SILESIAN UNIVERSITY OF TECHNOLOGY

FACULTY OF AUTOMATIC CONTROL, ELECTRONICS AND COMPUTER SCIENCE

Doctoral Dissertation

by

Anna Glodek

Deisotoping methods in MALDI ToF Mass Spectrometry Imaging

Supervisor: Professor Joanna Polańska, PhD, DSc
Co-supervisor: Marta Gawin, PhD

2023
Gliwice, Poland

ACKNOWLEDGEMENTS

Foremost, I would like to express my gratitude to my supervisor, **Professor Joanna Polańska,** for her constant support at every stage of my research and determining the direction for further research. I am also grateful to my co-supervisor **Marta Gawin, PhD,** for all the support, data provided, and expertise shared with me. I would also like to thank **Professor Monika Pietrowska** for expert knowledge shared with me. My sincere thanks also go to **Professor Jacek Łęski** for providing knowledge and expertise in fuzzy-inference systems. Moreover, I would like to thank **Katarzyna Frątczak** for providing the implemented methods of pre-processing and feature extraction.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Meaning
BU	bottom-up
BU-MS	bottom-up mass spectrometry
CE-MS	capillary-electrophoresis-mass spectrometry
CSI	Critical Success Index
DT	Decision Tree
E	envelope
FF	fresh frozen
FFPE	formalin-fixed paraffin-embedded
FMI	Fowlkes-Mallows Index
FN	false negative
FP	false positive
FT-ICR	Fourier Transform Ion Cyclotron Resonance
GC-MS, GC/MS	gas chromatography-mass spectrometry
GLCM	Gray-Level Co-Occurrence Matrix
HNC	head and neck cancer
IMS/MS, IMMS	ion mobility-spectrometry – mass spectrometry
ITO	indium tin oxide
LC-MS, LC/MS	liquid chromatography-mass spectrometry
LCQ, LTQ	Linear Ion Trap
m/z	mass-to-charge ratio
MA	Mamdani-Assilan
MALDI	Matrix-Assisted Laser Desorption/Ionization
MCC	Matthews Correlation Coefficient
MS	mass spectrometry
MS/MS, MS2	tandem mass spectrometry
MSI	Mass Spectrometry Imaging
NB	Naiive Bayes
nE	non-Envelope
OCT	optimal cutting temperature
SVM	Support Vector Machine
TD	top-down
TN	true negative
ToF	time of flight
TP	true positive

1. INTRODUCTION

1.1. Introduction and Statement of the Problem

One of the most meaningful steps towards protein identification is mass spectrometry which allows obtaining protein structural information, such as amino acid sequence, which can be the basis of protein identification via searching protein databases [1]. Due to the fact that the majority of chemical elements have isotopes of different masses, the isotopic mass of a molecule observed on a mass spectrum reflects the kind and number of atoms included in the measured (molecular) ion and the distribution of different isotopes [2]. Depending on the achievable resolving power of a mass spectrometer, molecular ions can be represented by either the monoisotopic mass, which considers only the masses of the most abundant naturally occurring stable isotope of each atom present in a molecule or the average mass, which considers the presence of both light and heavy isotopes. Theoretically, it can be considered as the sum of the average weights of all elements. For an atom, the difference between those two masses is insignificant. However, within a biomolecule like a peptide, the difference between the monoisotopic mass and the average mass increases with the number of atoms of which the biomolecule is constructed. Such mass variance causes a bias in results leading to the misidentification of peptides and the limitation of the number of quantified peptides. Therefore, accurate identification should consider the presence of natural isotopes in peptides. [2]

It has been proven in [2] that considering natural isotopes in the analysis results in precise peptide quantification and validation. The method for removing isotopic envelopes from a mass spectrum is called *deisotoping*.

A plethora of methods concerning deisotoping exist, but they have several limitations, which were thoroughly explained and compared in 2.6. Most of these methods are dedicated to different kinds of data – from different types of mass spectrometry experiments, various types of biomolecules-driven, suitable only for low- and high-resolution mass spectrometry, or not suitable for large MALDI-MSI-driven data. Hereby, the method for MALDI-TOF data-driven deisotoping, based on universal assumptions and expert knowledge in the field of mass spectrometry imaging, is proposed, which applies different approaches to finding a solution – fuzzy-inference system (4), machine learning, and statistics (5).

1.2. Purpose of the Study

The main goal of the research described in this dissertation is to identify isotopic envelopes for data from MALDI-TOF mass spectrometry imaging experiments by applying *in silico* methods. Due to the developed methodology, the accuracy of peptide detection can be improved since deisotoping is based on removing the peaks that are the members of an isotopic envelope. As a consequence, peptide detection is more accurate and reliable. The research goal is as follows: application of the method of initial preselection of isotopic envelope members using a combination of fuzzy-inference system with an evaluation of the peaks' spatial distribution. Such a method allows for effective isotopic envelopes identification in medium-resolution mass spectrometer data.

The goal will be achieved by:

1. development of the fuzzy-inference system for the preselection of the isotopic envelope members
2. verification of the membership of the predefined peaks into the isotopic envelopes based on the descriptors of the spatial distribution of the peaks' differential images.

2. REVIEW OF THE LITERATURE

2.1. Proteomics

Proteomics is a large-scale study of proteins, their structure, functions, and physiological role [3]. Proteins are crucial biomolecules since they are either structural or functional elements of a cell, their amino acid sequences determine their structure and, accordingly, their cellular function. It is worth mentioning that they can also function in the extracellular space, circulating via the bloodstream. Therefore the plethora of proteins (e.g. serum or urine proteins) serve as clinical biomarkers. [4]

Proteins determine the cellular structure, activity, and signalling between cells and tissues. Moreover, they support metabolism by catalysing chemical reactions. Generally, proteins are notable biomolecules from medicine's point of view, as they can be a root cause of a disease (such as Huntington's or Alzheimer's disease), but what is more, they can be used for curing – for instance, antibodies are used in therapies against bacterial or viral infections. [5]

Spatial proteomics is essential from the point of view of modern biology and medicine, since it enables detection of dozens of proteins across a tissue with their simultaneous spatial distribution, and preservation of tissue histology. [6]

Proteins are built from amino acids, linked by peptide bonds [7] [8]. Depending on the number of linked amino acids in the molecules, the following groups of biomolecules can be distinguished: oligopeptides (2 – 20 amino acids) [9], polypeptides (20 – 50 amino acids) [10], proteins (above 50 amino acids) [11].

The vast majority of proteins are too large to be analysed in a mass spectrometer as intact molecules. In order to perform their mass spectrometric measurement, they are usually digested to peptides using a specific proteolytic enzyme, e.g. trypsin [12] (*Figure 1*).

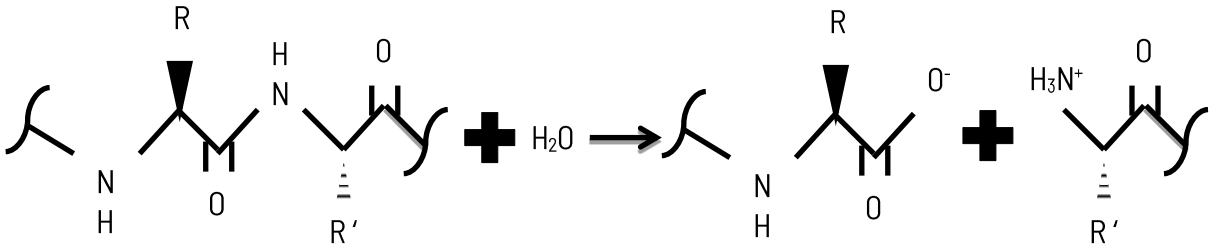

Figure 1. Scheme of trypsin digestion.

Several proteases are used in proteomics experiments (*Table 1*), but the most frequently used is trypsin, which cleaves proteins into peptides with an average size of 700-1500 Da (such a range is considered ideal for mass spectrometry) [13].

The cleavage site is the carboxyl side of Arginine and Lysine residues. As a result, a positive charge appears at the peptide C-terminus, which is an advantage from the MS (mass spectrometry) analysis point of view, and tryptic peptides become easily detectable by the mass spectrometer. [13] [14]

Table 1. Proteases used in typical shotgun proteomics experiments **[15]**.

Enzyme name	Cleavage site	Recommended digestion conditions		
		pH	temperature [◦C]	hours
ArgC	C-terminal of R	8	37	12
AspN	N-terminal of D	8	37	12
Chymotrypsin	C-terminal of F, Y, L, W and M	8	25	12
GluC	C-terminal of D	8	25	12
LysargiNase	N-terminal of R and K	-	-	-
LysC	C-terminal of K	8	37	12
LysN	N-terminal of K	8	37	12
Pepsin	C-terminal of Y, F and W	-	-	-
Trypsin	C-terminal of R and K	8	37	12
WaLP and MaLP	C-terminal of aliphatic amino acids	-	-	-

Mass spectrometry is an analytical technique for protein analysis [16] that allows accurate measurement of chemical substances' molecular masses. Moreover, it provides information on molecular composition and chemical structure. [17]

There are two approaches used in mass spectrometry to identify proteins and define their amino acid sequences and post-translational modifications: bottom-up (BU) and top-down (TD)

approach. Bottom-up proteomics allows for the analysis of peptides originating from enzymatically digested proteins. In this approach, the information on the protein is gathered based on its peptide fragments identification/analysis, which were created as a result of proteolytic cleavage. Consequently, the BU approach is based on creating the protein beginning from peptides and ending on the whole protein. In contrast, top-down proteomics allows the direct analysis of an intact proteoform (which arose from the same gene product via degradation and post-translational modifications) which cleaves into smaller fragments during the mass spectrometry experiment, [18][19]

however it requires application of dedicated techniques of ion activation. Thus, such an approach is more complex. Moreover, the protein should be pure and isolated. The analysed sample cannot be a mixture, as the obtained results will be hard to analyse. The bottom-up mass spectrometry (BU-MS) is the most widely used proteomic approach nowadays. It involves several steps, such as protein enzymatic digestion into peptides, ions separation according to their mass-to-charge ratio (m/z), selection of ions for fragmentation, controlled fragmentation of parent ions and detection of daughter/fragment ions. [5]

In mass-spectrometry-based proteomics, the accurate determination of a mass is challenging in many cases because of the high mass of the analytes and applied method - ESI (Electrospray Ionization, soft ionization technique favouring the protonation of many amino acid residues). As a result, different charge states and isotopes (in the case of a mass analyser with specified resolution) appear on the spectrum, which makes the signal complex. [18]

Accordingly, deconvolution and deisotoping are crucial steps in mass spectrometry data analysis, thoroughly explained in the subsequent chapters.

2.2. Mass spectrometry

An essential part of clinicopathological diagnostics is investigation of molecular and morphological tissue features. Understanding the molecular basis of diseases provides meaningful insights into their mechanism. One of various advanced technologies used for tissue-based research is mass spectrometry. Mass spectrometry (MS) has become a powerful tool to characterize large molecules [20], ranging from 100 Da to beyond 100 kDa [21]. Its advantages are high sensitivity, a wide range of detectable molecules, and molecular specificity. [22]

Due to the need for high-throughput techniques in cancer proteomics, mass spectrometry-based techniques are widely used in the oncology clinic since they are helpful for quantifying proteins from complex clinical samples [23].

Mass spectrometry analysis is based on ionising the molecules and determining the mass-to-charge (m/z) ratios of molecular ions [17]. A molecule needs to undergo the ionisation process, as mass spectrometers can only detect charged analytes (e.g. proteins, peptides) [24].

Gas-phase ions are split into characteristic fragments. As a result, the measured fragment masses define the original ion's molecular structure. [17]

There are a plethora of ionization techniques and various types of mass analysers and detectors. High- and low-resolution mass spectrometry instruments can be distinguished and many configurations of them exist. Still, they consist of three basic parts: the ion source, the mass analyser, and the detector [25] (*Figure 2*).

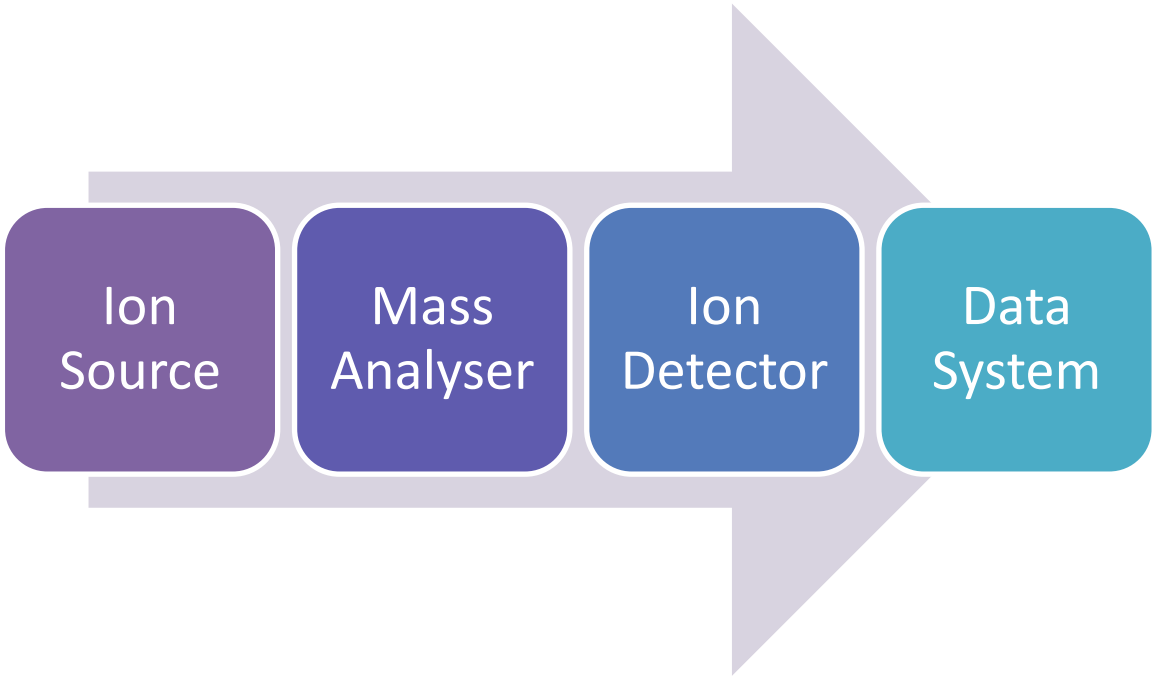

Figure 2. Flowchart of a mass spectrometer.

In an ion source, a sample is ionised. Then, the obtained ions are separated according to their mass-to-charge ratio in a mass analyser. The detector identifies ions and finally, obtained mass spectra are analysed and interpreted.

A mass spectrum is a plot of the relative abundance of ions (the numbers of counts from the ion detector [26]) versus their *m/z* values [27] (*Figure 3*). The higher the mass spectrometer

resolution is, the more peaks are visible in a mass spectrum, since an isotopic distribution in a given molecule gets visible in a mass spectrum.

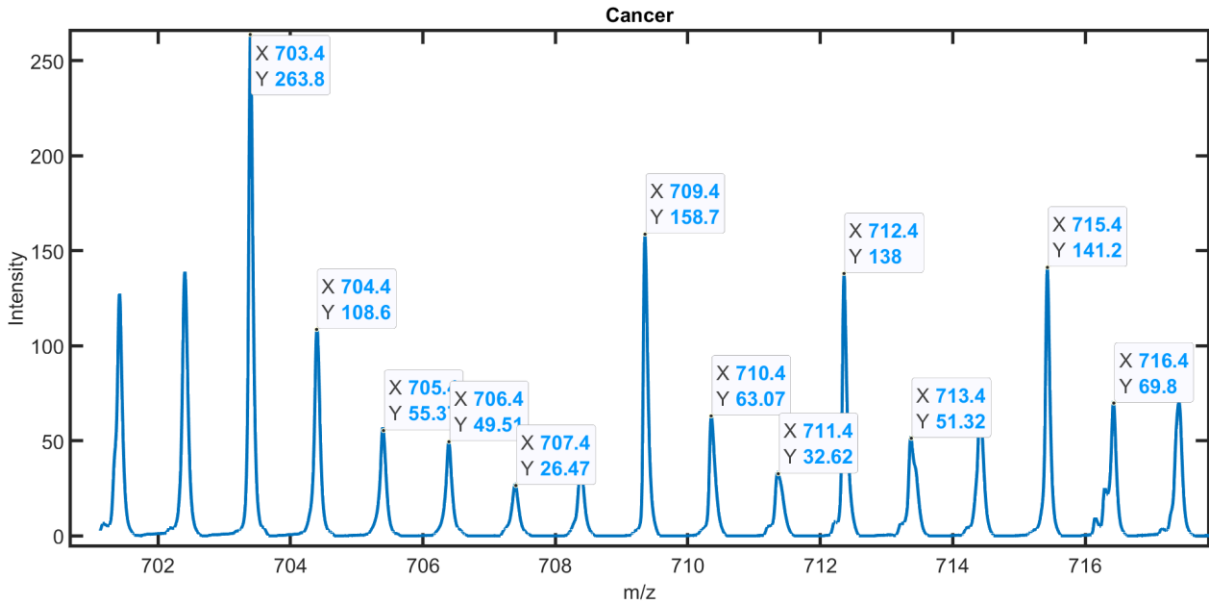

Figure 3. A MALDI-TOF MSI mass spectrum of tryptic peptides.

There are two types of ionisation techniques: soft and hard. The main product of the first method is a molecular ion, whilst the second method allows to detect the molecular ion and fragments of a given molecule. [28]

The most important and frequently used soft ionisation methods are MALDI and ESI:

1. Matrix-Assisted Laser Desorption/Ionization (MALDI) is based on the desorption and ionisation of analyte molecules by using a pulsed laser beam which is directed to co-crystals of matrix and the analyte [25];
2. In Electrospray Ionization (ESI) [17] [29] [30] high voltage is used in order to disperse and spray solvated analytes; multiply charged ions are created. An analyte is ionised in the liquid phase and released to the gas phase [25].

Other ionisation methods are as follows: Atmospheric Pressure Chemical Ionization (APCI) [25], Atmospheric Pressure Photoionization (APPI) [25], Ambient Desorption Ionization [25], Desorption Electrospray Ionization (DESI) [31], and Direct Analysis in Real Time (DART) [25].

Mass analysers allow the determination the mass-to-charge ratio of the ions by subjecting the ions to several magnetic or electric fields [17] [27].

The following types of mass analysers can be distinguished:

1. Quadrupole: generated ions pass through rods, and rotating magnetic field potentials are then applied to them in order to focus the ions and allow the ions to reach a detector [17] [32];
2. TOF (time of flight): the principle of operation is measuring the flight time for an ion [33], which is converted to an *m/z* value [17]. Generally, ions with the same kinetic energy (they are accelerated by an applied potential [17]) but different masses have different velocities – lighter ions arrive before the heavier ones. Thus, a mass spectrum can be recorded [33];
3. Ion trap: ions stabilise their trajectories because of hyperbolic-shaped electrodes inside an ion trap [17];
4. FT-ICR (Fourier Transform Ion Cyclotron Resonance): a trapped ion technique which is based on measuring the cyclotron frequency (cyclic oscillation of a charged ion within a magnetic field). Fourier Transform is used to transform the time-domain transient current to the frequency measurements [34];
5. Orbitrap: a trapped ion technique, ions are subjected to an electric field, cycle around the central electrode, and oscillate along the horizontal axis [17]. In order to measure ion frequencies, time-domain image current transients are acquired, and simultaneously Fast Fourier Transform is used to obtain a mass spectrum [35].

Separation techniques are combined with mass spectrometry in order to separate different compounds. The separation techniques are as follows:

1. Liquid chromatography–mass spectrometry (LC-MS or LC/MS) is a technique used for elucidating the composition of liquid samples. The sample contains analytes (molecules such as lipids or peptides), which have to be identified in the process called *identification*. Moreover, their quantity in the sample has to be determined in the process called *quantification*. [24]

   Generally, chromatography separates the components by distributing them between two phases: stationary (solid, gel or liquid) and mobile (liquid, gas or supercritical fluid), which moves in a definite direction [36];
2. Gas chromatography-mass spectrometry (GC-MS or GC/MS) [37];
3. Capillary-electrophoresis-mass spectrometry (CE-MS) [38];
4. Ion mobility spectrometry-mass spectrometry (IMS/MS or IMMS) [39].

Tandem mass spectrometry is based on controlled fragmentation of a selected ion and as a result, a mass spectrum of the resulting fragment ions is obtained. MS/MS (MS2) is simple, single-stage tandem mass spectrometry experiment, whereas other, higher-order stages are also possible and take place when a fragment ion (so-called "daughter ion") is further fragmented into a "grand-daughter ion", and so forth. [16] [40]

2.3. MALDI-TOF Mass Spectrometry

Generally, for proteins, the most effective ionisation methods are soft ionisation methods [17], such as MALDI and ESI [21] [41], as they are used for sensitive detection of large, labile molecules using mass spectrometry [16]. Several types of mass analysers can be used for MALDI ions analysis, with the time-of-flight mass analyser (TOF) as one of the most common [16].

A solution of an analyte, or a mixture of analytes, is mixed with a solution of a matrix and the solvent is evaporated, which results in co-crystallisation of an analyte with a matrix. Then, a pulsed laser beam irradiates the crystals, which results in the desorption of analyte and matrix molecules, followed with ionisation of analytes in the gas phase. TOF mass analyser is used for separation of the formed ions. Knowing the flight time (T), the mass-to-charge ratio can be calculated based on the following *Eq. 1*: [42]

$$T = C_1\sqrt{m/z} + C_2 \quad (1)$$

where:

C_1, C_2 are the instrumental constants.

2.4. MALDI Mass Spectrometry Imaging (MALDI MSI)

The basic principle of mass spectrometry imaging (MSI) is performing mass spectrometry directly on the sample surface. This technique is used for gathering information on the distribution of various molecules, such as lipids, peptides, proteins, and drugs. MALDI is one of the most commonly used ionisation methods. What differentiates MSI from other general imaging methods, such as optical microscopy, is the possibility of acquiring a variety of molecular distributions at once – especially in the case of TOF MS. Moreover, another distinct feature is that distribution of

molecules that have been ionised can be visualised in a form of molecular images generated for the analysed tissue. [43]

Sample preparation is one of the essential steps of the experiment because tissue preservation is critical from when the surgical resection takes place to the protein digestion stage in order to avoid sample degradation or contamination [21] [23]. Therefore, different kinds of preservation methods can be distinguished, each with its unique sets of advantages and disadvantages: fresh frozen (FF), formalin-fixed paraffin-embedded (FFPE), and optimal cutting temperature embedded (OCT) [23]. MALDI Imaging experiment differs for sample preparation of FF (*Figure 4*) tissue and of FFPE tissue (*Figure 5*). In both methods tissue specimens are sectioned [22] and mounted on conductive indium tin oxide (ITO)-coated slides [44]. The processing of FF tissue includes washing in order to remove all unwanted chemical species [22] – to remove lipids the most common methods are employed: "Carnoy's wash" [44] or washing in increasingly higher percentages of ethanol [22]. Finally, a tissue section is covered with a matrix in order to extract molecules from the tissue specimen into the matrix [22].

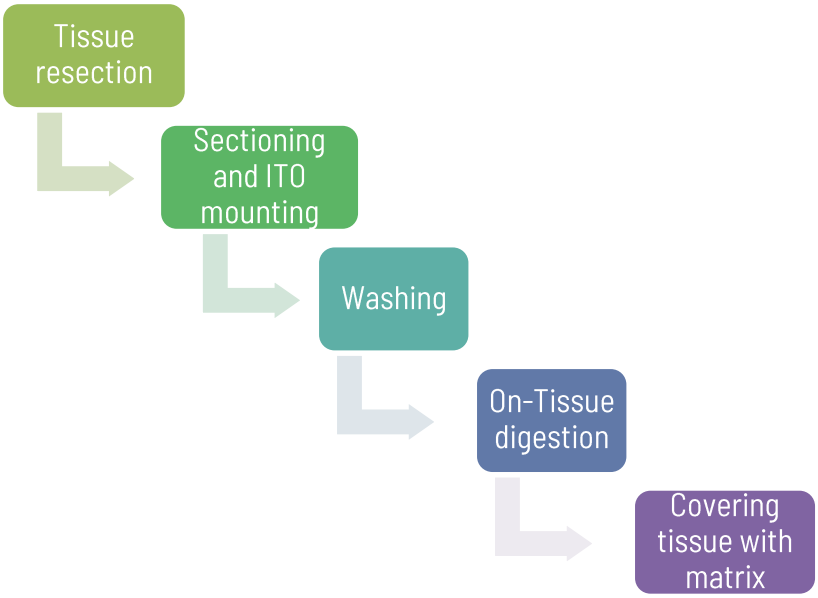

Figure 4. FF tissue collection and preservation for MALDI MSI **[21] [22] [43] [45].**

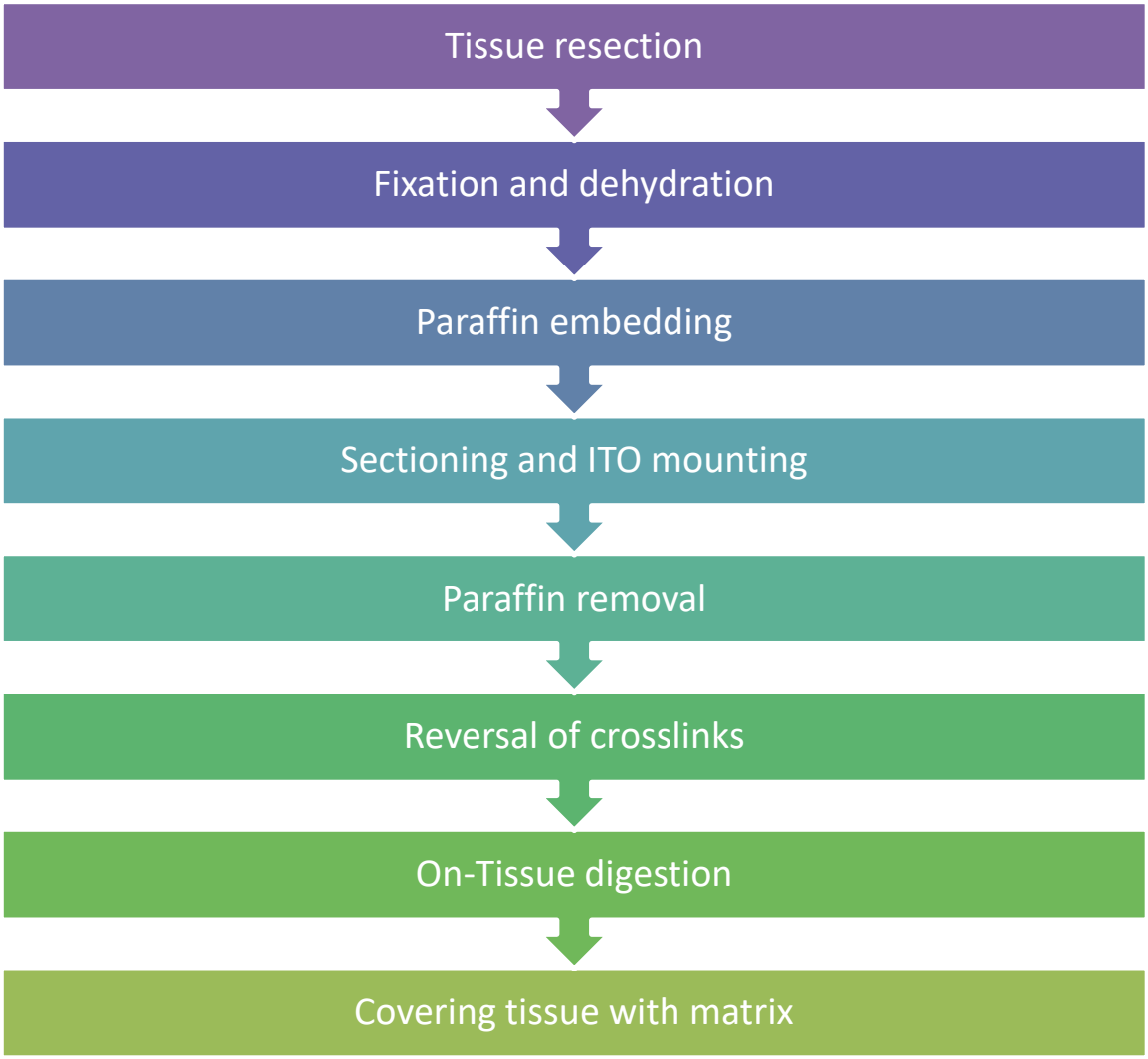

Figure 5. FFPE tissue collection and preparation for MALDI MSI [21][22][43][45].

Principle of MSI [21][22] (*Figure 6*):

After covering a tissue section with a matrix, a pulsed laser beam [46] shoots in the matrix layer, whereas the underlying tissue remains intact (it allows histological tissue examination in the same tissue section after the measurement). The matrix absorbs the laser energy, and the analytes are desorbed and transferred to the gas phase where they get ionised. The produced ions are then accelerated and analysed in the TOF mass analyser. The tissue is scanned with the laser, with a pre-defined raster width, so that a mass spectrum is acquired for each laser ablation position. After the MALDI Imaging experiment, the tissue section can be stained, for instance, using the H&E technique (Hematoxylin and eosin stain is widely used in histologic examination of human tissues [47] [48]) for further histopathological examination. Then, the detected m/z signals are visualised as colour intensity maps.

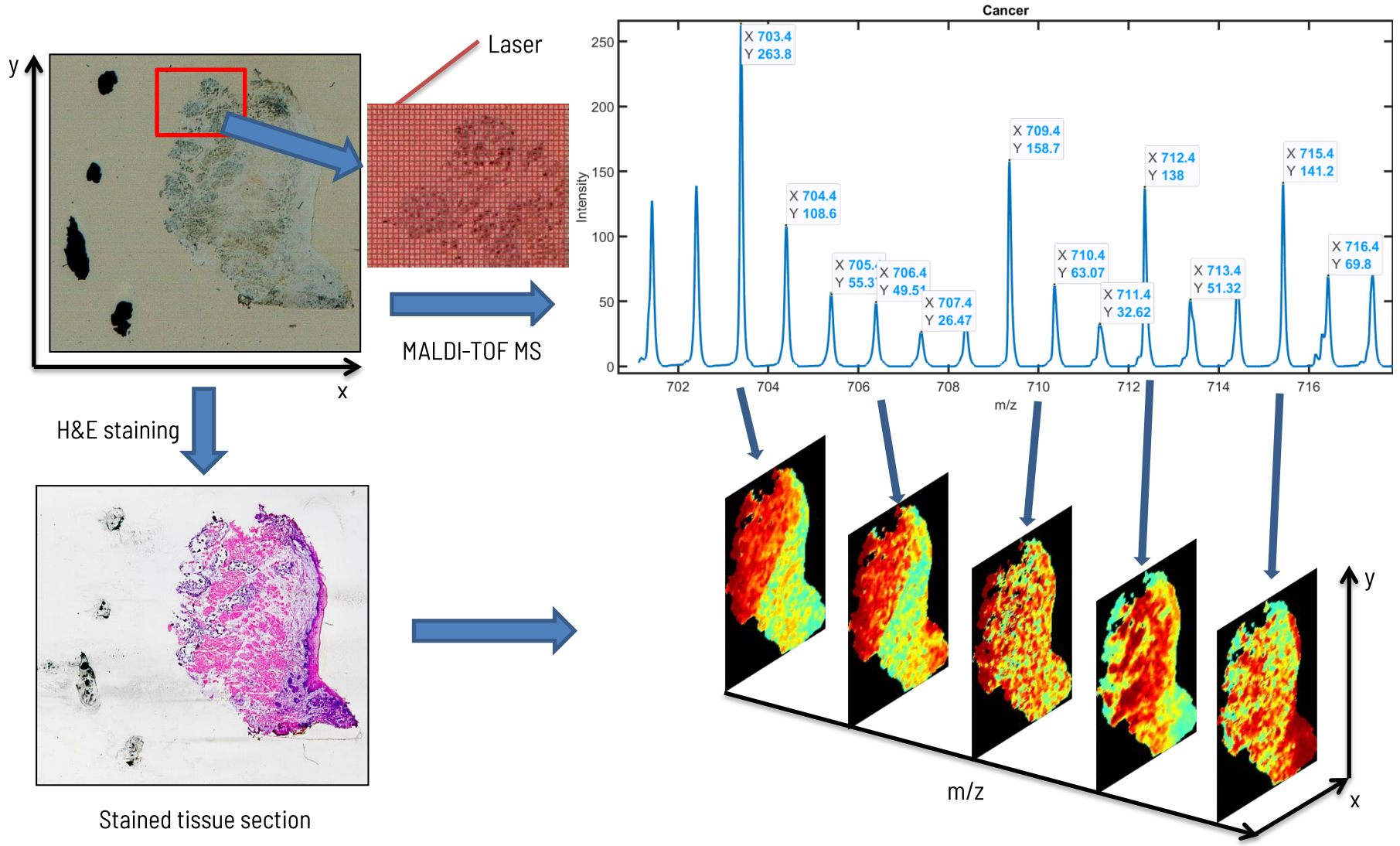

Figure 6. MALDI Imaging Mass Spectrometry workflow based on [22]. The tissue section image is from the head and neck cancer data published in [70], courtesy of Maria Skłodowska-Curie National Research Institute of Oncology, Gliwice Branch.

Imaging of proteins provides the ability to gather information about the distribution of proteins, protein isoforms, modifications and posttranslational modifications, degradation or cleavage [22].

2.5. Isotopic envelope

The vast majority of elements (carbon, hydrogen, oxygen, nitrogen, phosphorus, sulphur) create different isotopic forms in nature [49]. Mass distributions of the isotopically complex molecules result from the contributions from different isotopic combinations [50]. The same chemical properties characterise isotopes as their parental element, but they differ in the number of neutrons, which results in the difference in mass [2]. As a result, in the mass spectrum the isotopic envelope (isotopic pattern of peaks, also called *isotopomer envelope* [51]) is observed [49], which is a result of the presence of natural isotopes in the sample [40]. Peaks that are members of an isotopic envelope are composed of chemically similar compounds, which differ in the weight of particular isotopes [51]. Peak patterns are observed in intact proteins, digested proteins (peptides, for instance (*Figure 7*)), metabolites, and tandem mass spectra (MS/MS) of proteins, peptides, or metabolites [49].

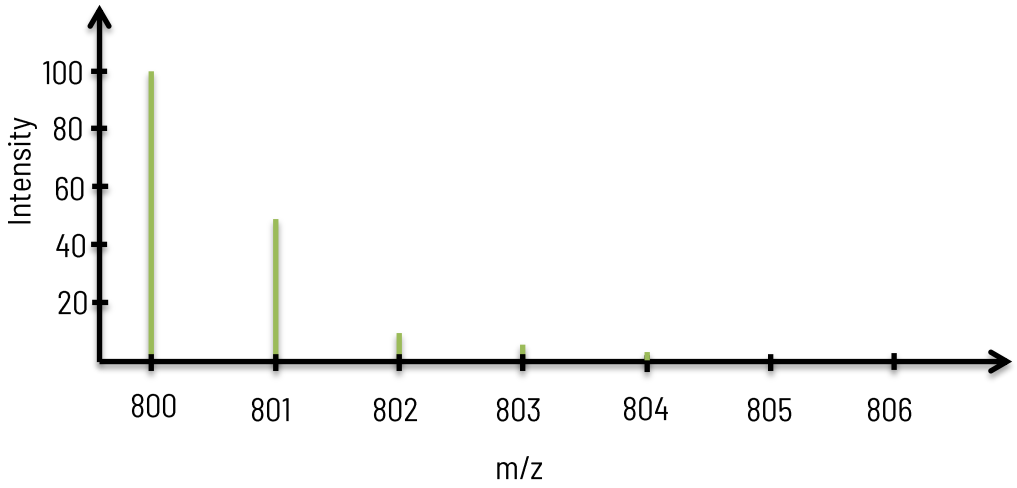

Figure 7. Theoretical isotopic envelope of the peptide YDLDFK.

