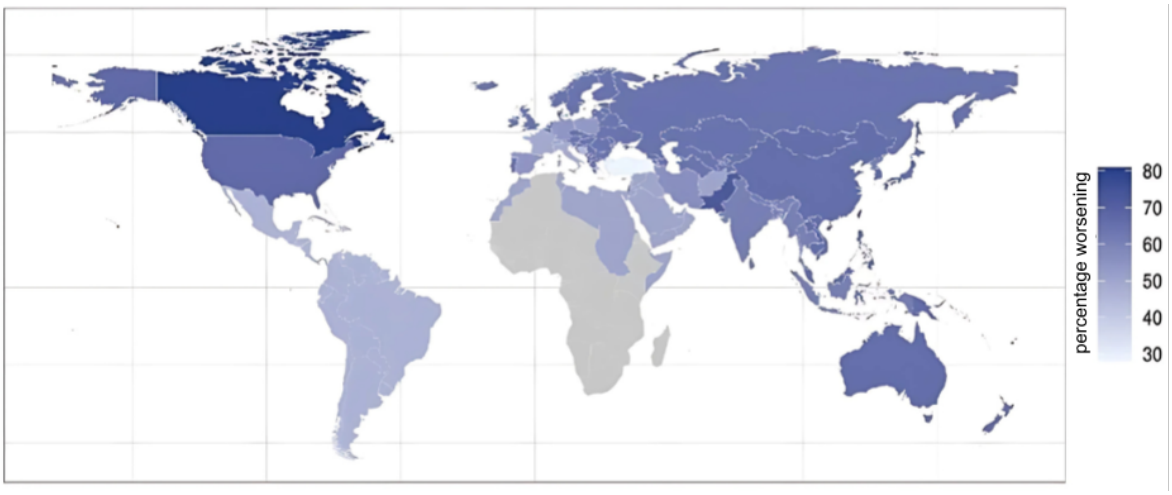

Figure 5.1: Geodemographic representation of the survey participants with pre-pandemic psychiatric conditions that reported worsening of their condition during lockdowns ($n = 2861$). The map shows the percentage of worsening preexisting psychiatric conditions separately for each of the countries (United States, Spain, Italy, France, Germany, Iran, Turkey, Switzerland, Canada, Poland, Bosnia and Herzegovina, and Pakistan) that had enough patients to consider them individually, remaining countries were aggregated into WHO regions.

The Texas clinical cohort provided crucial additional insights. Clinicians identified new psychiatric symptoms in 44% of patients, with sleep disturbances being the most frequently reported issue. Nearly half of the patients required treatment adjustments, including dose changes or the introduction of new medications. Women in this cohort were significantly more likely to require medication changes (OR: 2.22, $p < 0.05$), echoing broader trends observed in the survey data regarding the heightened vulnerability of women during the pandemic.

These findings highlight the significant impact of the COVID-19 pandemic on individuals with preexisting psychiatric conditions. By integrating large-scale survey data with clinical observations, the study provides a strong foundation for understanding global mental health challenges and emphasizes the need for expanded mental health services to support at-risk populations during global crises.

