2D Feature Detection for Ion Mobility Imaging Mass Spectrometry

Gautam Sinha

Master of Science Thesis

TU Delft Delft University of Technology

Delft Center for Systems and Control

2D Feature Detection for Ion Mobility Imaging Mass Spectrometry

Master of Science Thesis

For the degree of Master of Science in Systems and Control at Delft University of Technology

Gautam Sinha

May 14, 2024

Faculty of Mechanical Engineering (ME) · Delft University of Technology

The work in this thesis was supported by DCSC,TU Delft. Their cooperation is hereby gratefully acknowledged.

Table of Contents

List of Figures

List of Tables

Preface

The objective of this thesis is to design a feature detection algorithm that can be implemented for Ion-Mobility Imaging Mass Spectrometry. The first chapter discusses the importance of omics in the study of biological specimens and cellular metabolism and analytical techniques used for this study. This is followed by an introduction to Mass Spectrometry, one of the most popular analytical technique in the field of omics. The chapter then explores the instrumentation behind mass spectrometry and its various modifications. Special attention is given to Ion Mobility Imaging Mass Spectrometry and its applications. Chapter 2 focuses on the data analysis aspect of mass spectrometry. A brief outline on the various data processing procedures is presented. Out of the various procedures presented, the chapter goes in detail regarding feature detection, which is the leading light of the thesis. The chapter presents an overview on the different types of algorithms that are being used to carry out feature detection. This is followed by critical analysis of each type of algorithm. Chapter 3 lays down the foundation of feature detection algorithm. A brief introduction to wavelet transform and it's properties are provided. This is followed by an introduction to wavelet transform maxima, a technique commonly used for feature detection in real world signals. A brief literature study of the technique is provided with an emphasis on the parameters being used. Based on the results obtained from the test signal and the literature review, the research objectives for the thesis is formulated. Chapter 4 presents the 2D feature detection algorithm. The algorithm is broken down into four sections and the relevant theory around each section along with their implementation is discussed and presented. In Chapter 5, we evaluate the performance of the algorithm. This evaluation is performed on a synthetic and a real world data sample. A brief outline on the description of the synthetic data sample and real world data sample is presented. We then study the impact of the parameters used by the algorithm by performing various experiments. This helped us to establish about the parameters used in the algorithm. Lastly, using this knowledge, we compare the performance of the algorithm with an existing 2D feature detection algorithm. The final chapter concludes the research. It was found that while the performance on the synthetic data sample was good, more study is required to optimize the performance on the real world data sample. A brief sketch on the future work associated with the development of the algorithm is provided.

This thesis was undertaken under my thesis supervisor Dr. Ing. Raf Van de Plas. I would like to thank my supervisor for his excellent guidance and support during the process. I would also like to thank my supervisor Dr. Lukasz Migas, whose support and critical feedback were very essential for me to complete my literature study.

I would also like to thank all my colleagues in DCSC(Delft Center of Systems and Control) for keeping me motivated during such difficult circumstances. In the end, I would like to thank my parents and my sister for always supporting me and helping me to fullfil my dreams. Their wisdom and motivation have always served me well.

I hope you enjoy reading this report.

Gautam Sinha

April,2024

Chapter 1

Introduction to Omics and Mass Spectrometry

1-1 Introduction

The field of Omics is a collection of various biological disciplines that aim for characterization and quantification of biological molecules. The ending -ome is used for addressing the entities studied in various biological fields. For example, the term Proteome refers to the sum of all protein molecules present in a cell, tissue or an organism and Proteomics [5] is the science that studies these molecules with respect to their biochemical properties, functioning and structural changes in response to an internal or an external stimuli. Similarly, Transcriptome is the set of all RNA transcripts expressed by an individual or a population of biological cells and Transcriptomics [57] is the study of these RNA transcripts using high-throughput analytical methods. In all cases, the field of omics involve a detailed analysis of molecules in the most crucial biological processes. The field principally comprises of study of deoxyribonucleic acid (DNA) (genomics[2], epigenomics[98]), proteins (proteomics) and various other molecules (metabolomics[33]). Apart from these categorical platforms, various other omic based sub-disciplines have also emerged such as lipidomics[42] and metallomics[79] showing that the discipline is constantly evolving. Figure 1-1 provides a brief overview of the various types of omic platforms.

Omic based technologies play a crucial role in different biological applications. In clinical biology, omic technologies are used for characterization of diseases and to study the efficiency of existing clinical therapies. Omics based studies have been utilized in the field of food science (foodomics) [16], defence [25] where researchers aim to identify potential biomarkers of toxicity occuring within the warfighter as a pre-clinical indicator and in environmental science where the impact of toxic substances at all levels of biological organizations (from molecular level to community and ecosystem) is studied [97].

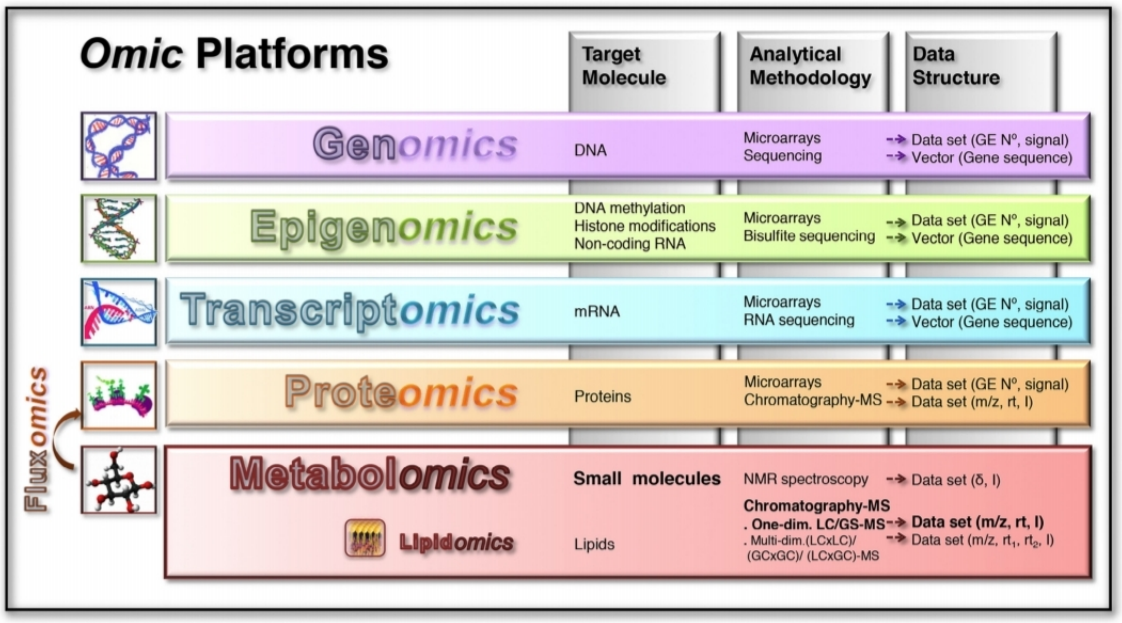


Figure 1-1: Overview of OMIC Platforms: Target Molecules, Analytical Methodologies used and the structure of Generated Data [38]

1-1-1 Metabolomics

Among the various disciplines, metabolomics is the study of chemical processes involving metabolites - the small molecular substrates, intermediates and by-products of cell metabolism. In a nutshell, metabolic profiling provides an instant picture of the physiology of the cell.

Characterization of metabolites have enabled the use of precision medicine at a number of levels, from characterization of metabolic changes caused due to presence of underlying disease to discovery and monitoring of new therapeutics [74]. Metabolomic based studies have been used in study of pancreatic cancer [51], type 2 diabetes [35] and various other diseases. Metabolomic based studies have also provided novel insights into the relationship between diet and diseases. For example, Rauchert et al. [76] have explored the relation between branched chain amino acids and obesity to insulin resistance. Thus, metabolomics have been widely used for studying the perturbations in the human cell caused due to presence of a disease, drugs or toxins in key biological processes.

In practice, metabolomics present a significant analytical challenge as it aims to study molecules that have varying physical properties (such as hydrophilic organic acids or hydrophobic non-polar lipids) [52]. As a result, metabolomic based technological platforms have taken the strategy of dividing the metabolome based on their molecular properties - compound polarity, functional groups, structural similarity, etc. The sample preparation and the analytical technique used for the study of these subset of metabolites are then devised accordingly. Figure (1-2) provides a brief outline for the various analytical procedures developed for the study of various molecules using Liquid Chromatography Mass Spectrometry (LC-MS) platforms. In addition to challenges faced in sample preparation, there are significant challenges present in the analysis of data produced by the instrument. Firstly, the diversity of technologies used

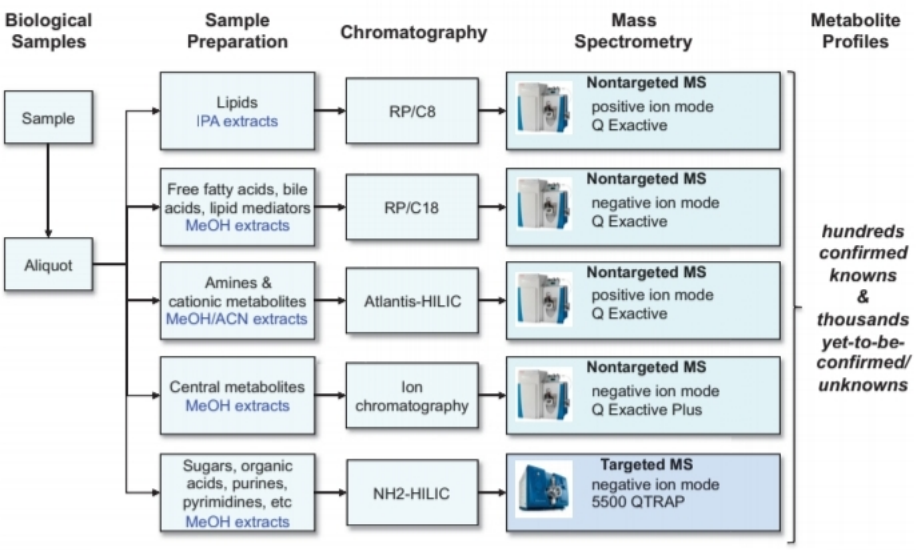


Figure 1-2: Illustration of the various sample preparation techniques, types of chromatography and types of analytical procedures used for LC-MS platforms[64]

in this field pose significant challenge when comparing results across laboratories due to various issues (e.g difference in precision of instrument). Secondly, the degree of identification of metabolites vary across methods, ranging from rigorously confirmed metabolites using standard references to 'unknown' signals which may or may not be a potential biomarker. These challenges have given researchers, the need to develop standard guidelines for reporting data [34], a detailed outline for testing procedures to evaluate different metabolites in order to obtain similar results [86] and the need to maintain open access repositories for modification and verification of result[43].