A single mass spectrum consists of mass-to-charge ratios of ions (m/z) and corresponding intensities (abundance) values [40][49]. Such a spectrum „consists of patterns of isotopic peak distributions for many different peptides, each with its charge and intensity" [49]. The isotopic distribution of the elements the peaks are composed of determines the spacing between observed peaks and also their relative heights [49]. Also, the operational aspects of the instrument, such as resolution, type of detector, etc., impact the distance between adjacent

peaks and their heights [49]. Isotopic distributions of peptides are dominated by carbon (it has the largest proportion of naturally abundant isotope to any other one). For MALDI, for a singly charged peptide, the average mass difference between peptides peaks is $1.003 \ / \ 1 = 1.003$ Da (the difference between the masses of 13-C and 12-C) (*Figure 8*). Peptides with a higher charge generate peaks with spacing ~1.003/charge. [49]

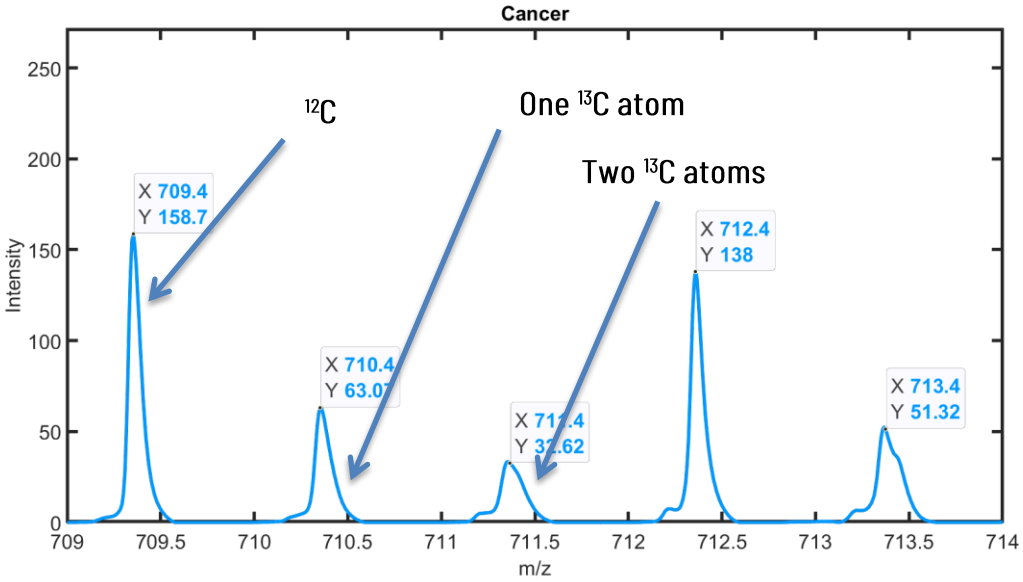

Figure 8. Isotopic envelope consisted of three peaks.

There are different isotopic combinations of elements, which are not resolvable by a mass spectrometer [49]. Therefore, there is an urgent need of automated interpretation of obtained mass spectra during the experiment [49]. In order to resolve multi-charged spectra and remove isotopic envelopes, decharging and deisotoping are applied. Decharging and deisotoping are together called spectra deconvolution [18]. Spectral deconvolution means grouping into isotopic envelopes [51]. As a result, the monoisotopic mass and the charge state of each isotopic envelope can be effectively determined [51]. The main goal of deisotoping is to collapse a complex mass spectrum into a representative set of peptide or metabolite masses, which in most cases means a monoisotopic mass and their abundance values, respectively [49]. The monoisotopic mass is the total abundance of a peptide: a summation of the intensities of all the isotopomers [52] or sum of the masses of the atoms using the most abundant (principal) isotope for each element [51].

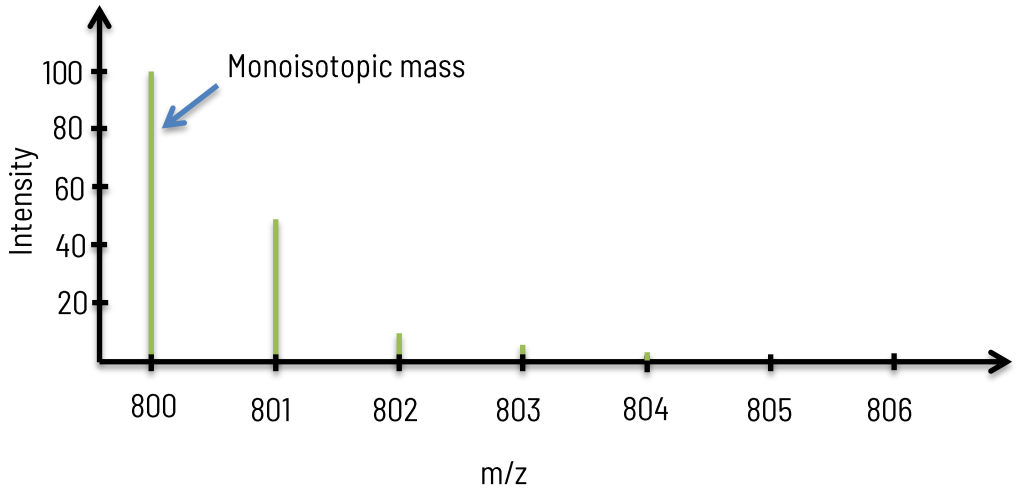

Figure 9. Isotopic envelope of the peptide YDLDFK, generated using *Compass IsotopePattern* by Bruker Daltonics.

For a small peptide, the most intense peak is the first one, but it is not the case for larger proteins – simultaneously with the mass range, the peak that is the most intense one, changes.

For instance, in the mass range ~800 to ~1800 Da, the abundance of the first peak is 100.00, and this is the monoisotopic mass (*Figure 9*), (*Table 2*).

Table 2. YDLDFK peptide m/z values and corresponding abundance.

m/z	Abundance
800.382497	100.000
801.385540	44.735
802.388205	12.245
803.390798	2.494
804.393326	0.414

Beginning from the mass range ~1900 Da, the second or even subsequent peaks are the most intense ones, with the highest abundance (*Figure 10*), (*Table 3*).

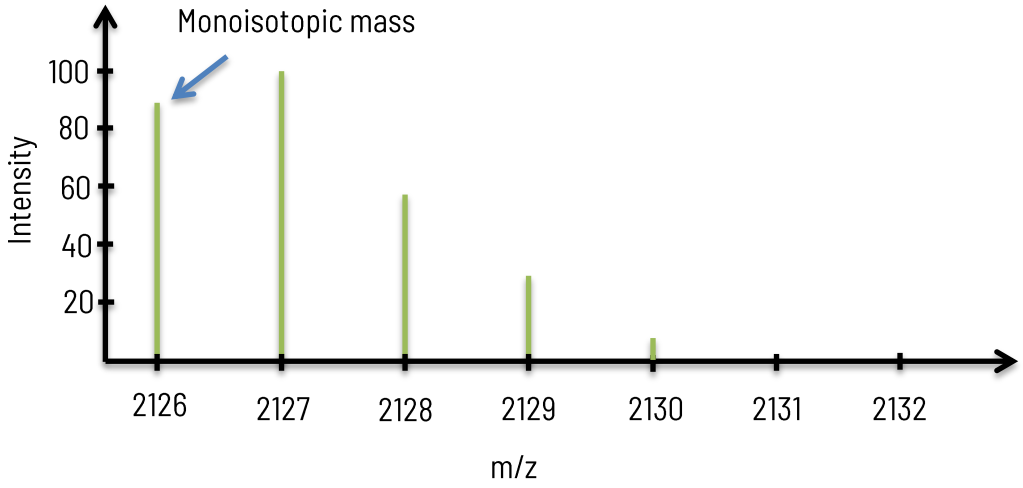

Figure 10. Isotopic envelope of the peptide ALPGQLKPFETLLSQNQGGK, generated using *Compass IsotopePattern* by Bruker Daltonics.

Table 3. ALPGQLKPFETLLSQNQGGK peptide m/z values and corresponding abundance.

m/z	Abundance
2126.1604	86.837
2127.1633	100.000
2128.1660	62.256
2129.1687	27.490
2130.1713	9.587

There are different types of isotopic envelopes that can be distinguished in a mass spectrum [53]:

1) **overlapping**

   Overlapping isotopic patterns can be observed from proteomics and metabolomics samples [49], because many isotopic peaks are observed in a narrow m/z range [54] (*Figure 11*). It is also possible that overlapping isotopic envelopes have a shared peak.

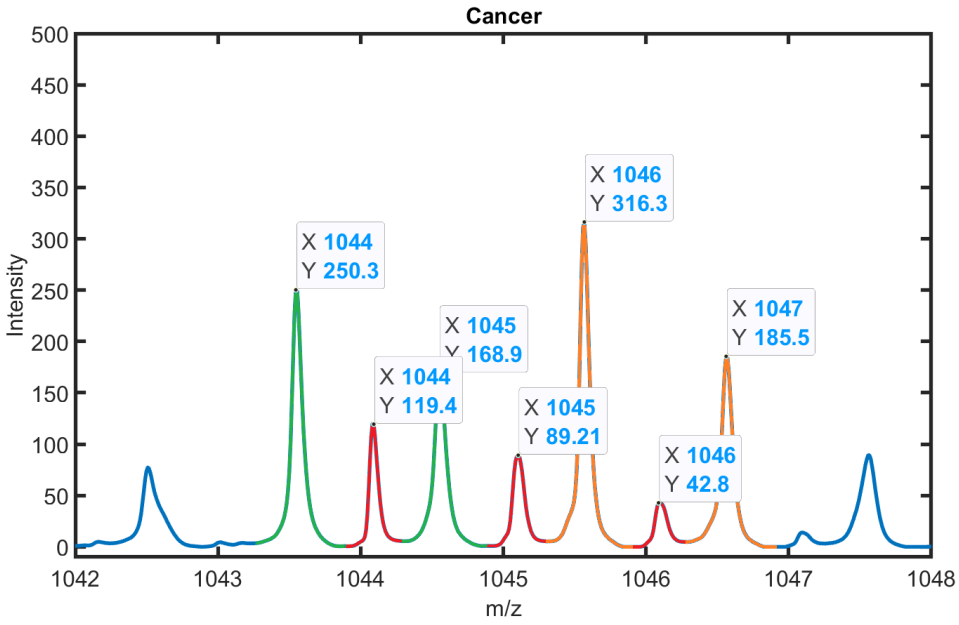

Figure 11. Overlapping isotopic envelopes in FFPE peptides data (green – first isotopic envelope, red – second, orange – the third one).

2) **non-overlapping** (*Figure 12*).

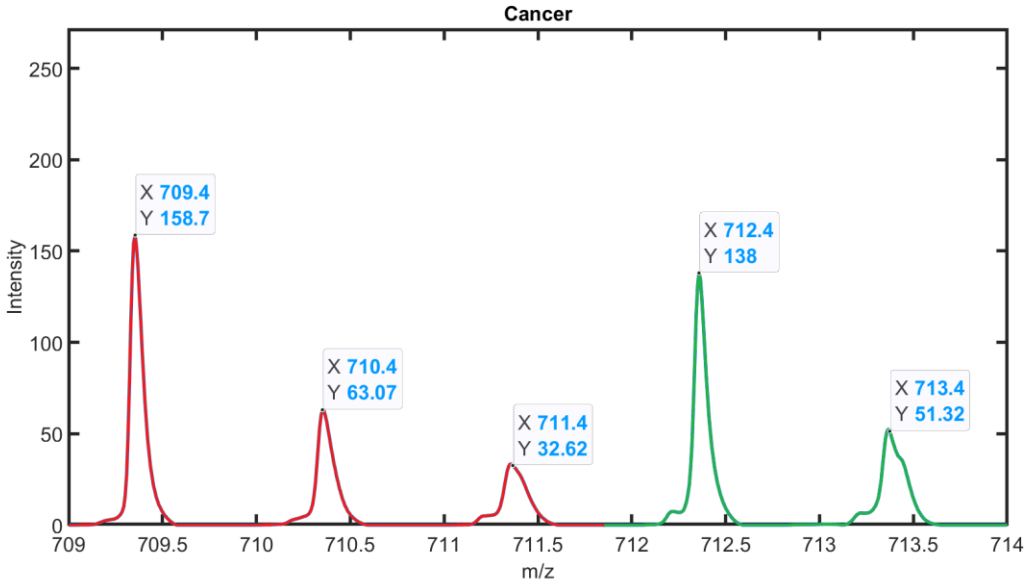

Figure 12. Non-overlapping isotopic envelopes in FFPE peptides data (red – first isotopic envelope, green – the second one).

2.6. Algorithms for isotopic envelope identification

Hereby, algorithms and methods used for deisotoping are described.

Envelope selection problems focus on selecting these isotopic envelopes that explain the spectrum most accurately – the most optimal ones, as isotopic envelopes often overlap and share peaks [51] [53]. A plethora of deisotoping algorithms have been described in the literature, which are described below.

There are two main approaches [53]:

- Graph theory
- Based on matching the theoretical isotopic distribution with the experimental.

The major disadvantage of the approach based on matching the theoretical and experimental isotopic envelope is that in the case of overlapping isotopic envelopes, relying solely on intensities of the theoretical and experimental isotopic envelopes is insufficient [53] [55]. Other algorithms that rely only on the information on the spectrum intensity are as follows: *BPDA* [56] and method based on LASSO [60]. Algorithm based on quadratic programming requires the knowledge of the parameters that need to be optimized [55] [57]. The aforementioned statistical approach decreases the false positives and false negatives because of selecting the simplest model with the least number of isotopic envelopes [53] [55] [58] . In order to accurately analyse complex mass spectra, the overlapping isotopic envelopes should be taken into consideration, but some existing algorithms do not take it into account [53], such as: probabilistic classifier [59], *Decon2LS* [60].

1. Graph theory-based
    a) Features-Based Deisotoping Method for Tandem Mass Spectra: the algorithm is based on constructing isotopic-cluster graphs. Firstly, it searches for possible isotopic clusters, where a space between pairs of adjacent isotopic peaks is $1.003/z$ ($z = 1, 2, 3$), then the isotopic-cluster graphs are constructed. Possible isotopic clusters are scored on the basis of four non-intensity features, such as losing a water or ammonia molecule by side chains of some amino acids residues of fragment ions. [55]
    b) MS-Deconv is a combinatorial algorithm for spectral deconvolution based on graph theory. Firstly, a large set of isotopic envelopes is generated and represented as a graph. Afterwards, envelopes that have the highest score (a heaviest path in the graph) are selected. [51]

c) MS2-Deisotoper identifies isotopic clusters consisting of two or more peaks. The cluster is based on comparing the mass and relative intensity of each peak to every other higher peak in the MS/MS spectrum. This method is solely dedicated to deisotoping high-resolution, centroided MS/MS spectra. [61]

2. **Based on comparing theoretical isotopic distribution to the experimental**

Envelope detection problem is based on using theoretical isotopic distributions to detect and evaluate potential isotopic envelopes. This problem has been well studied and a plethora of various metrics for evaluating the candidate isotopic envelope to its theoretical isotopic distribution has been proposed. [51]

Some of the algorithms are based on matching theoretical isotopic distribution with the experimental one [53]. If the matching of these two distributions is good, then these peaks are considered as an isotopic cluster. [55]

Knowing the average molecular mass, the monoisotopic mass of a molecule can be estimated [1]. For more accurate monoisotopic mass assignment, the *Averagine* method ($C_{4.9384}N_{1.3577}O_{1.4773}S_{0.0417}H_{7.7583}$) was introduced. It is based on comparing high resolution spectra with model isotopic distributions. Measured isotopic distribution is compared with the distribution for a model molecule of the same average molecular mass, based on a statistical test. It results in determining the monoisotopic mass. [53] [62]

a) THRASH

It is one of the widely used and cited algorithms. Possible isotopic clusters are found in the spectrum by using a subtractive peak finding routine. It determines the charge using Fourier transform / Patterson method and compares a peak cluster with an A*veragine* isotopic distribution by applying the least-squares method [60].

b) Decon2LS

Decon2LS algorithm is based on comparing the molecular formula generating by *Averagine* and the theoretical profile generated by Mercury. The authors stated that the fitness score is usually poor when overlapping isotopic envelopes occur [49]. It is based on THRASH algorithm [60].

c) OpenMS

It is the Python library *pyOpenMS.* It groups peaks of the same isotopic envelope charge state and uses a theoretical isotope pattern to find members of an isotopic envelope [63].

d) NITPICK

It is the *R* package. This approach Is an extension for the well-known averaging model [64].

e) DeconMSn deisotopes the high mass measurement accuracy precursor spectrum by comparing the theoretical profile with the observed one. Additionally, SVM-based (Support Vector Machine) charge detection is implemented in order to determine parent mass for low-resolution data (LCQ / LTQ) [65].

f) FLASHDeconv is dedicated to top-down mass spectrometry (TD-MS)-based proteomics studies. It transforms m/z within spectra into log m/z [18]. Deisotoping is based on „finding theoretical isotope patterns (derived from the *Averagine* model [62]) around the charge-determined peaks" [18].

g) mMass searches for isotopic envelopes by comparing an intensity of every isotope with its theoretical value. Due to the fact that it is mostly used for proteomic data, theoretical isotopic patterns are approximated using *the Averagine* [62] [66].

h) iMEF and ProteinGoggle 2.0 allow to determine an isotopic mass-to-charge ratio and experimental isotopic envelope fingerprinting by fingerprinting to the values from the corresponding theoretical isotopic envelope based on computing elemental composition of the product ions' amino acids [54].

i) RAPID presents a probabilistic model of an isotopic distribution, which is based on calculating intensity ratios of the adjacent peaks, which are then approximated as linear functions of peptide mass values, using the theoretical distributions of tryptic peptides [20].

3. **Another existing approaches for selecting a set of isotopic envelopes:**
    a) Probabilistic classifier with dynamic programming algorithm with a condition that the envelopes do not overlap. Dynamic programming has been also employed in order to predict the probabilities of potential isotopic distributions, taking into consideration length, shape, inter- and intradistribution distances. [59]
    b) Statistical approach based on variable selection based on non-negative sparse regression scheme [58]
    c) LASSO method for solving the statistical problem of variable selection [67]
    d) Quadratic programming: an approach called *Pepex* is based on spectra modelling by a linear mixture model [57].
    e) Approximation of isotopic patterns by a Poisson distribution [68]
    f) Bayesian approach for deisotoping and simultaneous deconvolution of mass spectra: *BPDA* (a Bayesian peptide detection algorithm) [56]
    g) Xtract is a top-down approach that includes also tandem mass spectra for protein identification, which automatically combines isotopic peaks to one monoisotopic peak mass [69].
    h) Zscore – algorithm for isotopic cluster identification based on a charge-scoring scheme [70].
    i) Isotopica is used for calculating and visualising isotopic distributions on the basis of molecular formulas, peptides or proteins, DNA or RNA or carbohydrate sequences using the Fast Fourier Transform [40].
    j) An algorithm dedicated to overlapping isotopic envelope identification, published in [54], that is not based on *Averagine* units that create the theoretical isotopic envelope was developed. It is based on computing isotopic envelopes "from the elemental composition of the product ions' actual amino acids" [53].

Table 4. Comparison of commonly used deisotoping algorithms [53].

Name	Mass spectrometry technique/proteomic strategy	Biomolecules
DeconMSn	LC-MS/MS	peptides
Decon2LS	LC-MS/MS	proteins, peptides, metabolites
FLASHDeconv	Top-down proteomics	proteins, peptides
MS-Deconv	Tandem mass spectra, Top-down proteomics	proteins, peptides
MS2-Deisotoper	High resolution bottom-up spectra, MS/MS	peptides
RAPID	LC-MS/MS	-
BPDA	MALDI-TOF-MS, LC-MS	peptides
Features-Based Deisotoping Method	Tandem mass spectra, bottom-up spectra	proteins, peptides
THRASH	MALDI-MS, ESMS	proteins, peptides, DNA, polymers
pyOpenMS	LC-MS/MS	proteins, peptides, metabolites
NITPICK	not limited	not limited
mMass	MALDI-TOF-MS, all others	proteins, peptides
iMEF & ProteinGoggle 2.0	Tandem mass spectra	proteins, peptides
Xtract	Top-down proteomics	proteins
Zscore	ESI-MS	proteins

According to *Table 4*, there is no standardisation across different instruments and experiments. There are only a few methods dedicated to MALDI-TOF-MS data. MALDI MSI experiments datasets are large (even 40 GB), which requires the methods to be efficient. The dataset can be comprised of over 608 000 spectra, ~108 000 m/z mass channels, and corresponding intensity values. Hence, there is a need to create a method that will be able to handle the such size of datasets.

3. MATERIALS

3.1. Data characteristics

The studies were carried out using four fresh frozen (FF) tissue datasets and four formalin-fixed paraffin-embedded (FFPE) tissue datasets. The workflow for handling both types of tissues mentioned above is presented in *Figure 13*. Both datasets were collected in Maria Skłodowska-Curie National Research Institute of Oncology Gliwice Branch, Poland, from patients who suffered from head and neck cancer (HNC) and consisted of peptide mass spectra acquired in MALDI-TOF MSI experiments [53].

HNC-FF

This dataset was collected for fresh frozen tissues in Maria Skłodowska-Curie National Research Institute of Oncology Gliwice Branch, Poland, and published in [71]. Four oral cavity squamous cell carcinoma patients (males with tumour located in the tongue and on the floor of the mouth) were involved in the study at age 36-59 years old. After surgical resection, the tissue specimens were frozen and stored at -80 °C, and then each sample was cut into a 10 µm section in a cryostat and placed onto ITO glass slide. One consecutive section was H&E stained for histopathological evaluation by an experienced pathologist. As a next step of sample preparation for the MALDI-MSI experiment, the sections were dried, and washed twice in 70% ethanol and once in 100% ethanol and dried again. Subsequently, the samples were coated with a trypsin solution, incubated, and coated with a matrix. For spectra analysis MALDI-TOF *ultrafleXtreme* mass spectrometer (Bruker Daltonik, Bremen, Germany) was used. The process of spectra acquisition was performed in positive reflectron mode within the mass range 800-4000 m/z. From each ablation point, 400 spectra were collected, and a 100 µm raster width was applied. After analysis in the mass spectrometer, the matrix was removed from the slides, H&E staining was done, and the slides were scanned for co-registration with MALDI images (*flexImaging 4.1* software, Bruker Daltonik, Bremen, Germany). As a result, the dataset consisting of 45 738 raw spectra with 109 568 mass channels was obtained. The acquired spectral data were pre-processed and the features were extracted by performing the following steps: spectrum resampling, baseline removal, TIC (total ion current) normalisation, alignment based on Fast Fourier Transform, and spectra modelling and peaks detection by using the Gaussian mixture model (GMM) approach. Then, pairwise convolution of the GMM components and individual spectra was done in order to estimate the peptide

abundance. After that, neighbouring peaks (a result of the right skewness of spectral peaks) were identified and merged (summation of their estimated abundance values). All of these steps resulted in the reduction of dimensionality to 3 714 Gaussian components. Such a data set was used for further analysis. [53] [71]

HNC-FFPE

The analysed tissue material comprised four patients suffering from oral squamous cell carcinoma (located on the tongue and the floor of the mouth). After surgical resection, the collected material was stored as FFPE tissue blocks. Then, the blocks were sectioned using a rotary microtome and placed on a conductive glass slide. The slides were dried (56 °C, 1h) and then stored at room temperature. Afterward, the tissue sections were heated (60 °C, 30 minutes), paraffin was removed from ITO slides, boiled (for reversal of protein crosslinking), and dried. Next, the trypsin was deposited onto a section, and the section was placed in a humid chamber (solution: 100 mM NH_4HCO_3M, 5% MeOH, 37 °C, 18h). After that, the section was coated with a matrix solution. MALDI MSI measurement was performed using the *ultrafleXtreme* MALDI-TOF mass spectrometer (Bruker Daltonik, Bremen, Germany) operated in positive reflectron mode. Mass spectra were recorded in the 700 – 3000 m/z range (raster width = 100 μm). The final datasets comprised 22 389, 22 267, 21 395, and 31 654 raw spectra with 200 704 mass channels, respectively. The spectra underwent the pre-processing and feature extraction pipeline that consisted of mass channels unification, baseline removal, spectra identification, peak alignment, TIC normalisation, and peak detection using GMM (Gaussian Mixture modelling of the average spectrum). The aforementioned steps resulted in a reduction to 1 776, 1 766, 1 697, 2 510 components (tryptic peptide species), respectively. [53] [72]

Data was collected in Maria Skłodowska-Curie National Research Institute of Oncology Gliwice Branch, Poland, and published in [72]. MSI components were identified by assigning a component location at m/z for the measured masses of tryptic peptides identified by LC-MS/MS. The +/- 0.05% mass tolerance was allowed. [72]

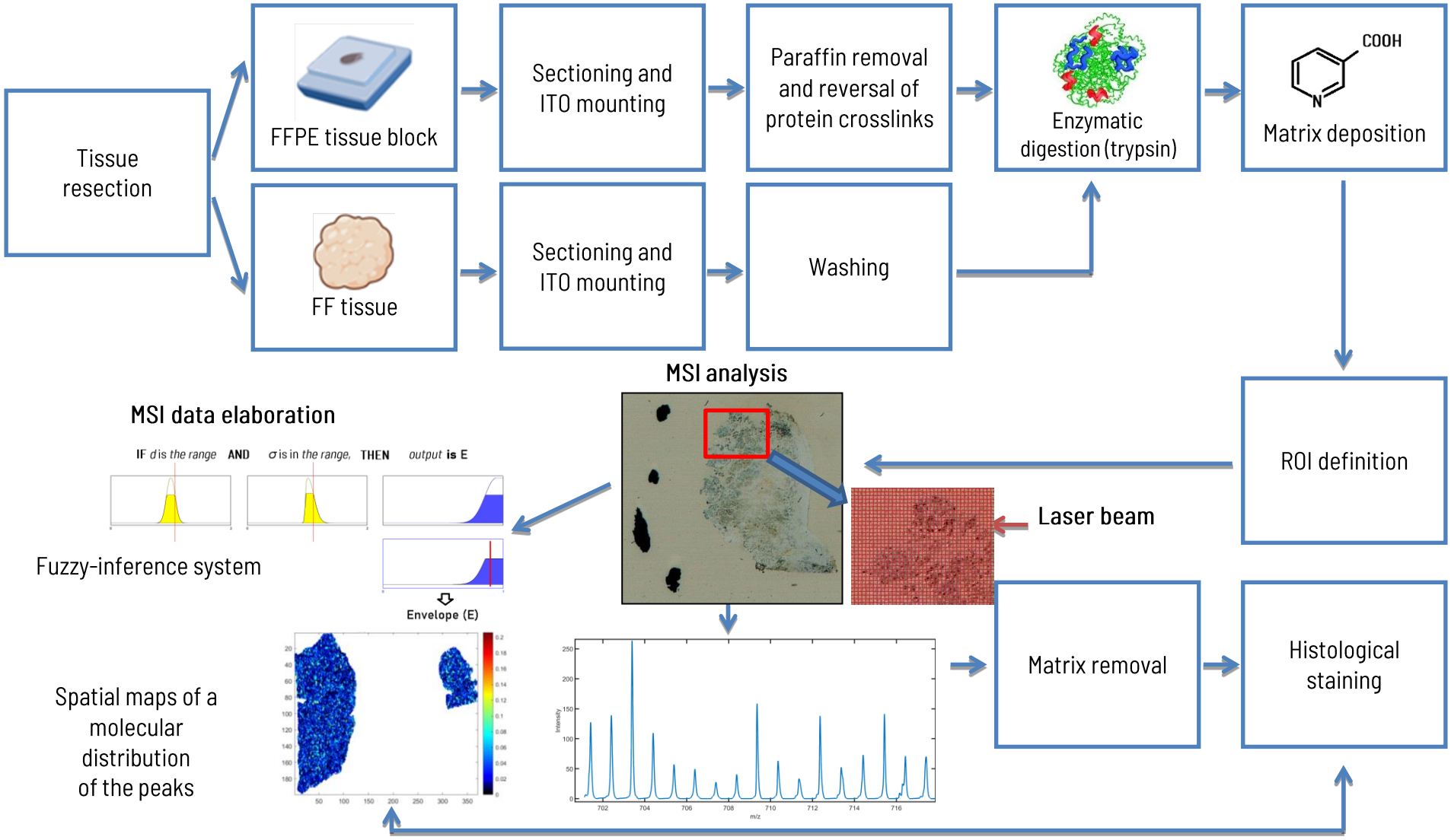

Figure 13. Workflow of MSI experiment for FFPE tissue block and FF tissue based on [45][53].

3.2. Data pre-processing methods

Determination of peptide mass fingerprint (experimental peptide masses) is a formidable challenge since distinguishing spectra peaks corresponding to digested peptides from the associated isotopomer peaks and spurious peaks (that come from sample contamination and noise) require complex pre-processing of the raw mass spectrum. [34]

The following pre-processing steps were employed [53] [59] [60] [73]:

1. Unification of mass channels (resampling)
2. Noise reduction with the use of the Savitzky-Golay filter
3. Adaptive baseline correction
4. Outlying spectra identification: outlying means that spectra are characterised by too small or too big TIC (total ion current) value; for such spectra, the Bruffaerts' criterion is used in order to cope with highly skewed distributions
5. Spectra alignment with the use of Fast Fourier Transform
6. Normalisation to the mean TIC.

As a result, the average spectrum is created (a mean intensity signal of all samples) [74].

3.3. Feature extraction

1. Average spectrum modelling and peak detection with the use of the Gaussian Mixture Modelling approach (GMM) [26] [53] [61] [62]

Signals from spectra registered by the mass spectrometer in mixtures of proteins or peptides are considered spectral peaks that reflect a specific protein or peptide species. In further analysis of proteomic mass profiles, they are used as features of the MS spectra [26].

Spectral signals can be modelled by mixtures of Gaussian distribution component functions, such as the univariate Gaussian mixture probability density function *(Eq. 2)*. The components of mixture models of MS spectra are characterised by position and shape (width) [26].

$$f(x_n) = \sum_{k=1}^{K} \alpha_k f_k(x_n, \mu_k, \sigma_k), \quad (2)$$

where:

x_n – m/z values of the registered ions,

K – the number of Gaussian components,

α_k – Gaussian component,

$k = 1, 2, ..., K$ – component weights that sum up to 1,

μ_k – means of the Gaussian components,

σ_k - standard deviations of the Gaussian components.

The probability density function of the Gaussian distribution is denoted by *(Eq. 3)* [26]:

$$f_k(x_k, \mu_k, \sigma_k) = \frac{1}{\sigma_k\sqrt{2\pi}} e^{-\frac{(x_n-\mu_k)^2}{2\sigma_k^2}} \quad (3)$$

The mixture model is scaled, and after that, the obtained scaled mixture model is fitted to spectral data by using expectation maximization (EM algorithm) on previously decomposed MS signal to smaller fragments [26].

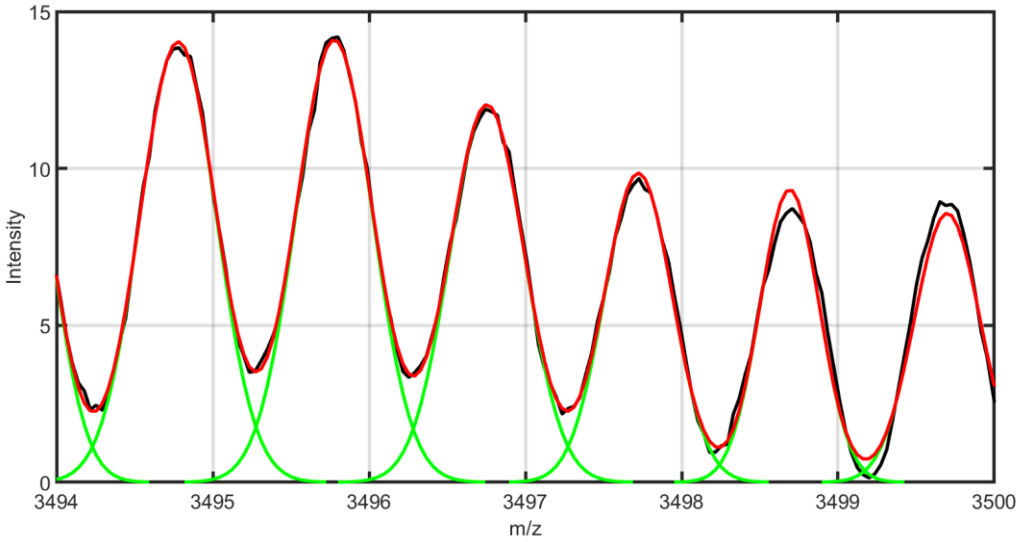

Figure 14. Zoomed out fragment of 3494-3500 Da mean spectrum from the proteomic dataset of head and neck cancer tissues: Fragment of the MS signal **(black)**, GMM of the signal (red), components of the Gaussian mixture model (green).

Finally, all the GMM components are aggregated into one set – a "whole-spectrum mixture model of the MS signal" [26]. Computations in that work are based on the components of the Gaussian mixture model, marked in green in *Figure 14* [26]. A constructed mixture model of the average

spectrum consists of components with defined location, spread, and weight parameters [74]. Each component "represents measured protein/peptide concentration in an analysed spot on the tissue" [74].

2. Spectrum filtering by removal of GMM components with high variance and/or low amplitude [53] [73]

3. Modelling the right-skewed spectrum peaks by GMM components and merging with the left-neighbouring major component; the estimated abundance of the same spectrum peaks is merged, and the dominant component has the location of a peptide ion m/z value [53] [73].

4. Pairwise convolution of GMM components and individual spectra to calculate peak intensity (abundance) [53] [73].

Every Gaussian component is defined by three parameters: location of a component, spread of a component, and a weight of the component [74]. In mass spectrometry, the term *peak* is used when referring to the mass spectrum. Thus, in subsequent chapters of the dissertation, the term peaks will be used to reflect the model Gaussian components since further analysis steps take into account the aforementioned model components parameters and calculations are performed based on those components parameters. [53]

3.4. Final data structure and general workflow

In order to measure the spatial distribution of peptides in a sample, MSI is applied. For every tissue coordinate a separate spectrum is acquired. The raster width (lateral resolution) of obtained MS images is 100 μm. The predominant charge of MALDI ions is the single one [16]. Thus, there is no need to apply deconvolution methods for the data.

In order to define which peaks are members of an isotopic envelope, the two-step algorithm has been proposed (*Figure 15*):

1. Potential isotopic envelope member peaks are identified using the Mamdani-Assilan fuzzy-inference system. For every peak pair, the possibility value of being an isotopic envelope member is assigned. Peak pairs with a possibility value greater than 89.66% are annotated as potential isotopic envelope members.
2. Then, the molecular spatial maps of every peak are created, and based on them, several descriptors related to the parameters of Gaussian components, image texture, and structure are calculated, which are used in the classification process. Naïve Bayes classifier denotes the peaks

as the potential isotopic envelope members (*E*) or non-included in isotopic envelopes (*nE*).

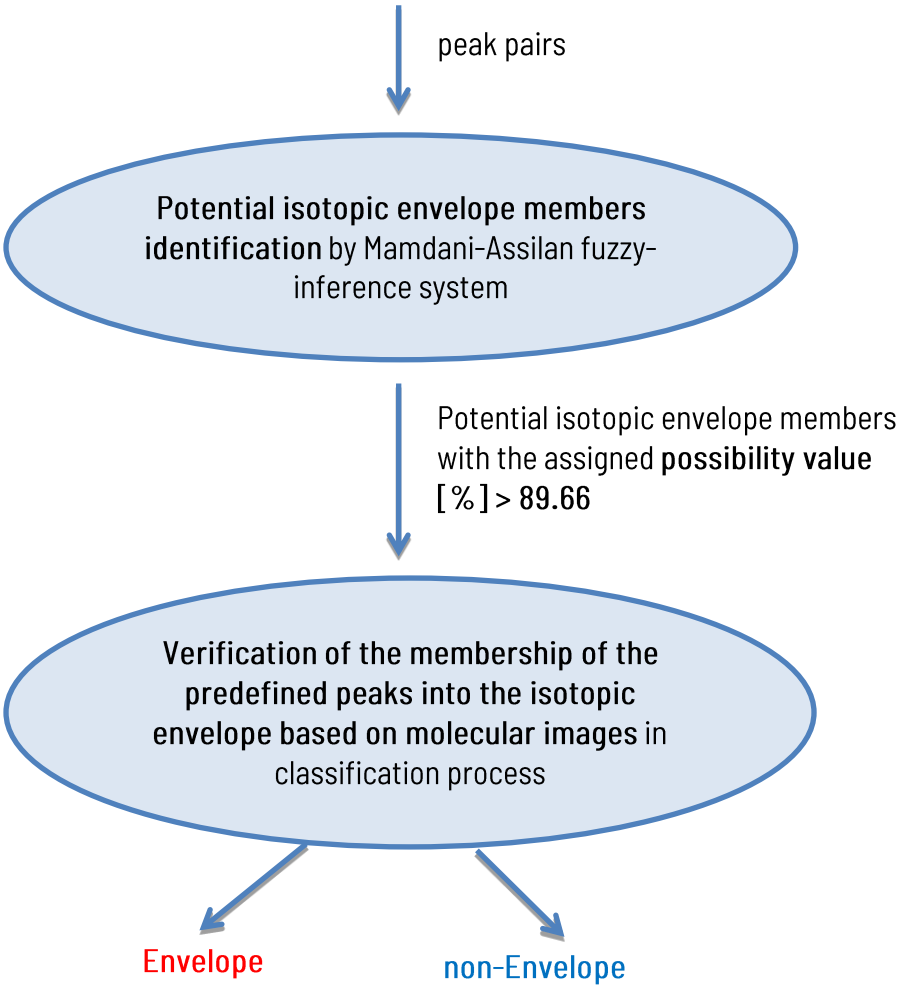

Figure 15. General workflow of the algorithm [53].

4. ISOTOPIC ENVELOPE IDENTIFICATION IN MALDI-TOF MOLECULAR IMAGING DATA

4.1. Fuzzy-inference systems

The term fuzzy set was introduced in 1965 by Lotfi Zadeh [75]. The idea of fuzzy sets arose as a generalisation of crisp sets.