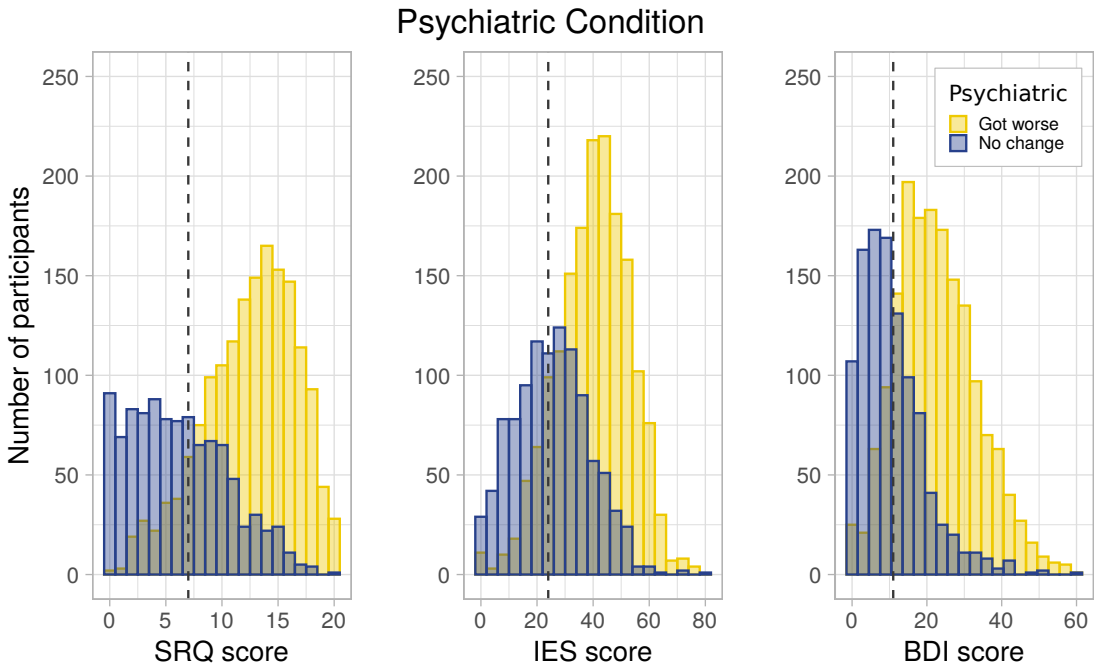

Figure 5.2: Population distribution of individuals with preexisting psychiatric conditions across SRQ, IES, and BDI scores, divided into subgroups reporting "worsening" or "no change". Vertical lines represent literature-based thresholds for diagnosing psychological disturbances, PTSD, and depression, respectively.

5.1.3. Childern's role in the pandemic spread

❞ Mańdziuk, J., Okarska-Napierała, M., Woźniak, W., Hryniewicka, A., Radziński, P., Gambin, A., Podsiadły, E., Demkow, U., Kuchar, E. (2023). *Monte Carlo Regression for Evaluating Children's Role in the Pandemic Spread on the Example of Delta COVID-19 Wave.* the Pediatric Infectious Disease Journal, 42(12), 1086-1092.

This study shifts focus away from mental health to the epidemiological aspects of SARS-CoV-2, specifically investigating the role of children in household transmission during the Delta variant wave. Data were analyzed from 515 families, comprising 883 children and 1060 adults, each with at least one SARS-CoV-2-positive household member. Children's infection status, risk-associated factors, and the role of childcare facilities in transmission were assessed using a detailed in-hospital survey. The primary aim was to characterize children as the first infected household member (so-called *index case*) and to explore how predictors varied across different pandemic waves, particularly between the Delta and non-Delta variant periods.

A key methodological challenge in the study was addressing dependencies inherent in the data, as the infection status of one household member influences others – only a single person can be the index case. To resolve this, we proposed a Monte Carlo simulation approach. This technique involved randomly selecting a single child from each family to create independent data subsets and performing linear regression to identify predictors. This process was repeated 10^4 times to ensure robustness, aggregating results across itera-

tions. To enhance reliability, we implemented a voting mechanism, counting the number of Monte Carlo runs in which each predictor was statistically significant ($p < 0.05$). Predictors were deemed significant if they exceeded a predefined threshold (80% of positive votes), ensuring that findings were consistent and not artifacts of random sampling.

The results of this study reveal that children were the index case in nearly 70% of households during the Delta variant wave, indicating that children often introduced SARS-CoV-2 into their families. Key predictors included attending childcare facilities, such as nurseries, kindergartens, and schools, significantly increasing the likelihood of being the first infected family member. For example, children attending nurseries had an estimate of 1.456 (p<0.001), while school attendance was also strongly associated (estimate = 1.230, p=0.001). Age was another significant factor, with an estimate of 0.106 (p=0.002), indicating older children are the index case more often. A selected part of the findings for the Delta wave is summarized in Table 5.1.

predictor	estimate	SE	z-value	p-value	$p < 0.05$ voting
intercept	-0.820	0.194	-4.228	<0.001	100%
sex (male)	0.015	0.204	0.075	0.891	0%
age	0.106	0.034	3.143	0.002	100%
attending nursery	1.456	0.409	3.565	<0.001	100%
attending kindergarten	0.899	0.296	3.038	0.003	100%
attending school	1.230	0.370	3.320	0.001	100%
fully vaccinated	1.090	0.744	1.451	0.155	0%
previous hospitalizations	0.284	0.453	0.627	0.532	0%

Table 5.1: Summary of Monte Carlo regression results conducted on the complete Delta wave dataset. The "Estimate" represents the average effect size of each predictor, indicating its strength and direction of association with the outcome. "SE" (standard error) reflects the variability of the estimate, while the z-value measures the predictor's statistical significance, with its corresponding p-value indicating the likelihood of observing such an effect by chance. The final column shows the percentage of Monte Carlo runs in which each predictor was deemed significant ($p < 0.05$).

This study directly debunks a widely circulated myth, especially common in Poland during the mid-stage pandemic, that children do not spread SARS-CoV-2. The finding that children were the index case in nearly 70% of households during the Delta wave clearly demonstrates their significant role in household transmission. Moreover, a considerable difference in predictors between the Delta and non-Delta variant waves highlights the dynamic nature of transmission patterns, influenced by both viral properties and social behaviours. These results underscore the importance of targeted infection prevention measures in childcare and educational settings to mitigate the spread of infectious diseases, particularly during waves driven by highly transmissible variants.

5.2. Defining endotype of sarcoidosis related to sTNF-α

❞ Goljan-Geremek, A., Radziński, P., Puścińska, E., Demkow, U. (2024). *Defining serum tumor necrosis factor α concentration-related endotype of sarcoidosis: a real-life, retrospective, observational Polish study.* Polish archives of internal medicine, 134(4), 16718.