1-1-2 Instrumentation

Various analytical techniques have been developed for the field of omics including RNA-based sequencing techniques [15], Nuclear Magnetic Resonance spectroscopy (NMR) [58] and Mass Spectrometry (MS) [27]. Among these methods, NMR and MS based techniques are the most used analytical platforms in the field of metabolomics. High Resolution Nuclear Magnetic Resonance spectroscopy (HR-NMR) is used for study of bio-fluids and intact tissues to produce a complete profile of metabolite signals without any separation or derivatization of the sample specimen [50]. MS based methods provide a comprehensive analysis of low molecular weight compounds present in biological systems. Both of these approaches complement each other and more information can be extracted with the integration of both of these technologies. For this research, we will focus on the analysis of MS based techniques.

1-2 Introduction to Mass Spectrometry

1-2-1 Mass Spectrometry

MS is based on the ability of electric field to influence the motion of charged atoms and molecules in relation to their mass and charge. Therefore, by controlling the electric field inside the spectrometer, one can deconvolve the motions of various molecules into distinct particles with their specific mass-to-charge ratios (m/z).

Results of MS based investigations are represented on a 2D axis called as mass spectrum where the x-axis represents the m/z ratio and y-axis represents the relative intensity of a chemical compound with a specific m/z. For Imaging Mass Spectrometry (IMS) experiments, the results obtained can be interpreted as spectral images, where each pixel is associated with its own mass spectrum. In this case, the data consists of three key axes, the spatial axes (x and y) and the spectral axis (m/z).

Instrumentation

The major components that govern the performance of Mass Spectrometers are :

1. **Ion Source** - The ionization step is the most critical step for characterization of analytes in MS. Analytes need to be vaporized from a solid or liquid phase to gaseous phase before being transferred into the vacuum system of the mass analyzer. This process (desputtering) is fairly energetic and is used to transform analytes before they are characterized.

   Some of the most commonly used ionization processes are :

   (a) Secondary Ion Mass Spectrometry (SIMS) [94]
   (b) Desorption Spray Ionization (DESI) [46]
   (c) Matrix Assisted Laser Desorption Ionization (MALDI) [45]
   (d) Nanostructure Initiator Mass Spectrometry (NIMS) [40]
   (e) Laser Desorption Ionization (LDI) [72]
   (f) Electrospray Ionization (Electrospray Ionization (ESI)) [95]

   The type of ionization process used is critical as it governs the: (i) spatial resolution of the observed sample specimen (ii) the type of analytes which are ionized efficiently and (iii) the sensitivity of the analysis. Different ionization processes require different sample preparation steps.

2. **Mass analyzer** - Mass analyzer plays an important role for discriminating ions according to their mass-to-charge ratio and their structure before these ions reach the detector. The process of transferring ions to the mass analyzer can be done immediately after the ionization process. Based on various specifications such as mass resolution, manipulation of analyte ions (continuous or pulsed), length and quality of the vacuum system being utilized, different types of mass analyzers can be used:

(a) Quadrupole instruments

(b) Time of flight instruments

(c) Ion traps

(d) Ion cyclotron resonance

3. **Detector** - After ions are separated in the mass analyzer, they need to be detected. Commonly used detectors have multi channel plates where ions strike the detector's surface inside an individual channel, thereby inducing secondary electron and photon emission. These secondary electrons now move towards the detector striking tunnels, again generating more electrons. This phenomenon continues inside a channel until an electron reaches the conductor which transmits the current generated to the amplifiers thus completing the process of detection. One of the important instrumental parameter for IMS is detector sensitivity. This parameter controls the detection of analyte molecules present in the specimen. Increase in sensitivity will lead to detection of low abundant molecules but at the cost of increased noise levels. Increased sensitivity also degrades the lifetime of the detector.

All these three components together produce an output in the form of measurements of m/z and intensity. Different ionization sources can be combined with different mass analyzers and detectors to detect different analytes present in the sample specimen. For example, Quadrupole mass analyzers are shown to be compatible with continuous ionization sources such as DESI. In this case, both the mass analyzer and the ionization source are suitable for analysis of analyte molecules of low molecular weights. Similarly, orbitrap mass analyzers combine well with ESI sources where multiply charged analyte ions are generated so that a broader mass range can be investigated. As a result, orbitrap mass analyzers are popular in proteomics investigations where intact protein molecules are analyzed. The resolution of the image obtained in IMS can also be increased with the combination of different components. For example, while MALDI-IMS instruments can achieve resolutions upto 1 μm, MALDI-LDI-IMS based instruments have shown to achieve a maximum resolution of 0.6 μm [82].

1-2-2 Imaging Mass Spectrometry

MS based instruments,with some modifications, can provide temporal and spatial localization of atoms and molecules with adequate resolution. IMS [63] is a powerful analytical platform that allows untargeted investigations into the underlying spatial distribution of the molecular species present in a target biological specimen. Over the years, IMS has seen tremendous development in instrumentation as well as in software which has made this sub-field an attractive avenue for study and analysis of various complex molecules present in different types of living organisms. The technique has the capability to capture the spatial as well as the chemical information of hundreds to thousands of molecules such as metabolites, lipids, peptides, proteins in a single experiment.

In IMS, tissue samples of the target specimen are sectioned into thin slices, mounted on conductive glass slides and coated with a light absorbing matrix which forms microcrystals around the target specimen. The prepared sample is then converted to gaseous state either by laser irradiation or by ESI [32]. A laser or an ion microprobe sequentially probes a discrete set

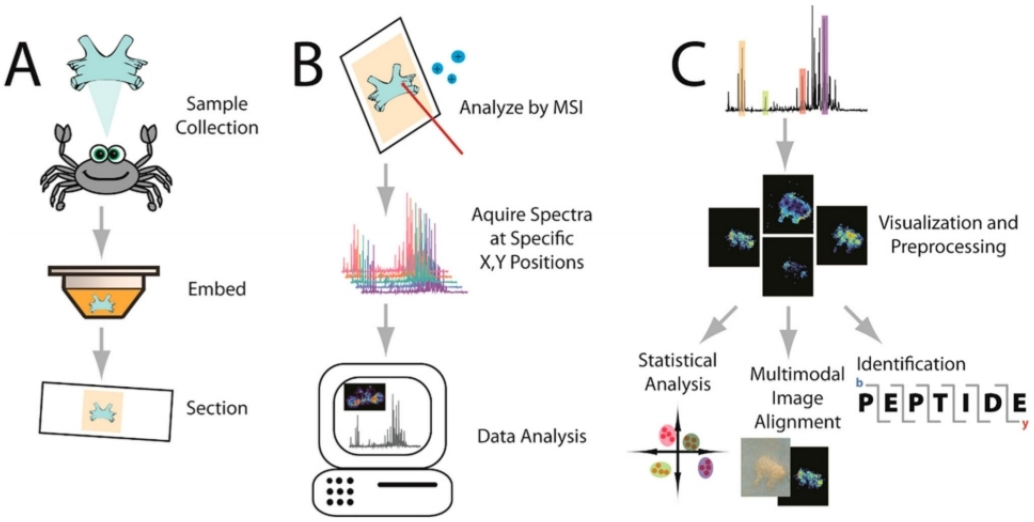


Figure 1-3: Visual Workflow for IMS analysis. (A)Sample Preparation. The brain of the target specimen(crustacean) is collected and embedded in a supporting medium for sectioning into slides. (B)Analysis. The mass spectrum for each grid point on the tissue sample is acquired using the instrument. (C) Data Processing. After preprocessing, distribution of each molecule present in the target sample is visualized and further statistical analysis procedures are conducted.[12]

of points on the surface of the target specimen in a raster pattern. The spatial resolution of the image obtained is governed by the size of the laser spot on the surface, spacing between the points on the surface of the target specimen and the type of sample preparation technique being used. The resolution of the image obtained vary from 1 μm - 5 μm in commercial instruments although higher resolutions of $\leq$ 1 μm can be achieved using advanced optical systems [41]. The individual mass spectra obtained for each point probed on the surface of the target specimen is stored digitally. Custom softwares help in selection of an analyte signal from an array of mass spectra and in plotting the intensity of the analyte signal present across sample surface. The intensity of the signal is represented by a color scale and as a result, an image of an analyte's ion distribution is generated (Figure (1-3)).

1-2-3 Liquid Chromatography Mass Spectrometry

In order to enhance the specificity and the sensitivity of the MS instrument, coupling of MS instruments with chromatographic detectors was found to be extremely desirable. LC-MS is one of the hyphenated analytical techniques that combines two techniques for analysis of mixtures of organic compounds. The breakthrough for LC-MS was via the development of ESI technique [ref]. The technique heavily improved the performance of LC-MS instruments and had a great impact in the field of proteomics [75] to an extent that laboratories were able to use this adapted technology.