One of the first use of fuzzy sets was fuzzy data grouping, described by Ruspini in 1969 [76] [77]. In 1973 L. A. Zadeh introduced the terms linguistic variable and fuzzy IF-THEN *rules* to represent human knowledge [78]. Based on this work, in 1975, Mamdani and Assilan designed a fuzzy regulator which controls a steam engine. Hence, this work enables the first practical usage of fuzzy sets theory. [76]

Fuzzy sets have become more and more popular over the years since they allow the expression of non-precise terms (not specific knowledge) in a formal way. They are applied in a plethora of different fields, especially in medicine, automation control, economics, information technology, and forensics. [79]

A fuzzy set A in the space $\mathbb{X}$ can be described directly either by the function $\mu A\,(x)$ or by the set of ordered pairs $(x, \mu_A\,(x))$, where $\mu_A\,(x)$ represents the degree (level) of membership of object x to the fuzzy set A [76] [79] *(Eq. 4)*:

$$\mathrm{A} = \{\,(\mathrm{x}, \mu_{\mathrm{A}}\,(\mathrm{x}))|\ \mathrm{x}\ \ [0, 1]\}. \qquad \textbf{(4)}$$

An element belongs to the crisp set (value 1) or not (value 0), whereas fuzzy sets are characterised by partial belonging to the set by an element [79].

An element $x \epsilon X$ can be a member of the fuzzy set A in the following ways [79]:

1. not included: $\mu_{\mathrm{A}}\,(\mathrm{x}) = 0$
2. partially included: $0 < \mu_{\mathrm{A}}\,(\mathrm{x}) < 1$
3. fully included: $\mu_{\mathrm{A}}\,(\mathrm{x}) = 1.$

Mutual relations between crisp and fuzzy sets are presented in (*Figure 16*):

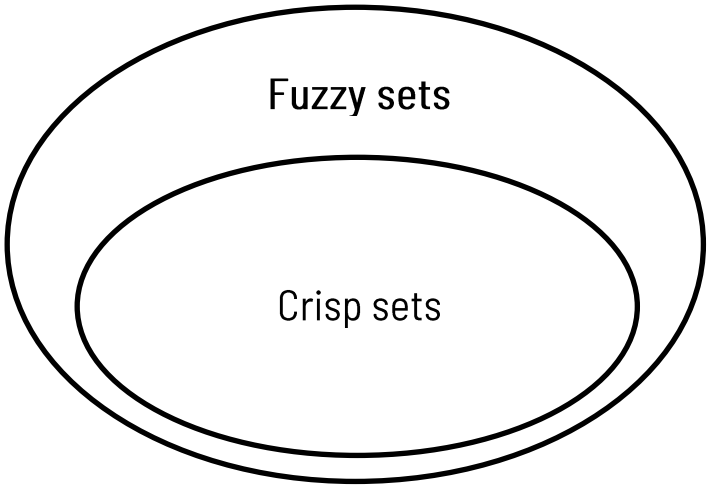

Figure 16. Relationship between fuzzy sets and crisp sets **[79].**

„A fuzzy set is a class of objects with a continuum of grades of membership. Such a set is characterized by a membership (characteristic) function which assigns to each object a grade of membership ranging between zero and one". [75]

The membership function takes values from the interval <0, 1> [76] [75] *(Eq. 5)*:

$$\mu_A : \mathbb{X} \rightarrow [0, 1]. \qquad (5)$$

The values of a membership function can be interpreted as [76] [80]:

1. a grade of membership of the element x to the fuzzy set A
2. a grade of preference – the set A presents the set of more or less preferred objects, and $\mu_A(x)$ represents the intensity of the preference for the object
3. a grade of uncertainty - $\mu_A(x)$ describes the degree of reliability that the variable X will get the value x.

Each fuzzy set is described by its membership function [76].

The most common membership function classes are as follows [76]:

1. Triangular *(Eq. 6)* (*Figure 17*):

$$\mu A\,(x;\ a, b, c) = \begin{cases} 0, & x \le a, \\ \dfrac{x-a}{b-a}, & a < x \le b, \\ \dfrac{c-x}{c-b}, & b < x \le c, \\ 0, x > c \end{cases} \qquad (6)$$

where:

a, b, c are the parameters ($a \le b \le c$).

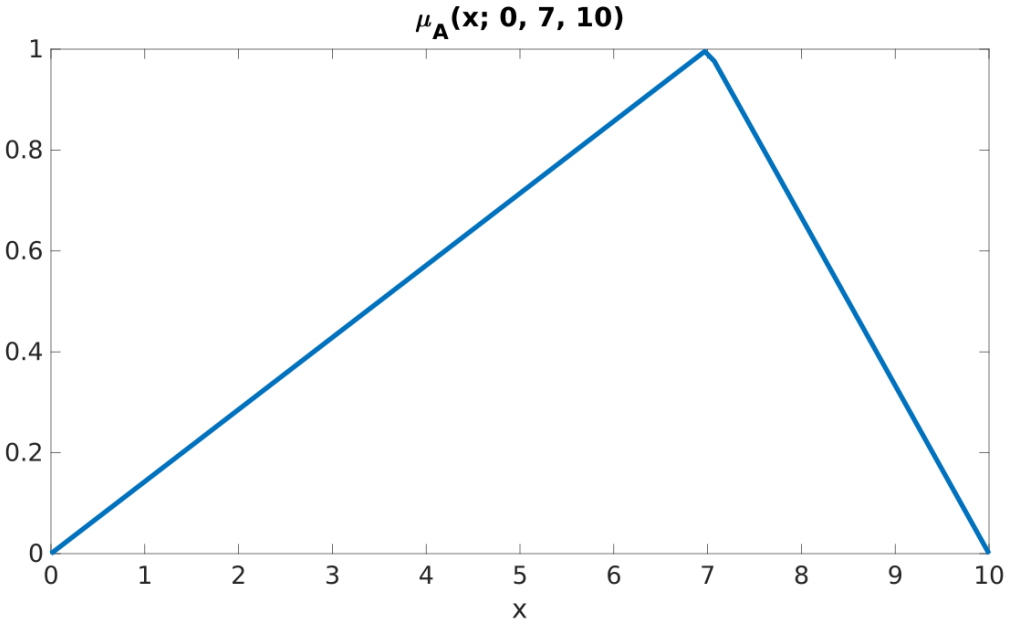

Figure 17. Example of triangular membership function.

2. Trapezoidal *(Eq. 7)* (*Figure 18*):

$$\mu_A\,(x;\ a, b, c, d) = \begin{cases} 0, & x \le a, \\ \dfrac{x-a}{b-a}, & a < x \le b, \\ 1, b < x \le c, \\ \dfrac{d-x}{d-c}, & c < x \le d, \\ 0, x > d \end{cases} \qquad (7)$$

where:

a, b, c, d are the parameters ($a \leq b \leq c \leq d$).

The rectangular membership function is obtained for $a = b$ and $c = d$.

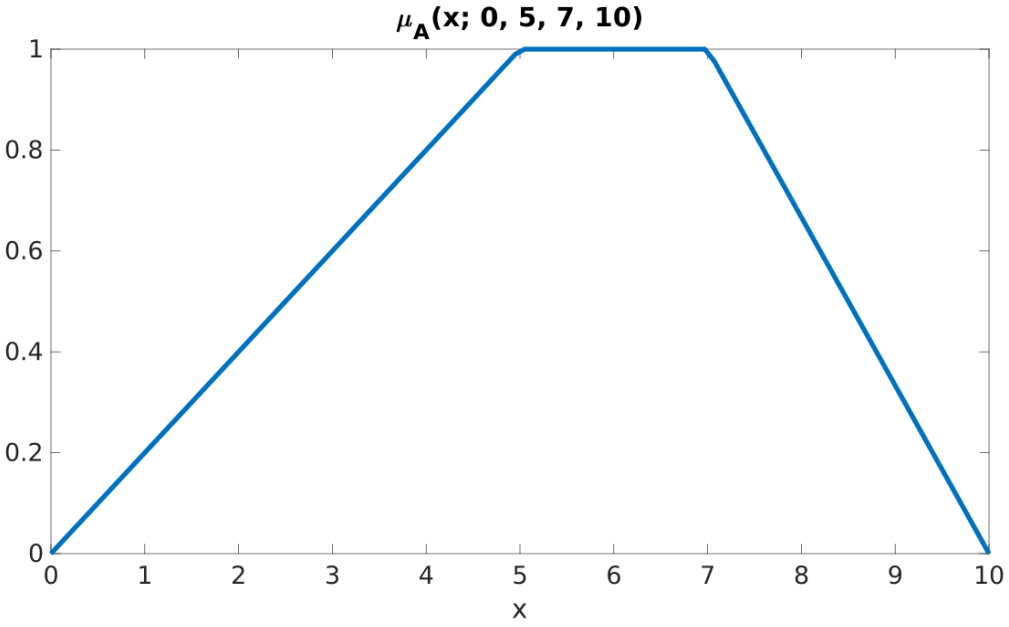

Figure 18. Example of trapezoidal membership function.

3. Gaussian *(Eq. 8)* (*Figure 19*):

$$\mu_A(x; m, \sigma) = e^{\frac{-(x-m)^2}{2\sigma^2}} \quad (8)$$

where m - mean, σ are the parameters.

$\sigma > 0$ determines the width of the fuzzy set.

The impact of σ on the membership function shape is presented in the following *Figure 19*:

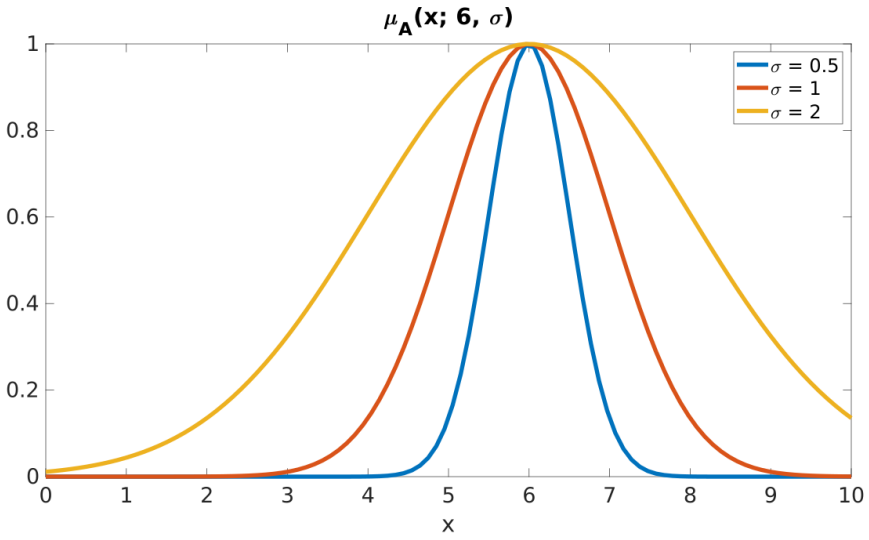

Figure 19. Example of Gaussian membership function.

4. Generalised bell shaped *(Eq. 9)*:

$$\mu_A(x; m, \sigma, \gamma) = \frac{1}{1 + \left|\frac{x-m}{\sigma}\right|^{2\gamma}} \quad (9)$$

where m, σ, γ are the parameters ($\sigma > 0$, $\gamma > 0$).

$\sigma > 0$ determines the width of the fuzzy set, whereas the parameters σ, γ influence the slope of its slopes.

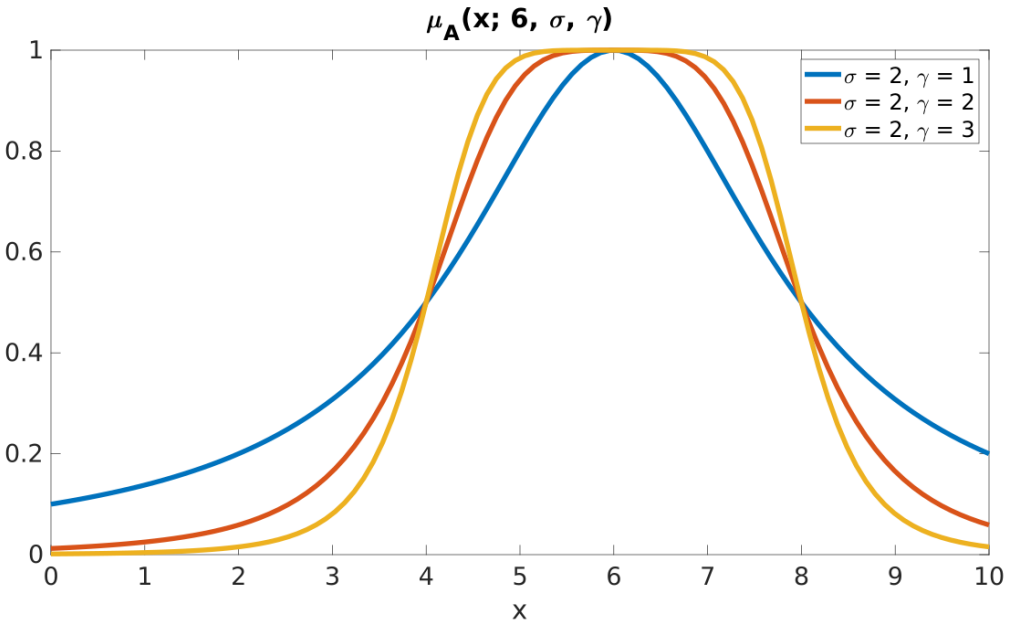

Figure 20. Example of Generalised bell shaped membership function with parameters $\sigma = 2$ and $\gamma = 1, 2, 3$.

The impact of the σ and γ values (for $\sigma = \gamma = a$) on the shape of the membership function is presented in the *Figure 20* and *Figure 21*:

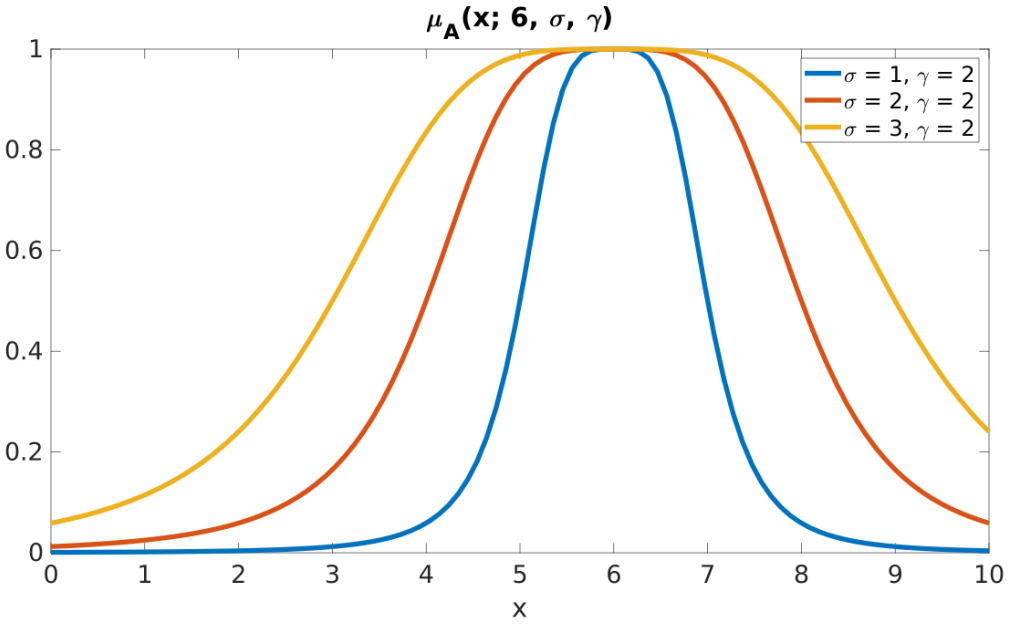

Figure 21. Example of Generalised bell shaped membership function with parameters $\sigma = 1, 2, 3$ and $\gamma = 2$.

5. Sigmoidal *(Eq. 10)*:

$$\mu_A(x; c, \beta) = \frac{1}{1+\exp[-\beta\,(x-c)]} \qquad (10)$$

where c, β are the parameters.

c defines the crossing point, whereas β affects the slope of the membership function.

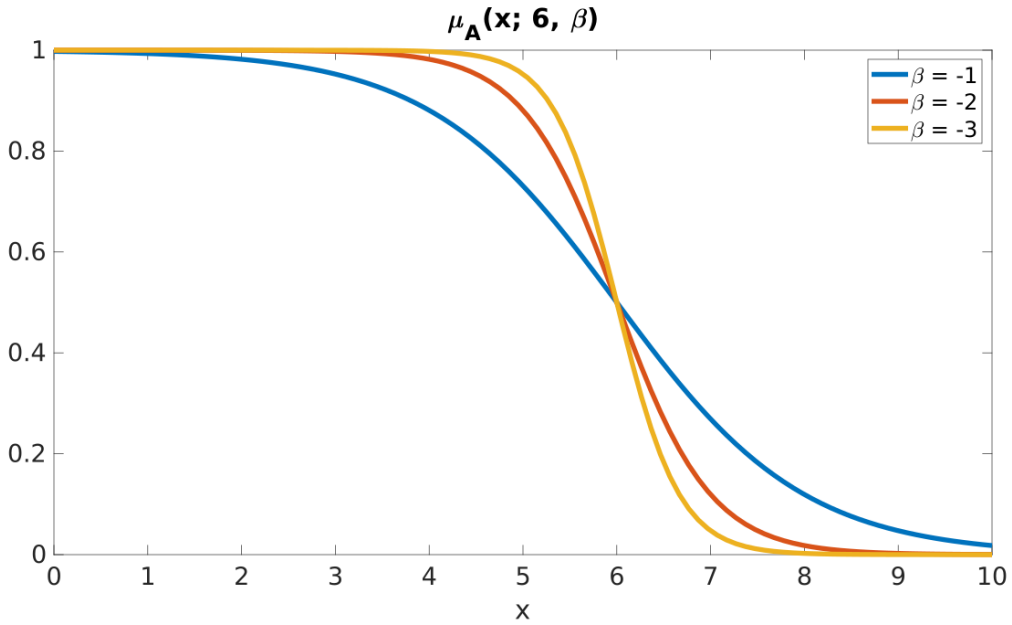

Figure 22. Example of sigmoidal membership function with positive β *values.*

Membership function shape dependence on the sign of the parameter value, and the value of β are presented in the *Figure 22* and *Figure 23*:

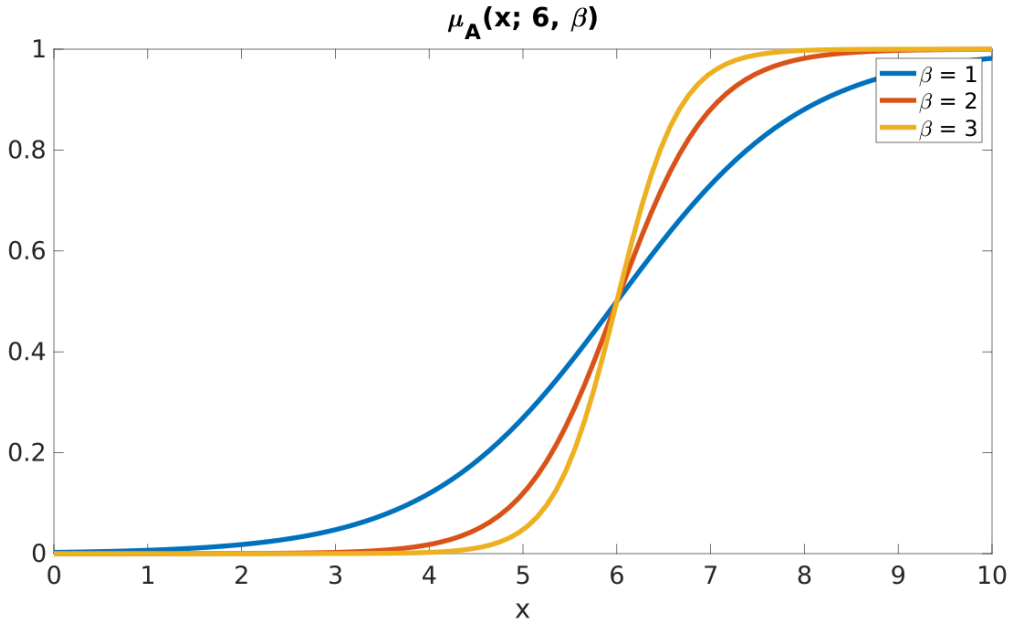

Figure 23. Example of sigmoidal membership function with negative β *values.*

Operations on fuzzy sets

Let A and B be fuzzy sets in a universe of discourse X, and μ_A and μ_B be the membership functions of A and B, respectively, then the operations of product and sum over fuzzy sets A and B are listed as follows [76] [81] *(Eq. 11, Eq. 12)*:

$$\mu_{A\cap B}(x) = \mu_A(x) \wedge \mu_B(x), dla\ x \in X \quad (11)$$

$$\mu_{A\cup B}(x) = \mu_A(x) \vee \mu_B(x), dla\ x \in X \quad (12)$$

In order to define the operations that can be performed on the fuzzy sets, norms (*t*-norms) and triangular conorms (*s*-norm) were introduced [79].

The *t*-norm (*s*-norm) is the function of two variables T:

$[0, 1]$ x $[0, 1] \rightarrow [0, 1]$, S: $[0, 1]$ x $[0, 1] \rightarrow [0, 1]$, with the following properties [76] [79]:

T1) boundaries: $T(x, 1) = x, \quad T(x, 0) = 0$

T2) monotonicity:

$x \leq u \quad T(x, y) \leq T(u, y)$

$y \leq r \quad T(x, y) \leq T(x, r)$

T3) commutativity: $T(x, y) = T(y, x)$

T4) associativity: $T(x, T(y, z)) = T(T(x, y), z)$

S1) boundaries: $S(x, 1) = 1, \quad S(x, 0) = x$

S2) monotonicity:

$x \leq u \quad S(x, y) \leq S(u, y)$

$y \leq r \quad S(x, y) \leq S(x, r)$

S3) commutativity: $S(x, y) = S(y, x)$

S4) associativity: $S(x, S(y, z)) = S(S(x, y), z)$,

$u, r, x, y, z \in [0, 1]$.

Exemplary *t*-norms and corresponding s-norms are presented in *(Table 5)* [76] [79]:

Table 5. Examples of *t*-norms and corresponding s-norms.

Name	*t*-norm	*s*-norm
Zadeh	$M(x, y) = \min(x, y)$	$M'(x, y) = \max(x, y)$
Algebraic / Probabilistic	$\prod(x, y) = x\,y$	$\prod'(x, y) = x + y - xy$
Łukasiewicz	$W(x, y) = \max(x + y - 1, 0)$	$W'(x, y) = \min(x + y, 1)$
Fodor	$\min_0(x, y) = \begin{cases} \min(x, y), & x + y > 1 \\ 0, & x + y \leq 1 \end{cases}$	$\max_1(x, y) = \begin{cases} \max(x, y), & x + y < 1 \\ 1, & x + y \geq 1 \end{cases}$
Drastic	$Z(x, y) = \begin{cases} \min(x, y), \max(x, y) = 1, \\ 0, otherwise \end{cases}$	$Z'(x, y) = \begin{cases} \max(x, y), \min(x, y) = 0, \\ 1, otherwise \end{cases}$
Einstein	$E(x, y) = \frac{xy}{2-(x+y-xy)}$	$E'(x, y) = \frac{x+y}{1+xy}$

Fuzzy-inference systems work in the following way (*Figure 24*):

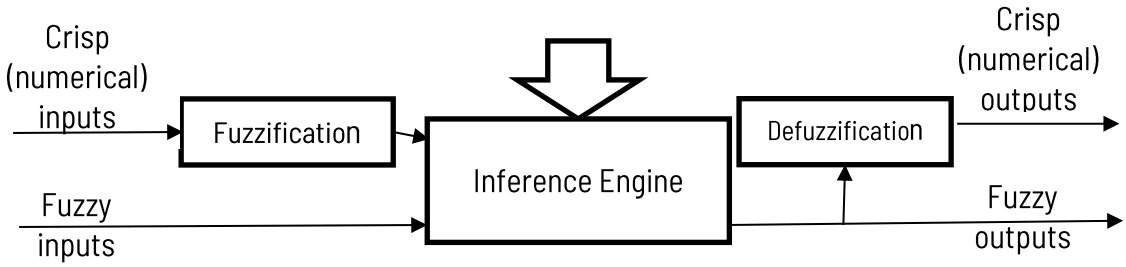

Figure 24. A structure of a fuzzy-inference system.

Both linguistic values (presented in the form of fuzzy sets) and numerical (crisp) values can be given for the inputs of fuzzy systems [76]. If crisp data are applied, then the inference process is preceded by fuzzification (the appropriate fuzzy set is assigned to the non-fuzzy input) [82] or these values are considered as the fuzzy singletons [76]. The fuzzy system contains a knowledge base written in fuzzy conditional IF-THEN rules and a fuzzy inference engine, an approximate inference mechanism based on the fuzzy set theory, and fuzzy inference [76]. If the numerical (crisp) data are required as the fuzzy system output, defuzzification methods need to be applied (the numerical data is assigned to the resultant output fuzzy set) [82].

The methods used for defuzzifying the fuzzy output functions are as follows [83]:

1. Max-membership principle
2. Centre of gravity method

3. Weighted average method
4. Mean-max membership
5. Centre of sums
6. Centre of largest area
7. First of maxima or last of maxima.

Fuzzy conditional rules follow the structure:

if X **is** A, **then** Y **is** B,

where A and B are the values of linguistic variables X and Y, respectively, defined by fuzzy sets with membership functions $\mu_A(x)$ and $\mu_B(y)$ [76].

The statement „X **is** A" is called the antecedent (premise), whereas the statement „Y **is** B" is called the conclusion (consequent). Each of the IF-THEN rules from the knowledge base represents a local dependency between input and output. The rule's premise defines the fuzzy area of its activity, whereas the conclusion determines the system's output for that area. [76][82] Linguistic variable values correspond to certain natural language categories and are represented by words or statements [76].

There are a plethora of fuzzy-inference systems, such as Mamdani-Assilan [84], Takagi-Sugeno-Kang [85][86], with moving consequents in IF-THEN rules [87], Tsukamoto [88], Baldwin [89].

The first fuzzy-inference system is the Mamdani-Assilan (MA) one [76][84].

This system is based on a set of conditional fuzzy IF-THEN rules in a canonical form, given by a human expert *(Eq. 13)* [76][82][84]*:*

$$\mathcal{R} = \{\mathcal{R}^{(i)}\}_{i=1}^{I} = \left\{ if \left(\bigvee_{n=1}^{N} X_n \; is \; A_n^{(i)} \right), then \; Y \; is \; B^{(i)} \right\}_{i=1}^{I} \quad (13)$$

where:

X_1, X_2, ..., X_N are the input linguistic variables of a system,

Y is an output linguistic variable of a system,

$A_1^{(i)}$, $A_2^{(i)}$, ..., $A_N^{(i)}$, $B^{(i)}$ represent the linguistic values for an *i*-th rule, defined directly on the universes of discourse $\mathbb{X}_1$, $\mathbb{X}_2$, ..., $\mathbb{X}_N$, .

E. H. Mamdani and S. Assilan used the minimum operation as the *t*-norm to model the conjunctive „AND" of the IF-THEN rules antecedents and the *s*-norm maximum to aggregate the results of

interference obtained based on individual rules. The singleton fuzzifier is used for mapping the numerical inputs into fuzzy sets. The centre of gravity method (COG) is used as a defuzzification method. [76][82]

4.2. Mamdani-Assilan fuzzy-inference system for isotopic envelope member peaks preselection

In order to perform a preselection of peaks that can be potential isotopic envelopes members, a Mamdani-Assilan fuzzy-inference system has been constructed.

Herein, the Mamdani-Assilan fuzzy-inference system for potential isotopic peaks (isotopic envelope members) is described, which has been published in [90], but since then, it has been modified. The reason for creating the fuzzy-inference system based on the Mamdani-Assilan one is that it is well-suited to human input (expert knowledge) and intuitive. The system is published in [53].

The system's structure is based on the knowledge of an experienced mass spectrometrist in the field of MALDI MS and is as follows: two inputs (distance between means of two adjacent model components and estimated variances ratio of two adjacent model components) and one output (*E*, which means 'envelope'):

1) Distance between means of two adjacent model components is approximately equal to one (*Figure 25*)[53]

The reason for that is typical MALDI data consist of single charged ions on the mass spectrum, therefore *(Eq. 14)*[53]:

$$\frac{1.003}{z} = \frac{1.003}{1} = 1.003\ [\mathrm{Da}] \qquad (14)$$

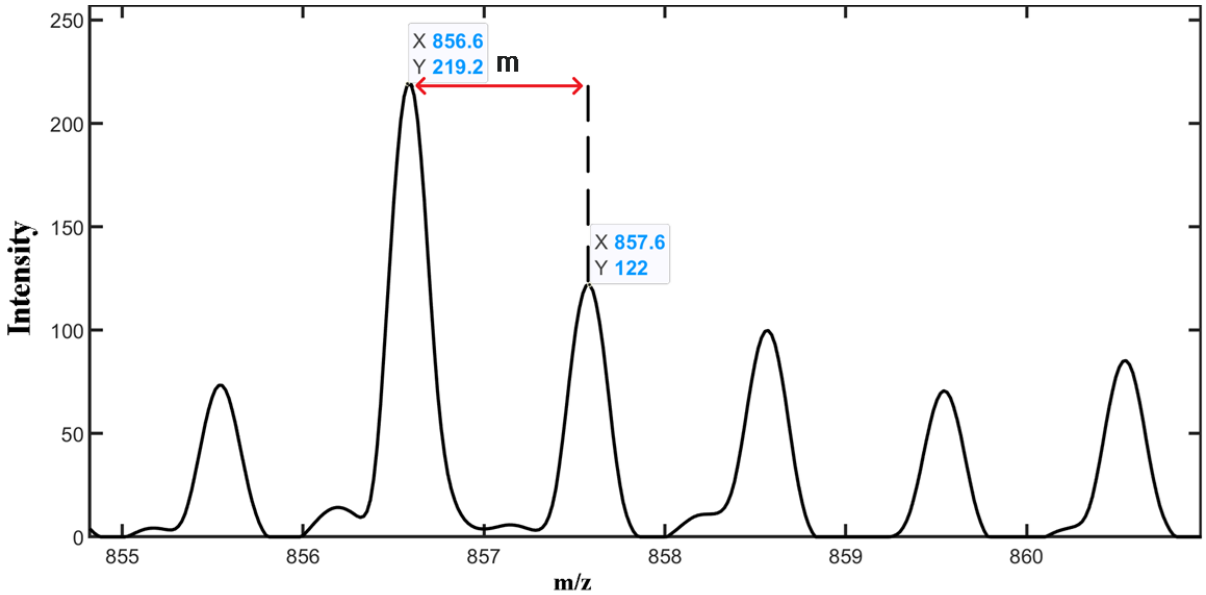

Figure 25. Distance (*m*) between means of two adjacent model components [53].

2) Ratio of estimated variances of adjacent model components is approximately equal to one *(Figure 26)* [53].

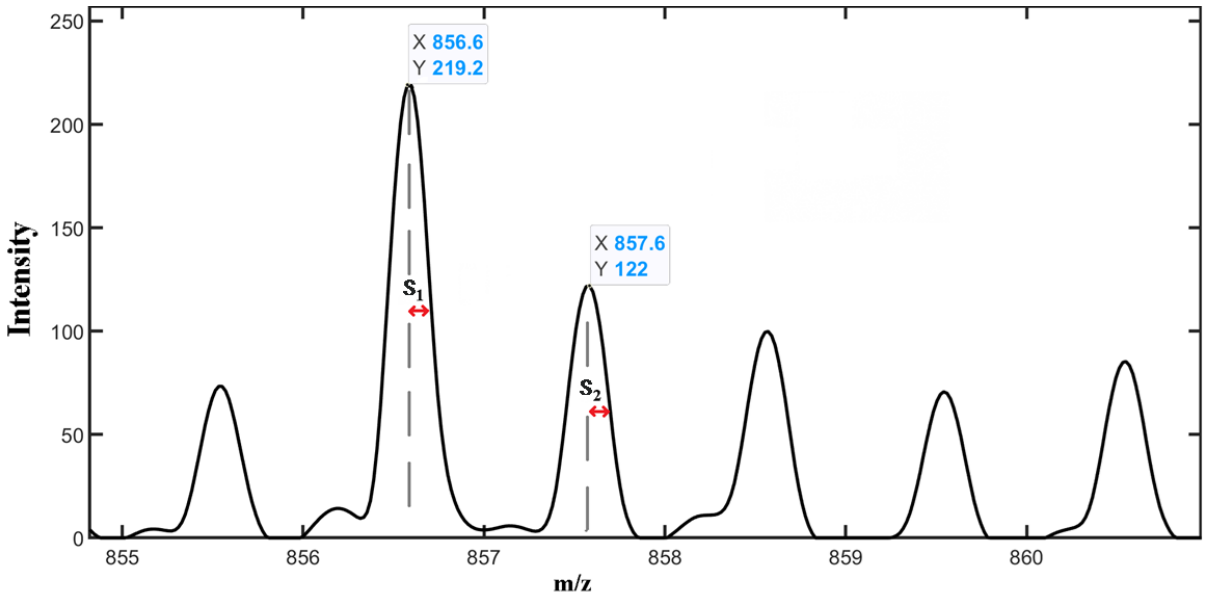

Figure 26. Ratio of estimated variances of two adjacent model components (s) [53].

Based on the histogram of distance values and ratio of variances, respectively, the modified Gaussian membership functions with the following parameters have been constructed:

$m_{11} = 0.99, \sigma_{11} = 0.0637$ $\quad$ $m_{21} = 0.99, \sigma_{21} = 0.0200$

$m_{12} = 1.01, \sigma_{12} = 0.0637$ $\quad$ $m_{22} = 1.01, \sigma_{22} = 0.1000$

Output:

$M_1 = 0.9405, \sigma_1 = 0.09216$

$M_2 = 1.06, \sigma_2 = 0.08710$

It was constructed in MATLAB with the knowledge base consisting of one conditional fuzzy rule *(Figure 27)*:

IF distance is **IN THE RANGE** and variances ratio is **IN THE RANGE**, then output is **E**.

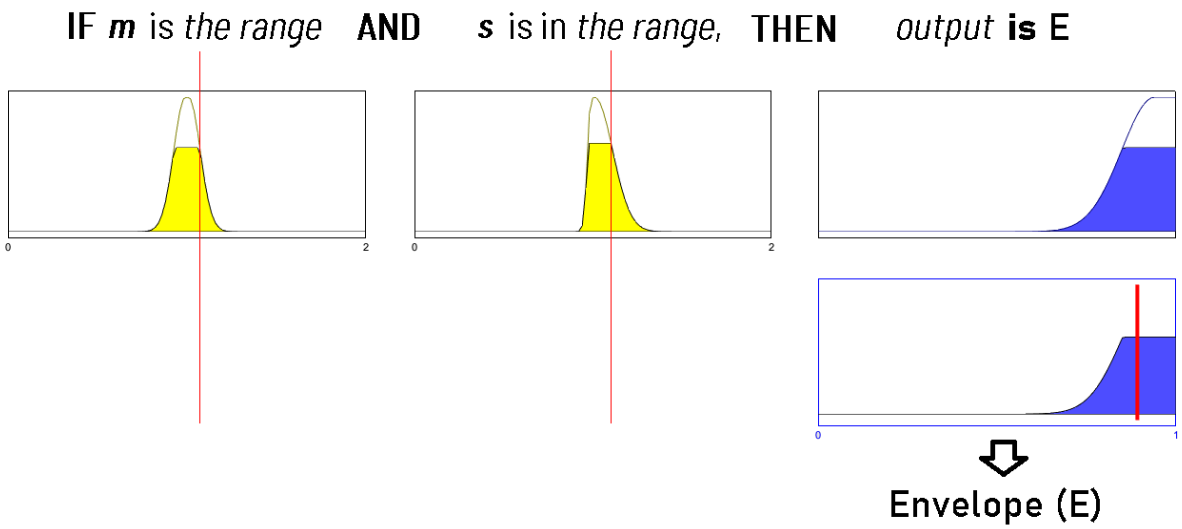

Figure 27. Implication and aggregation in fuzzy logic system [53].

It can be observed (*Figure 28*) that for $s = 1$ and $m = 1$, the summit of the surface plot is observed. The lower s is, the more significant PV changes can be observed, and the steep slope. PV is sensitive to the s changes. Below $s = 0.7$, PV values are small. PV decreases gradually for s in the range $s \in (1; 2)$,. The further from $m = 1$, the worse results are obtained.