Sarcoidosis is a complex multisystem disorder that predominantly affects the lungs and lymphatic system, although it can impact almost any organ in the body. Characterized by the formation of granulomas – tiny clumps of inflammatory cells – in affected tissues, sarcoidosis presents a broad spectrum of clinical manifestations, making its management a challenging endeavor. The disease is most commonly diagnosed in young adults, with a variable prognosis that ranges from spontaneous resolution to chronic progression leading to significant organ damage (Baughman and Lower, 2011). Despite extensive research, the exact cause of sarcoidosis remains elusive, though it is believed to involve a combination of genetic susceptibility and environmental triggers.

Our study focused on the exploration of serum Tumor Necrosis Factor Alpha (sTNF-α) levels as a potential biomarker for defining the endotype of sarcoidosis, specifically its active, severe and multiorgan manifestation that may possibly lead to irreversible organ damage. TNF-α, a critical cytokine in the inflammatory process, plays a pivotal role in the formation and maintenance of granulomas (Sharma *et al.*, 2008). Our investigation aimed to delineate whether sTNF-α concentrations could identify patients with a severe disease endotype characterized by worse radiological, clinical, and functional manifestation.

Statystical analysis

Statistical analysis revealed that sTNF-α levels were significantly associated with several markers of disease severity. We performed the Kruskal-Wallis test to investigate if distributions of serum TNF-α concentration differ between categories of qualitative parameters (H_0: $F_1 = F_2$, H_1: $F_1 \neq F_2$, with F_i standing for cumulative distribution function). For quantitative parameters, Kentall's τ correlation with sTNF-α concentration was computed. Then, the significance of the correlation was tested with the Z-test (H_0: $\tau = 0$, H_1: $\tau \neq 0$). Finally, we applied the Benjamini-Hochberg correction to control the False Discovery Ratio (FDR) for qualitative and quantitative variables separately. a threshold for the Benjamini-Hochberg procedure was set to 5%, indicating that we expect 95% of our discoveries to be actual ones. In contrast, we allow the possibility of some false positives that should remain at the level of 5% of all discoveries.

To distinguish which sTNF-α levels are considered high or low, we introduced *cut-off*

threshold. To compute the threshold, we analyzed all binary variables showing statistically significant differences in sTNF-α levels. For each, we determined a mean sTNF-α level for each of the two values. Afterwards, we established the cut-off as the average between the highest of the lower means and the lowest of the higher means. a graphical representation of the process is presented in Figure 5.3. This cut-off value allowed us to classify patients by sTNF-α concentration, offering insights into the disease's severity and potential progression in an individual based on their specific sTNF-α level.

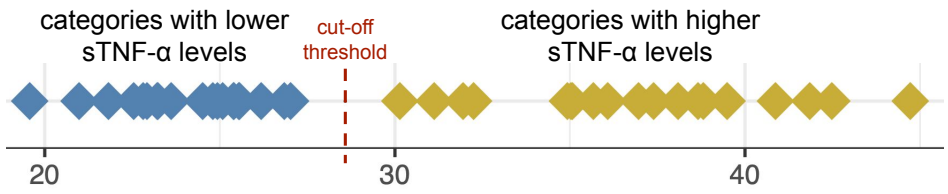

Figure 5.3: A diagram illustrating the selection process for the cut-off threshold. For every binary variable, an average sTNF-α level was calculated per category. Variables demonstrating a significant disparity between the two category averages were employed to establish the threshold. These averages were then divided into two clusters: one comprising the lower averages of significant variables (blue) and the other comprising the higher averages (yellow). The cut-off threshold was determined by finding the midpoint between these two clusters.

Finally, to investigate whether dependencies and probability distributions differ for patients with sTNF-α below and above 28.58 cut-off value, we performed the χ^2 independence test (H_0: there is no association between variables, H_1: otherwise) and Kruskal-Wallis test (again, H_0: $F_1 = F_2$, H_1: $F_1 \neq F_2$, with F_i standing for cumulative distribution function) for qualitative and quantitative variables, respectively. Again, we set the Benjamini-Hochberg procedure threshold to 5%.

Study results and implications

In the context of our study's focus on exploring sTNF-α levels in sarcoidosis patients, we have identified key insights into the biomarker's predictive capacity for disease progression. The cohort consisted of 125 individuals, with an average age of 43.3 years and a disease duration of 4.7 years. A significant number of patients were over 40 at both disease onset (62%) and study enrollment (70%).

Radiological evaluations showed fibrosis increase from 6% at diagnosis to 52% at enrollment, with 60% experiencing disease progression in retrospective analysis. Functionally, lung impairment was evidenced by a decrease in the Diffusing Capacity for Carbon Monoxide (DLco), an indicator of gas exchange efficiency, in 47% of patients, and by airflow limitation, reflecting obstruction in the airways, in 21%.