With the modification of the instrumentation, LC-MS found its application in various fields especially in the field of clinical biochemistry. LC-MS is extensively used in the field of Therapeutic drug monitoring [3] where the technique is used to study immunosupressants and anticancer drugs. In the field of toxicology, the instrument found its use for analysis

of toxic compounds via direct analysis of urine samples [96]. LC-MS also finds its use in drug development [54] as the instrument allows fast analysis and structural identification of different compounds thus speeding up the process of drug testing and development.

1-2-4 Ion Mobility Spectrometry

Ion Mobility Spectrometry (IM) refers to study of movement of ions present in gases under the influence of an electric field. The analytical technique gained traction in the 1960's when it was demonstrated for screening chemical vapors to check the presence of trace quantities of hazardous compounds. In order to enhance the sensitivity and selectivity of the instrument, research efforts were dedicated towards the instrumentation side of this technique. As a result, portable IMS instruments were developed and were utilized for detecting explosives and chemical warfare agents for military operations [59].

Principle of Operation

The main principle of IM is to separate ions in an inert gas under the influence of an electric field. The applied electric field (E) forces the analyte ions to move through the buffer gas with a velocity v_d which is specific to the analyte ions' mobility (K). This is represented as :

$$K = \frac{v_d}{E} \tag{1-1}$$

Depending on the IM method used, the technique separates the ions by their differences in mobility in either space or time. Mobility for each ion is also measured as a function of other external parameters such as temperature and pressure which are often normalized to standard pressure and standard temperature for the calculation of reduced mobility K_0. This is given as :

$$K_0 = K\frac{p}{p_0}\frac{T_0}{T} \tag{1-2}$$

K_0 is a valuable piece of information and is used in standalone IM systems as a means for identification of molecules. The mobility and the m/z values can be used to calculate the Collision Cross Section (CCS) or Ω which provides information about the conformation of the analyte ions travelling in the drift tube. For a particular analyte, the CCS value can be calculated using the Mason Schamp's Equation [48] as :

$$\Omega = \frac{\frac{3}{16}\left(\frac{2\pi}{\mu k_b T}\right)^{1/2} ze}{N_0 K_0} \tag{1-3}$$

where e = electron charge, z = ion charge, N_0 = buffer gas density, μ = reduced mass of collision parameters, k_b = Boltzmann's constant and T = drift region temperature. The parameters such as buffer gas density, temperature vary for different IM platform and are heavily dependent on the experiment. Figure (1-4) presents some of various the instrumentation involved in IM and highlights their specifications along with their commercial vendors.

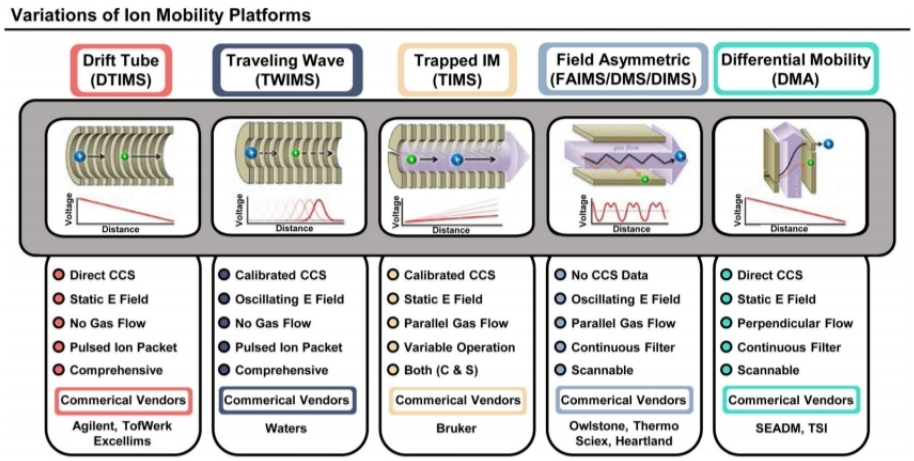


Figure 1-4: Illustration of the various IMS platforms.The figures illustrate the different methods of variation of electric field gradient used in different IMS platforms. The bullet points highlight the instruments' key attributes which includes : ability to measure CCS information, type of electric field used, type of gas flow, ion packet distribution and the footprint of the instrument. The last bullet presents the list of companies which manufactures the respective IMS method.[26]

Applications of Ion mobility mass spectrometry

Isomer Separation

While MS instruments provide high resolving power when differentiating between analytes of different masses, separation of isomeric (or isobaric) species require fragmentation methods or chromatographic techniques along with MS measurements. By utilizing the structural differences in the mobility dimension, IM-MS provides complementary information for separation of isomers. Using IM-MS, isomeric compounds have been identified and separated in various biological classes such as lipids [53], peptides [73]. IM-MS has also been successfully implemented in separation of compounds with different conformation (enantiomers) [67]. Thus, while isomeric separations in complex biological mixtures still remain a challenge, improvements in Ion Mobility Spectrometry-Mass Spectrometry (IM-MS) and chromatographic based separation techniques are starting to provide the necessary resolving power to characterize and separate these compounds.

Signal Filtering by IM-MS

By providing a mobility dimension for characterization of analyte ions, IM-MS based techniques have shown to increase the Signal-to-Noise Ratio (SNR) for specific analyte ions and decrease the background noise. IM-MS based techniques such as Differential Mobility Analyzer (DMA), which is used to detect large analyte particles, have been shown to operate as intrinsic mobility filters, where the instrumentation increases the SNR for a particular class of analyte molecules. Mobility based filtering is important for for standalone IM systems and has been used in the analysis of chemical vapors [65]. IM systems have also been used to separate contaminant ions from proteins and peptides of interest [9]. Drift Tube Ion Mobility

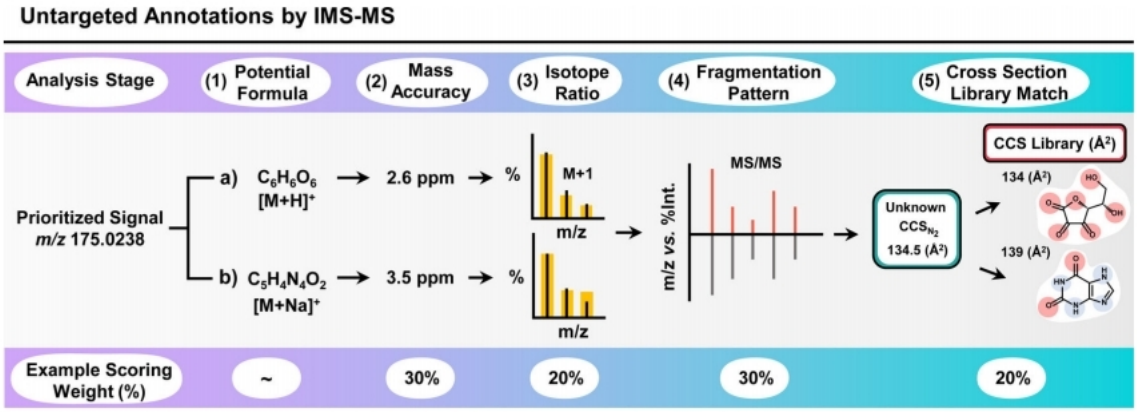


Figure 1-5: Workflow for Untargeted Analysis by IM-MS. After a molecule of specific molecular weight is detected by the instrument, additional information such as Mass Accuracy, Isotope Ratio, Fragmentation Pattern and Cross References with existing CCS Libraries are incorporated to increase the confidence on the spatial and formulaic structure of the molecule detected.[26]

Spectrometry (DTIMS) and Travelling Wave Ion Mobility Spectrometry (TWIMS) have been extremely significant in the proteomic and metabolomic analyses of complex samples such as soil, plant samples which possess high amounts of contaminant ions [14]. This is done by separating molecules of interest such from high concentrations of contaminant materials in the mobility dimension thereby increasing the proteome coverage of environmental samples.

Untargeted Analysis by IM-MS

One of the major advantages of IM-MS systems is its timescale of operation. Since IM-MS separates analyte molecules on a milisecond timescale, it can be easily nested into existing MS/LC-MS systems to increase structural confidence in detection of the prioritized feature (Figure (1-5)). Incorporating mobility based information as additional descriptors can increase the confidence of a molecule being correctly annotated in untargeted approaches. However, encorporating CCS values for untargeted modes would be currently called 'known-unknowns' as the target analyte molecule should have been previously characterized by a previous mobility experiment and uploaded in the CCS database. Characterization of 'unknown-unknowns' is much more challenging where there is no support provided by the CCS database. In these cases, ratio of mass to mobility(also called 'mass-mobility trendline') is very useful in characterizing unknown analyte molecules into a particular biological class [62]. These mass-mobility trendlines (Figure (1-6)) are established by previously calculated CCS values and are extrapolated to characterize unknown-unknowns. From a metabolomics' perspective, generation of high confidence and reproducible mobility measurements for establishing a base library is one of the important challenges in the IM community.