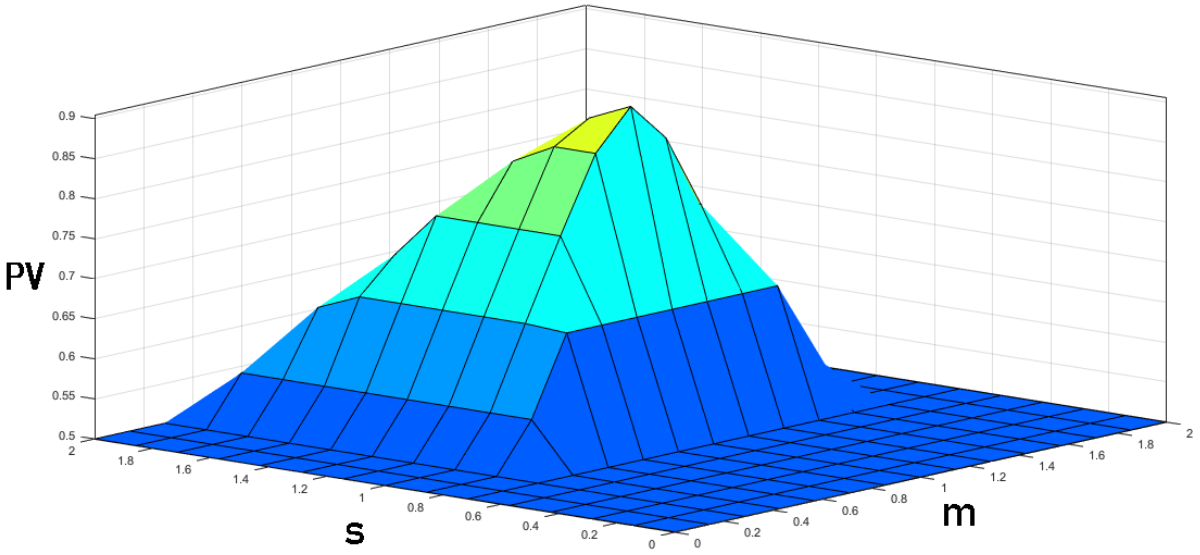

Figure 28. Surface plot of possibility value (*PV*), variance ratio (*s*) and distance between means of model components (*m*).

In order to find a threshold value of the output, Gaussian Mixture decomposition was applied [26] [91]. In order to find the number of components, BIC scores [92] with corresponding gradients were calculated. [53]

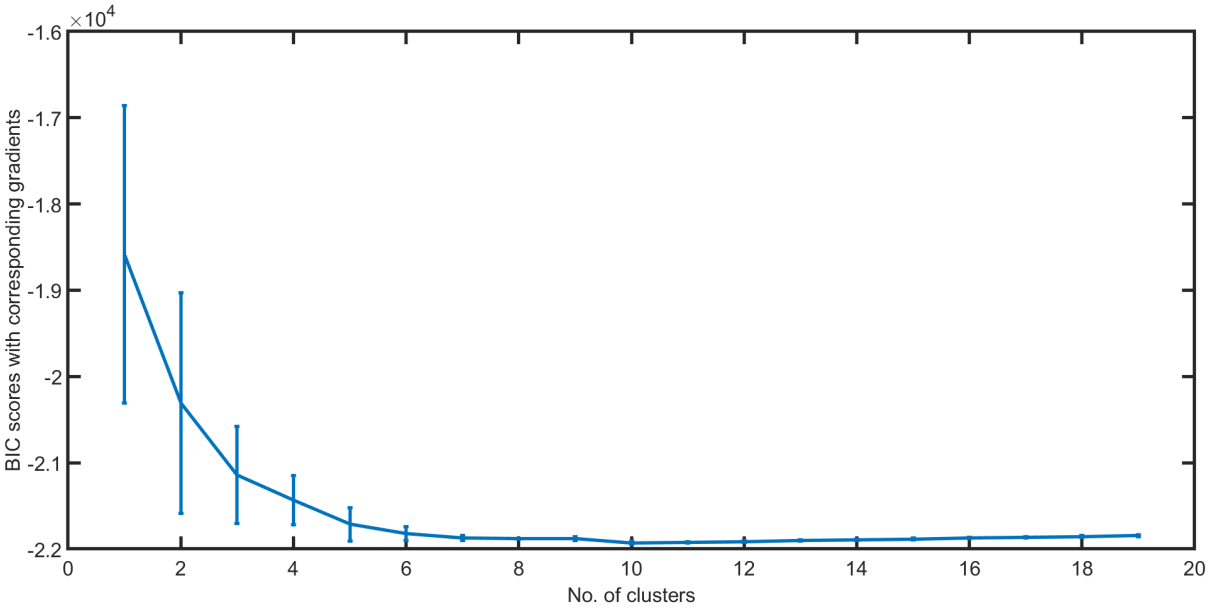

Figure 29. BIC scores with corresponding gradients vs. number of clusters **[53].**

Generally, the lower the BIC value is, the more accurate the model predictions are. According to the BIC scores gradients versus no. of clusters (*Figure 29*), it can be observed that beginning from the 6. cluster, the change in slope is not significant. Therefore, it is not worth taking into

consideration so big number of clusters. As a result, Gaussian Mixture decomposition has been performed with 5 components. [53]

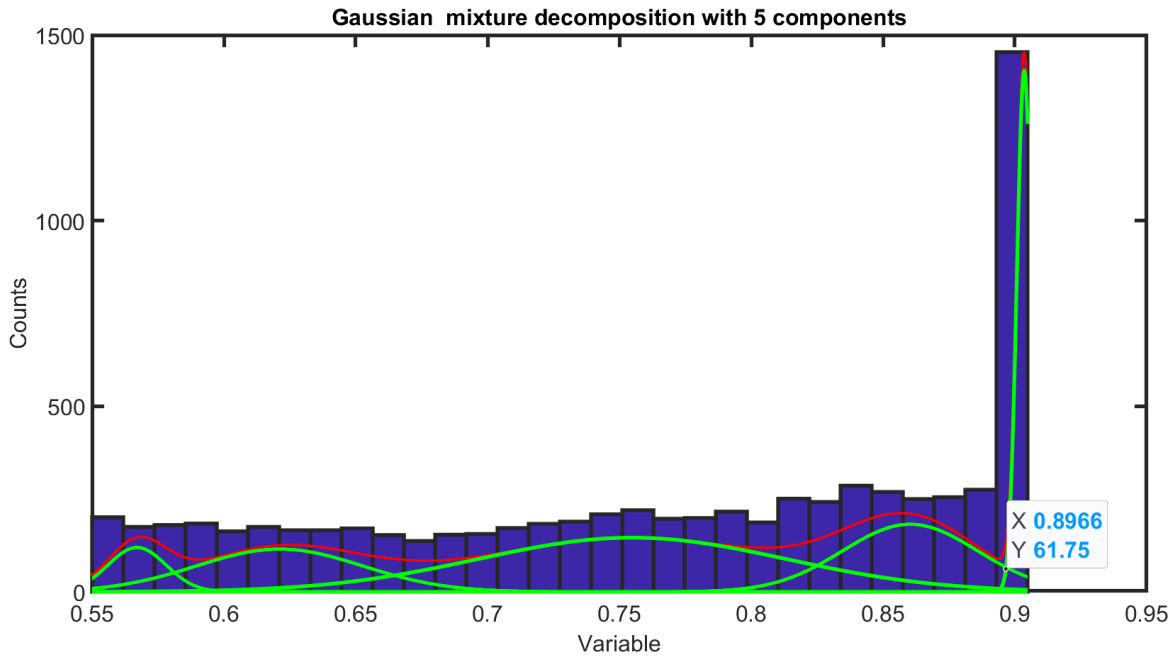

Figure 30. GMM decomposition with 5 components **[53]**.

Peaks with an output value over the threshold (0.8966) are considered the potential isotopic envelope members (*Figure 30*) [53]. Such a strict threshold was chosen in order to optimise the process and reduce the dimensionality of data.

The output of the fuzzy-inference system indicates whether the pair of peaks are the members of isotopic envelope by the possibility of an isotopic envelope membership [%].

Table 6. Exemplary results for *HNC-FF Dataset 1* mass spectrum **[53]**.

m/z_1	m/z_2	Possibility of isotopic envelope membership [%]
805.6	809.7	46.0 **(non-Envelope)**
808.7	809.7	74.7 **(non-Envelope)**
810.7	811.7	98.1 **(Envelope)**
810.8	897.6	15.3 **(non-Envelope)**
812.7	813.7	98.7 **(Envelope)**
812.7	897.6	25.1 **(non-Envelope)**
843.7	844.7	99.0 **(Envelope)**

It can be noticed in *Table 6* that isotopic envelope member peaks are characterised by possibility values bigger than 89.66%.

Table 7. Results of applying the fuzzy-inference system to different datasets [53].

Dataset Name	No. of input peak pairs	No. of potential isotopic envelope member peak pairs	↓ %of peak pairs reduction
HNC-FFPE Dataset 1	97 910	1 916	98.04
HNC-FFPE Dataset 2	55 030	1 457	97.35
HNC-FFPE Dataset 3	68 920	1 584	97.70
HNC-FFPE Dataset 4	97 840	1 945	98.01
HNC-FF Dataset 1	47 750	1 662	96.52
HNC-FF Dataset 2	46 350	1 610	96.53
HNC-FF Dataset 3	54 030	1 624	96.99
HNC-FF Dataset 4	55 070	1 603	97.09

In *(Table 7)* and (*Figure 31*) it can be observed that the number of peak pairs that should be taken into further analysis has significantly decreased for every dataset (over 96%) [53].

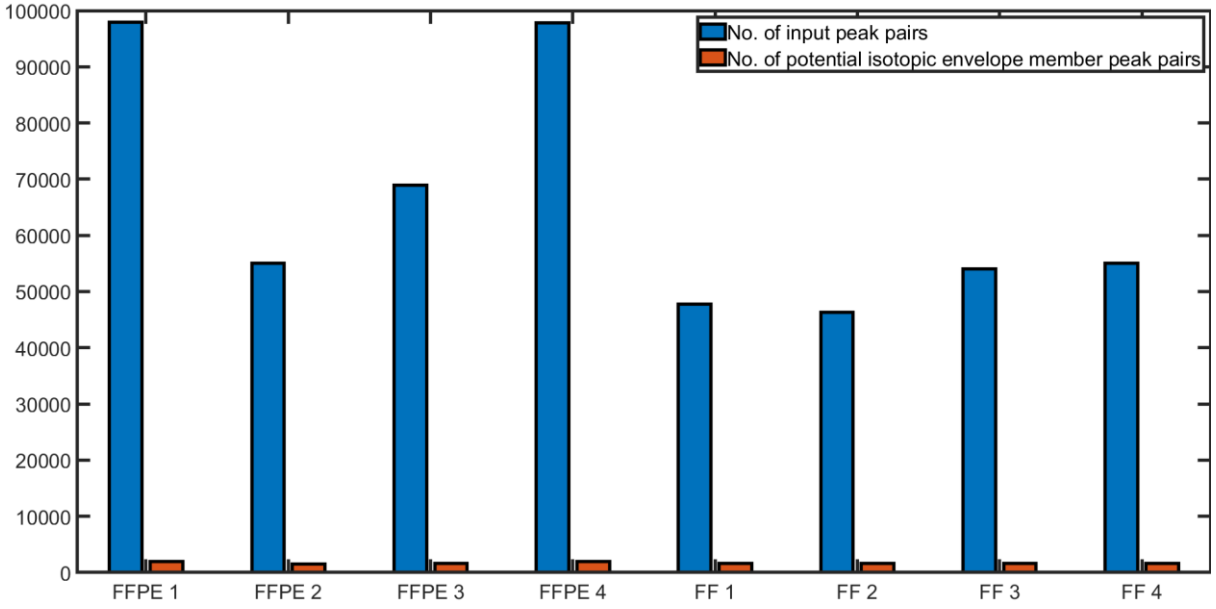

Figure 31. Reduction of input peak pairs after applying Mamdani-Assilan fuzzy-inference system [53].

For every dataset, the algorithm has reduced the number of peaks to over one thousand deisotoped peaks.

The primary purpose of employing fuzzy logic as a first step to the problem of isotopic envelope identification is to reduce the dimensionality of peak pairs that should undergo subsequent analysis.

5. VERIFICATION OF THE MEMBERSHIP OF THE PREDEFINED PEAKS INTO THE ISOTOPIC ENVELOPE

5.1. Peaks spatial distribution as a basis for further analyses

The peak intensity in the mass spectra of each spot (or pixel) is defined for each analyte. Therefore, it is possible to map the spatial distribution of the analyte. All peaks from the mass spectrum (with a given m/z value) are visualised as a map of intensities (creating the so-called image) that are shown in *Figure 32*. Such spatial map of a molecular distribution presents the intensities of peaks registered for m/z values across the whole tissue section. [53][93]

The idea has been published in [93].

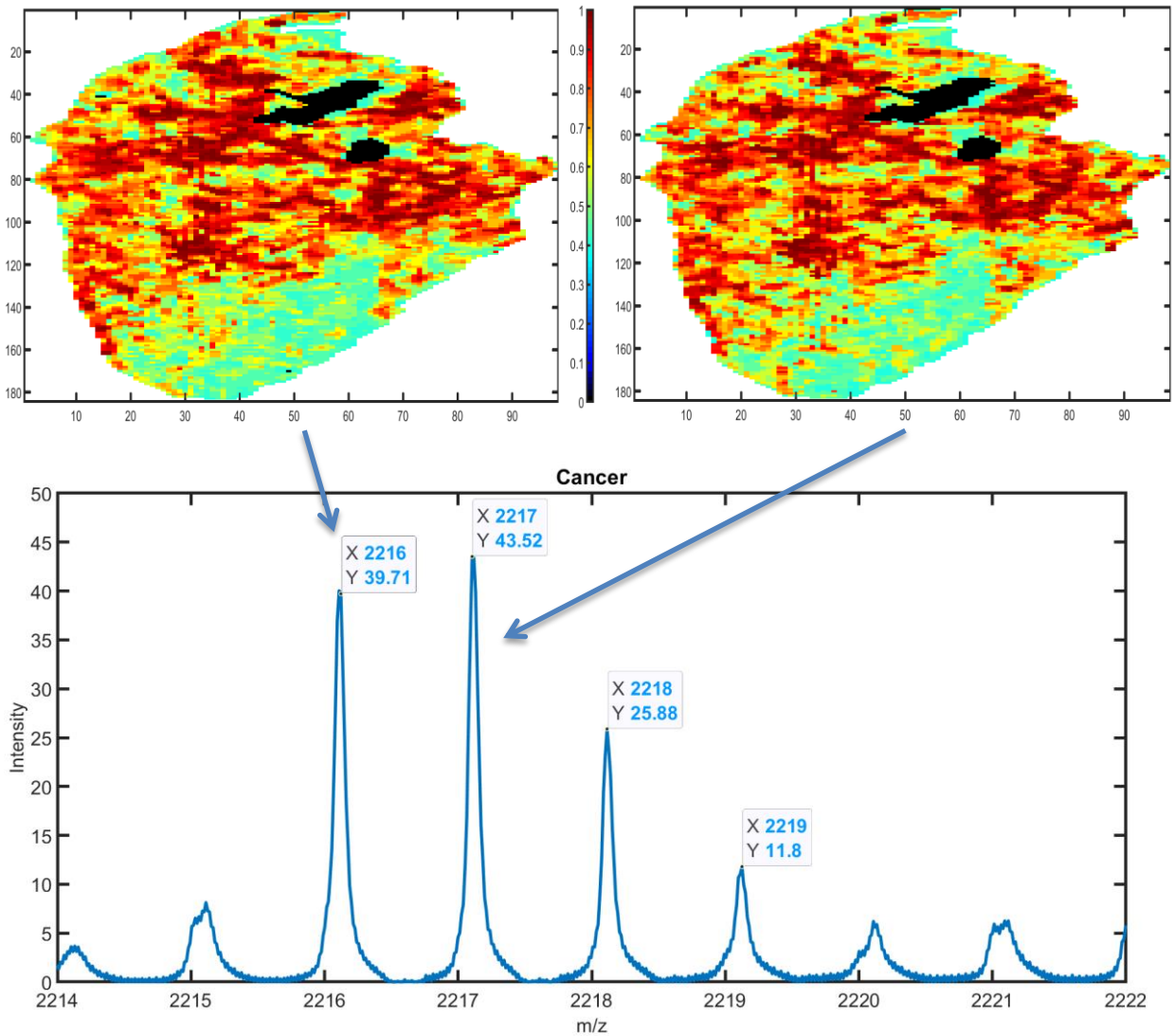

Figure 32. Visualisation of the peaks with given m/z values as the spatial maps of molecular distribution (maps of intensities) **[53]**.

The spatial maps of molecular distribution of peaks are constructed in the following way (*Figure 33*)[93]:

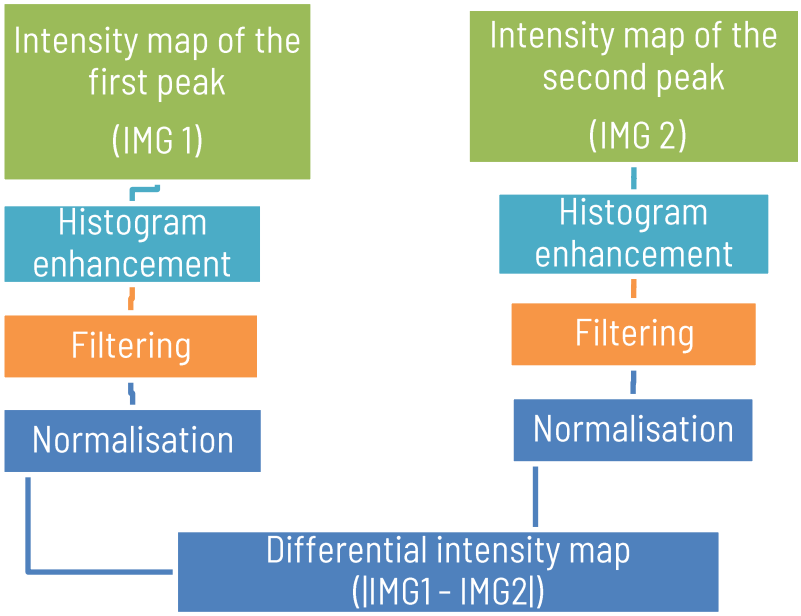

Figure 33. The pipeline of constructing the differential intensity map **[93]**.

a) Constructing a map of intensities for each peak from the mass spectrum with a given m/z value is based on presenting peak intensities registered for given m/z values across the whole tissue section at the original coordinates. Different colours reflect different peak intensity values.[93]

b) Histogram enhancement[94]

Histogram equalization was performed to enhance the contrast by widening the range of intensities of the image. Histogram equalization transformation is performed in the following way *(Eq. 15, Eq. 16, Eq. 17)*

$$s_k = K \times CDF(r_k) \quad \textbf{(15)}$$

$$CDF(r_k) = \sum_{j=0}^{k} p(r_j) \quad (16)$$

$$s_k = T(r_k) = K \sum_{j=0}^{k} p(r_j) \quad (17)$$

where:

r_k – specific random intensity value,

s_k – corresponding random intensity value,

r – random variable,

K – scaling constant.

c) Spatial filtering was performed in order to remove noise from the image using the median filter [94] *(Eq. 18).*

$$\hat{f}(x, y) = \text{median}\,\{g(s, t)\} \quad (18)$$
$$(s, t) \in S_{xy}$$

m – number of image rows spanned by the filter,

n – number of image columns spanned by the filter,

g – noisy image,

S_{xy} – *m x n* neighbourhood of the input noisy image; the neighbourhood is centred at spatial coordinates *(x, y),*

$\hat{f}(x, y)$ - estimate of *f* that denotes the filter response at coordinates *(x, y).*

d) Normalisation

For each m/z, a vector of maximum intensities for a given m/z is constructed. Then, the average from that vector is calculated, so it is the average of maximum intensities for all m/z.

e) Differential intensity map

After visualising all peaks from the mass spectrum as the map of intensities, pairwise differential intensity maps were created by subtracting the intensity maps of two peaks (*Figure 34* and *Figure 36*, *Figure 35* and *Figure 37*). The outcome is the absolute value of two spatial intensity maps (*Figure 38, Figure 39*). [93]

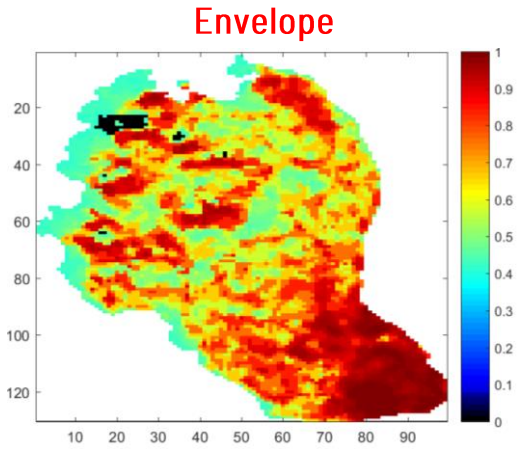

Figure 34. Image *A* [53].

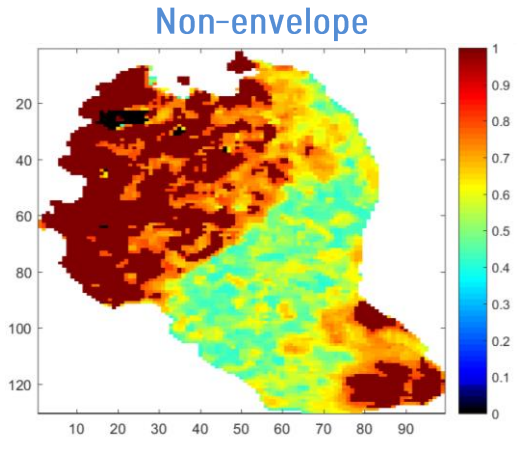

Figure 35. Image *C* [53].

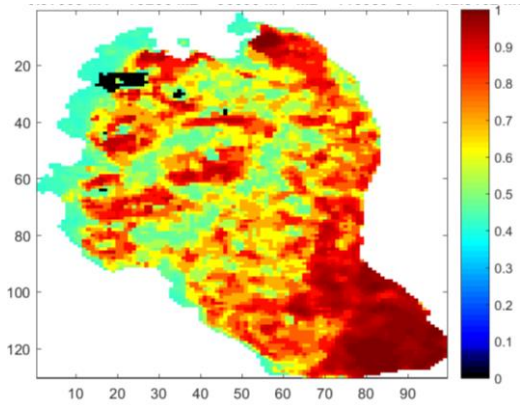

Figure 36. Image *B* [53].

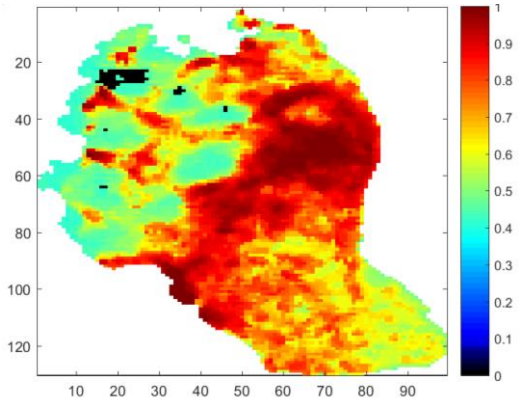

Figure 37. Image *D* [53].

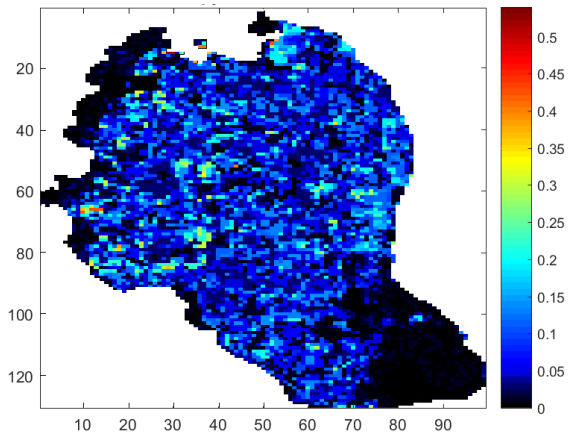

Figure 38. Differential image |*A* – *B*| [53].

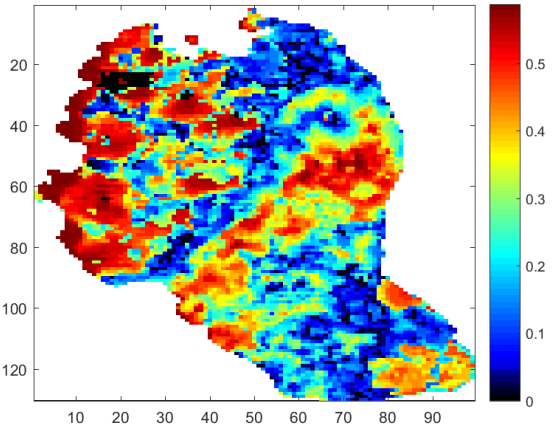

Figure 39. Differential image |*C* – *D*| [53].

The assumption is that for isotopic envelope peak members, the spatial distribution is similar. That means that the intensity structure and structural image properties are similar – there is no texture visible in the obtained differential intensity map for those peaks (*Figure 38*). Hence, the

differential image has a noise character, only uniform intensity distribution is visible or the intensities are at similar level. On the contrary, peaks not included in isotopic envelopes have no similar spatial distribution, resulting in clearly visible structure in the differential image (*Figure 39*). In order to turn this assumption into measurable values, several measures for image structure and texture analysis have been employed, described in detail in 5.2.1. [53] [93]

5.2. Classifier construction process

5.2.1. Descriptors selection

In order to assign peaks to the correct class: envelope (*E*) or non-envelope (*nE*), several metrics have been applied (*Figure 40*). They can be divided into two groups: based on parameters of the model components and image-based. The second ones are split into texture metrics (most of them are based on *Gray-Level Co-Occurrence matrix – GLCM*) and image structure. GLCM-based texture metrics are calculated based on the differential intensity maps, whereas those concerning the parameters of the model components, image structure, and non-GLCM-based texture metrics are calculated on the spatial intensity map of a single peak (model component). [53]

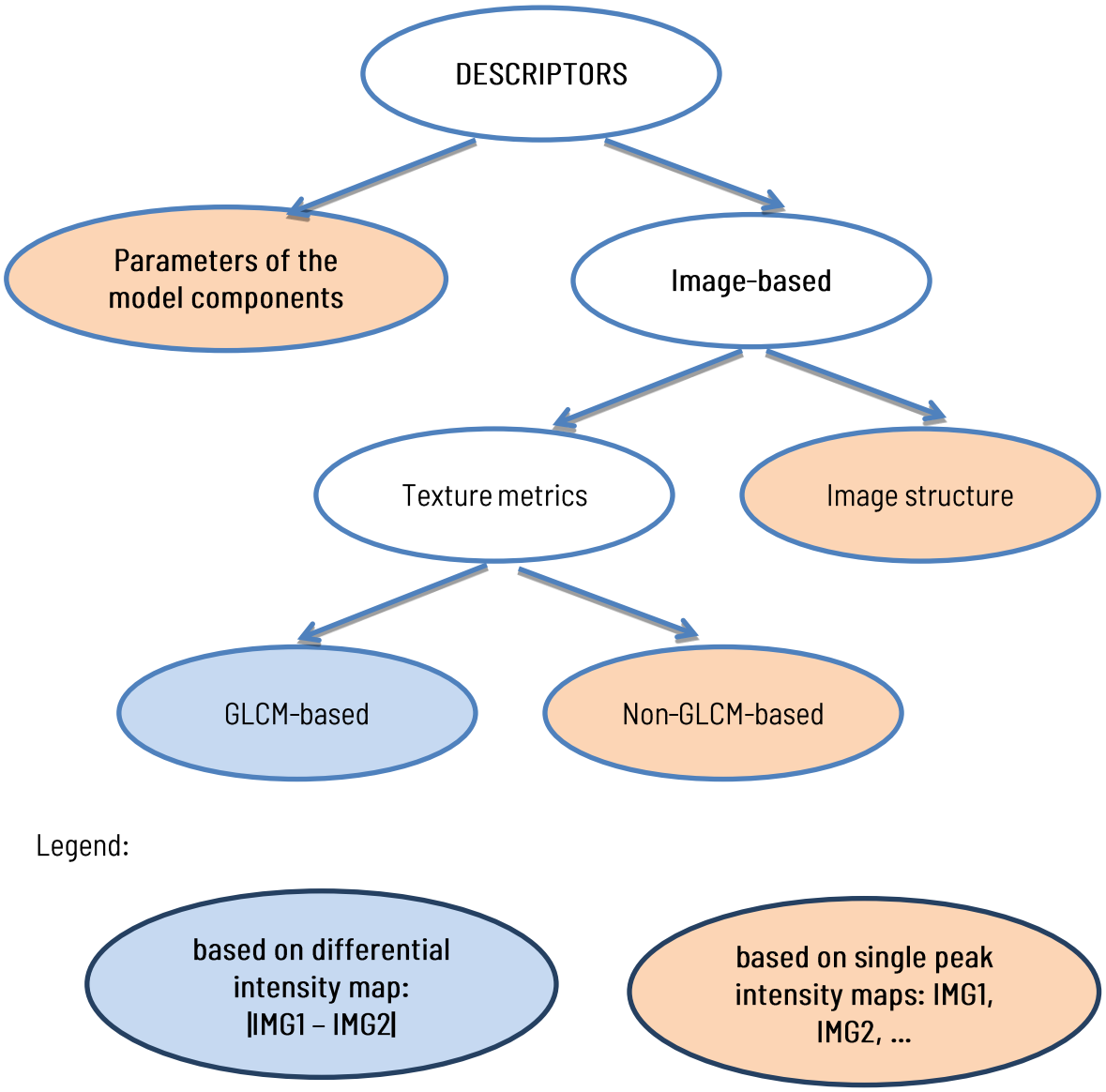

Figure 40. Descriptors divided into groups and distinguished based on calculations: differential image-based and based on separate peaks images **[53]**.

<u>I group of descriptors: parameters of the model components</u> [53]

The first group of descriptors is based on parameters of the model components [53].

1. The **distance** between the means of two adjacent model components is approximately equal to 1.003 Da (*Figure 41*) *(Eq. 19)* [53]:

$$m = \frac{1.003}{z}, \qquad (19)$$

where:

z – ion charge.

This assumption is based on the condition that in MALDI MS, there is a single peptide ionisation in most cases. It was aforementioned and explained in 2.5, based on [49].

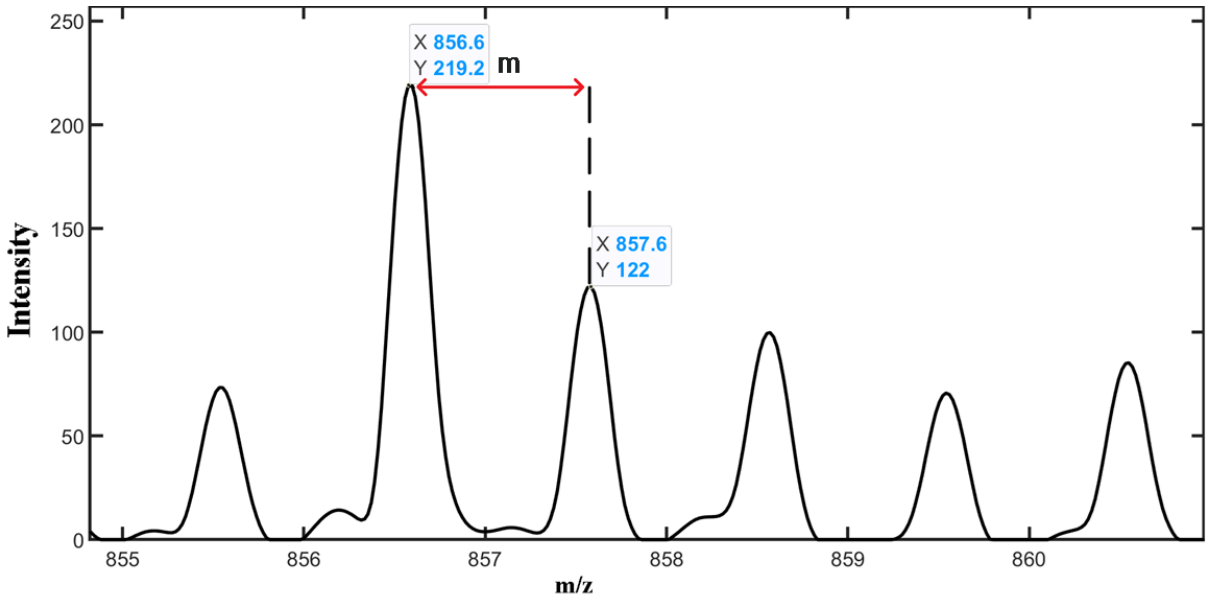

Figure 41. Distance between the means of two adjacent model components **[53]**.

2. **The estimated variances ratio** of two adjacent model components (s) that are members of the isotopic envelope is approximately equal to 1 (*Figure 42*) *(Eq. 20)* [53]:

$$s = \frac{s_1}{s_2} = 1 \qquad (20)$$

This assumption is based on expert knowledge in the field of mass spectrometry.

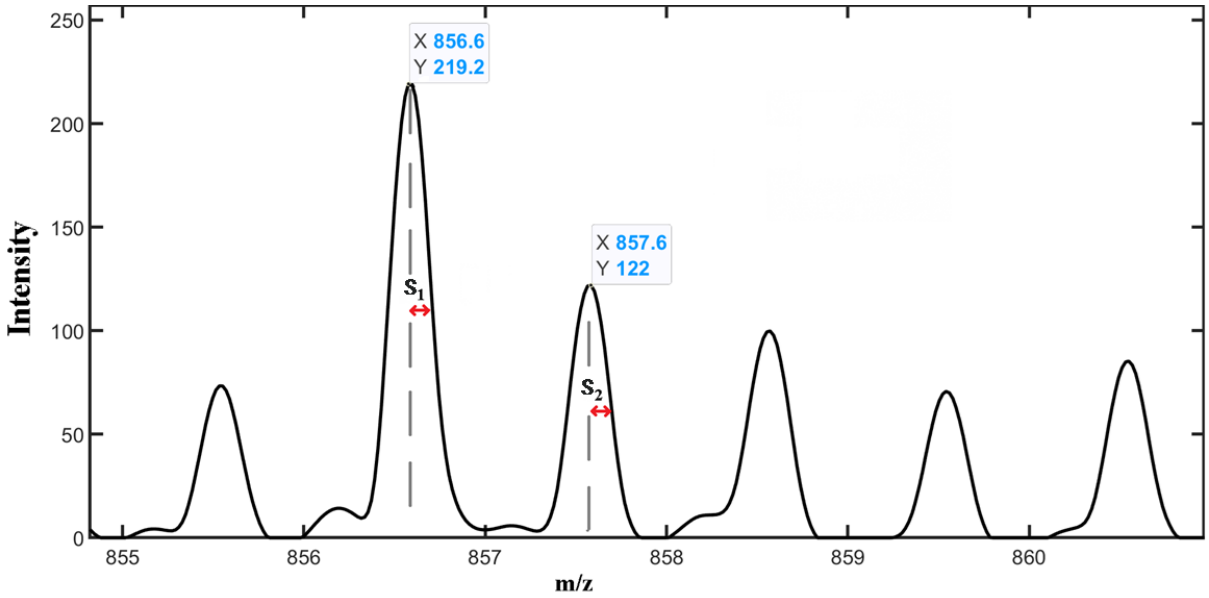

Figure 42. The estimated variances ratio of two adjacent model components [53].

II group of descriptors: texture metrics

"Texture is data variation at scales smaller than the scales of interest". It is a characteristic that splits the image into areas of interest. [95] In essence, texture represents the local spatial arrangement of reflectance values [96]. It is worth mentioning that texture carries valuable information about the structural arrangement of surfaces and how they are related to the surrounding environment [97].

Texture can be defined as distribution of patterns, spatial arrangement of textons, or specific spatial frequencies [98]. Several different approaches can be applied to assess the texture: statistical, structural, model-based, and transform-based methods [98]. Texture metrics can be characterised as first- or second-order metrics. Second-order texture measures take into consideration the relationships between pixels in an image, whereas the first-order are statistics that do not consider a relationship between the pixels [99].

A plethora of texture metrics is based on the *Gray-Level Co-Occurrence Matrix* (*GLCM*). GLCM can be defined as directional. In this case, it is parametrised by two parameters: distance *d* between pair of pixels (one is called the reference and the second – the neighbouring pixel [95]) and an angle $\emptyset$ formed by the line connecting the pair of pixels with the reference direction of the image [100] *(Eq. 21, Eq. 22)*:

$$C(k,l;d,\emptyset) = \sum_i \sum_j \delta\left(k - g(i,j)\right)\delta\left(l - g(i + d\cos\emptyset, j + d\sin\emptyset)\right), \quad (21)$$

where:

$g(i, j)$ – grey value of pixel (i, j),

$C(k, l; d, \emptyset)$ – total number of pixels pairs at distance d and angle $\emptyset$ from each other,

k, l – grey values of the first and second pair of pixels, respectively.

$$p\,(m,n) = \frac{1}{All\ pairs\ of\ pixels\ used}\ C_d\,(m,n) \qquad (22)$$

In all the above definitions, G is the total number of grey levels that were used. The co-occurrence matrix corresponds to a joint probability distribution function ("a given pair of gray-levels co-occur at a specific relative distance in the image" [98]). [95] [100]

In essence, the GLCM presents the frequency of occurrence of different combinations of grey levels (pixel brightness values) in an image [99]. Haralick [97] proposed 14 texture measures based on GLCM and tonal differences between pixel pairs [96]. Generally, statistical texture metrics are based on evaluating of the spatial distribution of grey values. The evaluation process is as follows: local characteristics in the image are calculated at each stage, and then, from the local characteristic distribution, a collection of statistics is extracted. [95]

The most commonly used features computed from the GLCM were calculated in order to differentiate envelope from non-envelope differential images. The statistical analysis, including contrast probability estimation and empirical cumulative distribution function, is based on the training data.