Critically, sTNF-α levels, irrespective of gender, showed no correlation with the disease's duration, symptoms, or treatment. However, age stood out, with patients over 40 displaying higher sTNF-α levels ($\tau = 0.21$, $p < 0.01$ at onset; $\tau = 0.2$, $p < 0.01$ at enrollment), pointing to an age-associated inflammatory response. Importantly, elevated sTNF-α

levels were significantly associated with radiological disease progression ($p < 0.01$). The inverse correlation has been shown with all lung function parameters except with Forced Expiratory Volume in 1 second to Forced Vital Capacity ratio (FEV_1/FVC). Its concentration was also significantly higher in patients experiencing radiological progression and those with clinical markers indicating a more severe disease course, underscoring the biomarker's value in identifying patients at risk for a disabling form of sarcoidosis. Moreover, exercise capacity, assessed through the 6-Minute Walk Test (6MWT), showed a significant relationship with sTNF-α; lower exercise tolerance correlated with higher sTNF-α levels ($\tau = -0.15, p = 0.03$), especially in patients experiencing a notable decrease in oxygen saturation during exertion ($\tau = -0.25, p < 0.01$ for SpO2max, maximal oxygen saturation).

This detailed exploration underscores sTNF-α's significant role as a biomarker for identifying sarcoidosis patients at risk of severe clinical manifestation of the disease. By highlighting the association between elevated sTNF-α levels and key indicators of disease severity, especially in older patients and those with radiological evidence of progression, it emphasizes the biomarker's utility in refining diagnostic and therapeutic strategies. Integrating sTNF-α level monitoring into clinical protocols can aid healthcare providers in making informed decisions, potentially directing treatments to prevent irreversible damage and facilitating personalized management plans, thereby improving patient outcomes in the complex landscape of sarcoidosis care.

6

Conclusions and future research

"Leading-edge" in the title of this thesis is intended to emphasize that the computational methods presented address critical issues that have hindered the development of science and technology. These solutions have been anticipated by chemical experts, who expected them to boost their work significantly. Thus, at the beginning of this chapter, we wish to summarize once again the achievements of this thesis and how they impact current challenges in mass spectrometry.

6.1. Impact of the thesis on scientific progress

In our research on monoisotopic mass prediction, we focused on addressing the crucial issue of off-by-one dalton errors. To this end, we developed `Envemind` algorithm, a sophisticated method designed explicitly for protein analysis. Unlike `MIND`, which relies on the most abundant mass and offers only a probability distribution across multiple candidates, `Envemind` assumes access to higher-quality mass spectra, enhancing its capability to determine monoisotopic mass accurately. Additionally, our model is trained on a much broader mass range, enabling it to process significantly larger proteins. Importantly, its modular design allows `Envemind` to be easily adapted as advancements in data resolution, computational power, and spectra similarity measures evolve. In conclusion, `Envemind` finds its application in scenarios where prediction errors can be costly, and it is beneficial

to ensure high data quality beforehand.

Compact oligonucleotides, up to approximately 3-4 kDa or equivalently 9-12 nucleobases, have monoisotopic masses that are visibly distinct in mass spectra "by eye". Beyond this limit, prediction becomes necessary, yet even for the shortest DNA and RNA chains, automating the process of monoisotopic mass determination proves beneficial. Thus, the development of `MIND4OLIGOS` algorithm emerged as a straightforward solution to an issue that had not been previously addressed. Consequently, the algorithm has rapidly been integrated into several pipelines, including those at Janssen Pharmaceuticals, part of Johnson & Johnson Pharmaceutical Research and Development. Moreover, again, due to its proposed alternative methodology, the algorithm is well-positioned for adaptation to future advancements in mass spectra quality.

In the field of mass spectrometry imaging, we have introduced a highly efficient compression algorithm based on an encoder-decoder neural network architecture. As demonstrated, applying the encoder to an MS image significantly reduces its memory footprint and enables computations that would otherwise be too CPU-intensive on raw images. Most importantly, encoding facilitates image segmentation tasks without concerns about memory limitations. Additionally, segmentation algorithms like k-means benefit from the regular distribution ensured by the encoder, resulting in higher accuracy compared to segmentation performed on raw images.

Furthermore, we would like to emphasize another important application of our encoding algorithm: efficient storage of MS images. Storing all acquired data remains a significant challenge for laboratories conducting large-scale mass spectrometry imaging experiments. By using our algorithm, only the encoded images and their corresponding trained models – both of which are lightweight in terms of memory requirements – need to be preserved. Thanks to the encoder-decoder architecture, encoded MS images can be easily decoded whenever needed. It is also important to note that the previously mentioned computational times refer to training the encoder, not the encoding process itself. Once the model is trained, encoding and decoding entire MS images take only a few seconds.

At the beginning of Chapter 4, we provided an introduction to the MC-ICP-MS, which, we hope, helps readers understand both the fundamental capabilities and limitations of this field. Beyond the strictly chemical results, which were excluded from this dissertation, we conducted a detailed uncertainty analysis that offered new insights into currently used methods. Our findings revealed that widely adopted approaches, such as DerSimonian-Laird and Hartung-Knapp-Sidik-Jonkman, can produce unreliable estimates when in-study variance is difficult to distinguish from random noise. To address this issue, we developed a straightforward Monte Carlo-based solution that provides a more conservative and reliable framework for uncertainty estimation, ultimately enabling a better assessment of

measurement quality.