Ion Mobility Spectrometry-Imaging Mass Spectrometry (IM-IMS)

While IM in itself is a powerful standalone device, the interfacing of IM-IMS enhanced the analytical prowess of the MS instruments as both techniques provide complementary infor-

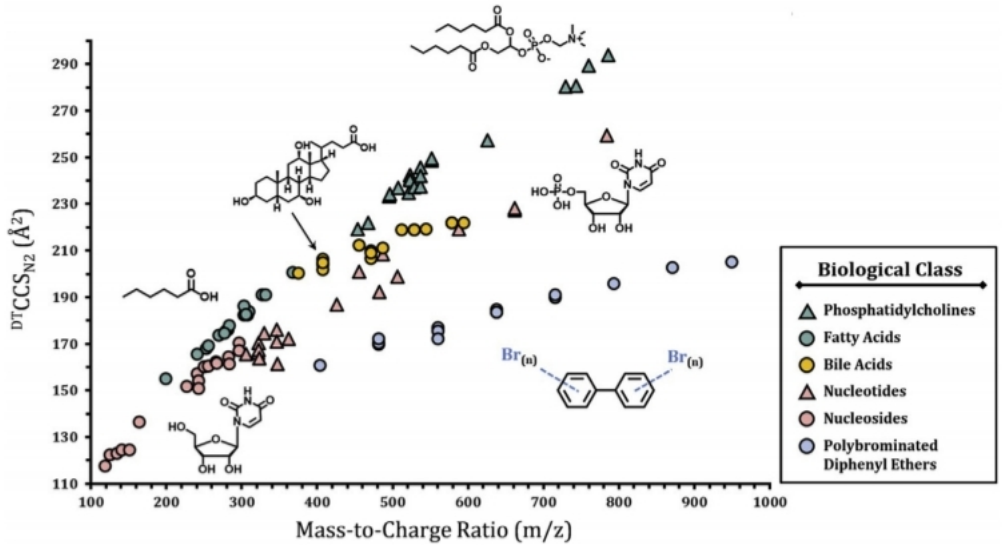


Figure 1-6: CCS Trend lines. The distinct trend lines observed for different class of molecules using DTIMS and N_2 as the drift gas.[14]

mation resulting in multidimensional characterization of the sample analytes. The interfacing results obtained are of high resolution in the chemical space as the complementary information of mobility and m/z provide very high level of selectivity and specificity [62]. Spraggins et al. [83] demonstrated the use of IM-IMS in the field of molecular imaging. The instrument utilizes the application of MALDI based IM technique along with Trapped Ion Mobility Spectrometry (TIMS) (Figure (1-7)). The instrument was shown to achieve 10 μm spatial resolution with m/z error of less than 5 parts per million (ppm). Here, IM based techniques proved to be highly important as the instrumentation provided rapid separations of analyte molecules (in the order of $\mu s - ms$) compared to chromatography based separation techniques(in the order of min-hr) thus making IM based systems suitable for imaging application. The instrumentation was also shown to resolve isomeric and isobaric metabolites in the low molecular weight region while maintaining high spatial resolution. This combination of high resolution imaging combined with IM separations can be used to address many of the challenges currently present in molecular imaging based applications.

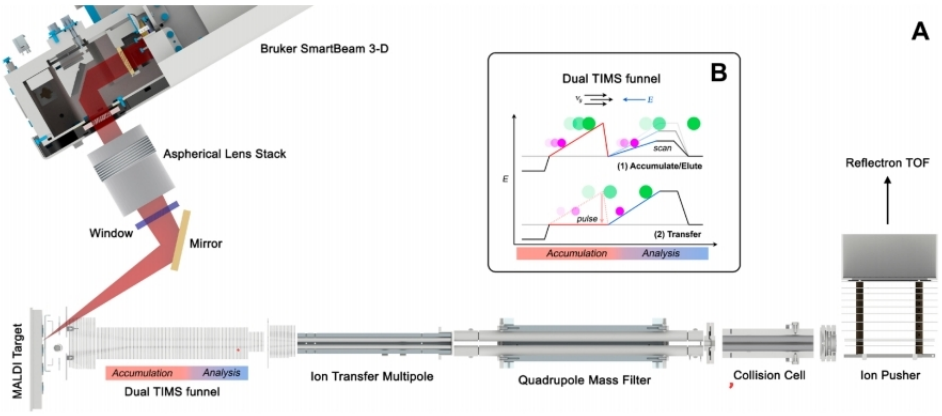


Figure 1-7: (A) Illustration of the instrumentation of MALDI timsTOF mass spectrometer. The instrument is able to provide high-spatial resolution for the sample specimen using the MALDI source for ionization. This is coupled with TIMS for mobility based separation of molecular species. (B) Ion Mobility separation in TIMS. The electric field inside the TIMS funnel is adjusted to allow separation of molecules based on their mobility. The electric field is first raised for accumulation of molecules in the funnel. This is followed by lowering of the electric field in a pulsed manner where the accumulated ions in the funnel are released based on their mobility.[83]

l

Chapter 2

Data Analysis for Mass Spectrometry

MS data is difficult to process for a number of reasons. With increasing resolution of the instrument as well as the increasing dimensions, the technique has experienced an exponential growth in data file sizes. This emphasizes the need for key software developments to ensure that effective analysis can be done without the loss of valuable information in the process. We present some of the key data processing steps that are used for MS data analysis. As an example, Figure (2-1) shows the complete data processing pipeline for LC-MS platforms.

2-1 Data Analysis Pipeline for MS

2-1-1 Visualization

IMS requires a visualization of distribution of various molecules throughout the tissue. Each pixel of an image, produced by the instrument, contains an entire spectrum of chemical information. Therefore, special software is needed to handle these kind of spectral images. Recent efforts have been made to design open-source visualization tools that are user-friendly and are applicable for multiple instruments. Out of the many softwares developed, MSIReader [77] is one of the famous open-source software for visualization of IMS datasets that provides a visual graphical interface as well as a MATLAB open source code for users. Other famous softwares include MassImager [44] and LipostarMSI [88]. Apart from accessibility of the software, efforts have been made in the direction of 3D visualization of MALDI Imaging datasets. Patterson et al. [71] have developed an open source software for 3D reconstruction of IMS datasets using multivariate segmentation.

One of the key points in visualization of IMS data is that the software should ensure that the image shown is an accurate representation of the distribution of molecules. Cropping images to remove chemical/background artefacts is not encouraged as this may lead to skewed representation of distribution of molecules. Therefore, background details should be preserved to capture the correct distribution of molecules [70]. With mass spectrometry, making an

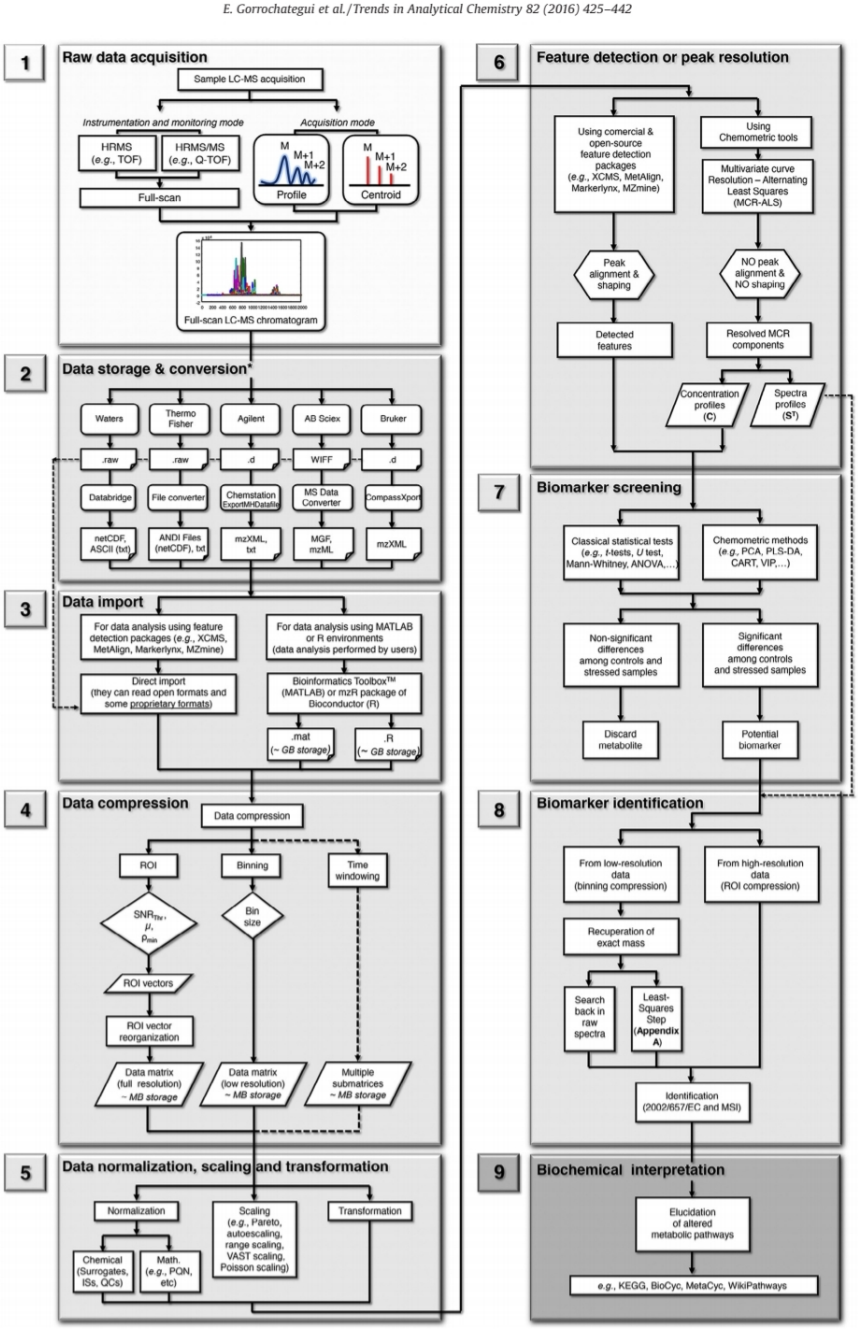


Figure 2-1: Flowchart for the LC-MS Data Analysis. The flowchart listing consists of nine different steps for untargeted analysis of target specimen. These steps can be grouped into four namely : Raw Data Acquisition, Data Processing, Feature Detection and Biomarker Identification. Parallelograms indicate data matrices. Rectangles indicate processing steps, Diamonds indicate key choices, corners indicate file format choices and rounded rectangles indicate vendors and their choice of software.[38]

increased presence in the case of biomedical applications as a diagnostic tool, appropriate visualisation tools are critical for accurate diagnosis.