According to [99], texture metrics can be divided into four groups *(Figure 43)*. Despite calculating texture features based on GLCM, another image texture metric has been analysed, which is not based on GLCM: autocorrelation function .

All image texture metrics were calculated on the differential images. In order to avoid classifier overfitting, only some descriptors were chosen for further analysis in the process of feature selection, described in details in 5.2.1. Probability density estimate and empirical cumulative functions based on the training data are presented only for those descriptors. [53]

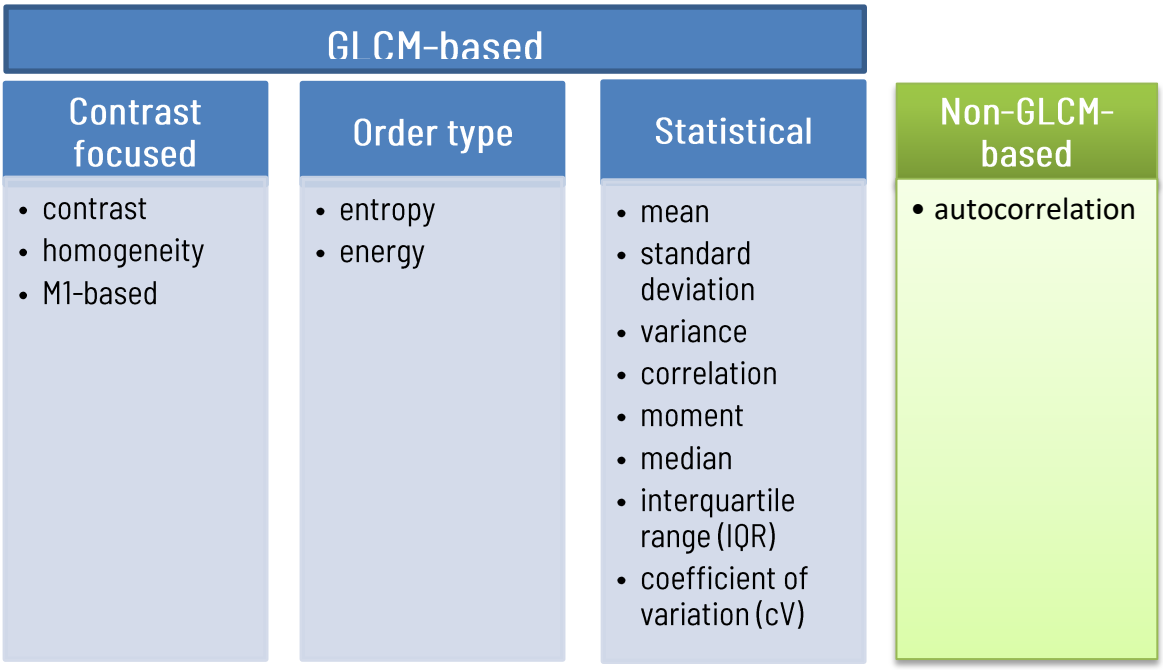

Figure 43. Groups of texture metrics [53].

1. **Contrast focused**

   These metrics are based on calculating the weights related to the distance from the GLCM diagonal [99].

   a) Contrast compares a pixel-neighbour intensity over the entire image [95]. It reflects the depth of the texture grooves and sharpness of an image [101] and represents the grey level variation in GLCM [102]. It carries information for two neighbouring pixels about the linear dependency of grey levels [102]. This feature is represented by *(Eq. 23)* [100]:

$$Contrast = \frac{1}{(G-1)^2} \sum_{m=0}^{G-1} \sum_{n=0}^{G-1} (m-n)^2 \, p(m,n) \qquad (23)$$

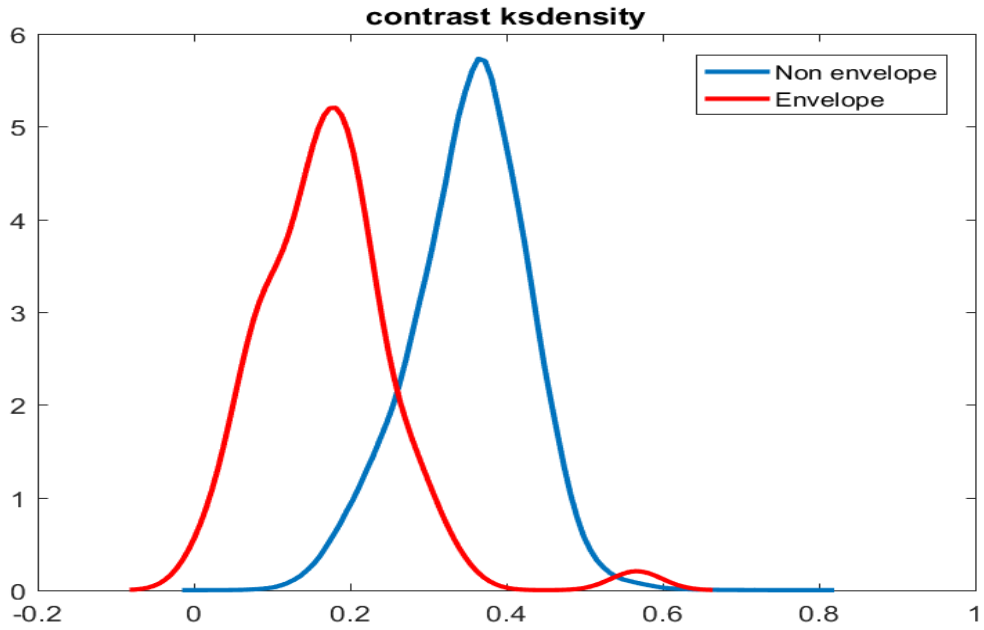

Figure 44. Contrast probability density estimate for the training data (*E* and *nE* peaks) [53].

It can be observed that over 95% of envelopes are in the range (0; 0.3), whereas non-envelopes in the same range are about 15% (*Figure 44*), (*Figure 45*) [53].

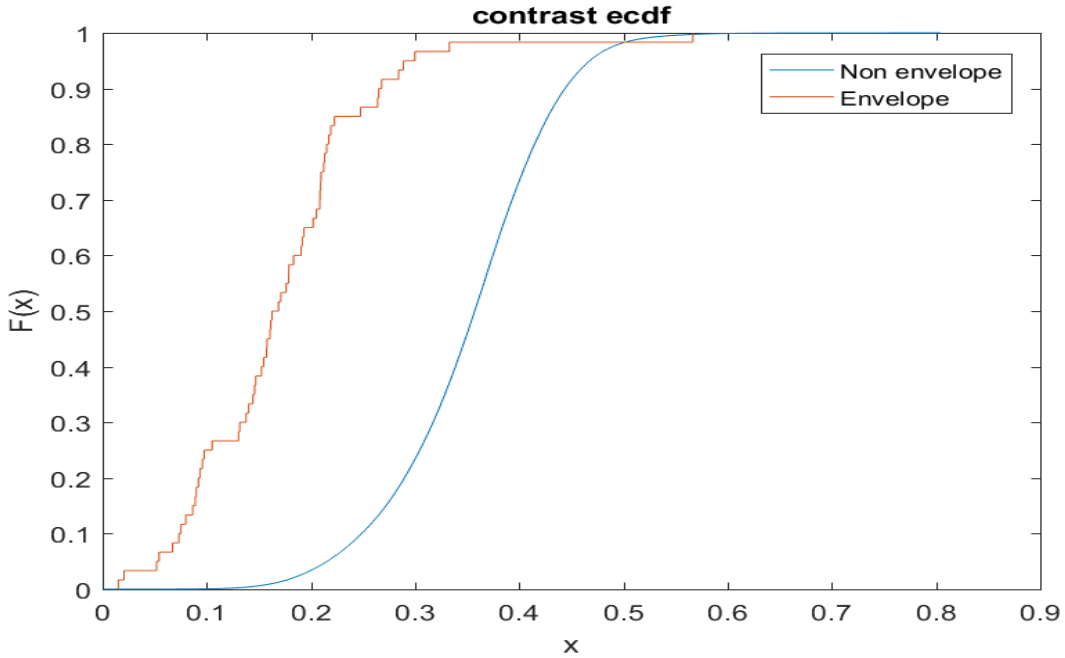

Figure 45. The empirical cumulative distribution function of contrast metric for the *E* and *nE* peaks [53].

b) Homogeneity measures how close is the distribution of the components in GLCM to the GLCM diagonal [95]. It denotes how homogenous the image textures are and scaled the local changes of image texture [101]. This feature is represented by *(Eq. 24)* [100]:

$$Homogeneity = \sum_{m=0}^{G-1}\sum_{n=0}^{G-1}\frac{p\,(m,n)}{1+|m-n|} \qquad (24)$$

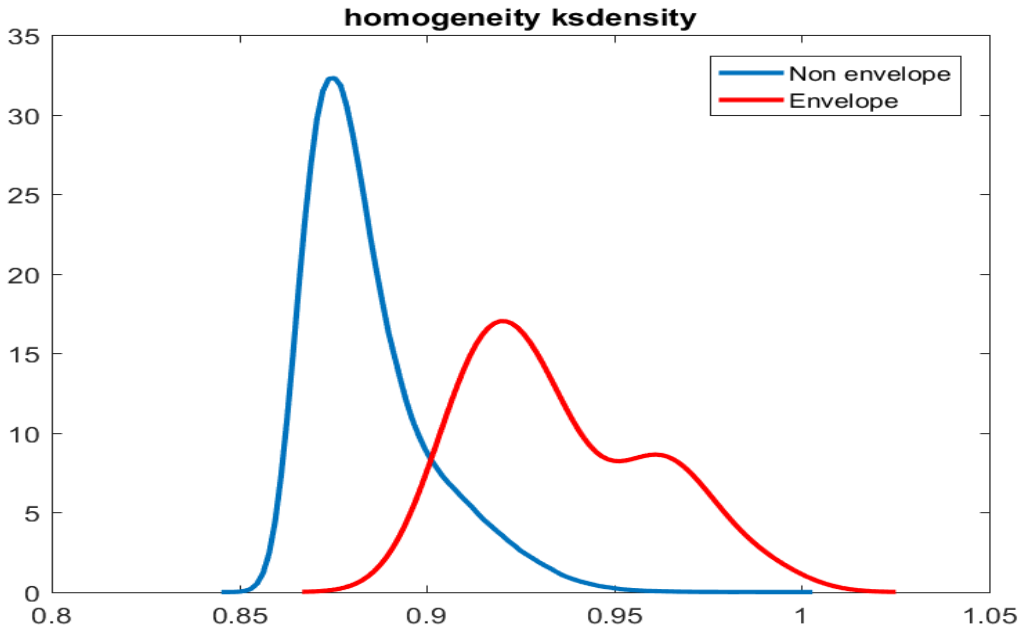

Figure 46. Homogeneity probability density estimate for the training data (*E* and *nE* peaks).

It can be observed that below 0.9 homogeneity value ~90% are the peaks that are not included in isotopic envelopes (*Figure 46*), (*Figure 47*).

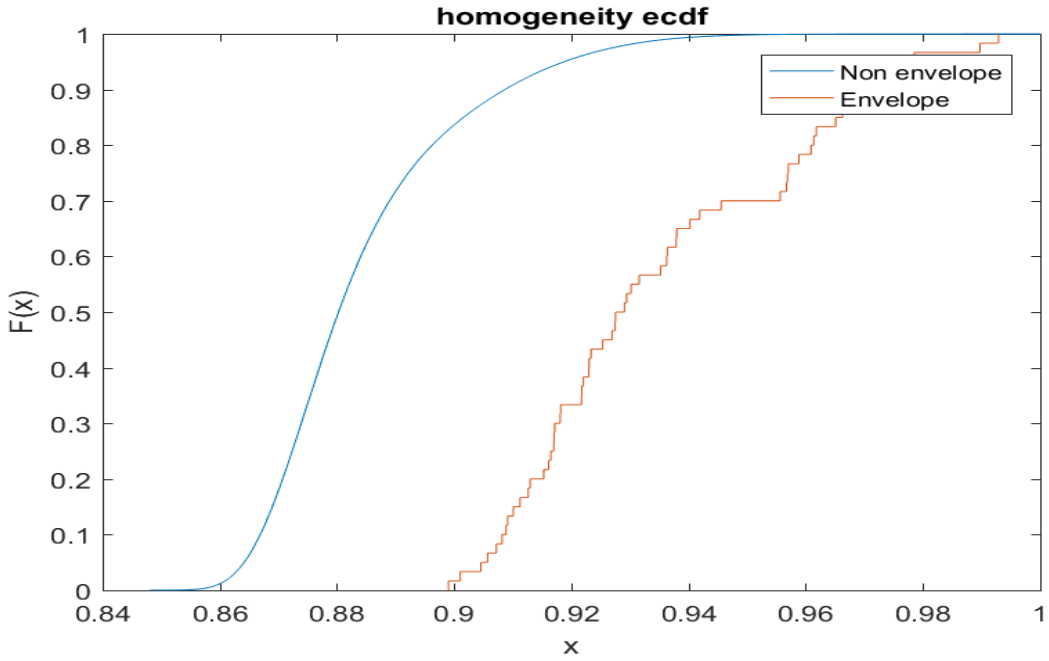

Figure 47. The empirical cumulative distribution function of homogeneity metric for the *E* and *nE* peaks.

c) M1 is a GLCM-based measure used for denoting the structured molecular images. The assumption is that structured images contain many co-occurring low-intensity and high-intensity pixel pairs since the contrast between those two groups of intensity values shows a clear structure. Based on that assumption, two weight vectors are created in order to assign higher weights to the pairs of the aforementioned pixels of interest and lower to other values *(Eq. 25)*.[103]

$$\mathrm{M1} = 4 \times \mathrm{GLCM}(1,1) + 2 \times \{\mathrm{GLCM}(1,2) + \mathrm{GLCM}(2,1) + \mathrm{GLCM}(2,2)\} + \{\mathrm{GLCM}(1,3) + \mathrm{GLCM}(2,3) + \mathrm{GLCM}(3,1) + \mathrm{GLCM}(3,2) + \mathrm{GLCM}(3,3)\} \quad (25)$$

2. Order type

The second group of texture features is related to the order – how regular are the differences of pixel values in an image are [99].

a) Entropy reflects the randomness of intensity distribution and complexity and the non-uniformity of image texture [100][101] *(Eq. 26)*.

$$Entropy = \sum_{m=0}^{G-1} \sum_{n=0}^{G-1} p(m,n) \log p(m,n) \quad (26)$$

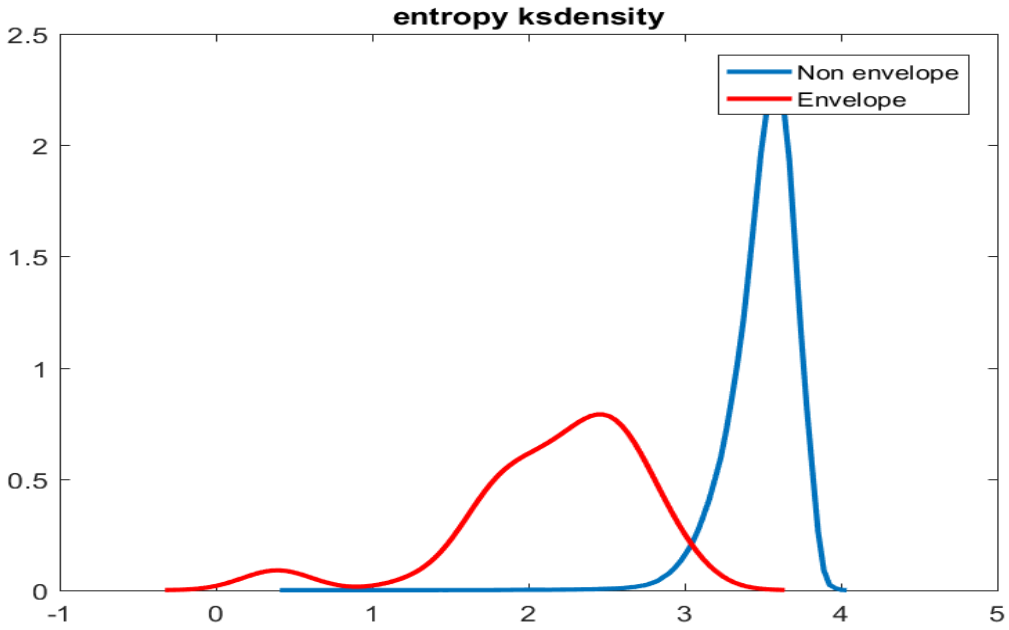

Figure 48. Entropy probability density estimate for the training data (*E* and *nE* peaks).

It can be observed (*Figure 48*, *Figure 49*) that below the 2.6 entropy value, ~90% are the peaks that members of isotopic envelopes.

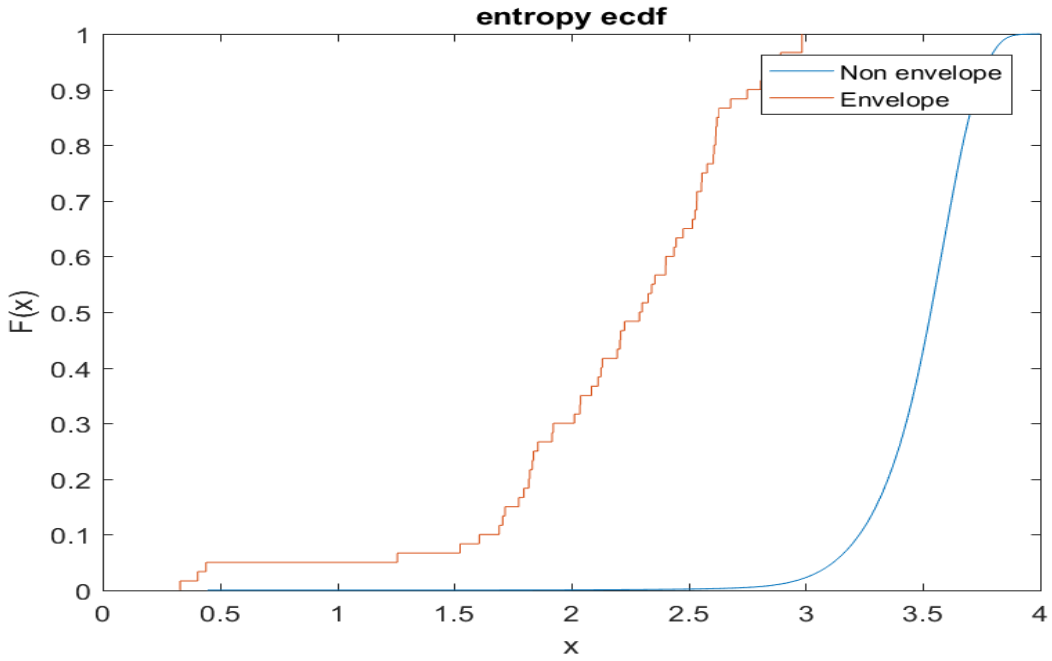

Figure 49. The empirical cumulative distribution function of entropy feature for the *E* and *nE* peaks.

b) Energy carries an information about the grayscale distribution homogeneity of an image and the crudeness of texture [100] [101] *(Eq. 27).*

$$\text{Energy} = \sum_{m=0}^{G-1} \sum_{n=0}^{G-1} p(m,n)^2 \qquad (27)$$

3. Statistical

These are the descriptive statistics descriptors based on histogram analysis. Their regions' textures are described by higher-order moments of their grayscale histogram. [99]

- Mean informs about the mean intensity level of the image or texture.

  Eq. 28. is related to the mean of reference pixels, whereas *Eq. 29* refers to the mean based on the neighbour pixels. Both values are identical in the case of the symmetrical GLCM. [95]

$$\mu_x = \sum_{m=0}^{G-1} m \sum_{n=0}^{G-1} p(m,n) \quad (28)$$

$$\mu_y = \sum_{n=0}^{G-1} n \sum_{m=0}^{G-1} p(m,n) \quad (29)$$

to

- Standard deviation *(Eq. 30, Eq. 31):*

$$\sigma_x = \sum_{m=0}^{G-1} (m-\mu_x)^2 \sum_{n=0}^{G-1} p(m,n) \quad (30)$$

$$\sigma_y = \sum_{n=0}^{G-1} (n-\mu_y)^2 \sum_{m=0}^{G-1} p(m,n) \quad (31)$$

- Variance (σ_x^2, σ_y^2) denotes the dispersion of the pixel values around the mean (difference between the reference and neighbour pixels)[99].

- Correlation *(Eq. 32)* indicates if there is a linear and predictable relationship between the neighbouring pixels[99] and if an image texture is consistent[101].

$$\text{Correlation} = \frac{\sum_{m=0}^{G-1}\sum_{n=0}^{G-1} mnp(m,n) - \mu_x\mu_y}{\sigma_x\sigma_y} \quad (32)$$

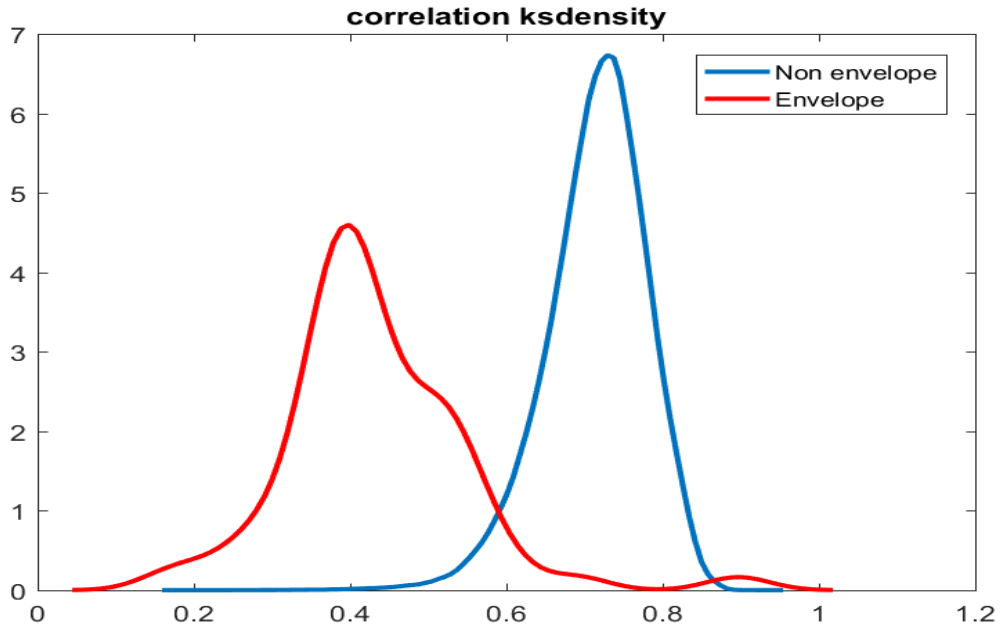

Figure 50. Correlation probability density estimate for the training data *E* and *nE* peaks).

In 90% of cases, the envelope peaks are in the range (0; 0.55), whilst the peaks not included in the isotopic envelopes are over 0.55 correlation values (*Figure 50, Figure 51*).

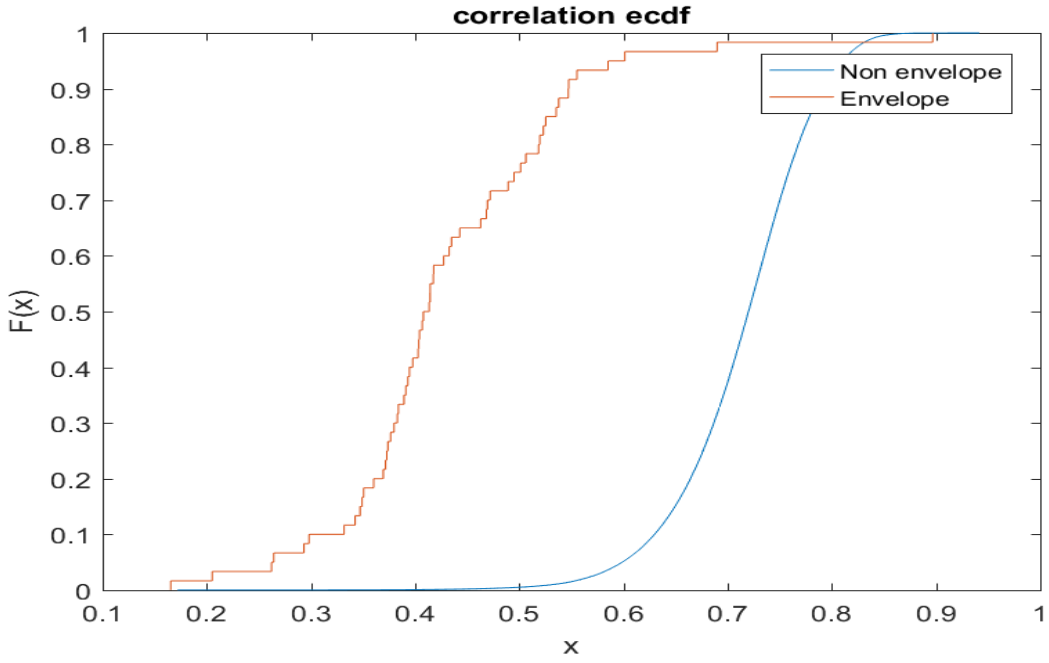

Figure 51. The empirical cumulative distribution function of correlation metric for the *E* and *nE* peaks.

- The central second moment (moment about the mean) is used to describe an image histogram's shape *(Eq. 33)*[94].

$$\mu_2 = \sum_{i=0}^{L-1} (z_i - m)^2 \, p(z_i) \qquad (33)$$

where:

z_i – discrete random variable denoting intensity levels in an image,

$p(z_i)$, $I = 0, 1, 2, \ldots, L\text{-}1$ – corresponding normalised histogram,

L – number of possible intensity levels,

$p(z_i)$ – a histogram component (estimation of the probability of the occurrence of intensity value z_i),

m – mean.

- Median is the middle measurement (pixel intensity) in an ordered set of image pixels [104].

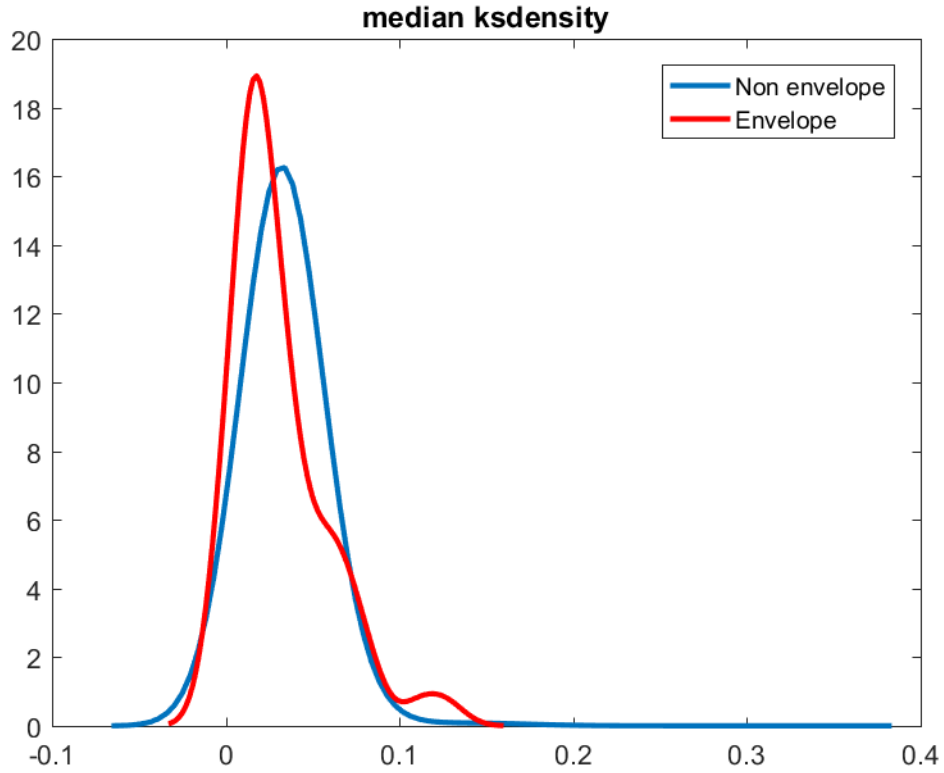

Figure 52. Median probability density estimate for the training data *E* and *nE* peaks).

Median seems to be the descriptor that differentiates the least between envelope and non-envelope peaks (*Figure 52, Figure 53*).

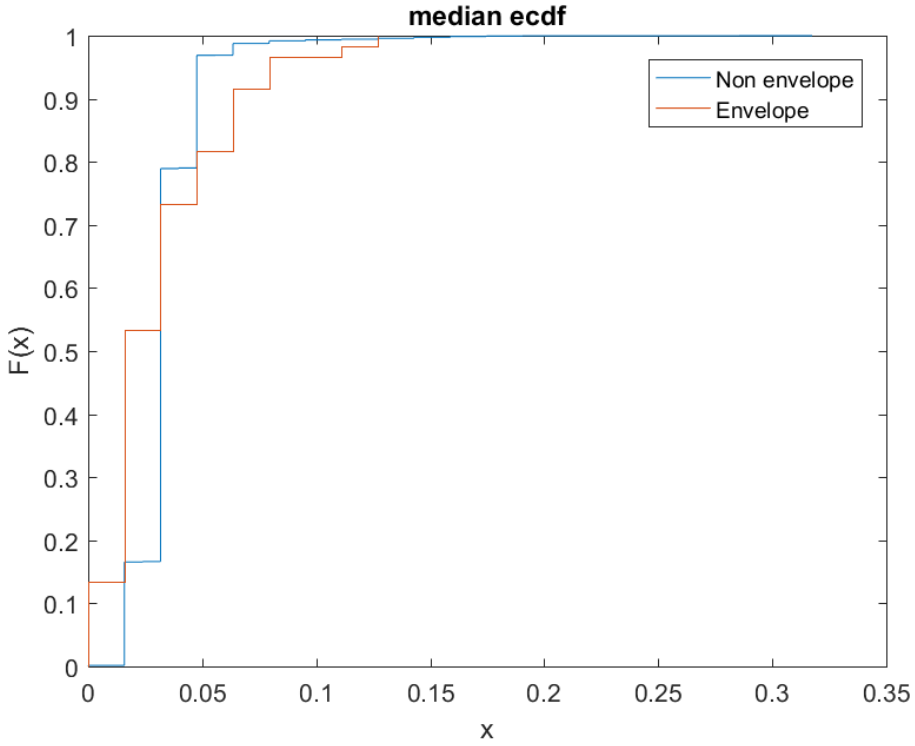

Figure 53. The empirical cumulative distribution function of median for the *E* and *nE* peaks.

- Interquartile range (IQR) is the distance between the first and third quartiles of pixel intensities *(Eq. 34)*[104].

$$IQR = Q_3 - Q_1 \quad (34)$$

- Coefficient of variation (cV) *(Eq. 35)* is a measure of variability. The variance and the standard deviation have magnitudes dependent of the magnitudes of the data. [104]

$$c_V = \frac{\sigma}{\mu} \times 100\% \quad (35)$$

In *Figure 54*, an analysis of image texture metrics is presented. What is more, isotopic envelopes membership was also considered. Below every arrow, it is shown whether low and high values determine the peaks that are members of isotopic envelopes (*E*) or non-Envelope peaks (*nE*). [53]

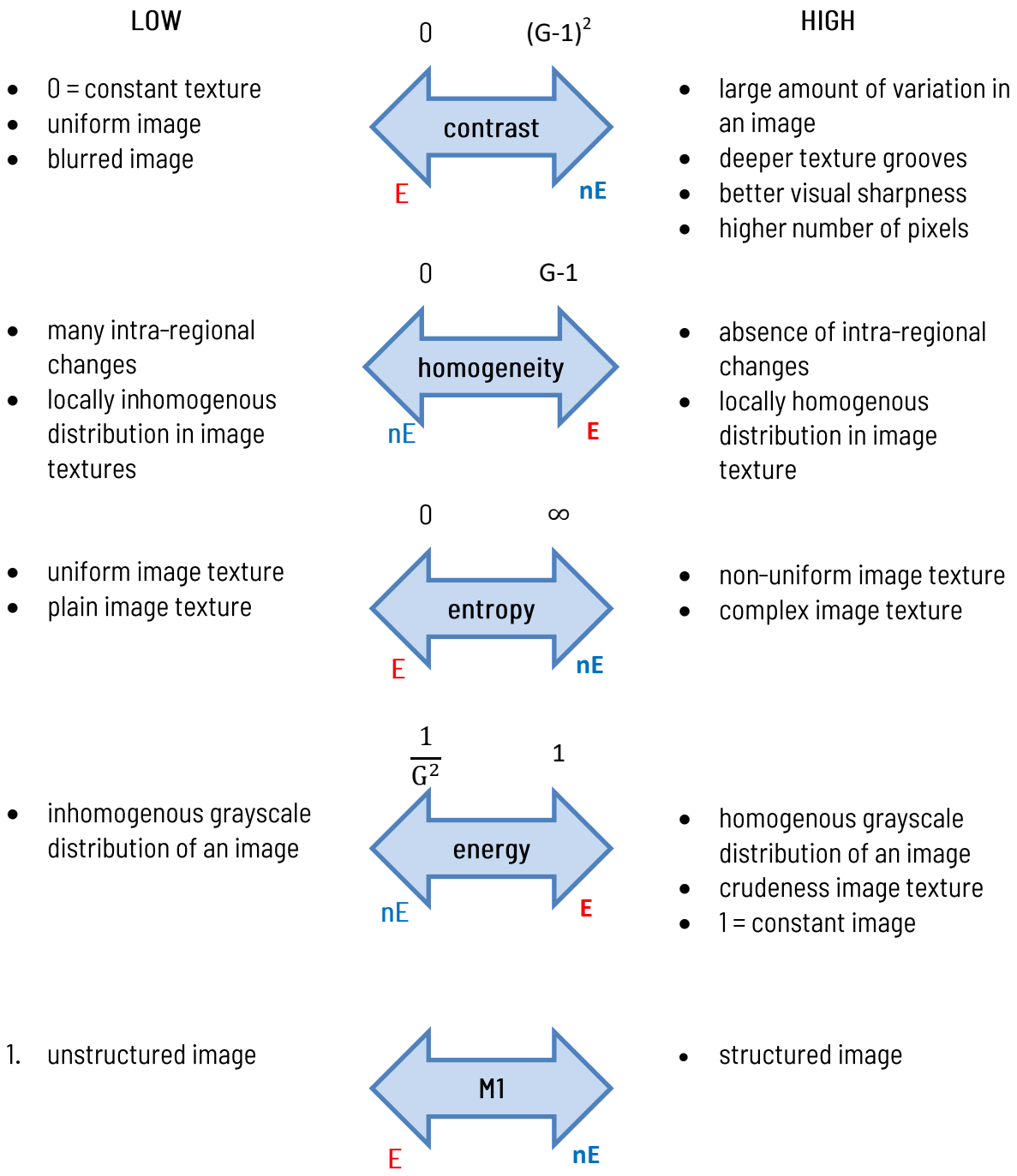

Figure 54. Image texture metrics from contrast- and order-type groups interpretation in relation to isotopic envelopes membership. Interpretation based on [53][95][96][101][103].

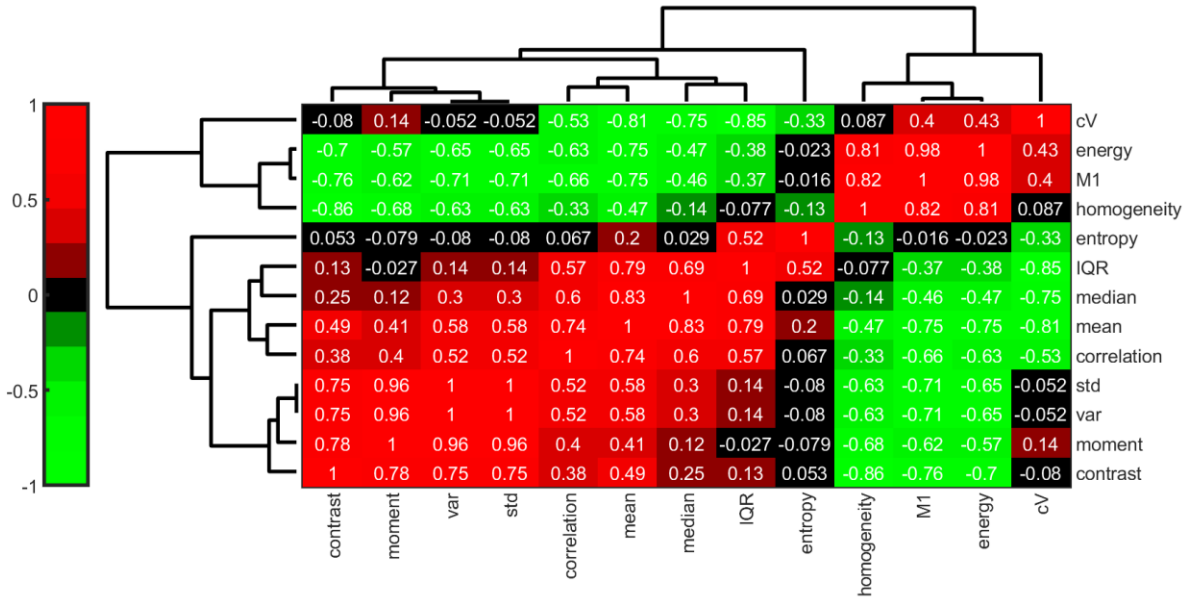

Figure 55. Clustergram of Spearman's rank correlation coefficient of the image texture metrics [53].