In closing, we remain committed to upholding best practices in scientific publishing. Accordingly, all of our publications are available in open access journals, ensuring broad dissemination of our findings. Furthermore, the algorithms and codes used in our analyses are freely accessible to the public. We have also made a concerted effort to ensure that all data first reported in our works are publicly available, supporting transparency and fostering further research in the field. The thesis was written with the assistance of OpenAI's ChatGPT-4 and 4o to refine the language, enhance the overall fluency, and ensure clarity and coherence throughout the text.

6.2. Atomic MS seems to be a key to the future, yet much remains to be done

As our future research primarily focuses on the MC-ICP-MS, we first discuss current points of interest in the field, which can be categorized into three primary objectives.

The most time-consuming aspect of the measurement procedure is sample preparation. Therefore, the first objective focuses on improving these processes. However, this presents a significant challenge, as each element and sample type (e.g., biological or geological, liquid or solid) requires an individual approach. An example of such efforts is our study on selenium separation improvement.

The second objective focuses on the overall improvement of isotopic fractionation correction methods and the development of new ones. Yet, to make progress in that matter, a simultaneous deepening of our understanding of the fractionation nature is required. Moreover, as part of this objective, we also examine the underlying assumptions of these methods to determine whether a given element (or its specific isotopes), with its unique chemical properties, can be accurately measured using a particular correction method. In other words, beyond developing these methods, we aim to characterize their strengths and limitations.

The final objective hinders the most painful thorn in the field – something best illustrated through an example. In 2003, interlaboratory comparisons were conducted to assess differences in a $^{11}B/^{10}B$ isotope ratio (Gonfiantini *et al.*, 2003). The results revealed δ differences between laboratories reaching 11‰, which can be considered exceedingly high. These discrepancies undermine the credibility of the entire isotope ratio measurement field, as each laboratory adheres to its own inconsistent results. In the following years, the experiment was repeated several times, with laboratories given the freedom to choose sample preprocessing techniques, mass spectrometers involved, standards used, and, ultimately, the correction methods applied. a decade later, the differences were reduced to approxi-

mately 1.5–3‰, which represents significant improvement, yet still falls short of being fully satisfactory (Aggarwal *et al.*, 2009; Tirez *et al.*, 2010; Foster *et al.*, 2013).

As stated by Malinovsky *et al.* (2020), the concept of "tested once, accepted everywhere" is becoming increasingly important, yet, as suggested by the title of this section, much remains to be done. Therefore, the third objective focuses on metrology; its goal is to ensure results' traceability and interlaboratory comparability, which can be achieved by validating existing measurement procedures and unification of reference standards for specific isotopic pairs and estimating the expected level of uncertainty.

6.3. The future of our research

First, we aim to develop an online app that provides an easily accessible tool for estimating the uncertainty of the SSB method, while also expanding it with additional functionalities. In particular, we plan to incorporate the inverse-σ method for estimating aggregated uncertainty, for which our preliminary results suggest there may be fewer limitations compared to the DSL and HKSJ estimators (Huang, 2019).

Our next point of interest is the increasingly popular ORM correction method. The key research question here is how to assess whether a fitted regression line will provide reliable results. Currently, the assessment is primarily done using the coefficient of determination, R^2. However, as our recent validations indicate, a high R^2 value is necessary for measurements to provide accurate results, but it is not sufficient. It is a common scenario where the calibration line exhibits nearly perfect R^2, yet the results are inaccurate. Moreover, no suitable alternative is currently known. Therefore, we aim to investigate an assessment method for ORM regression lines, including various approaches involving estimating uncertainty.

While the above research tasks primarily align with the second objective, our final study falls under the third and is related to the Internal Standard method. Significant limitations in the field are of a financial nature; standards are required for every measurement, yet they are costly. Notably, due to the rigorous certification process, standards used in MC-ICP-MS often cost between 2000 and 3000 USD (with limited availability), whereas those used in ICP-MS concentration measurements rarely exceed 200 USD. Thus, the research question is whether a mixture of monoisotopic elements can serve as a substitute for such standards. Our preliminary results, involving the measurement of $^{87}Sr/^{86}Sr$ isotope ratio calibrated using a Nb/Y mixture, suggest that this approach is feasible yet challenging. We hope that success in this study may make the development of more affordable standards possible.

Bibliography

Aggarwal, J., Böhm, F., Foster, G., Halas, S., Hönisch, B., Jiang, S.-Y., Kosler, J., Liba, A., Rodushkin, I., Sheehan, T. *et al.* (2009). How well do non-traditional stable isotope results compare between different laboratories: results from the interlaboratory comparison of boron isotope measurements. *Journal of Analytical Atomic Spectrometry*, **24** (6), 825–831.