2-1-2 Data Compression

Multidimensional MS acquisitions tend to create large data files. As a result, data processing becomes difficult and requires demanding computational methods. To alleviate this problem, several data compression strategies have been implemented. Among them, 'binning' and Region Of Interest (ROI) are two of the most successful strategies that require the least amount of computational strain for 2D MS datasets.

1. **Binning**

   Binning in 2D LC-MS datasets involves the transformation of raw data into a two dimensional matrix representation with m/z values in one dimension and retention time values in the other dimension. Conversion of this high resolution data into a matrix representation requires division of the m/z axis into predefined (equidistant or custom) sections. These predefined sections are referred to as bins and they determine the resolution of the instrument in the m/z dimension. Thus, by dividing the m/z dimension in predefined bins, data compression as well as a compact matrix representation of the data are obtained simultaneously. However, a drawback of the binning procedure is finding the appropriate bin size. This is a very critical parameter and needs to be chosen carefully, as the bin size is strongly correlated with the chromatographic peak profile. If the bin size is too small, it may lead to generation of spurious chromatographic peak profiles. If the bin size is too large, it may lead to merging of multiple peak profiles which will lead to loss of spectral resolution in the m/z dimension.

2. **ROI Compression**

   ROI based compression is an alternative compression technique to binning. The idea was proposed by Stolt et al. [85] and is based on the concept of considering analytes located in highly dense regions of data points surrounded by sparse regions of data points ('data voids'). These highly dense region of data points contain interesting mass traces and have significantly higher intensity than the established Signal-to-Noise Ratio (SNR) threshold. In order to be classified as an ROI, the region must contain a minimum number of consecutive data points within the mass deviation range, usually based on the mass accuracy of the spectrometer. This condition prevents noise to be classified as an ROI. However, additional filtering may be needed to prevent chemical/instrumental artefacts present in the data sample to be classified as an ROI. The identified ROIs are stored in a list and then later reorganized in the form of a matrix with retention time as one dimension and m/z mean value of the ROI as the other dimension.
   ROI based compression, circumvents the problem of defining bin size presented in the binning procedure without the loss of spectral information.

2-1-3 Normalization

Normalization is used to remove systematic artefacts that are present in the mass spectra. The artefacts present maybe due to matrix application, ion suppression or differential ionization efficiencies in complex samples and can influence the intensity of peaks detected in the mass spectra. These random effects can be reduced by various techniques present for the normalization of the mass spectra. A brief outline of the various normalization techniques is presented below:

1. **Normalization to the Total Intensity Count (TIC)**

   Normalization to the TIC is the most commonly implemented normalization method in MS data analysis [24]. The total intensity of the particular mass spectrum is calculated and the intensity of each m/z of that mass spectrum is divided by this quantity. Normalization to the TIC ensures that all signals have the same integrated area and the underlying assumption is that there are comparable number of signals in each spectra of the image pixel. However, this assumption fails when the selection of the sample area is variable run-to-run leading to uneven distribution of molecules in each spectra.

2. **Normalization to matrix related peaks**

   In addition to normalization to the TIC, the data sample can be further normalized to matrix related peaks for MALDI imaging experiments to check for uneven application of the matrix coating [37]. Different types of matrix coating on the sample lead to different distribution of matrix ions after ionization and so matrix ion signals can be used as a reference signal for normalization.

3. **Normalization to Internal Standards**

   For samples with different tissue types, an externally applied Internal Standards (IS) similar to compound of interest should be applied before or during matrix application. In this case, normalization of each spectrum is done with respect to the intensity of reference molecule. Normalization to an IS [11] reduces the impact of ion suppression that arises from tissue inhomogeneity and improves the pixel-to-pixel variability. Other options include normalization to an endogenous molecule that is consistently present throughout the whole tissue.

2-1-4 Feature Detection

Feature detection is the task of searching for peaks which can be defined as bounded two-dimensional MS signals with a local maxima and a relatively high SNR (Figure (2-2)). These bounded two-dimensional signals indicate the presence of an analyte and can be in the range of few hundreds to thousands depending on the complexity of the sample. There are various methods proposed for carrying out feature detection in the LC-MS dataset. Centwave [87] utilizes the concept of ROI for data compression and then look for features in the chromatographic profile of the compressed data. Trevino et al. [89] proposed a grid-based method for

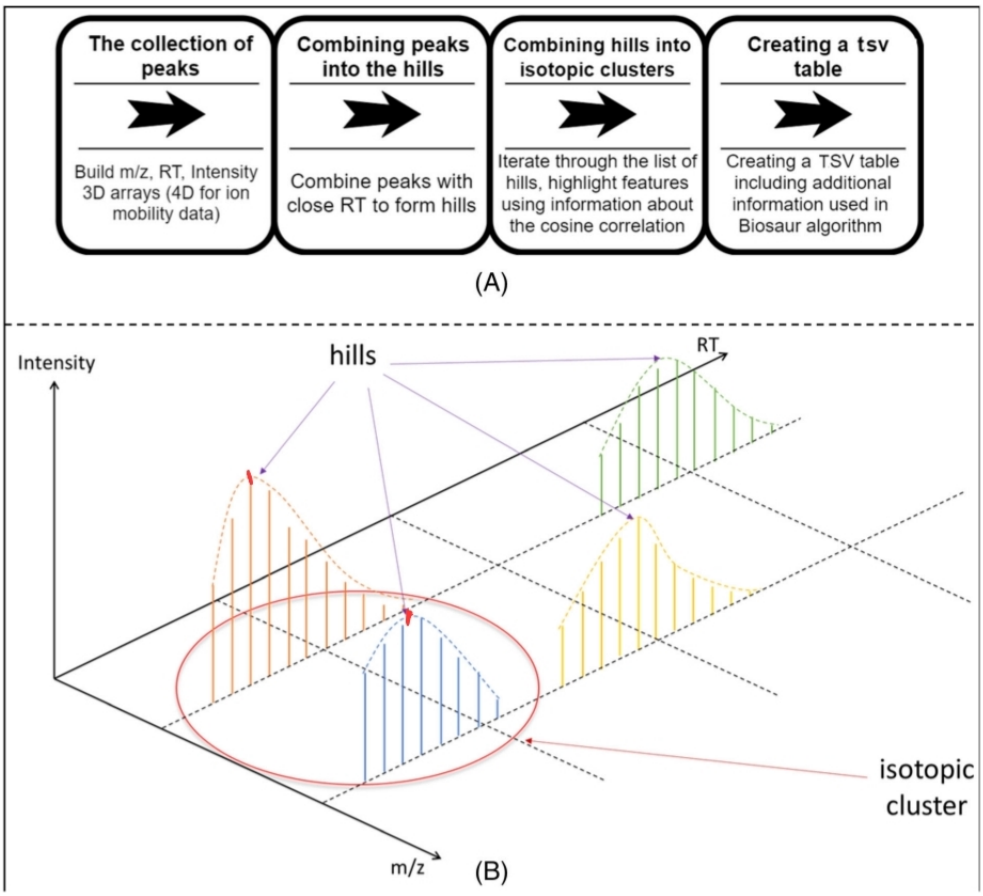


Figure 2-2: Feature Detection. (A) Schematic Workflow of a Feature Detection algorithm is presented. (B) Diagrammatic Representation of peaks present in an LC-MS data sample. The presence of hills indicate the possibility of a feature in the data sample. Based on the m/z values of the hills, as well as the Retention Time (RT) and the peak intensity information, the algorithm combines several hills to form isotopic clusters.[1]

locating features where the algorithm directly operates on the raw 2D data sample. Some algorithms incorporate prior information about the shape of the feature [93] and then use this information to look for features in the data sample. Most of these algorithms require prior pre-processing such as feature alignment and peak shaping prior to feature detection. On the other hand, chemometric techniques [80] have also been proposed to resolve LC-MS datasets. These techniques have the advantage that they allow feature detection without applying any pre-processing techniques. An additional goal of feature detection is to distinguish analytes present in the data sample from false positives.

2-1-5 Biomarker Identification

Following feature detection, the next step involves identification of isolated metabolites from a referential database. Identification of target metabolites is an active area of research [21] and involves hierarchical strategies for correct identification. According to Sumner et al. [86],

four levels of identification of molecules can be defined starting from definitive identification (level 1), which refers to matching of at least two orthogonal molecular properties of metabolite found in the data sample with an authentic chemical standard. Levels 2 and 3 comprise of putative metabolite identification, which provides metabolite-specific or class-specific identification and can be compared against various datasets. The final level (level 4) consists of identification of unknown compounds in which case the method and platform used for identification of these compounds should be presented.

Biomarker identification is a complex task and the task becomes even more difficult when untargeted metabolomic profiles are generated. The review by Dunn et al. [31] provides an extensive study of all the computational tools available for untargeted metabolomic studies. The review concludes that in the past decades, the number of unknown metabolites discovered, due to enhanced resolution of mass spectrometers as well as addition of other chemical dimensions (e.g. chromatography profile), have increased. However, the proportion of identified metabolites, with respect to unknown metabolites, still remain low (around 50%). Therefore, development of efficient computational and identification strategies is a widely pursued research interest.

2-1-6 Biochemical Interpretation

The overall process of data analysis concludes with the biological interpretation of the results linked with the identified biomarkers. The final result of the analysis is the confirmation or rejection of an altered candidate biomarker driven by an initial biological hypothesis. These results are then deciphered with the help of online databases. Various online databases allow interpretation and cross-validation of altered metabolic pathways present in the sample specimen, such as KEGG [47] and BioCyc [49]. With the help of these databases, alteration to metabolic pathways can be studied in detail which can potentially contribute to identification of new metabolic checkpoints.