In order to assess the correlation between the descriptors, Spearman's rank correlation coefficient was calculated [105] (*Figure 55).* The vast majority of texture descriptors within the same group are strongly correlated because of the similar way of calculation [96]. Nonetheless, variance is closely correlated to the measures included in the contrast group [96] (contrast r = 0.7487, homogeneity r = -0.6327 and M1 r = 0.7068). Between contrast and homogeneity, a negative correlation is expected [96], and it turns out that homogeneity is strongly correlated to contrast (r = -0.8623). Entropy is more independent of other texture measures since $r < 0.5194$ for every other descriptor. Those descriptors can be used profitably in combination with each other. For classification purposes, at least one texture descriptor from each group should be chosen for further analysis. [53] [96]

4. Autocorrelation function

An autocorrelation function can be used directly for textures similarity comparison [100].

The image's autocorrelation function applies to the texture's coarseness of fineness evaluation in the image. It is also used to identify periodic textures in the image, indicating some repeated element models of texture in the image. [95]

Images with a rough texture are characterised by progressively falling autocorrelation function (*Figure 56*), whereas the shape of the function for images with no rough texture falls hastily (*Figure 57*) *(Eq. 36)*[95].

$$\rho(x,y) = \frac{\frac{1}{(N_i - |x|)(N_j - |y|)} \sum i \sum j \, I(i,j) I(i+x, j+y)}{\frac{1}{N_i N_j} \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} I(i,j)^2} \qquad (36)$$

where:

I(i,j) – the grey value of pixel *(i,j)*

$N_i \; x \; N_j$ – size of the image

x, y – shifts.

In order to find high-frequency variations of the data and to avoid any drift caused by spatially variable illumination, the mean of the image should be removed before performing the calculations [100].

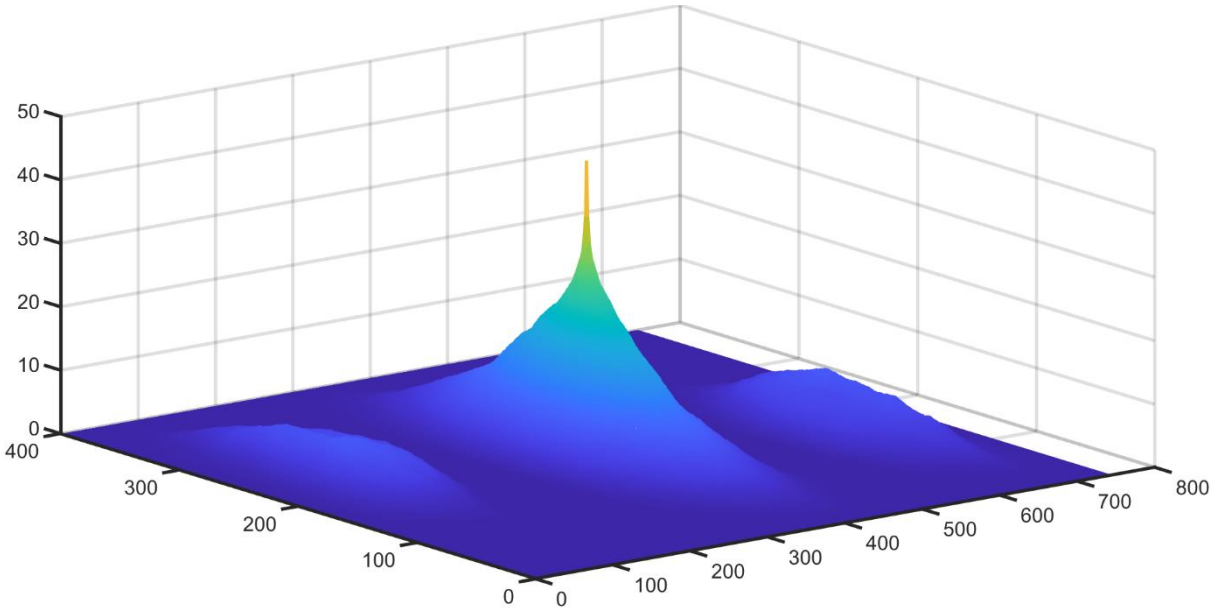

Figure 56. Normalised autocorrelation function for exemplary envelope peak.

In order to find similarities between the autocorrelation functions of two adjacent model components (peaks), the correlation coefficient was employed [100].

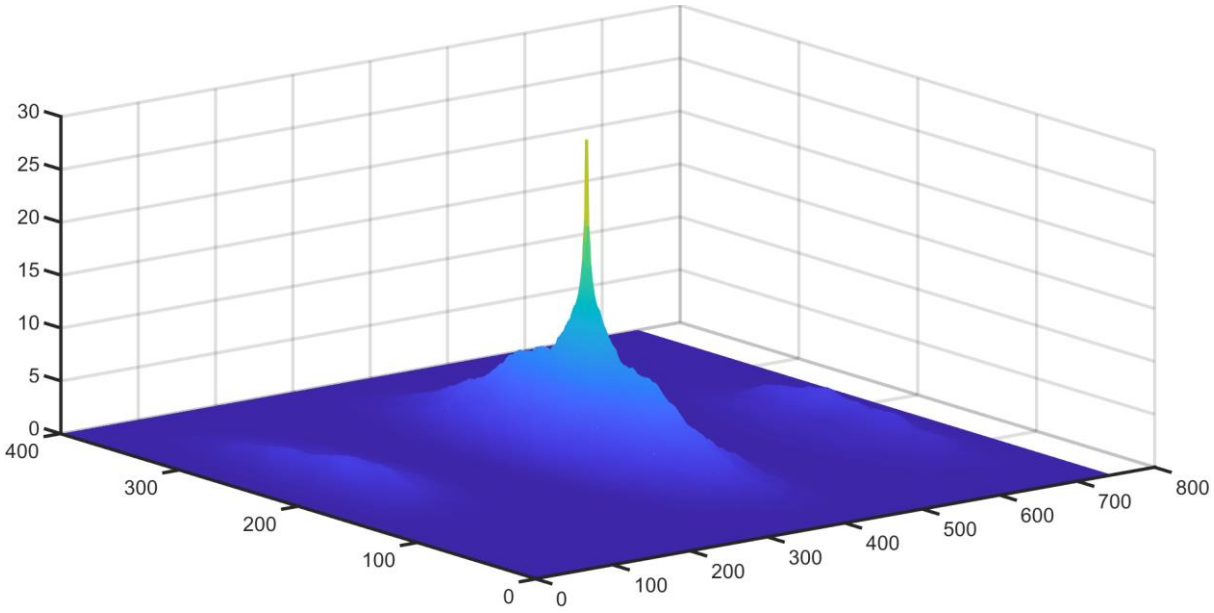

Figure 57. Normalised autocorrelation function for exemplary non-envelope peak.

Structural similarity index (SSIM)

SSIM is the only metric that is not calculated based on the differential image but on the separate peaks' images. It measures the two images' local brightness (luminance), structure, and contrast separately. Then, all those local assessments are combined into one overall measure [106]. It ranges from 0 (no similarity between compared images - non-envelope peaks) to 1 (perfect match between two images - isotopic envelope members) *(Eq. 37).*

$$\mathrm{SSIM}(x,y) = [l(x,y)]^{\alpha}[c(x,y)]^{\beta}[s(x,y)]^{\gamma} \quad (37)$$

where:

α, β, γ – exponents,

l – luminance,

c – contrast,

s – structure.

Descriptors selection

In order to avoid redundancy in descriptors, a descriptor selection method was employed based on the wrapper approach. In such an approach for descriptor evaluation, the clustering algorithm is used. This approach incorporates a search component (sequential forward selection) wrapped around Naïve Bayes model clustering. As a descriptor evaluation criterion, the best accuracy was taken into consideration. [107]

One of the commonly used algorithm for feature selection is *forward selection* (also called *forward stepwise selection*) [53] [108] [109] [110]. This method starts with the empty set of descriptors, and add descriptors one by one greedily [108].

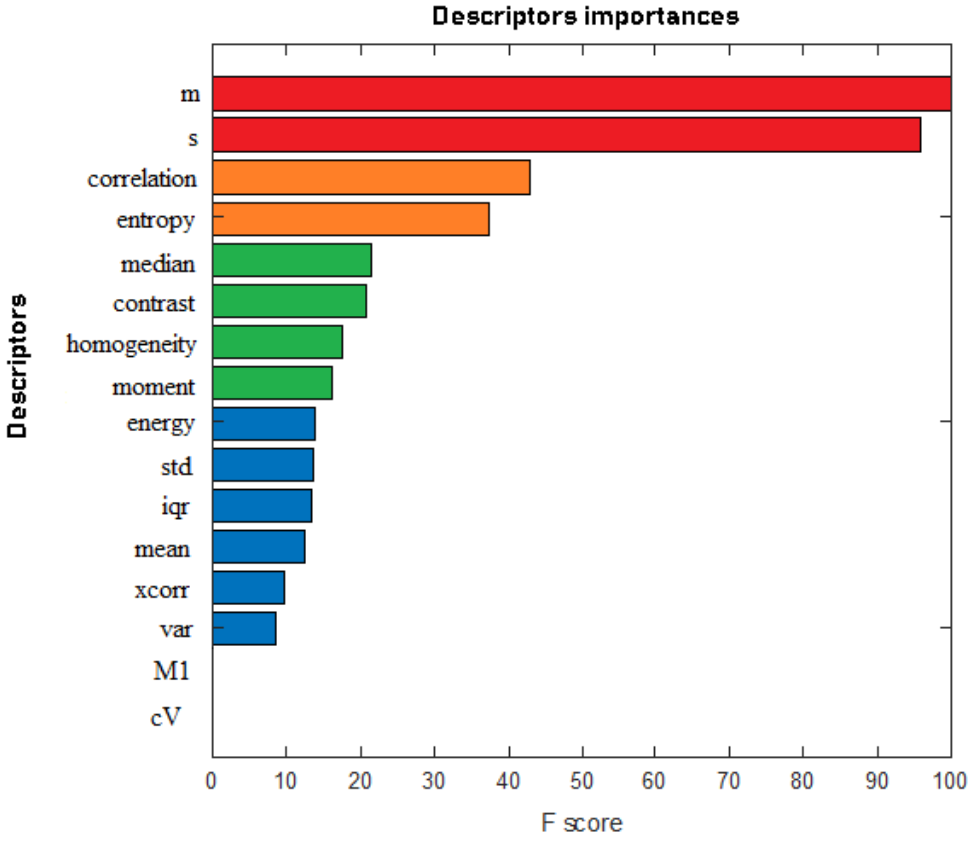

Figure 58. Descriptors importances [53].

Eighty models were randomly generated, and the feature selection method was applied to each. According to *Figure 58*, the descriptors are divided into four clusters (red, orange, green, and blue). Once this step was carried out, only eight descriptors out of 17 were selected for further analysis [53]:

1) distance of the means of adjacent model components *(m)*
2) estimated variances ratio of adjacent model components (*s*)
3) correlation
4) entropy
5) median
6) contrast
7) homogeneity
8) moment.

5.2.2. Classifier construction

Supervised classification is essential from a data analysis point of view. Generally, a classifier assigns a class label to instances that are described by variables. [111]

In order to classify peaks into the envelope and non-envelope classes, a supervised learning approach was employed. Three types of classifiers were tested: Naïve Bayes (NB) with Epanechnikov kernel function, Support Vector Machine (SVM) with cubic kernel function, and Decision Tree (DT). All of them were tested on the testing data set, which was not used in the training process. NB is the generative model, whereas SVM and DT are discriminative. All the metrics were calculated after the completion of the fuzzy-inference decision process. As a consequence, the values of the metrics are only related to the classification step, resulting in not considering non-envelope peaks removed after the fuzzy-inference decision step.

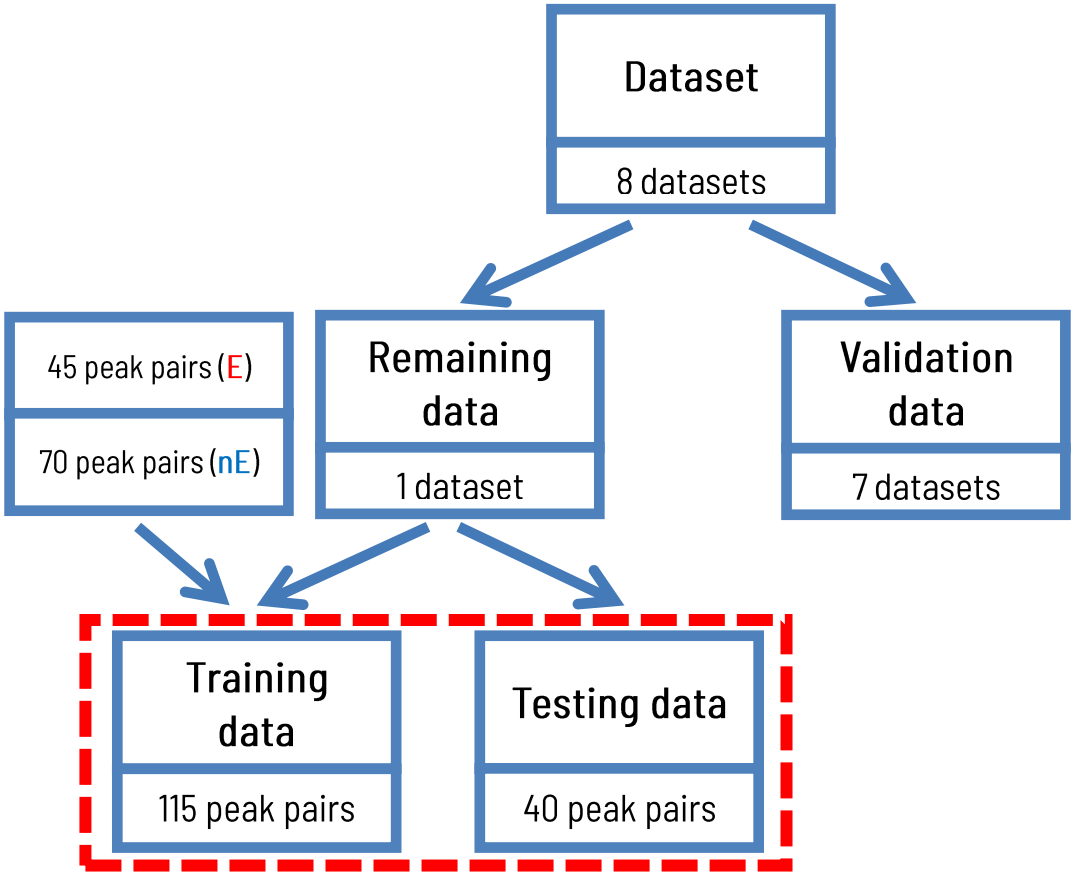

Figure 59. Dataset division scheme [53].

The process of building the model includes a training phase and a testing phase. Training and testing data are involved in the creation of the model. The dataset was divided into three sets: testing data, training data, and validation data (*Figure 59*). The testing data is a sample of data used to evaluate the final model fitted to the training data. Validation data was used to evaluate the final model fit. It is worth noting that the validation dataset does not affect the model creation process.

Training data was used to fit a model. This dataset was created in the following way: the matrix of differential images was created. Each differential image was based on subtracting two peaks from each other. Once an expert had annotated the peaks, it is possible, at this stage, to assign exactly one label to each cell in the matrix, whether or not the differential image came from peaks that were members of an isotopic envelope (E) or not (nE). To validate the model, cross-validation, with the division of the entire dataset into five subsets, was performed. Then, the model was trained 100 times on each subset. The performance measure (accuracy) of overall training subsets was calculated.

In order to make the calculations less complex, differential images that contain peaks that are not so far apart from each other were only included in the further calculations (working range in yellow – *Figure 60*) because it is implausible that these peaks are members of an isotopic envelope if the distance between the means of adjacent model components is considerable. In nature, the distance between peaks enclosed in the same isotopic envelope is approximately 1 Da. Then, the descriptors described in section II are calculated for the differential images obtained in the previous step.

M/z	799.7757	799.8721	800.5334	...	842.6263	843.6202	...
799.7757	-	nE	nE	...	nE	nE	...
799.8721	nE	-	nE	...	nE	nE	...
800.5334	nE	nE	-	...	nE	nE	...
...	...	...	...	...	...	...	...
842.6263	nE	nE	nE	...	-	E	...
843.6202	nE	nE	nE	...	E	-	...
...	...	...	...	...	...	...	...

Figure 60. Matrix of a differential image.

- **Naïve Bayes**

The Naive Bayes classifier is a probabilistic classifier based on the Bayes' theorem. The reason for choosing this classifier to decide which peaks are members of isotopic envelopes was that it is advantageous when the dimensionality of the inputs is high. Moreover, it can be successfully applied to large datasets due to its ability to perform with high accuracy. [112]

In the case of this work, objects (peaks) should be classified into one of two classes: envelope (C_1) and non-envelope (C_2). Objects $X = (x_1, x_2, \ldots, x_n)$ are characterised by their descriptors (distance of the means of adjacent model components, estimated variances ratio of adjacent model components, median, entropy, contrast, correlation, homogeneity, moment). The classifier predicts that the object is included in the category because of the highest value of the posterior probability conditioned on that object. [113]

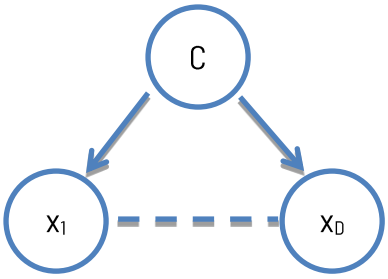

Figure 61. A graphical representation of the naive Bayes model for classification **[53][114].**

In *Figure 61*, the key assumption of the naïve Bayes model can be observed: conditioned on class C, the distributions of the input variables (peaks) $x_1, \ldots, x_n$ are independent [114].

Object X is classified into category C_i if and only if *(Eq. 38)* [113]:

$$P(C_i|X) > P(C_j|X) for\ all\ j \neq i \qquad (38)$$

It can be expressed by Bayes' theorem in the following way *(Eq. 39)* [113]:

$$P(C_j|X) = \frac{P(X|C_j)P(C_j)}{P(X)} \qquad (39)$$

which means *(Eq. 40)*:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{the probability of X being observed}} \qquad (40)$$

Due to this assumption, the classifier is called 'naïve', since this is an assumption that the features (predictors) are independent [113].

In order to represent the prior probability, in case the predictors are non-binary and no assumptions of normality of the distribution can be made, a kernel density estimator (KDE) is used *(Eq. 41)*[115].

$$\hat{f} = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad (41)$$

where:

n – sample size,

h – bandwith.

The true density of variables is estimated using Epanechnikov kernel function *(Eq. 42)*[116][117][118].

$$K(x) = \frac{3}{4}(1 - x^2), |x| \leq 1 \quad (42)$$

The main formula of the naïve Bayes is an attribute conditional independence *(Eq. 43)*[113]:

$$P(X|C_j) = \prod_{k=1}^{n} P(x_k|C_j) \quad (43)$$

The essential part of the Bayesian model formulation is $P(X|C_j)$ determination (time-consuming and computationally expensive), which in that classifier is simplified by using the equation [113].

The process of classification peaks to the classes is as follows [113]:

1. Training data is used for determining the probabilities $P(C_j)$
2. Category conditional probabilities for discrete variables are based on training data
3. For each $C_j, P(X|C_j)P(C_j)$ are calculated in order to classify the unclassified object (peak) to one of the class
4. The unclassified object (peak) is assigned to the category which attains the largest score.

- **Support Vector Machines (SVM)**[112][119]

SVM is based on finding a hyperplane between the data points in *N*-dimensional space (*N* is the number of features) to separate the data points into two classes. Separability implies the existence of margins delimited by two parallel hyperplanes, inside which not a single data item lies. The margin should be as wide as possible. [112][119]

That works aims to denote the peaks included in isotopic envelopes. Thus, the number of classes is two: *E* and *nE*. SVM with cubic kernel function was created. The Cubic SVM hyperplane is based on a three-order polynomial [112] [119].

Given the data in the form of label-feature pairs *(Eq. 44)* [112] [119],

$$(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x}_n, y_n), y_i \in \{-1, 1\}, \boldsymbol{x_i} \in R^p \quad (44)$$

where the classes (*E* and *nE*) were denoted as 1 and -1.

The goal is to find a decision function which correctly predicts the label y of an input feature $\boldsymbol{x}$ *(Eq. 45)* [112] [119]:

$$f(\boldsymbol{x}) = \begin{cases} +1 \; when \; h(\boldsymbol{x}) > 0, \\ -1 \; when \; h(\boldsymbol{x}) < 0. \end{cases} \quad (45)$$

The separating surface is defined as follows *(Eq. 46)* [112] [119]:

$$H = \{\boldsymbol{x} \,|h(\boldsymbol{x}) = 0\}. \quad (46)$$

- Decision Tree

This kind of classifier is based on binary selections that correspond to the traversal of a tree structure [114]. From the mathematical point of view, a *tree* is defined as an undirected acyclic connected graph. A decision tree is a directed tree with an apex (called 'a root'), which is the initial apex of the tree. [119]

Generally, the decision tree partitions the input space (data) hierarchically until it reaches a subspace associated with a class label (envelope or non-envelope) by traversing the tree beginning at the root [112]. The whole data sample is concentrated in the tree root, and then the subsequent data items (peaks) are moved along the branch, from top to bottom, through the nodes, in each of which a decision is made to choose the branch along which the data item will move. In this way, at each node (which is not a leaf), the data items reaching that node are divided into subgroups. Under each node, there is a criterion for dividing the subgroup reaching this node into smaller subgroups reaching the child nodes based on the data's attributes. The splitting rule made in a given node is the same for all elements of the sample that are in that node. The sample elements are moved to an end node, the tree's leaf, which is usually labelled with the class of the analysed discrimination problem, from which most data items that reached this leaf originate. [112] [119]

In order to denote the isotopic envelope member peaks, the 2-class (*E* and *nE*) decision tree of medium flexibility with a smaller number of leaves has been constructed (*Figure 62*). It was

constructed based on 115-element data sample, and the prior chosen features describe each data item: distance of the means of adjacent model components, estimated variances ratio of adjacent model components, median, entropy, contrast, correlation, homogeneity, and moment.

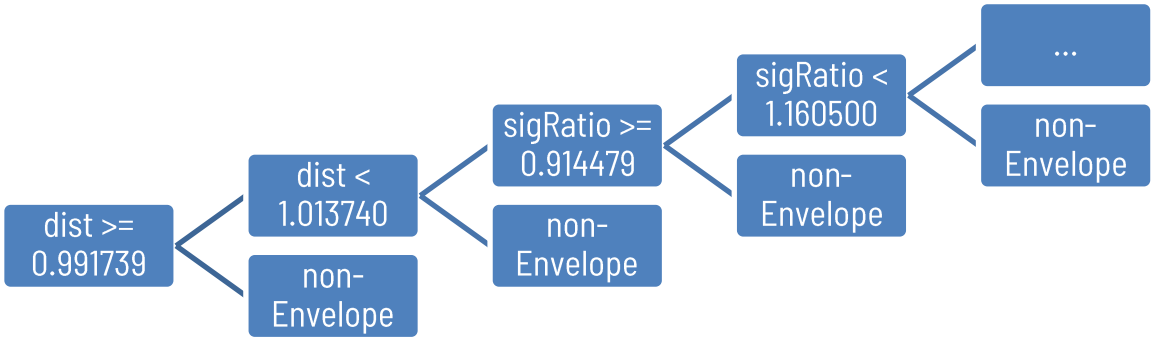

Figure 62. 2-class decision tree for determining the envelope member peaks.

In order to compare the three aforementioned classifiers, several statistics metrics have been calculated and gathered in *Table 8:*

1. True Positive (TP)
2. True Negative (TN)
3. False Negative (FN)
4. False Positive (FP)
5. Specificity = $\frac{TN}{TN+FP}$
6. Precision = $\frac{TP}{TP+FP}$
7. Recall = $\frac{TP}{TP+FN}$
8. Balanced Accuracy = $\frac{Recall+Specificity}{2}$
9. Critical Success Index (CSI): CSI = $\frac{TP}{TP+FN+FP}$
10. Matthews Correlation Coefficient (MCC) [120]: $MCC = \frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
11. Fowlkes-Mallows Index (FMI) [121]: $FMI = \sqrt{Precision\ x\ Recall}$

Table 8. Confusion matrix-based metrics.

	Naïve Bayes	SVM	Decision Tree
TP	128	110	29
TN	1437	126	1100
FN	5	23	104
FP	33	1344	370
Specificity [%]	97.76	8.57	74.83
Precision [%]	79.50	7.57	7.27
Recall [%]	96.24	82.71	21.81
Balanced Accuracy [%]	97.00	45.64	48.32
Critical Success Index (CSI) [%]	77.11	7.45	5.77
Matthews Correlation Coefficient [%]	86.26	-8.29	-2.15
Fowlkes-Mallows Index (FMI) [%]	87.47	25.01	12.59

It can be observed (*Figure 63*) that SVM has significant compliance with the results compared to the expert in annotating members of isotopic envelopes since the recall value is 82.71%. Nevertheless, it annotates many peaks as the envelope ones, whereas they are not included in isotopic envelopes – the value of specificity is notably low in comparison to other investigated classifiers (8.57%). Decision Tree classifier has a substantial compliance of non-envelope peaks annotation with the expert (specificity = 74.83%), but unlike the SVM classifier – it has low envelope peaks detection ability (recall = 21.81%). Both SVM and DT classifiers are characterized by low values of precision (7.57% and 7.27%, respectively). Therefore they cannot accurately classify peaks that are members of isotopic envelopes. Moreover, balanced accuracy values obtained for the two classifiers mentioned above are comparable (below 50%), which means that only less than 50% of peaks were classified correctly to the envelope or non-envelope classes.

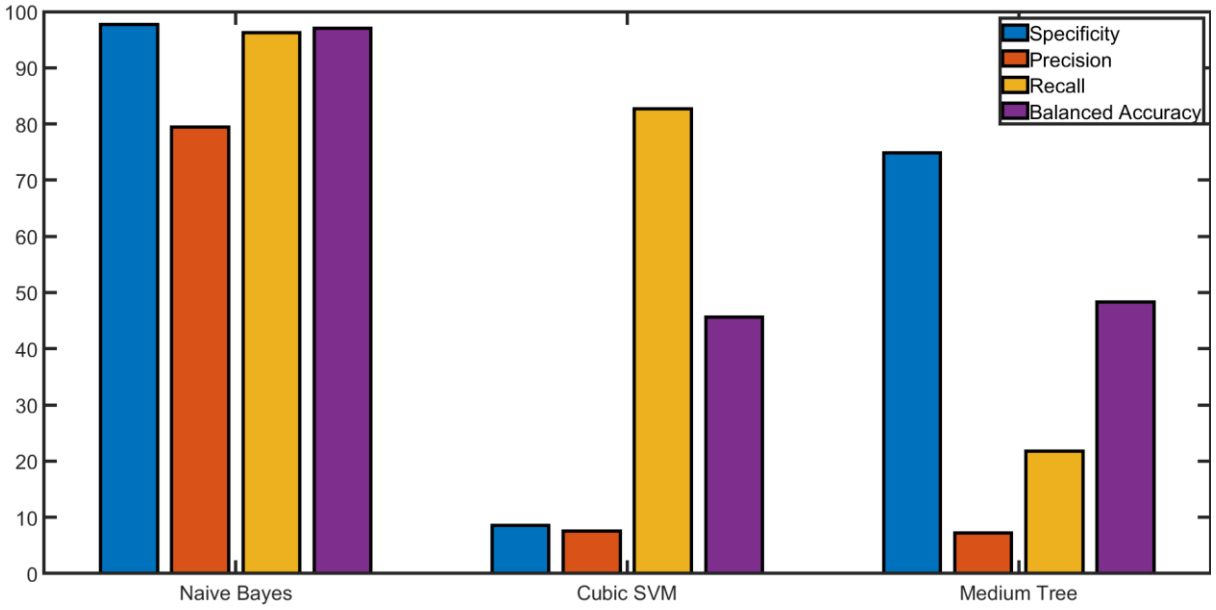

Figure 63. Comparison of four confusion matrix metrics for Naïve Bayes, Cubic SVM and Medium Tree classifiers.

Concerning the results of the metrics gathered in *Table 8*, the Naïve Bayes classifier has the best performance compared to SVM and DT classifiers due to the fact that all of the metrics are significantly high. The CSI value of the NB model is 77.11%, which is much better than that of SVM and DT. Additionally, Matthews Correlation Coefficient indicates that the quality of the NB model (regardless of the differences in the quantity of envelope and non-envelope classes) is remarkably better than those of SVM and DT. Moreover, the FMI of NB is notably higher than of the aforementioned two classifiers. Thus, this classifier seems to classify peaks as members of isotopic envelopes and non-envelopes correctly. The low performance of SVM and DT could be that DT is sensitive to outliers and likely overfit the data, whereas SVM classifies the data based on geometry. Moreover, because the weights of the variables are not constant in SVM, the contribution of each variable to the output is variant [122]. Generally, for many classification problems, the probability-based approach gives better results. Hence, the Naïve Bayes is further considered a classifier for defining envelopes' member peaks.

6. RESULTS

The proposed two-step method based on Mamdani-Assilan fuzzy-inference system potential isotopic envelope members preselection and verification based on the Naïve Bayes classifier was tested on eight datasets comprised of peptides data, mentioned in *3.1*. Results evaluation in *6.1* and *6.2* comprises of the combined results obtained in both method steps. The results were published in [53].

6.1. Results for Head and Neck Cancer – Fresh Frozen tissues peptide datasets

Four datasets of peptide-related data collected from patients suffering from head and neck cancer were used to test the proposed algorithm. After performing the first step of the method – preselection of the input peak pairs by the Mamdani-Assilan fuzzy-inference system, the number of peaks that should be considered as the potential isotopic envelope members decreased significantly – for these four datasets number of input peak pairs ranges from 46 350 to 55 070 and was diminished to the values ranging from 1 662 to 1 603 number of peaks.

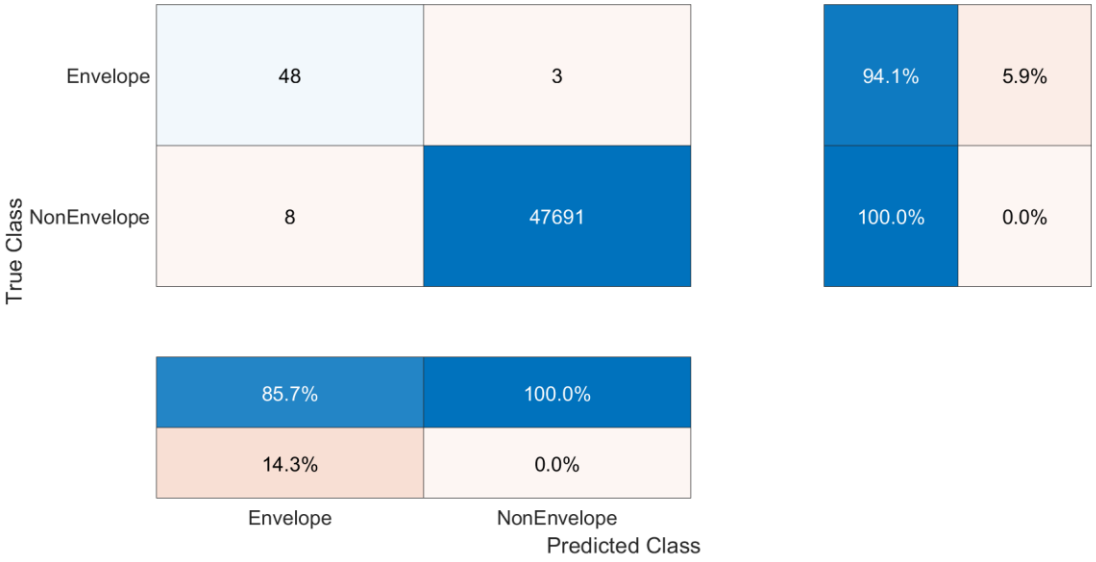

Figure 64. Confusion matrix for *HNC-FF Dataset 1* results [53].

In *Figure 64*, the confusion matrix for *HNC-FF Dataset 1* is presented, which contains combined results of the two steps of the method, as the significant number of peak pairs were removed from further analysis based on the first step of the method (Mamdani-Assilan fuzzy-inference system). The numbers included in the confusion matrix reflect the peak pairs – 48 peak pairs were classified as the members of isotopic envelopes by the expert and by the proposed two-step

method. Eight peak pairs were incorrectly classified as the isotopic envelope members. 47 691 peak pairs were correctly classified as not included in any isotopic envelope. Three peak pairs were incorrectly classified as the *non-Envelope* ones. Further steps of the analysis of the results are adjusted to the data characterised by the considerable number of negative predictions, since a significant number of peaks is not included in any isotopic envelope.

Table 9. Confusion matrix-based metrics for 4 *HNC-FF* peptide datasets **[53]**.

	HNC-FF Dataset 1	HNC-FF Dataset 2	HNC-FF Dataset 3	HNC-FF Dataset 4
TP	48	45	89	128
TN	47 691	46 289	53 896	54 900
FN	3	6	12	9
FP	8	10	33	33
Specificity [%]	99.98	99.98	99.94	99.94
Precision	85.71	81.82	72.95	79.50
Recall	94.12	88.24	88.12	93.43
Balanced Accuracy	97.05	94.12	94.03	96.69
Critical Success Index	81.36	73.77	66.42	75.29
Matthews Correlation Coefficient	89.81	84.95	80.14	86.15
Fowlkes-Mallows Index	89.82	84.97	80.18	86.19

In order to evaluate obtained results, several confusion matrix-based metrics were calculated and gathered in *Table 9*. According to every dataset's recall and precision values, the envelope-member peaks were classified accurately, ranging from 88.12% to 94.12% and from 72.95% to 85.71%. In comparison, the peaks not included in isotopic envelopes were classified with an accuracy of over 99% (specificity). Matthews Correlation Coefficient carries information on the model quality regardless of the differences in the number of elements included in two classes: *Envelope* and *non-Envelope*. It can be observed that a correlation between predicted values and those annotated by the expert ranges from 80.14% to 89.81%. In order to evaluate the obtained

results, Balanced Accuracy was calculated instead of the standard accuracy measure due to the fact that the sets of Envelope and non-Envelope peak pairs are imbalanced, as the number of *nE* is notably higher than the *E* set size. Balanced accuracy changes from 94.03% to 97.05%. Moreover, the Fowlkes-Mallows Index indicates a significant similarity (over 80%) between predicted values and those assessed by the expert. [53]

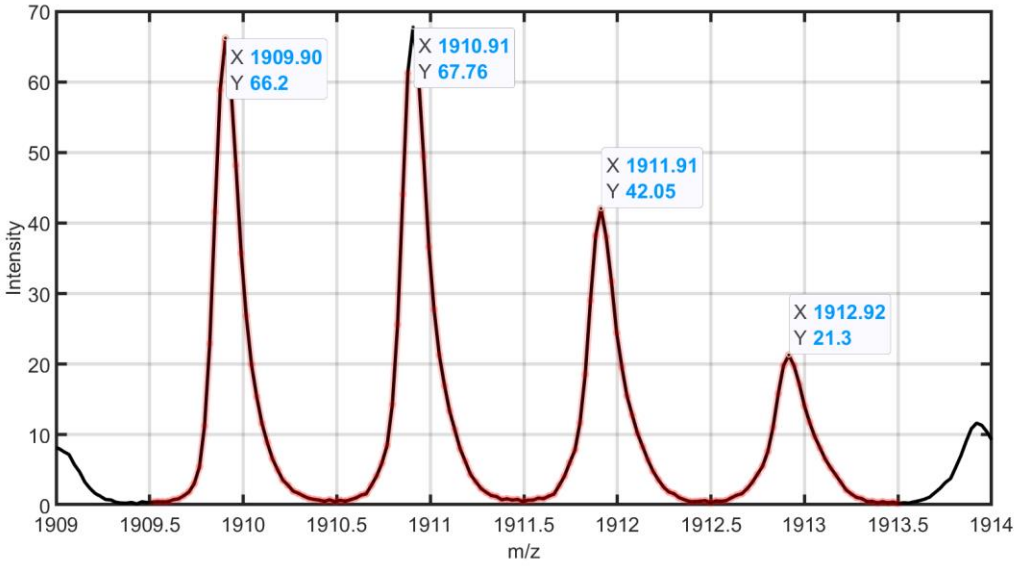

Figure 65. Exemplary isotopic envelope consisted of four member peaks.

In *Figure 65* exemplary isotopic envelope correctly classified by the proposed method is presented. This isotopic envelope comprises four peaks, where the second peak has an abundance higher than the first one. Nonetheless, the distance between adjacent model components is approximately equal to one, which was one of the assumptions that should be met if two peaks belong to the isotopic envelope.