Agten, A., Prostko, P., Geubbelmans, M., Liu, Y., De Vijlder, T. and Valkenborg, D. (2021). A compositional model to predict the aggregated isotope distribution for average DNA and RNA oligonucleotides. *Metabolites*, **11** (6), 400.

Alexandrov, T. (2020). Spatial metabolomics and imaging mass spectrometry in the age of artificial intelligence. *Annual review of biomedical data science*, **3**, 61–87.

— and Kobarg, J. H. (2011). Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. *Bioinformatics*, **27** (13), i230–i238.

Barr, A. J. (2018). The biochemical basis of disease. *Essays in biochemistry*, **62** (5), 619–642.

Basu, A., Schilling, K., Brown, S. T., Johnson, T. M., Christensen, J. N., Hartmann, M., Reimus, P. W., Heikoop, J. M., Woldegabriel, G. and DePaolo, D. J. (2016). Se isotopes as groundwater redox indicators: detecting natural attenuation of Se at an in situ recovery U mine. *Environmental science & technology*, **50** (20), 10833–10842.

Baughman, R. P. and Lower, E. E. (2011). Who dies from sarcoidosis and why? *American journal of respiratory and critical care medicine*, **183** (11), 1446–1447.

Beuque, M., Martin-Lorenzo, M., Balluff, B., Woodruff, H. C., Lucas, M., de Bruin, D. M., van Timmeren, J. E., de Boer, O. J., Heeren, R. M., Meijer, S. L. *et al.* (2021). Machine learning for grading and prognosis of esophageal dysplasia using mass spectrometry and histological imaging. *Computers in biology and medicine*, **138**, 104918.

Chen, Y.-F., Chang, C. A., Lin, Y.-H. and Tsay, Y.-G. (2013). Determination of accurate protein monoisotopic mass with the most abundant mass measurable using high-resolution mass spectrometry. *Analytical biochemistry*, **440** (1), 108–113.

Chughtai, K. and Heeren, R. M. (2010). Mass spectrometric imaging for biomedical tissue analysis. *Chemical reviews*, **110** (5), 3237–3277.

Ciach, M. A., Miasojedow, B., Skoraczyński, G., Majewski, S., Startek, M., Valkenborg, D. and Gambin, A. (2020). Masserstein: Linear regression of mass spectra by optimal transport. *Rapid Communications in Mass Spectrometry*, p. e8956.

Claesen, J., Lermyte, F., Sobott, F., Burzykowski, T. and Valkenborg, D. (2015). Differences in the elemental isotope definition may lead to errors in modern mass-spectrometry-based proteomics. *Analytical chemistry*, **87** (21), 10747–10754.

Costas-Rodríguez, M., Delanghe, J. and Vanhaecke, F. (2016). High-precision isotopic analysis of essential mineral elements in biomedicine: natural isotope ratio variations as potential diagnostic and/or prognostic markers. *TrAC Trends in Analytical Chemistry*, **76**, 182–193.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, **7** (3), 177–188.

Dexter, A., Race, A. M., Steven, R. T., Barnes, J. R., Hulme, H., Goodwin, R. J., Styles, I. B. and Bunch, J. (2017). Two-phase and graph-based clustering methods for accurate and efficient segmentation of large mass spectrometry images. *Analytical chemistry*, **89** (21), 11293–11300.

Dittwald, P., Claesen, J., Burzykowski, T., Valkenborg, D. and Gambin, A. (2013). Brain: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry. *Analytical chemistry*, **85** (4), 1991–1994.

Egli, M. and Manoharan, M. (2023). Chemistry, structure and function of approved oligonucleotide therapeutics. *Nucleic Acids Research*, **51** (6), 2529–2573.

Far, J., Bérail, S., Preud'Homme, H. and Lobinski, R. (2010). Determination of the selenium isotopic compositions in Se-rich yeast by hydride generation-inductively coupled plasma multicollector mass spectrometry. *J. Anal. At. Spectrom.*, **25** (11), 1695–1703.

Fortunato, G., Ritter, A. and Fabian, D. (2005). Old masters' lead white pigments: investigations of paintings from the 16th to the 17th century using high precision lead isotope abundance ratios. *Analyst*, **130** (6), 898–906.

Foster, G. L., Hönisch, B., Paris, G., Dwyer, G. S., Rae, J. W., Elliott, T., Gaillardet, J., Hemming, N. G., Louvat, P. and Vengosh, A. (2013). Interlaboratory comparison of boron isotope analyses of boric acid, seawater and marine $CaCO_3$ by MC-ICPMS and NTIMS. *Chemical Geology*, **358**, 1–14.

Glazier, D. A., Liao, J., Roberts, B. L., Li, X., Yang, K., Stevens, C. M. and Tang, W. (2020). Chemical synthesis and biological application of modified oligonucleotides. *Bioconjugate Chemistry*, **31** (5), 1213–1233.