2-2 Feature Detection

One of the initial steps of multidimensional MS Data analysis of complex biological samples is to separate the information from the noise. This procedure is carried out in feature detection. In a multidimensional MS data sample, a feature refers to bounded signal of interest that indicates the presence of a molecule present in the biological sample. Now, due to the complex nature of the data, with challenges such as varying feature widths, presence of chemical and background artefacts, increasing size of data samples, irregular sampling, the procedure of finding and identifying features becomes a complex task. In the past decades, different open source algorithms and software pipelines have been developed to tackle these challenges.We broadly these algorithms based on their method of operation.

2-2-1 Types of Algorithms

1. **ROI based methods**

The algorithm first performs compression of the data sample based on the application of ROI. This is then followed by peak deconvolution where the identified region of interests are then deconvolved into separate 2D peaks, with each peak indicating the presence of a molecule or an information of interest. This method was first implemented in Centwave [87] and has been widely modified and adapted for other algorithms. Abdrakhimov et al. [1] and Navarro et al. [68] provide two adaptations of this algorithm for LC-MS and IM-MS based instruments. One of the main reasons for the method being widely popular is that it requires the least amount of computational efficiency for the compression technique and can separate information from noise effectively.

2. **Grid Based Methods**

   These algorithms treat 2D MS data as an image and use image based algorithms to separate features from noise. Trevino et al. [89] proposed a grid based method for feature detection in Liquid Chromatography Mass Spectrometry (LC-MS) data samples. The method has the advantage that it attacks the 2D data space directly instead of processing each dimensions separately.

3. **Model Based Approaches**

   These algorithms assumes a prior model on the shape of the feature and then use fitting procedures to identify features present in the data sample. Literature suggests that an ideal feature present in the data sample should have a Gaussian or a Lorentzian shape [93]. Cox et al. [20] and Samanipour et al. [78] have proposed feature detection algorithms that assume features to have a Gaussian shape and then perform Gaussian fitting to look for Gaussian features present in the LC-MS dataset.

Having given a brief outline of the feature detection algorithms that are present, we briefly present the working mechanism of some of the algorithms that are already being used in the case of 2D Mass Spectrometry (MS) data sets

2-2-2 XCMS : Highly Sensitive feature detection for high resolution LC/MS

Tautenhahn et al. [87] describe the development of a feature detection algorithm CentWave for high-resolution LC-MS datasets using region of interest algorithm and Continuous Wavelet Transform (CWT).

The mechanishm of the algorithm can be broken down into two major steps :

1. The algorithm uses a density based approach [85] to identify regions of potential mass traces(ROI). These ROIs are then filtered using an intensity based prefilter. Figure 2-3 shows an identified ROI present in a test sample.

2. This is then followed by application of continuous wavelet transform techniques to deconvolve the peaks present in the identified region of interests. CWT techniques have been actively used [55] for detection of 'peaks' (1D features) for MALDI-time of flight mass spectrometry. Figure 2-4 demonstrates the use of CWT for deconvolution of chromatographic peaks present in the signal.

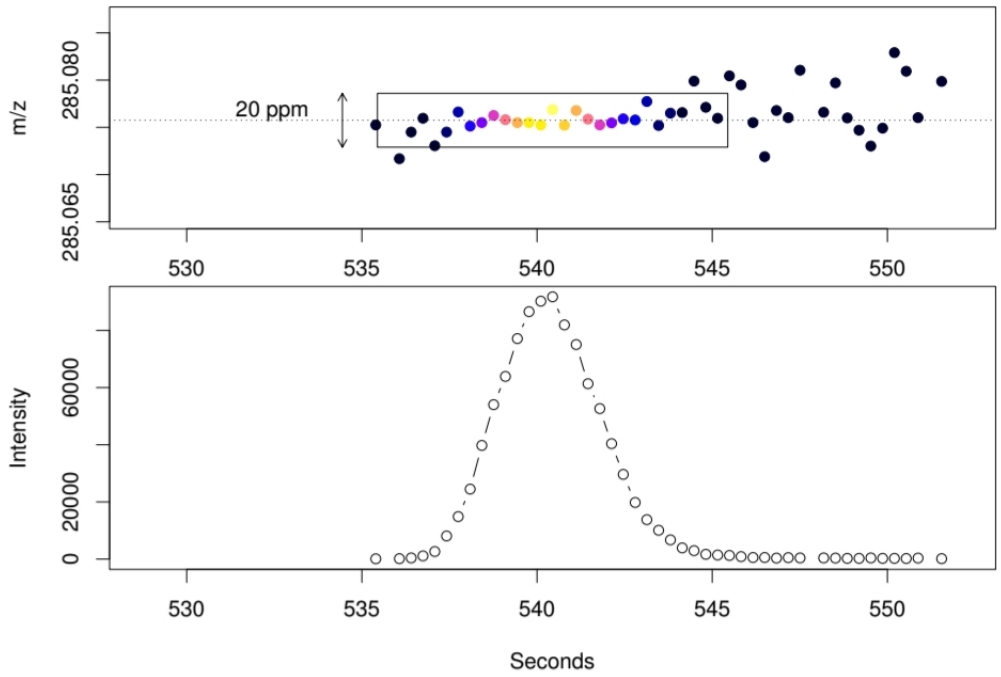


Figure 2-3: ROI Detection. The upper panel shows the mass trace of the mass signal with color coded intensities. The corresponding chromatographic peak is displayed below.[87]

Parameters

The algorithm identifies three important parameters. These are :

1. Mass Deviation (μ) in ppm - typically set to be the multiple of the mass accuracy of the mass spectrometer.

2. Prefilter (I) - ROIs are only retained if they contain at least k peaks with intensity >= I.

3. Chromatographic Peak width range - e.g w_{min}, w_{max} in seconds for UPLC separation.

4. SNR threshold (SNR_{thr}) - Threshold Signal/Noise ratio. Signal/Noise ratio is defined as (maxo - baseline)/sd, where maxo is the maximum peak intensity,baseline the estimated baseline value and sd the standard deviation of local chromatographic noise.

Advantages and Disadvantages

1. **Advantages**

   (a) The ROI based compression method is one of the least computationally expensive methods for capturing interesting mass traces thereby drastically reducing the computational time of the algorithm.

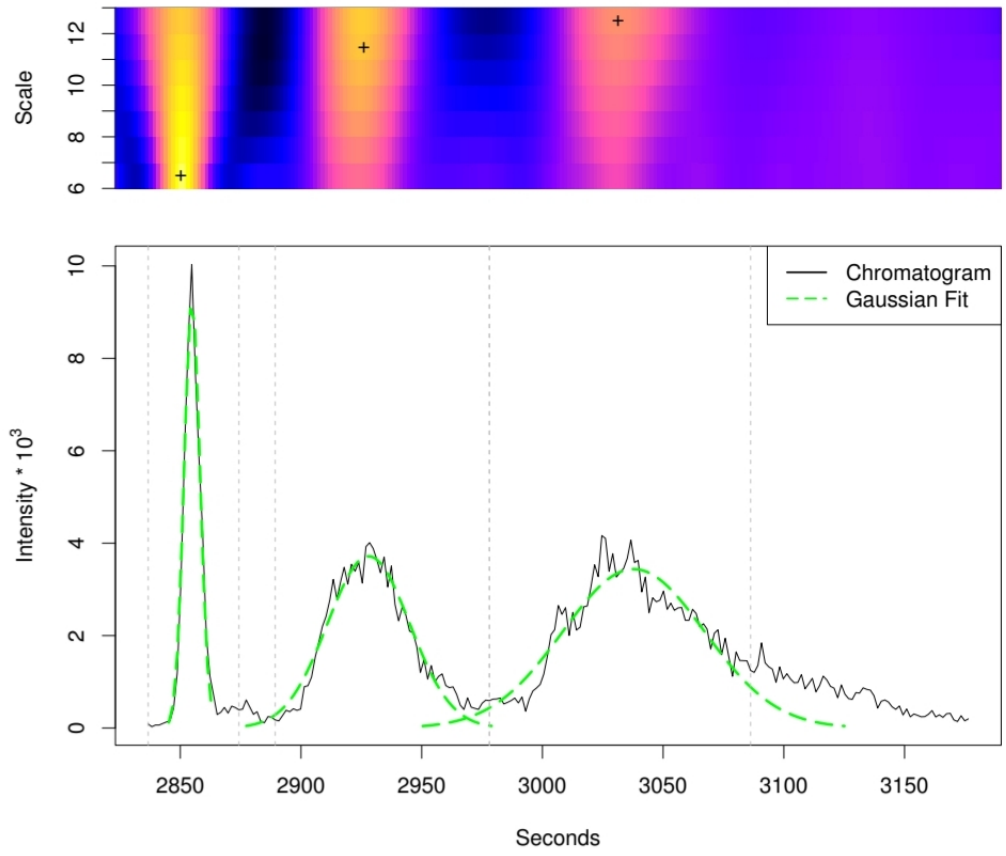


Figure 2-4: Chromatogram Peak Detection using wavelet transforms. The lower panel shows the extracted ion chromatogram and the various gaussian peaks observed in the chromatogram. The upper panel shows wavelet coefficients at different scales for the same chromatogram. The cross mark indicates the scale at which the specific peak is optimally localized.[87]

(b) CWT outperforms standard filters of fixed width in the case of peak deconvolution. This is because CWT coefficients are able to capture peaks of varying widths at different scales and is therefore excellent for peak deconvolution.

2. **Disadvantages**

(a) ROI based compression is customized for centroid mode acquisitions and the results are relatively poor when the algorithm is adapted for profile mode acquisitions.