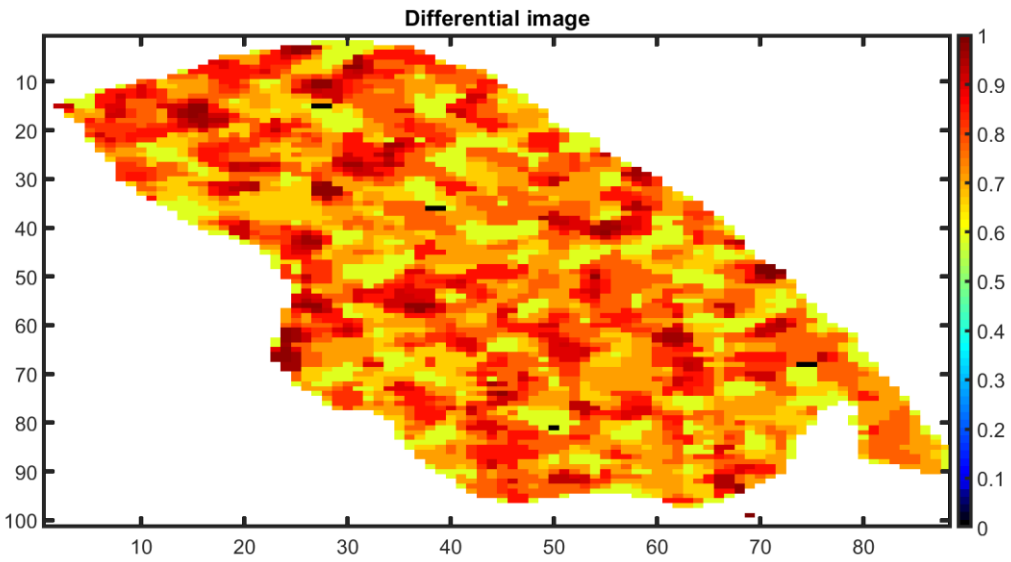

Figure 66. Differential image: abs | $m/z_1 - m/z_2$ | , where m/z_1 = 1909.90, m/z_2 = 1910.91.

In *Figure 66, Figure 67,* and *Figure 68,* the differential image (differential map of a spatial distribution) of the two adjacent peaks is presented.

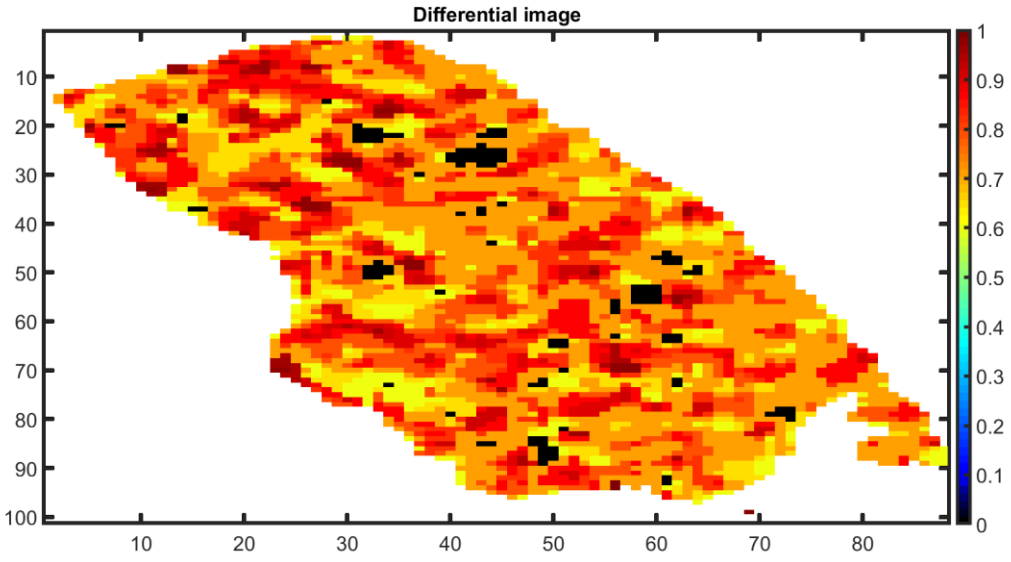

Figure 67. Differential image: abs | $m/z_1 - m/z_2$ | , where m/z_1 = 1910.91, m/z_2 = 1911.91.

It can be observed that there is no structurality visible in the aforementioned images that prove that from the peaks' spatial distribution point of view, those peaks should be considered as members of the isotopic envelope.

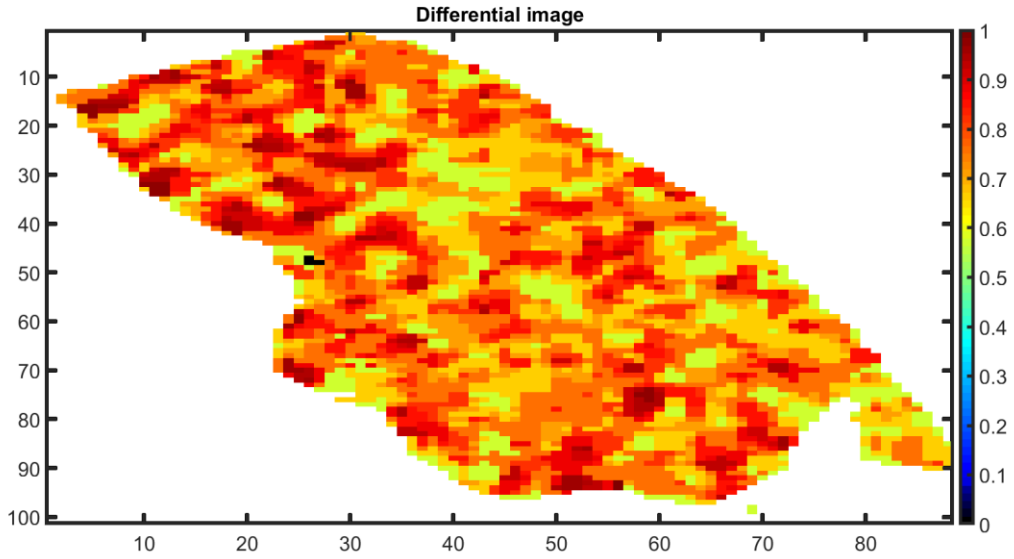

Figure 68. Differential image: abs | $m/z_1 - m/z_2$ | , where $m/z_1 = 1911.91$, $m/z_2 = 1912.92$.

6.2. Results for Head and Neck Cancer – Formalin-Fixed Paraffin Embedded peptide datasets

Four *HNC-FFPE* datasets of peptide-related data collected from patients suffering from head and neck cancer were used to test the proposed algorithm. After performing the first step of the method – preselection of the input peak pairs by the Mamdani-Assilan fuzzy-inference system, the number of peaks that should be considered as the potential isotopic envelope members decreased significantly – for these four datasets number of input peak pairs ranges from 55 030 to 97 910 and was diminished to the values ranging from 1 945 to 1 457 number of peaks.

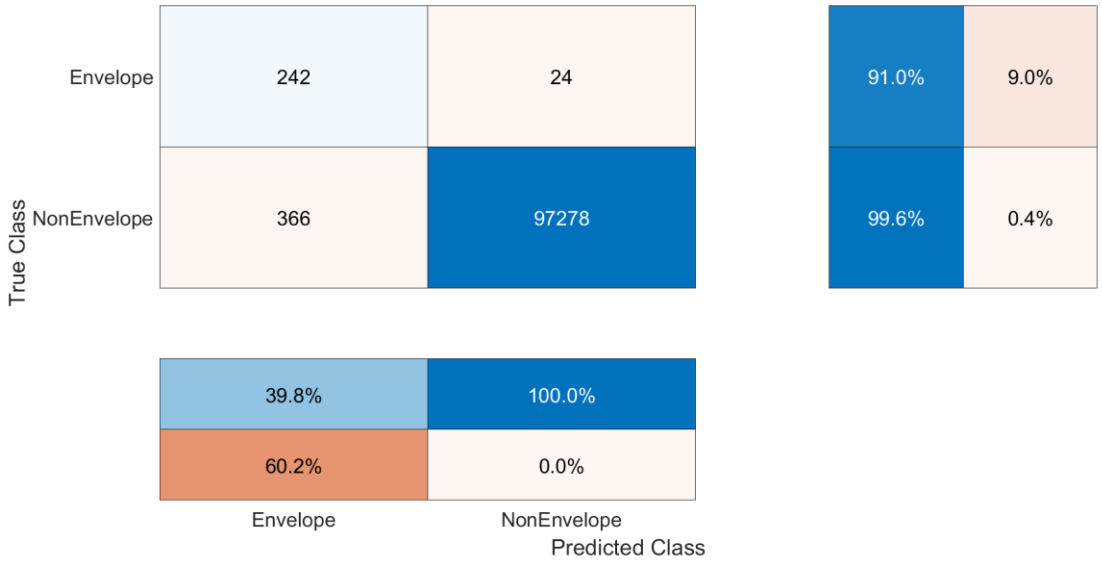

Figure 69. Confusion matrix for *HNC-FFPE Dataset 1* results [53].

In *Figure 69*, the confusion matrix for *HNC-FFPE Dataset 1* is presented, which contains combined results of the two steps of the method, as the significant number of peak pairs were removed from further analysis based on the first step of the method (Mamdani-Assilan fuzzy-inference system). The numbers included in the confusion matrix reflect the peak pairs – 242 peak pairs were classified as the members of isotopic envelopes by the expert and by the proposed two-step method. 366 peak pairs were incorrectly classified as the isotopic envelope members. 97 278 peak pairs were correctly classified as not included in any isotopic envelope. 24 peak pairs were incorrectly classified as the *non-Envelope* ones. Further steps of the analysis of the results are adjusted to the data characterized by the considerable number of negative predictions, since a significant number of peaks is not included in any isotopic envelope.

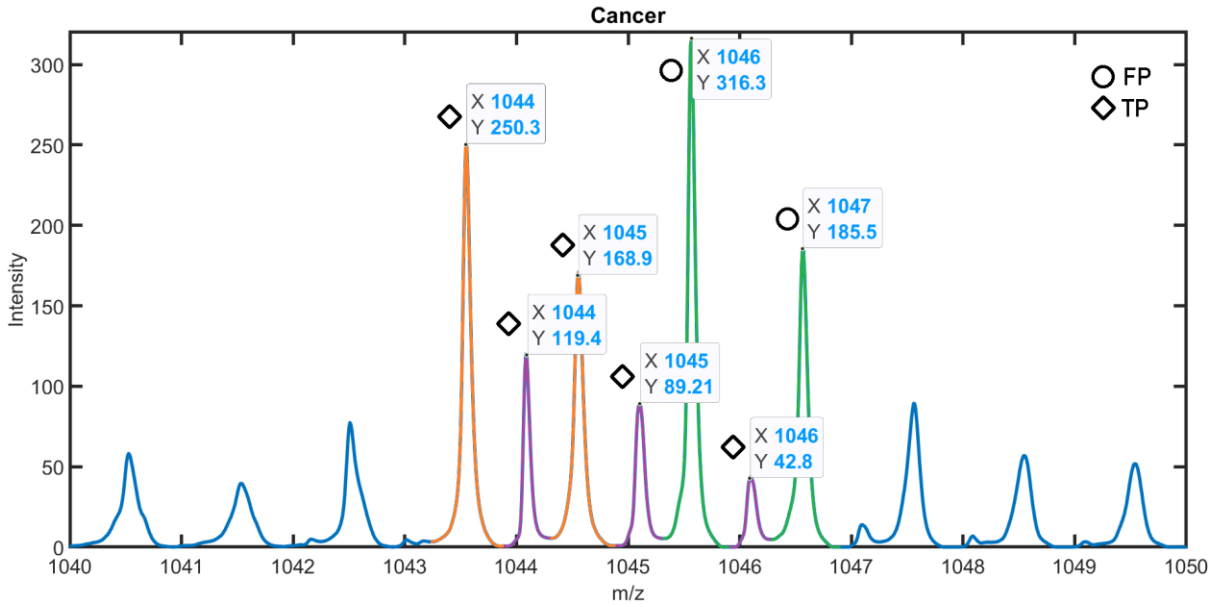

Figure 70. Exemplary overlapping isotopic envelopes in the m/z 1040-1050 mass range (orange – first isotopic envelope, violet – second isotopic envelope, green – third isotopic envelope). The reference for **FP** (False Positive) and **TP** (True Positive) is the expert annotation [53].

Figure 69 presents exemplary overlapping isotopic envelopes detected by the proposed method. It can be noticed that the expert did not annotate the green envelope. In *Figure 71, Figure 72, Figure 73,* and *Figure 74,* differential images (maps of spatial distribution) are pairwise presented for all peaks shown in *Figure 69.*

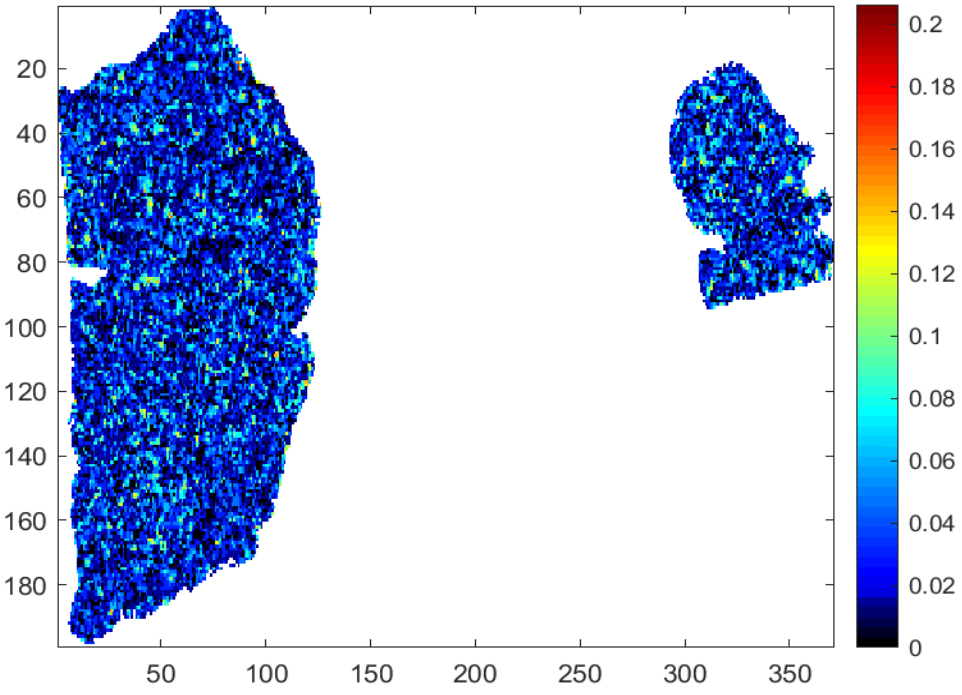

Figure 71. Differential image of two peaks marked in orange: $\mathrm{abs} \mid m/z_1 - m/z_2 \mid$, where $m/z_1 = 1044$, $m/z_2 = 1045$ [53].

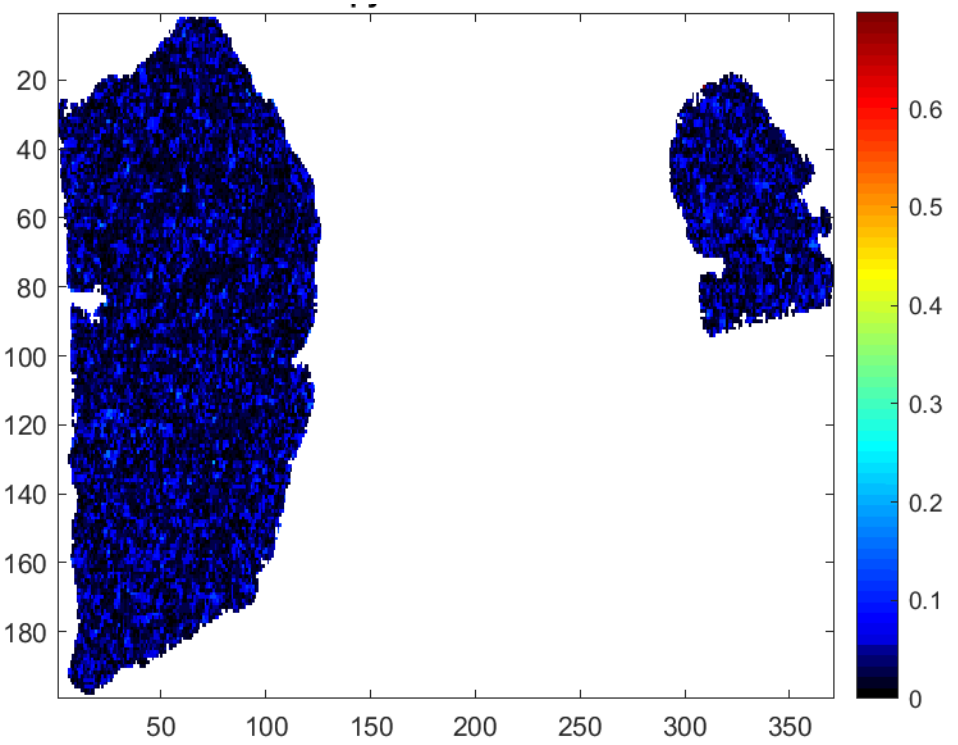

Figure 72. Differential image of two peaks marked in violet: abs | $m/z_1 - m/z_2$ |, where $m/z_1 = 1044$, $m/z_2 = 1045$ [53].

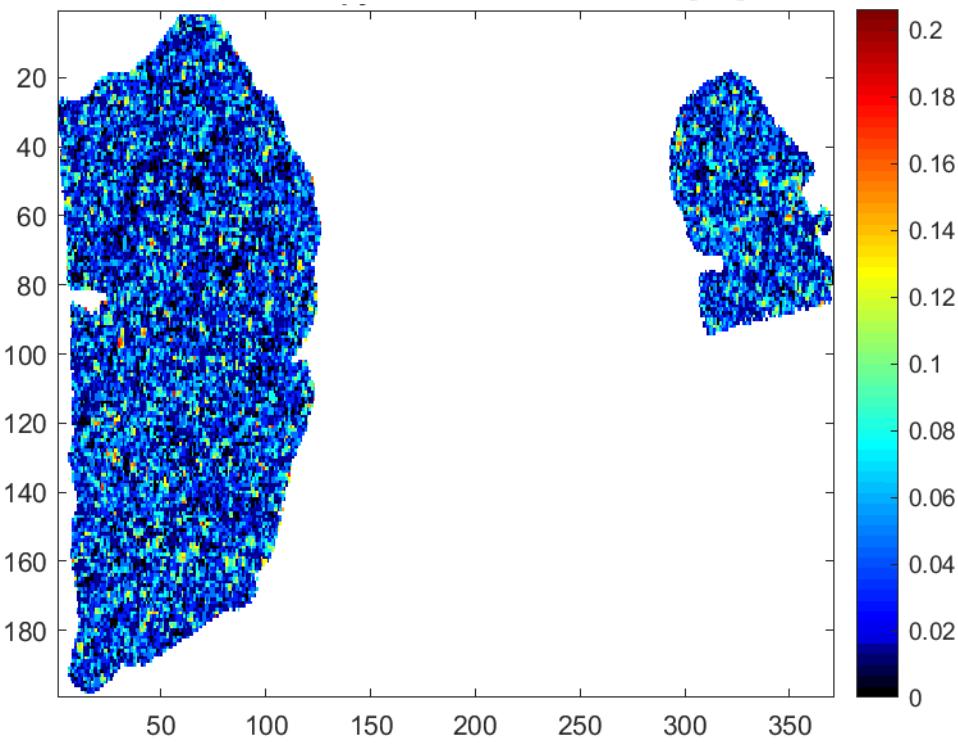

Figure 73. Differential image of two peaks marked in violet: abs | $m/z_1 - m/z_2$ |, where $m/z_1 = 1045$, $m/z_2 = 1046$ [53].

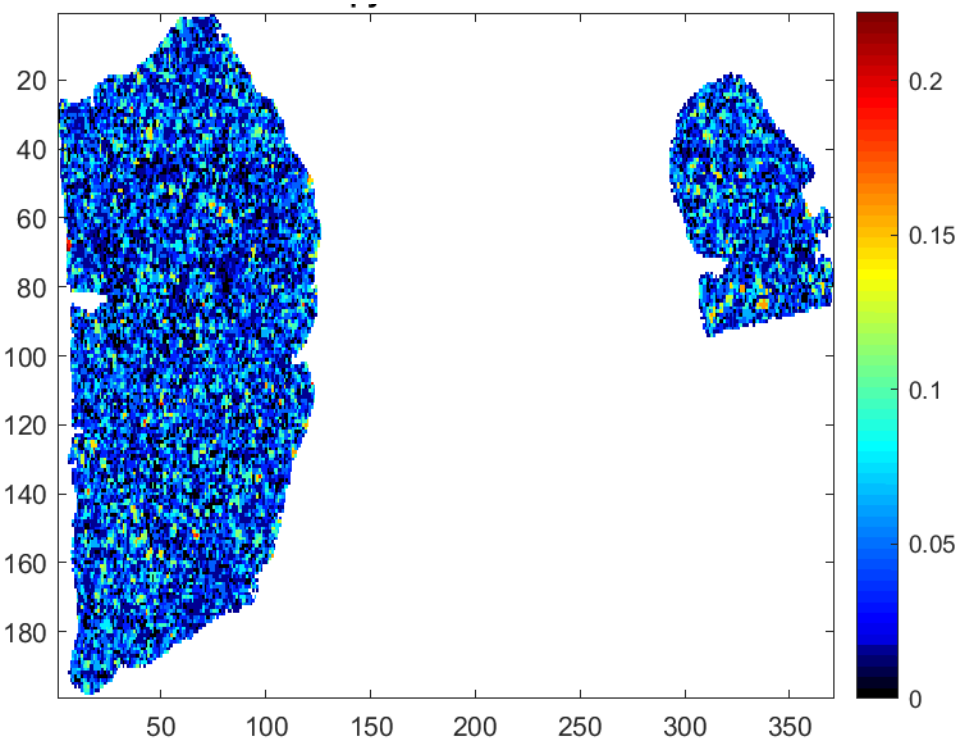

Figure 74. Differential image of two peaks marked in green: abs | $m/z_1 - m/z_2$ |, where $m/z_1 = 1046$, $m/z_2 = 1047$ [53].

It can be observed that all of the differential images prove that for every peak pair included in the isotopic envelopes mentioned above, after pairwise subtracting the spatial distribution maps, solely uniform intensity distribution is visible (or the intensities are at similar level), which indicates that a pair of peaks is included in the isotopic envelope.

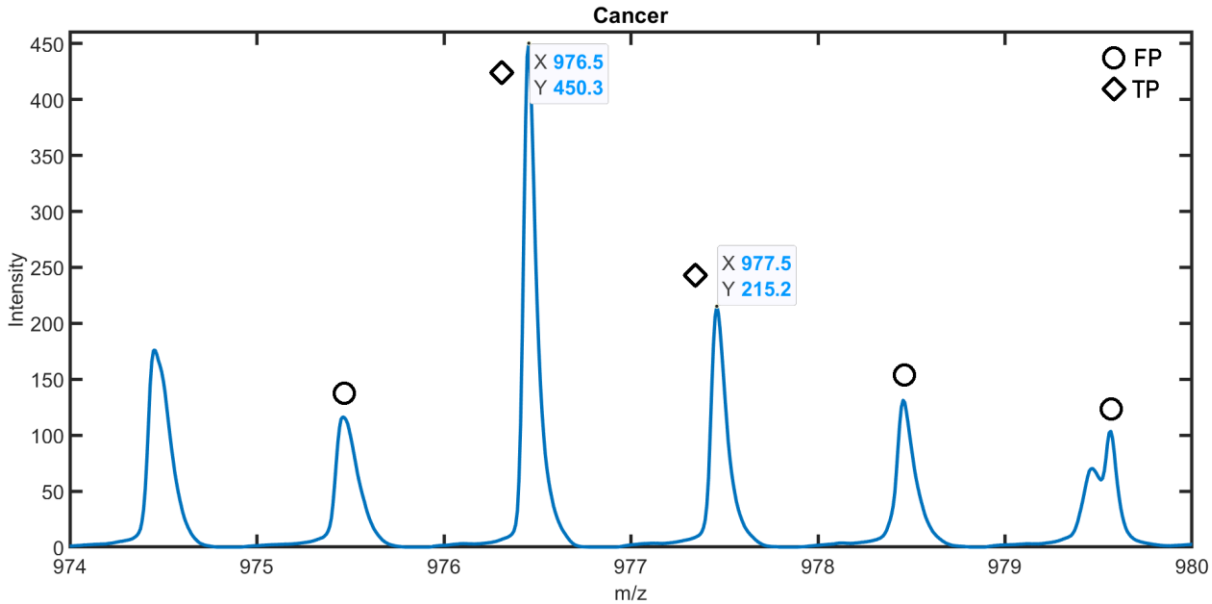

Figure 75. Exemplary isotopic envelope in the *m/z* 974-980 mass range.

In some cases, the isotopic envelopes were partially classified, as in *Figure 75.* At *m/z* 976.5 and 977.5, the complete agreement between the expert and proposed method is observed, unlike at *m/z* 975.5, 978.5, and 979.5, where the peaks were wrongly detected as the members of an isotopic envelope since not all requirements for being an isotopic-envelope member were met (abundances difference between adjacent peaks at *m/z* 975.5 and 976.5 and variance ratio of the peaks located at *m/z* 978.5 and 979.5; probably at *m/z* 979.5 after applying the pre-processing method of components merging, two components were merged which resulted in changing the shape of the peak). According to the analysis performed concerning the LC-MS experiment briefly described in 3.1, after comparing the list of MSI components with the list of identified peptides in the LC-MALDI-MS/MS measurement (briefly described in 3.1), it can be assumed that this isotopic envelope is probably derived from the peptide sequence K.AGFAGDDAPR.A.

Table 10. Number of detected isotopic envelopes with a given length in *HNC-FFPE Dataset 1* [53].

No. of peaks included in an isotopic envelope	No. of detected isotopic envelopes
2	295
3	64
4	38
5	11
6	2
7	1
11	1

In *HNC-FFPE Dataset* 1, the number of identified isotopic envelopes is 412. *Table 10* presents the number of detected isotopic patterns versus their length. It can be noticed that the vast majority of isotopic envelopes are comprises two-member peaks.

6.3. Discussion

This chapter discusses the advantages and limitations of the introduced two-step method for deisotoping based on a fuzzy-inference system and molecular spatial distribution of peaks. Presented results obtained for 4 *HNC-FF* datasets and 4 *HNC-FFPE* datasets show that regardless of the type of datasets – a way of the tissue procurement, preparation or preservation, and size of the dataset, their accuracy, and sensitivity of identifying peaks included in isotopic envelopes are similar.

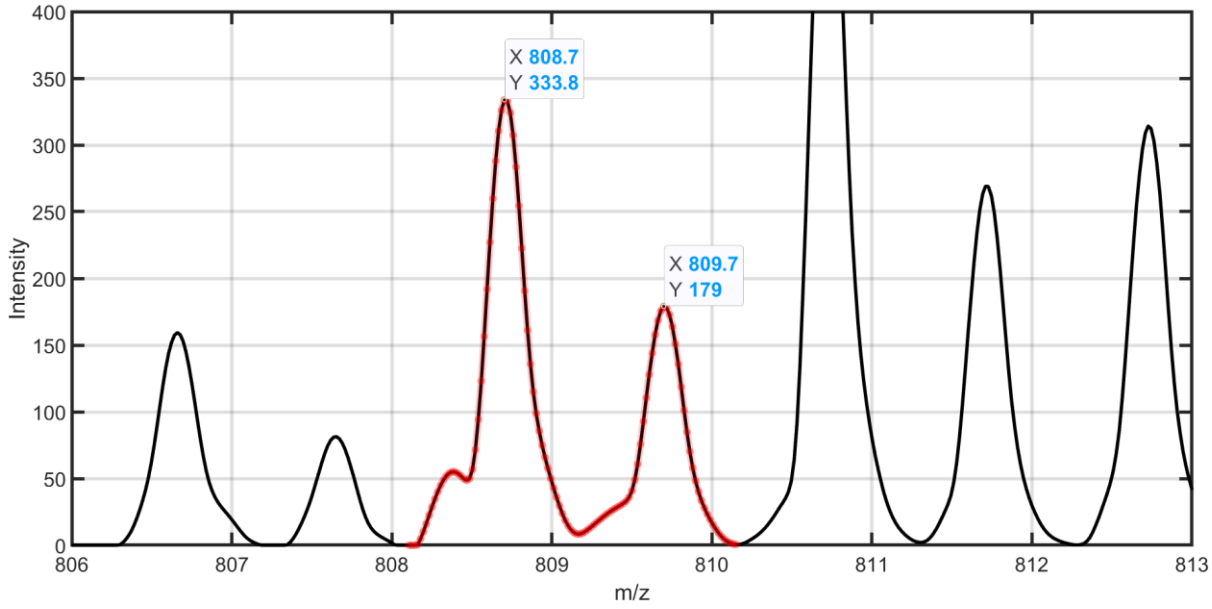

Figure 76. Deisotoping outcome (zoom) on the *m/z* 806-813 mass range.

In *Figure 76*, the exemplary correctly identified 2-element-isotopic envelope is presented.

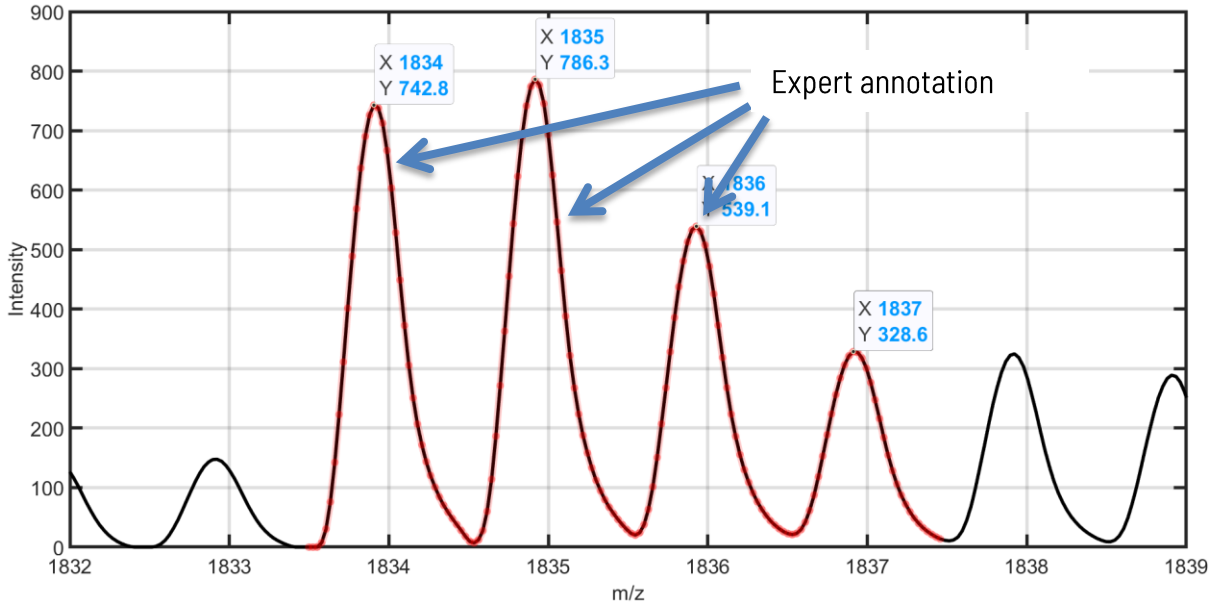

Figure 77. Deisotoping outcome (zoom) on the *m/z* 1832-1839 mass range. An example of isotopic envelope partially correctly identified by the proposed algorithm.

In *Figure 77,* the expert manually annotated the isotopic envelope comprised of four peaks, whereas the algorithm identified 3 out of 4 peaks as the envelope members. The most probable reason is that in that mass range, the fourth peak in the isotopic envelope, according to the theoretical isotope pattern calculated using *Compass IsotopePattern* (Bruker Daltonik), should be approximately 20.49% height (abundance) of the first peak, whilst in the case presented in *Figure 77* the abundance of the fourth peak is over 44.23%.

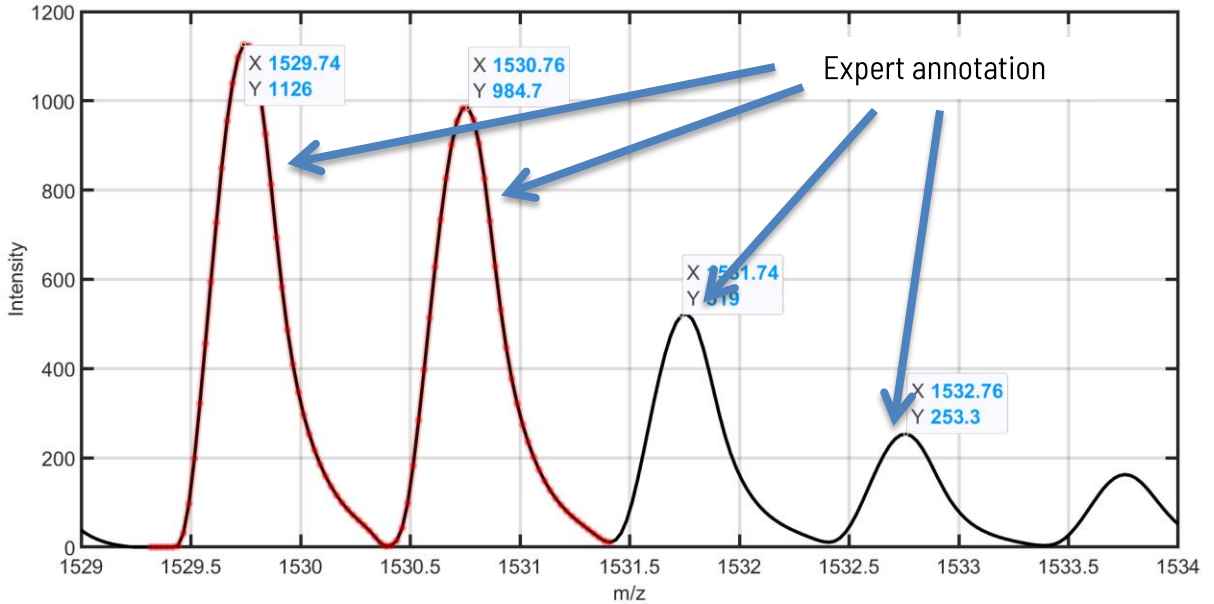

Figure 78. Deisotoping outcome (zoom) on the *m/z* 1529-1534 mass range. An example of isotopic envelope partially identified by the proposed algorithm.

In *Figure 78,* an exemplary isotopic envelope partially correctly detected is presented. Despite the unfavourable variance ratio of the adjacent peaks at *m/z* 1531.74 and *m/z* 1532.76 ($\sigma_1 / \sigma_2 = 1.4078$), the proposed method correctly identified 2 out of 4 peaks included in the isotopic envelope.

6.4. Comparative results with selected existing algorithms

The proposed two-step method called *DeisoLAB* was compared with three algorithms that perform deisotoping: *mMass, pyOpenMS,* and *MS2-Deisotoper,* described in detail in *2.6.*

HNC-FFPE Dataset 1, which comprises of 200 704 peaks, was used for comparative testing. Such raw data was the input for every algorithm since each algorithm has its pre-processing step embedded.

In order to compare results obtained by the aforementioned algorithms, a number of peaks classified by the *Envelope* and *non-Envelope* ones obtained by each algorithm have been compared with the expert's annotations.