Gonfiantini, R., Tonarini, S., Gröning, M., Adorni-Braccesi, A., Al-Ammar, A. S., Astner, M., Bächler, S., Barnes, R. M., Bassett, R. L., Cocherie, A. *et al.* (2003). Intercomparison of boron isotope and concentration measurements. Part II: evaluation of results. *Geostandards Newsletter*, **27** (1), 41–57.

Guo, D., Föll, M. C., Bemis, K. A. and Vitek, O. (2023). A noise-robust deep clustering of biomolecular ions improves interpretability of mass spectrometric images. *Bioinformatics*, **39** (2), btado67.

Han, J., Permentier, H., Bischoff, R., Groothuis, G., Casini, A. and Horvatovich, P. (2019). Imaging of protein distribution in tissues using mass spectrometry: An interdisciplinary challenge. *TrAC Trends in Analytical Chemistry*, **112**, 13–28.

Hou, X. and Roos, P. (2008). Critical comparison of radiometric and mass spectrometric methods for the determination of radionuclides in environmental, biological and nuclear waste samples. *Analytica chimica acta*, **608** (2), 105–139.

Hsieh, E. J., Hoopmann, M. R., MacLean, B. and MacCoss, M. J. (2010). Comparison of database search strategies for high precursor mass accuracy MS/MS data. *Journal of proteome research*, **9** (2), 1138–1143.

Hu, H., Bindu, J. P. and Laskin, J. (2022). Self-supervised clustering of mass spectrometry imaging data using contrastive learning. *Chemical Science*, **13** (1), 90–98.

Huang, H. (2019). A unified formula for uncertainty estimation in interlaboratory studies and key comparisons. *Cal Lab the International Journal of Metrology*, **26** (4), 22–29.

Huynh, D. and Elhamifar, E. (2020). Interactive multi-label cnn learning with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9423–9432.

Karasiński, J., Tupys, A., Yang, L., Mester, Z., Halicz, L. and Bulska, E. (2020). Novel approach for the accurate determination of Se isotope ratio by multicollector ICP-MS. *Anal. Chem.*, **92** (24), 16097–16104.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kumar Kulabhusan, P., Hussain, B. and Yüce, M. (2020). Current perspectives on aptamers as diagnostic tools and therapeutic agents. *Pharmaceutics*, **12** (7), 646.

Laeter, J. R. D., Böhlke, J. K., Bièvre, P. D., Hidaka, H., Peiser, H., Rosman, K. J. R. and Taylor, P. D. P. (2003). Atomic weights of the elements: Review 2000. *Pure and Applied Chemistry*, **75**, 683–800.

Larsen, K., Wielandt, D. and Bizzarro, M. (2018). Multi-element ion-exchange chromatography and high-precision MC-ICP-MS isotope analysis of Mg and Ti from sub-mm-sized meteorite inclusions. *Journal of Analytical Atomic Spectrometry*, **33** (4), 613–628.

Le-Khac, P. H., Healy, G. and Smeaton, A. F. (2020). Contrastive representation learning: A framework and review. *IEEE Access*, **8**, 193907–193934.

Lee, E. S. and Forthofer, R. N. (2005). *Analyzing complex survey data*. Sage Publications.

Lermyte, F., Dittwald, P., Claesen, J., Baggerman, G., Sobott, F., O'Connor, P. B., Laukens, K., Hooyberghs, J., Gambin, A. and Valkenborg, D. (2019). MIND: A double-linear model to accurately determine monoisotopic precursor mass in high-resolution top-down proteomics. *Analytical chemistry*, **91** (15), 10310–10319.

Little, J. L., Williams, A. J., Pshenichnov, A. and Tkachenko, V. (2011). Identification of "known unknowns" utilizing accurate mass data and chemspider. *Journal of the American Society for Mass Spectrometry*, **23** (1), 179–185.

Łącki, M. K., Startek, M., Valkenborg, D. and Gambin, A. (2017). IsoSpec: Hyperfast fine structure calculator. *Analytical chemistry*, **89** (6), 3272–3277.

Longuespée, R., Casadonte, R., Kriegsmann, M., Pottier, C., Picard de Muller, G., Delvenne, P., Kriegsmann, J. and De Pauw, E. (2016). Maldi mass spectrometry imaging: A cutting-edge tool for fundamental and clinical histopathology. *PROTEOMICS – Clinical Applications*, **10** (7), 701–719.

Ma, V. P.-Y. and Salaita, K. (2019). DNA nanotechnology as an emerging tool to study mechanotransduction in living systems. *Small*, **15** (26), 1900961.

Majewski, S., Ciach, M. A., Startek, M., Niemyska, W., Miasojedow, B. and Gambin, A. (2018). The Wasserstein distance as a dissimilarity measure for mass spectra with application to spectral deconvolution. In *18th International Workshop on Algorithms in Bioinformatics (WABI 2018)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Malinovsky, D., Dunn, P. and Goenaga-Infante, H. (2020). Calibration of boron isotope ratio measurements by MC-ICP-MS using normalisation to admixed internal standards. *Journal of Analytical Atomic Spectrometry*, **35** (11), 2723–2731.