(b) The parameters need to optimized with respect to the data sample. When the algorithm was compared with other feature detection algorithms, it was found that the parameter 'intensity threshold' had to be optimized continuously in order to discover more features.

(c) The algorithm does not pose any constraint on ridge lines constructed using continuous wavelet transform [66]. Ridge line detection is the process of detecting and connecting local wavelet maximum coefficients across all scales. Centwave dose not pose any constraint on the minimum number of scales at which the the local maximum coefficient should be present. As a result, the chance of detecting false positive features increases.

(d) While calculating SNR, the algorithm establishes a box (region) around the detected ROI and uses this box to estimate the baseline value and the standard deviation of the local noise for the detected ROI. The size of this box depends on the mass resolution of the spectrometer and the chromatographic peak width range, both of which are instrument dependent parameters of which the user may not be aware of.

2-2-3 Gridmass: a fast two-dimensional feature detection method for LC/MS

Trevino et al.[89] present another feature detection algorithm for High Performance Liquid Chromatography coupled to Mass Spectrometry (HPLC/MS) experiments. The paper describes feature detection as the procedure to detect boundaries of a putative molecule within the mass and time domains. In order to improve the computational efficiency of the process, the paper proposes a direct two dimensional approach to feature detection rather than performing a two-step peak detection. In this study, we briefly summarize the feature detection pipeline implemented and discuss the advantages and disadvantages related with this method.

HPLC/MS data sample can be thought of as a 2D image with m/z and retention time acting as the two dimensions of the image and the intensity (amount of molecules detected) being the color of the image. The algorithm first generates a set of equally spaced probes that span the entire m/z-retention time dimension. The probes are allowed to explore a small rectangular region in its vicinity to look for local maxima (intensity) within the rectangular region. The probe location is then shifted to the local maxima found in the region. This process is repeated until no higher values exist within the exploring rectangle. All the probes converging to the same maxima provides an estimate of the boundary of the feature detected. The local maxima found contains information about the intensity, m/z and the retention time of the feature detected. Figure (2-5) shows the overview of the implementation of the algorithm as described above. Considering the implementation of the algorithm on a noisy data sample, the algorithm is highly sensitive and specific for smooth surfaces and therefore

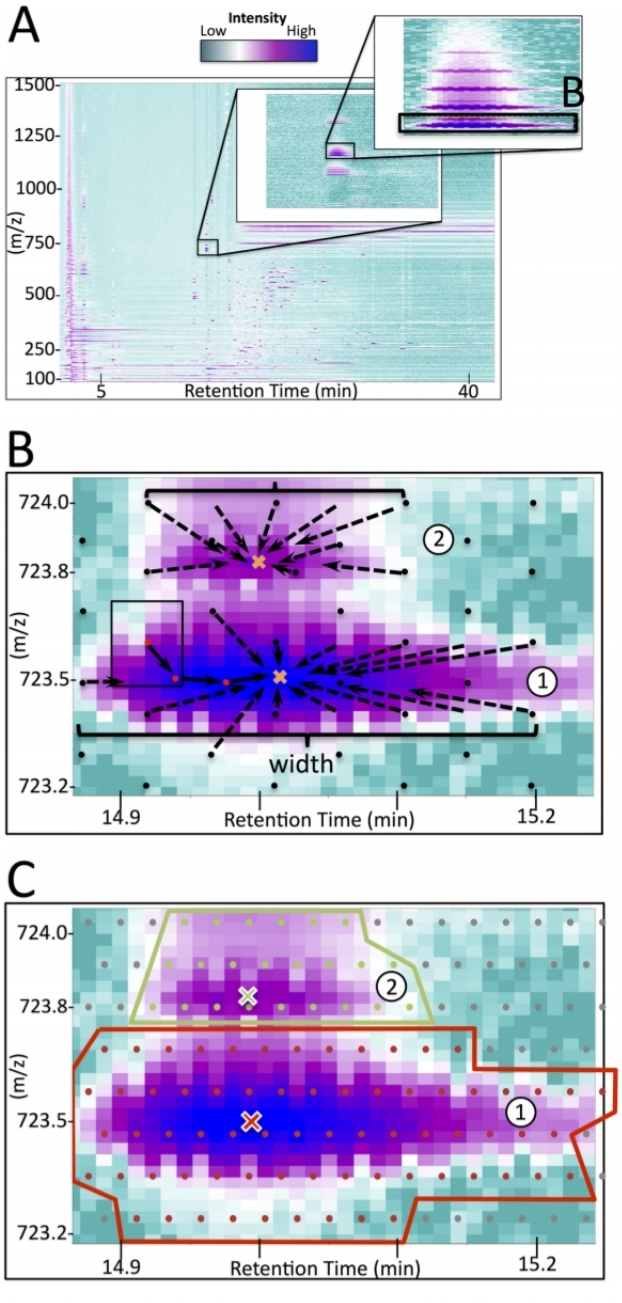


Figure 2-5: Feature Detection Gridmass.(A) The image representation of the LC-MS Data Space is presented. The intensity is drawn in log_{10} scale. (B) Depiction of the Gridmass Algorithm for two peaks found in the data space. Black dots represent the probes that will move towards there local optimum. Dashed arrows show the movement of the probes after one iteration. The red dots show the movement of one specific probe to the optimum location after a set of iterations. The area explored by each probe is limited by a rectangle. (C) Depiction of Detection of two features using the algorithm. The minimum height(intensity) threshold = 50 for this case. The green and red polygons shows the boundary estimation of those two features. The corresponding center of those features are represented by a cross mark.[89]

additional criteria(feature width, ignore times, etc.) needs to be implemented to provide some filtering in the features detected.

Parameters

The algorithm takes the following parameter as inputs:

1. Minimum Height Threshold - Intensities lower than this values will be ignored.
2. Width - Threshold width of the feature (chromatographic domain). In retention time dimension, this is equal to 4 times gap between probes.
3. m/z tolerance - Threshold width of the feature (m/z dimension). In m/z dimension, this is taken as 2 times gap between probes.
4. Intensity similarity ratio - For detecting artifact features having similar intensity and mass.
5. ignore times - list of time ranges to be ignored.
6. smoothing times - The time interval considered for smoothing the feature by averaging(retention time dimension).
7. smoothing m/z - m/z range for smoothing the feature by averaging (m/z dimension).

Advantages and Disadvantages

1. **Advantages**
   (a) The algorithm is computationally fast. By fixing the grid points, the algorithm essentially performs 1D (intensity) optimization in a local region defined by those grid points. This is relatively fast as the search space is limited.
   (b) The algorithm is highly sensitive. In principle, the algorithm in its default mode, looks only for locations based on the change in intensity. As a result, the algorithm picks up every part of the data space where change in intensity is observed. This is good for identifying 'potential features' with minimal change in intensity. These detected 'potential features' can then be filtered out based on additional filtering parameters, which is better than not getting detected at all.
   (c) The algorithm can operate on profile mode acquisition as well as centroid mode acquisition.
2. **Disadvantages**
   (a) There are many parameters that are needed to be optimized in the case of GridMass. This is because convergence to a feature using the optimization procedure is not guaranteed. This makes the software very unfriendly for users who have no information as to how the data was generated.

(b) The parameter 'Intensity Threshold' requires the user to know the SNR of the data sample. However, no such method or idea was implemented which demonstrates how to calculate the SNR for a given data sample.

(c) The parameter 'ignore times' depends on the sample and the chromatographic conditions. This means that the experimental conditions must be known by the user before operating the algorithm.

2-2-4 Self Adjusting Algorithm for the Nontargeted Feature Detection of High Resolution Mass Spectrometry Coupled with Liquid Chromatography Profile Data

Samanipour et al. [78] propose a self-adjusting feature detection algorithm for High Resolution Mass Spectrometry coupled with Liquid Chromatography (LC-HRMS) profile data. The idea proposed by the paper is to perform 2D Gaussian fitting in the profile data (generated by LC-HRMS technique) to detect features. The algorithm is self adjusting in the sense that it only requires user defined parameters as only the first guess in an adaptive process. The algorithm does not require optimization of parameters such as peak widths in the mass and time domain as in the case of previous methods. The working mechanics of the algorithm is presented below.

The algorithm is an iterative process where each point in the data space is processed individually starting with the point with the highest intensity. Once the presence of a feature is established in the chromatogram, the algorithms sets the intensity value of the feature to zero and moves to the next most intense point in the data sample. For a single feature detected, the algorithm goes through 9 steps during each iteration. These steps are :

1. Maximum Detection and Half-Height Placement - Figure (2-6) and Figure (2-7)
2. Signal Smoothing - Figure (2-8)
3. Signal Interpolation - Figure (2-9)
4. Gaussian Fit - Figure (2-10)
5. Baseline Tracing - Figure (2-11)
6. Gaussian Fit in the time Domain - Figure (2-12)
7. Signal Removal

Parameters

The algorithm takes four types of parameters :

1. **Importing Parameters** - The importing parameters include path to the file, the file format and finally the mass range limit.
2. **Stopping Parameters** - The stopping parameters consist of four thresholds namely:

(a) R^2 i.e. of regression coefficient fit - This parameter determines the quality of the recorded signal and acts as a decision making parameter as whether the signal is peak.

(b) Maximum Signal Increment - This parameter prevents grouping of overlapping features.

(c) Minimum Intensity of Peaks - Lower Bound for Peak Detection.

(d) Maximum number of Iterations - This parameter determines number of iterations to be performed for a given data sample.

3. **Filtering parameters** - These parameters are used for filtering out detected features based on their properties i.e. minimum peak width (2s), maximum peak width (300s) in time domain. The parameters are mainly used for removing time domain features that could be considered as noise/background.