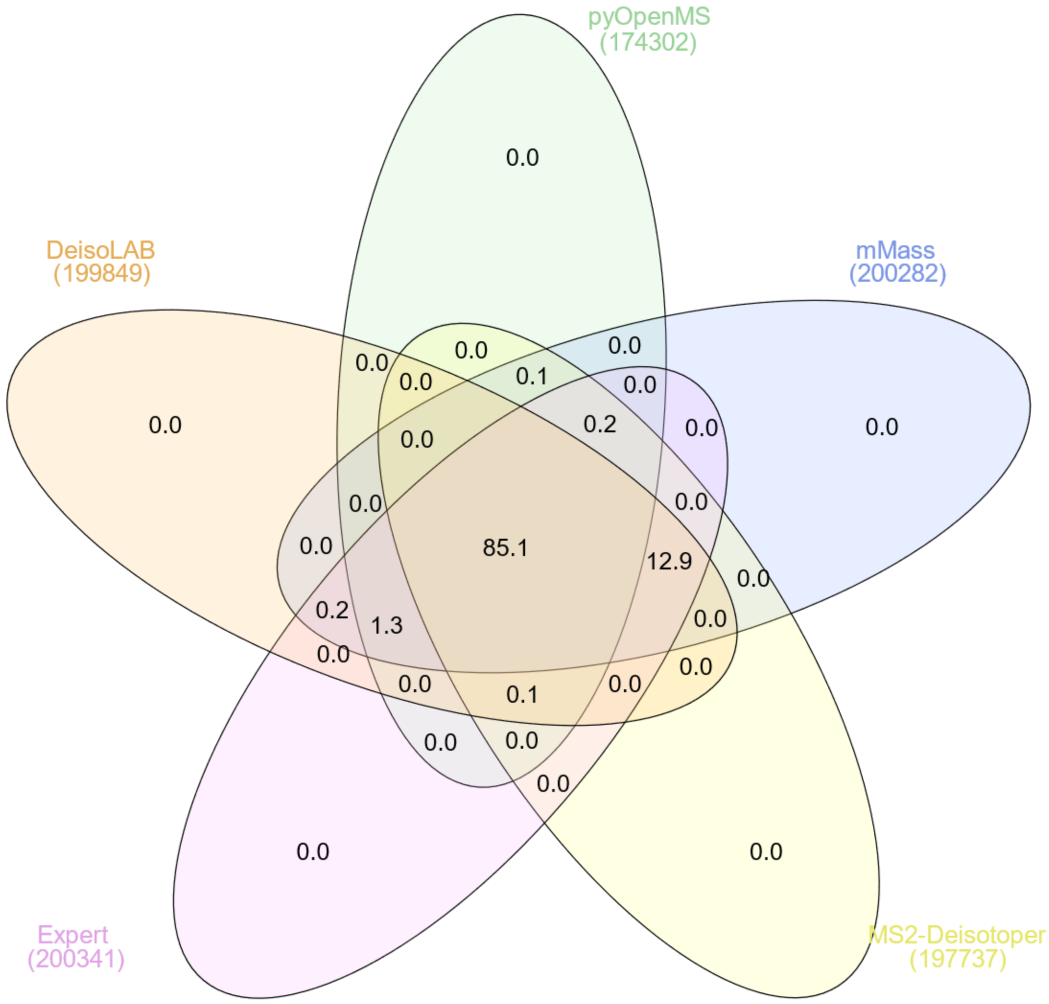

Figure 79. Venn diagram constructed using [123] - the intersection between *non-Envelope* peaks obtained by the expert, *DeisoLAB*, *mMass*, *pyOpenMS* and *MS2-Deisotoper* [DOI].

Figure 79 shows that the intersection of *non-Envelope* peaks between the expert and four mentioned algorithms is 85.1%. "*mMass* slightly outperforms *DeisoLAB* for true negatives (*TNs*) but simultaneously reveals notably smaller values of true positives *(TPs)*" [53].

Table 11. Confusion matrix-based metrics for *DeisoLAB*, *mMass*, *MS2-Deisotoper* and *pyOpenMS* [53].

	DeisoLAB	mMass	MS2-Deisotoper	pyOpenMS
TP	342	89	29	55
TN	199 828	200 008	197 403	173 994
FN	21	274	334	308
FP	513	333	2 938	26 347
Specificity [%]	99.74	99.83	98.53	86.85
Recall [%]	94.21	24.52	7.99	15.15
Balanced Accuracy [%]	96.98	62.18	53.26	51.00

It can be noticed that *DeisoLAB* outperforms other algorithms when taking into consideration the number of correctly classified peaks to the isotopic envelope. A high value of negative accuracy (*Figure 80*) is observed for all the investigated algorithms (*nE* peaks were correctly classified as the *nE* ones with specificity in the range of: 86.85% - 99.83%) (*Table 11*). Despite *DeisoLAB*, the algorithms have low values of balanced accuracy (from 51.00% to 96.98%). It provides information on how accurately isotopic envelope member peaks and non-Envelope peaks were correctly classified. [53]

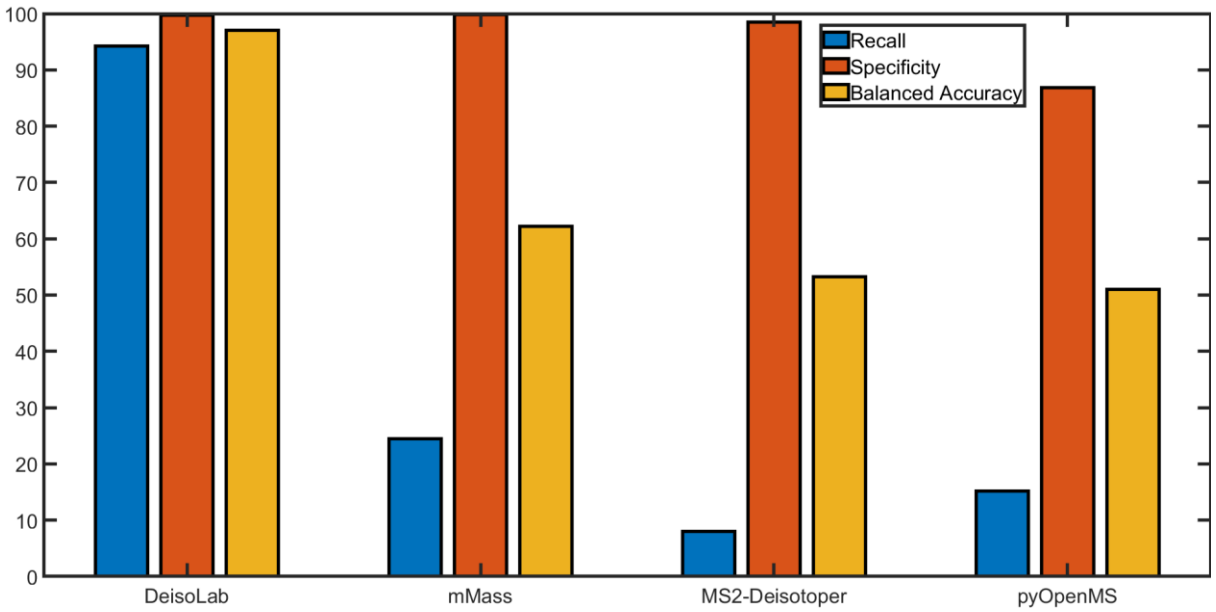

Figure 80. Comparison of three confusion matrix metrics for *DeisoLAB*, *mMass*, *MS2-Deisotoper* and *pyOpenMS* [53].

To sum up, in comparison with *mMass, MS2-Deisotoper, and pyOpenMS, DeisoLAB* exhibits better results with respect to all statistical measures gathered in *Table 11* and presented in *Figure 80*. It is undeniable that all three mentioned algorithms suffer from a severe decrease in recall values. Moreover, *pyOpenMS* suffers from a notable increase in false positives (*FPs*), yielding significant decreases in specificity and balanced accuracy. [53]

Hence, one can conclude that the *DeisoLAB* method classifies members of isotopic envelopes more accurately than other investigated algorithms. Moreover, it simultaneously copes with non-isotopic envelope members with also good accuracy.

7. SUMMARY AND CONCLUSION

This PhD dissertation copes with deisotoping methods used for MALDI-TOF MSI experiments. The primary purpose of the dissertation was to discuss the current challenges and perspectives in the field of a part of mass spectrometry pre-processing, which is deisotoping, in reference to the subject of MALDI-TOF experiments. The prominent gaps in the knowledge within the field of MALDI-TOF MSI deisotoping were comprehensively analysed.

For this purpose, the two-step methodology was introduced to handle large MALDI MSI datasets. The method incorporates two approaches: fuzzy reasoning and spatial map of a molecular distribution of peaks included in a mass spectrum. As it has been proven, such a novel approach is unique and allows for applying expert knowledge in algorithm creation and method development. The method has been tested on eight datasets obtained from different samples, which were also preserved in different ways – fresh frozen tissues versus formalin-fixed paraffin-embedded tissues. The results were presented and thoroughly analysed. It turned out that the method's first step based on the Mamdani-Assilan fuzzy-inference system significantly diminished the number of peaks that should undergo further analysis. As a consequence, even large datasets can be analysed. However, the method's limitations have been stated, and a few examples have been shown. Owing to the fact that the second part of the method is based on the spatial distribution of peaks and pairwise differential spatial distribution maps are created, the outcome is the peak pairs included in isotopic envelopes, which causes a formidable challenge to handle overlapping isotopic envelopes. A pairwise approach for identifying member peaks of isotopic envelopes leads to accurately detecting overlapping isotopic envelopes. Every dataset's recall and precision values, the envelope-member peaks were classified accurately, ranging from 88.12% to 94.12% and from 72.95% to 85.71%. In comparison, the peaks not included in isotopic envelopes were classified with an accuracy of over 99% (specificity). Matthews Correlation Coefficient and Fowlkes-Mallows Index values are over 80% for each dataset. Therefore, it can be stated that a correlation between predicted values by the proposed method and those annotated by the expert is high. What is worth mentioning is that the proposed method detects more isotopic envelope members than the expert, which is an advantage because, in many cases, there is no opportunity to annotate isotopic envelopes due to the signal density. Thanks to the combination of the fuzzy-inference system and classification based on the spatial distribution of peaks, isotopic envelopes can be effectively identified with the accuracy 96.98% (when compared to other algorithms).

Moreover, due to the pairwise approach for isotopic envelope peak members identification, the method allows to identify not only the simplest non-overlapping isotopic envelopes, but also the overlapping ones. Generally, the application of the proposed method allows for identifying potential isotopic envelope members that might be required to be proven experimentally. The proposed pipeline can be employed to other mass spectrometers that are characterised by similar properties of the isotopic envelope.

The dissertation theses were justified, as the new methodology for isotopic envelope identification in MALDI-TOF MSI data based on the spatial distribution of adjacent peaks that takes expert knowledge into consideration was developed. Furthermore, the significant data reduction was confirmed in the first step of the proposed methodology (after applying the fuzzy-inference system) since the number of peaks decreased by 96% for every dataset. Presented results show that regardless of the type of datasets – a way of the tissue procurement, preparation or preservation, and size of the dataset, their accuracy, and sensitivity of identifying peaks included in isotopic envelopes are similar. Hereby, considering expert knowledge in the field of mass spectrometry and information on the spatial distribution of spectral peaks, a novel approach towards isotopic envelope identification was proposed, leading to optimisation of a pre-processing step, which results in more accurate peptide/protein identification.

BIBLIOGRAPHY

[1] P. R. Graves and T. A. J. Haystead, "Molecular Biologist 'Guide to Proteomics'," *Microbiology and Molecular Biology Reviews,* vol. 66, p. 39–63, March 2002.

[2] S. E. Jadid, R. Touahni and A. Moussa, *The Importance of Considering Natural Isotopes in Improving Protein Identification Accuracy,* 2019.

[3] A. Misra, Ed., Challenges in delivery therapeutic genomics and proteomics, 1. ed. ed., Amsterdam [u.a.]: Elsevier, 2011.

[4] K.-H. Liang, *Bioinformatics for biomedical science and clinical applications,* 2013.

[5] W. Timp and G. Timp, "Beyond mass spectrometry, the next step in proteomics," *Science Advances,* vol. 6, January 2020.

[6] G. Guo, M. Papanicolaou, N. J. Demarais, Z. Wang, K. L. Schey, P. Timpson, T. R. Cox and A. C. Grey, "Automated annotation and visualisation of high-resolution spatial proteomic mass spectrometry imaging data using HIT-MAP," vol. 12, 2021.

[7] I. W. Hamley, Introduction to Peptide Science, Wiley, 2020.

[8] D. L. Nelson, M. M. Cox and A. L. Lehninger, Eds., Principles of biochemistry, 4. ed., 5. print. ed., New York: Freeman, 2005.

[9] B. Reddy, T. Jow and B. M. Hantash, "Bioactive oligopeptides in dermatology: Part I," *Experimental Dermatology,* vol. 21, p. 563–568, June 2012.

[10] K. K. Jessica Forbes, "Biochemistry, Peptide," *StatPearls,* January 2022.

[11] J. A. L. J. e. a. Alberts, Molecular Biology of the Cell. 4th edition, Garland Science, 2002.

[12] M.-L. Huynh, P. Russell and B. Walsh, "Tryptic Digestion of In-Gel Proteins for Mass Spectrometry Analysis," in *Methods in Molecular Biology*, Humana Press, 2009, p. 507–513.

[13] Ü. A. Laskay, A. A. Lobas, K. Srzentić, M. V. Gorshkov and Y. O. Tsybin, "Proteome digestion specificity analysis for rational design of extended bottom-up and middle-down proteomics experiments.," *Journal of proteome research,* vol. 12, no. 12, p. 5558–5569, December 2013.

[14] T. Dau, G. Bartolomucci and J. Rappsilber, "Proteomics Using Protease Alternatives to Trypsin Benefits from Sequential Digestion with Trypsin," *Analytical Chemistry,* vol. 92, p. 9523–9527, July 2020.

[15] P. Giansanti, L. Tsiatsiani, T. Y. Low and A. J. R. Heck, "Six alternative proteases for mass spectrometry–based proteomics beyond trypsin," vol. 11, pp. 993-1006, 2016.

[16] M. Karas and F. Hillenkamp, "Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons," vol. 60, pp. 2299-2301, 1988.

[17] E. V. Romanova, S. P. Annangudi, H.-C. Tai and J. V. Sweedler, "Mass Spectrometry of Proteins," in *Reference Module in Neuroscience and Biobehavioral Psychology,* Elsevier, 2017.

[18] K. Jeong, J. Kim, M. Gaikwad, S. N. Hidayah, L. Heikaus, H. Schlüter and O. Kohlbacher, "FLASHDeconv: Ultrafast, High-Quality Feature Deconvolution for Top-Down Proteomics," vol. 10, pp. 213-218.e6, 2020.

[19] Y. Z. W. Andy Tao, Mass Spectrometry-Based Chemical Proteomics, Wiley, 2019.

[20] K. Park, J. Y. Yoon, S. Lee, E. Paek, H. Park, H.-J. Jung and S.-W. Lee, "Isotopic Peak

Intensity Ratio Based Algorithm for Determination of Isotopic Clusters and Monoisotopic Masses of Polypeptides from High-Resolution Mass Spectrometric Data," vol. 80, pp. 7294-7303, 2008.

[21] J. L. Norris and R. M. Caprioli, "Analysis of Tissue Specimens by Matrix-Assisted Laser Desorption/Ionization Imaging Mass Spectrometry in Biological and Clinical Research," vol. 113, pp. 2309-2342, 2013.

[22] M. Aichler and A. Walch, "MALDI Imaging mass spectrometry: current frontiers and perspectives in pathology research and practice," vol. 95, pp. 422-431, 2015.

[23] A. Macklin, S. Khan and T. Kislinger, "Recent advances in mass spectrometry based clinical proteomics: applications to cancer research," vol. 17, 2020.

[24] R. Smith, J. T. Prince and D. Ventura, "A coherent mathematical characterization of isotope trace extraction, isotopic envelope extraction, and LC-MS correspondence," vol. 16, 2015.

[25] H. Awad, M. M. Khamis and A. El-Aneed, "Mass Spectrometry, Review of the Basics: Ionization," *Applied Spectroscopy Reviews,* vol. 50, p. 158–175, September 2014.

[26] A. Polanski, M. Marczyk, M. Pietrowska, P. Widlak and J. Polanska, "Signal Partitioning Algorithm for Highly Efficient Gaussian Mixture Modeling in Mass Spectrometry," *PLOS ONE,* vol. 10, p. e0134256, July 2015.

[27] A. M. Haag, "Mass Analyzers and Mass Spectrometers," in *Modern Proteomics – Sample Preparation, Analysis and Practical Applications*, Springer International Publishing, 2016, p. 157–169.

[28] Y. Wang, J. Sun, J. Qiao, J. Ouyang and N. Na, "A Soft and Hard Ionization Method for Comprehensive Studies of Molecules," *Analytical Chemistry,* vol. 90, p. 14095–14099, November 2018.

[29] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong and C. M. Whitehouse, "Electrospray Ionization for Mass Spectrometry of Large Biomolecules," vol. 246, pp. 64-71, 1989.

[30] M. Mann, C. K. Meng and J. B. Fenn, "Interpreting mass spectra of multiply charged ions," vol. 61, pp. 1702-1708, 1989.

[31] Z. Takáts, J. M. Wiseman, B. Gologan and R. G. Cooks, "Mass Spectrometry Sampling Under Ambient Conditions with Desorption Electrospray Ionization," vol. 306, pp. 471-473, 2004.

[32] P. H. Dawson, *Quadrupole Mass Spectrometry and Its Applications,* Burlington: Elsevier Science, 1976.

[33] G. Kaklamanos, E. Aprea and G. Theodoridis, *Mass Spectrometry: Principles and Instrumentation,* 2016, pp. 661-668.

[34] M. A. G. Wallace and J. P. McCord, *High-resolution mass spectrometry,* 2020, pp. 253-270.

[35] Q. Hu, R. J. Noll, H. Li, A. Makarov, M. Hardman and R. G. Cooks, "The Orbitrap: a new mass spectrometer," *Journal of Mass Spectrometry,* vol. 40, p. 430–443, 2005.

[36] R. E. Ardrey, Liquid chromatography-mass spectrometry : an introduction, John Wiley & Sons, Ltd, 2003.

[37] *Basic Gas Chromatography – Mass Spectrometry,* 1988.

[38] J. Cai and J. Henion, "Capillary electrophoresis-mass spectrometry," *Journal of Chromatography A,* vol. 703, p. 667–692, June 1995.

[39] G. Verbeck, B. Ruotolo, H. Sawyer, K. Gillig and D. Russell, "A fundamental introduction to ion

mobility mass spectrometry applied to the analysis of biomolecules," *Journal of Biomolecular Techniques,* vol. 13, pp. 56-61, 2002.

[40] J. F. de-Cossio, "Isotopica: a tool for the calculation and viewing of complex isotopic envelopes," vol. 32, pp. 3779-3779, 2004.

[41] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," vol. 422, pp. 198-207, 2003.

[42] R. Peter, "MALDI-TOF mass spectrometry in protein chemistry," *Proteomics in Functional Genomics,* 2000.

[43] S. Shimma, "Mass Spectrometry Imaging," vol. 11, pp. A0102-A0102, 2022.

[44] J. Hermann, H. Noels, W. Theelen, M. Lellig, S. Orth-Alampour, P. Boor, V. Jankowski and J. Jankowski, "Sample preparation of formalin-fixed paraffin-embedded tissue sections for MALDI-mass spectrometry imaging," *Analytical and Bioanalytical Chemistry,* vol. 412, p. 1263–1275, January 2020.

[45] E. Scifo, G. Calza, M. Fuhrmann, R. Soliymani, M. Baumann and M. Lalowski, "Recent advances in applying mass spectrometry and systems biology to determine brain dynamics," *Expert Review of Proteomics,* vol. 14, p. 545–559, June 2017.

[46] A. Römpp and B. Spengler, "Mass spectrometry imaging with high resolution in mass and space," vol. 139, pp. 759-783, 2013.

[47] D. Wittekind, "Traditional staining for routine diagnostic pathology including the role of tannic acid. 1. Value and limitations of the hematoxylin-eosin stain," *Biotechnic & Histochemistry,* vol. 78, p. 261–270, October 2003.

[48] J. K. C. Chan, "The Wonderful Colors of the Hematoxylin–Eosin Stain in Diagnostic Surgical Pathology," *International Journal of Surgical Pathology,* vol. 22, p. 12–32, January 2014.

[49] N. Jaitly, A. Mayampurath, K. Littlefield, J. N. Adkins, G. A. Anderson and R. D. Smith, "Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data," vol. 10, 2009.

[50] A. L. Rockwood, S. L. V. Orden and R. D. Smith, "Rapid Calculation of Isotope Distributions," vol. 67, pp. 2699-2704, 1995.

[51] X. Liu, Y. Inbar, P. C. Dorrestein, C. Wynne, N. Edwards, P. Souda, J. P. Whitelegge, V. Bafna and P. A. Pevzner, "Deconvolution and Database Search of Complex Tandem Mass Spectra of Intact Proteins," vol. 9, pp. 2772-2782, 2010.

[52] I. Bogdan, D. Coca, J. Rivers and R. J. Beynon, "Hardware acceleration of processing of mass spectrometric data for proteomics," vol. 23, pp. 724-731, 2007.

[53] **A. Glodek,** J. Polanska and M. Gawin, "Isotopic envelope identification by analysis of the spatial distribution of components in MALDI-MSI data," *arXiv,* 2023.

[54] K. Xiao, F. Yu, H. Fang, B. Xue, Y. Liu and Z. Tian, "Accurate and Efficient Resolution of Overlapping Isotopic Envelopes in Protein Tandem Mass Spectra," *Scientific Reports,* vol. 5, October 2015.

[55] Z. Yuan, J. Shi, W. Lin, B. Chen and F.-X. Wu, "Features-Based Deisotoping Method for Tandem Mass Spectra," *Advances in Bioinformatics,* vol. 2011, p. 1–12, January 2011.

[56] Y. Sun, J. Zhang, U. Braga-Neto and E. R. Dougherty, "BPDA - A Bayesian peptide detection algorithm for mass spectrometry," vol. 11, 2010.

[57] X.-j. Li, E. C. Yi, C. J. Kemp, H. Zhang and R. Aebersold, "A Software Suite for the Generation and Comparison of Peptide Arrays from Sets of Data Collected by Liquid Chromatography-Mass Spectrometry," vol. 4, pp. 1328-1340, 2005.

[58] J. Samuelsson, D. Dalevi, F. Levander and T. Rognvaldsson, "Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting," vol. 20, pp. 3628-3635, 2004.

[59] S. McIlwain, D. Page, E. L. Huttlin and M. R. Sussman, "Using dynamic programming to create isotopic distribution maps from mass spectra," vol. 23, pp. i328-i336, 2007.

[60] D. M. Horn, R. A. Zubarev and F. W. McLafferty, "Automated reduction and interpretation of," vol. 11, pp. 320-332, 2000.

[61] A. P. Tay, A. Liang, J. J. Hamey, G. Hart-Smith and M. R. Wilkins, "MS2-Deisotoper: A Tool for Deisotoping High-Resolution MS/MS Spectra in Normal and Heavy Isotope-Labelled Samples," *PROTEOMICS,* vol. 19, p. 1800444, August 2019.

[62] M. W. Senko, S. C. Beu and F. W. McLaffertycor, "Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions," vol. 6, pp. 229-233, 1995.

[63] H. L. Röst, U. Schmitt, R. Aebersold and L. Malmström, "pyOpenMS: A Python-based interface to the OpenMS mass-spectrometry algorithm library," vol. 14, pp. 74-77, 2014.

[64] B. Y. Renard, M. Kirchner, H. Steen, J. A. J. Steen and F. A. Hamprecht, "NITPICK: peak identification for mass spectrometry data.," *BMC bioinformatics,* vol. 9, p. 355, August 2008.

[65] A. M. Mayampurath, N. Jaitly, S. O. Purvine, M. E. Monroe, K. J. Auberry, J. N. Adkins and R. D. Smith, "DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra," vol. 24, pp. 1021-1023, 2008.

[66] M. Strohalm, D. Kavan, P. Novák, M. Volný and V. Havlíček, " mMass 3: A Cross-Platform Software Environment for Precise Analysis of Mass Spectrometric Data," *Analytical Chemistry,* vol. 82, p. 4648–4651, May 2010.

[67] P. Du and R. H. Angeletti, "Automatic Deconvolution of Isotope-Resolved Mass Spectra Using Variable Selection and Quantized Peptide Mass Distribution," vol. 78, pp. 3385-3392, 2006.

[68] E. J. Breen, F. G. Hopwood, K. L. Williams and M. R. Wilkins, "Automatic Poisson peak harvesting for high throughput protein identification," vol. 21, pp. 2243-2251, 2000.

[69] V. Zabrouskov, M. W. Senko, Y. Du, R. D. Leduc and N. L. Kelleher, "New and automated MSn approaches for top-down identification of modified proteins," vol. 16, pp. 2027-2038, 2005.

[70] Z. Zhang and A. G. Marshall, "A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra," vol. 9, pp. 225-233, 1998.

[71] P. Widlak, G. Mrukwa, M. Kalinowska, M. Pietrowska, M. Chekan, J. Wierzgon, M. Gawin, G. Drazek and J. Polanska, "Detection of molecular signatures of oral squamous cell carcinoma and normal epithelium - application of a novel methodology for unsupervised segmentation of imaging mass spectrometry data," *PROTEOMICS,* vol. 16, p. 1613–1621, April 2016.

[72] A. Kurczyk, M. Gawin, P. Paul, E. Chmielik, T. Rutkowski, M. Pietrowska and P. Widłak, "Prognostic Value of Molecular Intratumor Heterogeneity in Primary Oral Cancer and Its Lymph Node Metastases Assessed by Mass Spectrometry Imaging," vol. 27, p. 5458, 2022.

[73] K. Bednarczyk, M. Gawin, M. Chekan, A. Kurczyk, G. Mrukwa, M. Pietrowska, J. Polanska and P. Widlak, "Discrimination of normal oral mucosa from oral cancer by mass spectrometry

imaging of proteins and lipids," *Journal of Molecular Histology,* vol. 50, p. 1–10, November 2018.

[74] M. Marczyk, G. Drazek, M. Pietrowska, P. Widlak, J. Polanska and A. Polanski, "Modeling of Imaging Mass Spectrometry Data and Testing by Permutation for Biomarkers Discovery in Tissues," *Procedia Computer Science,* vol. 51, p. 693–702, 2015.

[75] L. A. Zadeh, "Fuzzy Sets," *Information and Control,* vol. 8, p. 338–353, 1965.

[76] J. Łęski, Systemy neuronowo-rozmyte, Wydawnictwo Naukowo-Techniczne, 2008.

[77] E. H. Ruspini, "A new approach to clustering," vol. 15, pp. 22-32, 1969.

[78] L. A. Zadeh, "Outline of a New Approach to the Analysis of Complex Systems and Decision Processes," Vols. SMC-3, pp. 28-44, 1973.

[79] J. Chojcan and J. Łęski, Zbiory rozmyte i ich zastosowania, Wydawnictwo Politechniki Śląskiej, 2001.

[80] D. Dubois and H. Prade, "The three semantics of fuzzy sets," vol. 90, pp. 141-150, 1997.

[81] M. Mizumoto, "Fuzzy sets and their operations, II," vol. 50, pp. 160-174, 1981.

[82] R. Czabanski, M. Jezewski and J. Leski, *Introduction to Fuzzy Systems,* 2017, pp. 23-43.

[83] S. N. Sivanandam, S. Sumathi and S. N. Deepa, *Introduction to Fuzzy Logic using MATLAB,* 2007.

[84] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," vol. 7, pp. 1-13, 1975.

[85] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," Vols. SMC-15, pp. 116-132, 1985.

[86] M. Sugeno and G. T. Kang, "Structure identification of fuzzy model," vol. 28, pp. 15-33, 1988.

[87] J. M. Leski and E. Czogala, "A new artificial neural network based fuzzy inference system with moving consequents in if-then rules and selected applications," *Fuzzy Sets Syst.,* vol. 108, p. 289–297, 1999.

[88] Y. TSUKAMOTO, *AN APPROACH TO FUZZY REASONING METHOD,* 1993, pp. 523-529.

[89] J. F. Baldwin, "A new approach to approximate reasoning using a fuzzy logic," vol. 2, pp. 309-325, 1979.

[90] **A. Glodek** and J. Polańska, "Method for mass spectrometry spectrum deisotoping based on fuzzy inference systems," vol. 46.

[91] A. Polanski, M. Marczyk, M. Pietrowska, P. Widlak and J. Polanska, "Initializing the EM Algorithm for Univariate Gaussian, Multi-Component, Heteroscedastic Mixture Models by Dynamic Programming Partitions," vol. 15, p. 1850012, 2018.

[92] D. F. Findley, "Counterexamples to parsimony and BIC," *Annals of the Institute of Statistical Mathematics,* vol. 43, p. 505–514, September 1991.

[93] **A. Glodek**, "Fuzzy-inference system for isotopic envelope identification in Mass Spectrometry Imaging data," *In: Rojas, I., Valenzuela, O., Rojas, F., Herrera, L.J., Ortuño, F. (eds) Bioinformatics and Biomedical Engineering. IWBBIO 2022. Lecture Notes in Computer Science, vol 13347. Springer, Cham.,* 2022.

[94] R. C. González, R. E. Woods and S. L. Eddins, Digital image processing using MATLAB, Pearson-Prentice-Hall, 2020.

[95] J. Chaki and N. Dey, *Texture Feature Extraction Techniques for Image Recognition,* 2020.

[96] M. Hall-Beyer, "Practical guidelines for choosing GLCM textures to use in landscape classification tasks over a range of moderate spatial scales," *International Journal of Remote Sensing,* vol. 38, p. 1312–1338, January 2017.

[97] R. M. Haralick, K. Shanmugam and I. Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics,* Vols. SMC-3, no. 6, p. 610–621, 1973.

[98] A. Hafiane, G. Seetharaman and B. Zavidovique, "Median Binary Pattern for Textures Classification," in *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2007, p. 387–398.

[99] M. Hall-Beyer, "GLCM Texture: A Tutorial v 3.0," *PRISM: University of Calgary's Digital Repository,* 2017.

[100] M. Petrou and P. G. Sevilla, *Image Processing,* 2006.

[101] Q. Zhao, C.-Z. Shi and L.-P. Luo, "Role of the texture features of images in the diagnosis of solitary pulmonary nodules in different sizes," *Chin J Cancer Res,* vol. 26, pp. 451-458, 2014.

[102] S. Singh, D. Srivastava and S. Agarwal, *GLCM and its application in pattern recognition,* 2017.

[103] C. D. Wijetunge, I. Saeed, B. A. Boughton, J. M. Spraggins, R. M. Caprioli, A. Bacic, U. Roessner and S. K. Halgamuge, "EXIMS: an improved data analysis pipeline based on a new peak picking method for EXploring Imaging Mass Spectrometry data," vol. 31, pp. 3198-3206, 2015.

[104] J. H. Zar, Biostatistical analysis, 5. ed., Pearson new internat. ed ed., Harlow: Pearson Education Limited, 2014.

[105] C. Spearman, "The Proof and Measurement of Association between Two Things," vol. 15, p. 72, 1904.

[106] I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli and L. Vanneschi, "Structural similarity index (SSIM) revisited: A data-driven approach," *Expert Systems with Applications,* vol. 189, p. 116087, March 2022.

[107] H. M. Huan Liu, Computational Methods of Feature Selection, Chapman & Hall/CRC, 2008.

[108] G. H. John, R. Kohavi and K. Pfleger, *Irrelevant Features and the Subset Selection Problem,* 1994, pp. 121-129.

[109] N. R. Draper and H. Smith, Applied regression analysis, 2nd ed. ed., New York [u.a.]: Wiley, 1981.

[110] J. Neter, W. Wasserman and M. H. Kutner, Applied linear statistical Models, 3. ed ed., Homewood, Ill.: Irwin, 1990.

[111] A. Pérez, P. Larrañaga and I. Inza, "Bayesian classifiers based on kernel density estimation: Flexible classifiers," *International Journal of Approximate Reasoning,* vol. 50, p. 341–362, February 2009.

[112] C. C. Aggarwal, Ed., Data classification. Algorithms and applications, Boca Raton, FL: CRC Press, 2015.

[113] R. R. Yager, "An extension of the naive Bayesian classifier," *Information Sciences,* vol. 176, p. 577–588, March 2006.

[114] C. M. Bishop, Pattern recognition and machine learning, Corrected at 8th printing 2009 ed., New, York: Springer, 2009.

[115] J. I. Haijin, S. HUANG, L. V. Xuewei, W. U. Yaning and F. E. N. G. Yuntian, "Empirical Studies of

a Kernel Density Estimation Based Naive Bayes Method for Software Defect Prediction," vol. E102.D, pp. 75-84, 2019.

[116] V. A. Epanechnikov, "Non-Parametric Estimation of a Multivariate Probability Density," *Theory of Probability & Its Applications,* vol. 14, p. 153–158, January 1969.

[117] J. N. K. Liu, Y.-L. He, X.-Z. Wang and Y.-X. Hu, *A comparative study among different kernel functions in flexible naiive Bayesian classification,* 2011.

[118] C.-Y. Chu, D. J. Henderson and C. F. Parmeter, "On discrete Epanechnikov kernel functions," *Computational Statistics &math; Data Analysis,* vol. 116, p. 79–105, December 2017.

[119] J. Ć. J. Koronacki, Statystyczne systemy uczące się, Akademicka Oficyna Wydawnicza EXIT, 2015.

[120] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) - Protein Structure,* vol. 405, p. 442–451, October 1975.

[121] E. B. Fowlkes and C. L. Mallows, "A Method for Comparing Two Hierarchical Clusterings," *Journal of the American Statistical Association,* vol. 78, p. 553–569, September 1983.

[122] B. Akkaya and N. Çolakoğlu, "Comparison of Multi-class Classification Algorithms on Early Diagnosis of Heart Diseases.," *ISBIS Young Business and Industrial Statisticians Workshop on Recent Advances in Data Science and Business Analytics,* 2019.

[123] H. Heberle, G. V. Meirelles, F. R. da Silva, G. P. Telles and R. Minghim, "InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams," vol. 16, 2015.

ABSTRACT

Mass spectrometry is one of the essential steps toward protein identification due to the fact that it provides information on the proteins' structure. Since most chemical elements have isotopes of different masses, the isotopic mass of a molecule observed in a mass spectrum reflects the type and number of atoms in the ion being measured and the distribution of the different isotopes. Depending on the resolution of the mass spectrometer, molecular ions can be represented either by the monoisotopic mass (taking into account only the mass of the most abundant stable isotope of each atom present in the molecule) or by the average mass (taking into account the presence of both light and heavy isotopes). For an atom, the difference between the two masses is insignificant. However, in molecules such as proteins, the difference between them increases with the number of atoms that make up the molecule. Such a discrepancy leads to the misidentification of peptides, which is why it is vital to remove isotope peaks from the mass spectrum by performing a deisotoping procedure. Several existing algorithms include the deisotoping step. However, most of them are dedicated to different mass spectrometry experiments and have limitations depending on the kind of data from those experiments. The MALDI-ToF technique provides high-dimension data. This dissertation introduces a method for isotopic envelopes identification in MALDI-ToF MSI data. It is based on combining the Mamdani-Assilan fuzzy-inference system with analysing the spatial molecular distribution of the peaks (model components). The spatial molecular distribution is evaluated by several image texture metrics. The proposed method was tested on eight MALDI-ToF MSI datasets provided by the National Institute of Oncology in Gliwice from patients who suffered from head and neck cancer. Obtained results were compared with three existing deisotoping algorithms. The method presented in this research is based on the pairwise approach for isotopic envelopes member identification. Such an approach enables the identification of overlapping isotopic envelopes in large MALDI-ToF MSI-driven datasets.

STRESZCZENIE

„Metody identyfikacji obwiedni izotopowych w danych z obrazowania molekularnego MALDI TOF"

Jednym ze znaczących etapów procesu prowadzącego do identyfikacji białek jest spektrometria masowa, która pozwala na pozyskanie informacji o strukturze białek. Ze względu na to, że większość pierwiastków chemicznych ma izotopy o różnej masie, masa izotopowa cząsteczki obserwowana na widmie masowym odzwierciedla rodzaj i liczbę atomów wchodzących w skład mierzonego jonu oraz rozmieszczenie różnych izotopów. W zależności od zdolności rozdzielczej spektrometru masowego, jony cząsteczkowe mogą być reprezentowane albo przez masę monoizotopową (uwzględniającą jedynie masy najliczniej występującego stabilnego izotopu każdego atomu obecnego w cząsteczce), albo przez średnią masę (uwzględniająca obecność zarówno lekkich, jak i ciężkich izotopów). Dla atomu różnica między tymi dwiema masami jest nieznaczna. Jednak w cząsteczkach takich jak białka, różnica między nimi wzrasta wraz z liczbą atomów, z których zbudowana jest cząsteczka. Taka rozbieżność prowadzi do błędnej identyfikacji peptydów, dlatego tak ważne jest, usunięcie z widma masowego pików izotopowych w procesie nazywanym deizotopingiem. Istnieją różne algorytmy umożliwiające przeprowadzenie deizotopingu, natomiast mają swoje ograniczenia, są dedykowane do różnych metod spektrometrii masowej. Dane pochodzące z eksperymentów wykonanych techniką MALDI-TOF cechują się dużą wymiarowością. W niniejszej pracy przedstawiono metodę identyfikacji obwiedni izotopowych w danych z obrazowania molekularnego MALDI-TOF opartą na systemie rozmytym Mamdaniego-Assilana oraz przestrzennych mapach dystrybucji molekularnej pików wchodzących w skład obwiedni izotopowej. Do oceny przestrzennych map dystrybucji molekularnej zastosowano szereg miar tekstury obrazu. Algorytm przetestowano na ośmiu zbiorach danych otrzymanych wskutek przeprowadzenia eksperymentu techniką MALDI-TOF na próbkach pochodzących z Narodowego Instytutu Onkologii im. Marii Skłodowskiej-Curie w Gliwicach od pacjentów cierpiących na nowotwór regionu głowy i szyi. Dane zostały poddane przetwarzaniu wstępnemu oraz ekstrakcji cech. Wyniki zebrano i porównano z trzema istniejącymi algorytmami do deizotopingu. Analiza otrzymanych wyników wykazała, iż zaproponowana w niniejszej pracy metoda do identyfikacji obwiedni izotopowych umożliwia wykrycie obwiedni nakładających się dzięki zastosowania podejścia zorientowanego na badanie par pików. Ponadto, zaproponowany algorytm umożliwia analizę dużych zbiorów danych.