Maréchal, C. N., Télouk, P. and Albarède, F. (1999). Precise analysis of copper and zinc isotopic compositions by plasma-source mass spectrometry. *Chemical geology*, **156** (1-4), 251–273.

Moczulski, M. (2024). *Segmentation of mass spectrometry images by self-supervised contrastive learning and classification head*. Master's thesis, University of Warsaw.

Nord, A. G. and Billström, K. (2018). Isotopes in cultural heritage: present and future possibilities. *Heritage Science*, **6**, 1–13.

Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., Kundu, D. J., Prakash, A., Frericks-Zipper, A., Eisenacher, M. *et al.* (2022). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic acids research*, **50** (D1), D543–D552.

Reitsema, L. J. (2013). Beyond diet reconstruction: stable isotope applications to human physiology, health, and nutrition. *American Journal of Human Biology*, **25** (4), 445–456.

Roberts, T. C., Langer, R. and Wood, M. J. (2020). Advances in oligonucleotide drug delivery. *Nature Reviews Drug Discovery*, **19** (10), 673–694.

Römpp, A., Guenther, S., Schober, Y., Schulz, O., Takats, Z., Kummer, W. and Spengler, B. (2010). Histology by mass spectrometry: label-free tissue characterization obtained from high-accuracy bioanalytical imaging. *Angewandte chemie international edition*, **49** (22), 3834–3838.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, **20**, 53–65.

Rouxel, O., Ludden, J., Carignan, J., Marin, L. and Fouquet, Y. (2002). Natural variations of Se isotopic composition determined by hydride generation multiple collector inductively coupled plasma mass spectrometry. *Geochim. Cosmochim. Acta*, **66** (18), 3191–3199.

Röver, C., Knapp, G. and Friede, T. (2015). Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC medical research methodology*, **15**, 1–7.

Rubner, Y., Tomasi, C. and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, **40** (2), 99–121.

Sarkari, S., Kaddi, C. D., Bennett, R. V., Fernández, F. M. and Wang, M. D. (2014). Comparison of clustering pipelines for the analysis of mass spectrometry imaging data. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, pp. 4771–4774.

Schwamborn, K. (2012). Imaging mass spectrometry in biomarker discovery and validation. *Journal of proteomics*, **75** (16), 4990–4998.

Senko, M. W., Beu, S. C. and 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.

Sharma, S., Ghosh, B. and Sharma, S. (2008). Association of TNF polymorphisms with sarcoidosis, its prognosis and tumour necrosis factor (TNF)-α levels in asian indians. *Clinical & Experimental Immunology*, **151** (2), 251–259.

Sidik, K. and Jonkman, J. N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in medicine*, **26** (9), 1964–1981.

Sindern, S. (2017). Analysis of rare earth elements in rock and mineral samples by ICP-MS and LA-ICP-MS. *Physical Sciences Reviews*, **2** (2), 20160066.

Stüeken, E., Foriel, J., Nelson, B., Buick, R. and Catling, D. (2013). Selenium isotope analysis of organic-rich shales: advances in sample preparation and isobaric interference correction. *J. Anal. At. Spectrom.*, **28** (11), 1734–1749.

Suárez-Criado, L., Queipo-Abad, S., Rodríguez-González, P. and Alonso, J. I. G. (2024). Comparison of different mass bias correction procedures for the measurement of mercury species-specific isotopic composition by gas chromatography coupled to multicollector ICP-MS. *Journal of Analytical Atomic Spectrometry*, **39** (2), 508–517.

Sullivan, K., Layton-Matthews, D., Leybourne, M., Kidder, J., Mester, Z. and Yang, L. (2020). Copper isotopic analysis in geological and biological reference materials by MC-ICP-MS. *Geostand. Geoanal. Res.*, **44** (2), 349–362.

Tirez, K., Brusten, W., Widory, D., Petelet, E., Bregnot, A., Xue, D., Boeckx, P. and Bronders, J. (2010). Boron isotope ratio (δ^{11}B) measurements in water framework directive monitoring programs: Comparison between double focusing sector field ICP and thermal ionization mass spectrometry. *Journal of Analytical Atomic Spectrometry*, **25** (7), 964–974.

UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic acids research*, **43** (D1), D204–D212.

Valkenborg, D., Mertens, I., Lemière, F., Witters, E. and Burzykowski, T. (2012). The isotopic distribution conundrum. *Mass spectrometry reviews*, **31** (1), 96–109.

Vaysse, P.-M., Heeren, R. M., Porta, T. and Balluff, B. (2017). Mass spectrometry imaging for clinical research–latest developments, applications, and current limitations. *Analyst*, **142** (15), 2690–2712.

Villani, C. (2008). *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg.

Zubarev, R. A., Håkansson, P. and Sundqvist, B. (1996). Accuracy requirements for peptide characterization by monoisotopic molecular mass measurements. *Analytical Chemistry*, **68** (22), 4060–4063.