4. **Performance Parameters** - These include Minimum peak width in the mass domain and minimum peak width in the chromatographic domain. This is taken as an initial guess and the algorithm will automatically adjust it according to the peak detected in the data-sample.

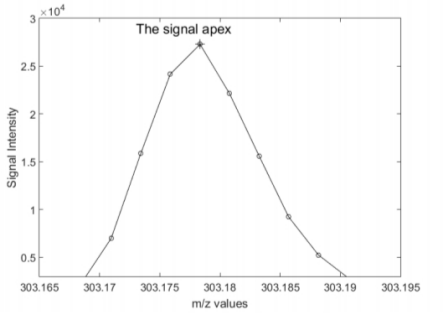


Figure 2-6: Step 1 : Maximum intensity detection in the m/z dimension for a feature present in the wastewater influent sample.[78]

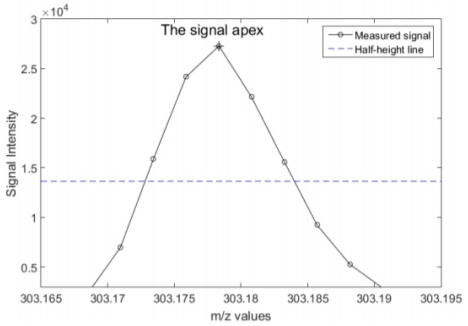


Figure 2-7: Step 2 : Detection of the half-height of a peak in the m/z dimension.[78]

Advantages and Disadvantages

1. **Advantages**

(a) The algorithm is extensive. For every intensity point, the algorithm calculates the peak width, peak shape, baseline in the m/z domain and then in the time domain. Thus every feature, along with all of their properties, gets characterized.

(b) The algorithm is self adaptive. Apart from few initial parameters, the algorithm adapts itself to the data provided without any interference.

(c) The algorithm can operate on profile mode acquisition as well as centroid mode acquisition. This is because the algorithm first operates in the m/z dimension, calculates all the necessary properties, and then moves to the retention time dimension.

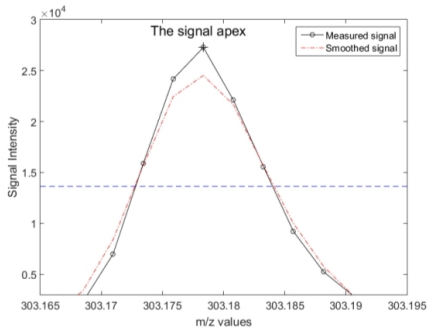


Figure 2-8: Step 3 : Smoothing of the peak using moving average filter.[78]

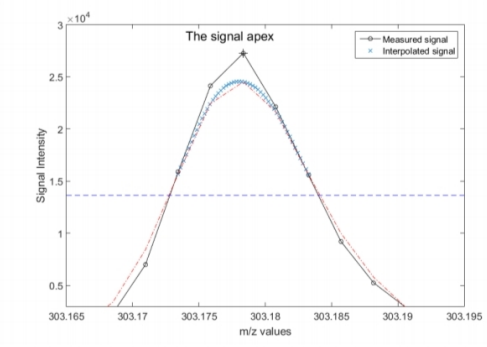


Figure 2-9: Step 4 : Interpolation of the smoothed signal using the Spline function.[78]

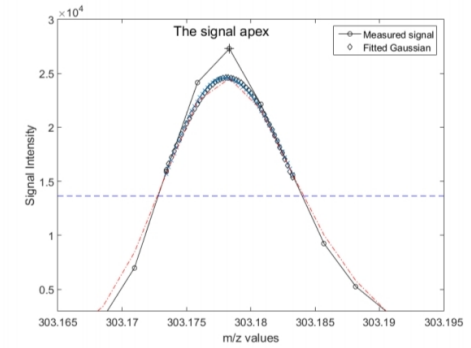


Figure 2-10: Step 5 : Fitting of the interpolated signal by a gaussian function using least squares method.[78]

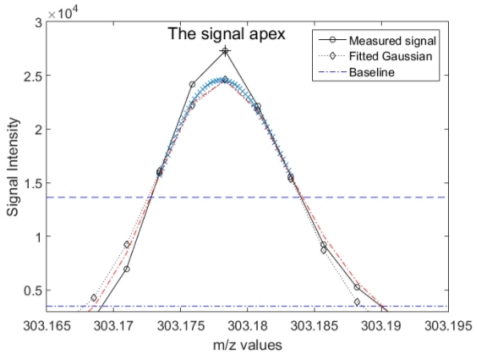


Figure 2-11: Step 6 : Tracing the baseline in the real signal through the fitted gaussian function.[78]

2. **Disadvantages**

(a) The algorithm is time consuming. The algorithm first performs gaussian fitting in the m/z dimension and then in the retention time dimension. For a particular data sample tested in their paper, the algorithm took 7 hours to generate a list of features with all of their properties.

(b) The algorithm parameter regression coefficient fit needs to be specified by the user. Using a lower value may lead to detection of high number false positive and a higher value will make the algorithm to noisy peaks.

(c) The parameter 'Maximum number of iterations' requires the user to have an estimate of the number of features that are roughly present in the data set. This might be difficult to guess, if the sample is complex, a low guess will lead to less number of features detected and a high guess will lead to increased computation time.

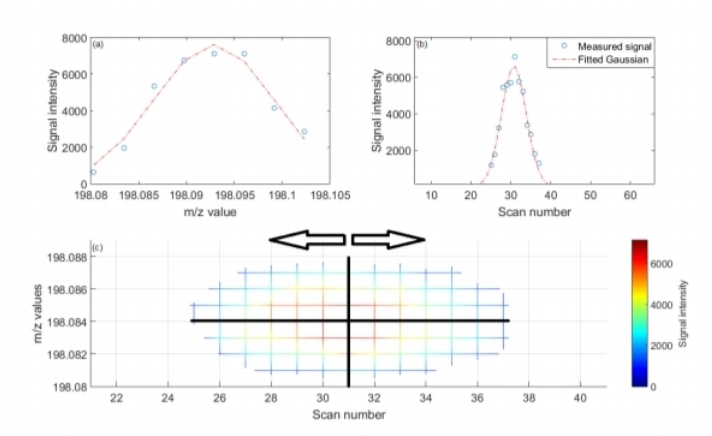


Figure 2-12: Step 7,8 : (a) The fitted Gaussian on the base peak in the m/z domain. (b) Fitted Gaussian in the Time domain. (c) a 3d overview of the algorithm moving from base peak in the m/z domain to the neighbouring scans in both direction(black arrow). [78]

(d) Although the half-height placement method bypasses the calculation of SNR by only considering the upper half of the peak, the technique might fail in the case of complex molecules where 'shoulder peaks (features)' are present in the upper half portion which might get smoothed out in Step 3.

2-2-5 Benchmarking feature detection algorithms

Performance evaluation on an annotated data sample

The performance evaluation of all the feature detection algorithms reviewed had a similar framework. The algorithms were first tested on an experimental setup with known compounds. In the case of GridMass and Centwave, the performance of the algorithms were tested on a standard MM14 data sample. The MM14 compounds consisted of a mixture marker compounds at a concentration of 20 μM which was analyzed using Electrospray Ionization based Mass Spectrometry coupled with Ultra High Performance Liquid Chromatography (UPLC/ESI-MS). Due to ESI technique being used, there were total of 296 features generated (21 features per compound). These feature consisted of adducts, fragments as well as their isotopic peaks. These features were first manually annotated before evaluating the performance of the algorithms. In case of Self Adjusting Feature Detection algorithm, the experimental setup consisted of a total of 55 samples with 4 equilibration injections, 3 internal standard injections and 44 composite wastewater influent samples. In this case, the IS were used for evaluation of the true positive and false negative detection while the equilibration samples were used for false positive detection.

After the experimental setup was prepared, the performance of the algorithms were then compared. The parameters of comparison were:

1. **No. of features detected** - The feature is said to correctly detected if the reported m/z and retention time by the algorithms are found to be ≤ 0.1 with respect to the theoretical

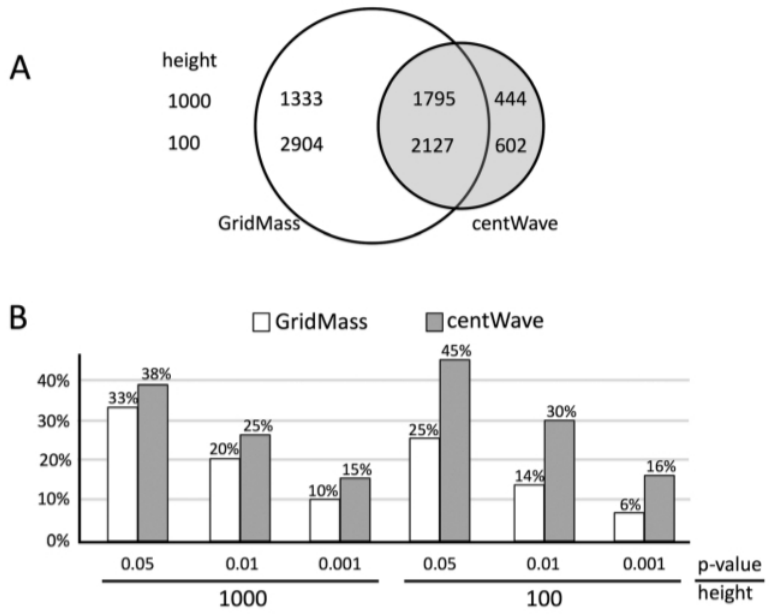


Figure 2-13: Comparison of features detected by Gridmass and Centwave from the Habanero samples. (a) Venn-Diagram Representation of features detected by the algorithms after de-isotoping using two height thresholds. (B) Percentage of false positives detected by the algorithms for the two height thresholds at three p-values that determines the false calls.[89]

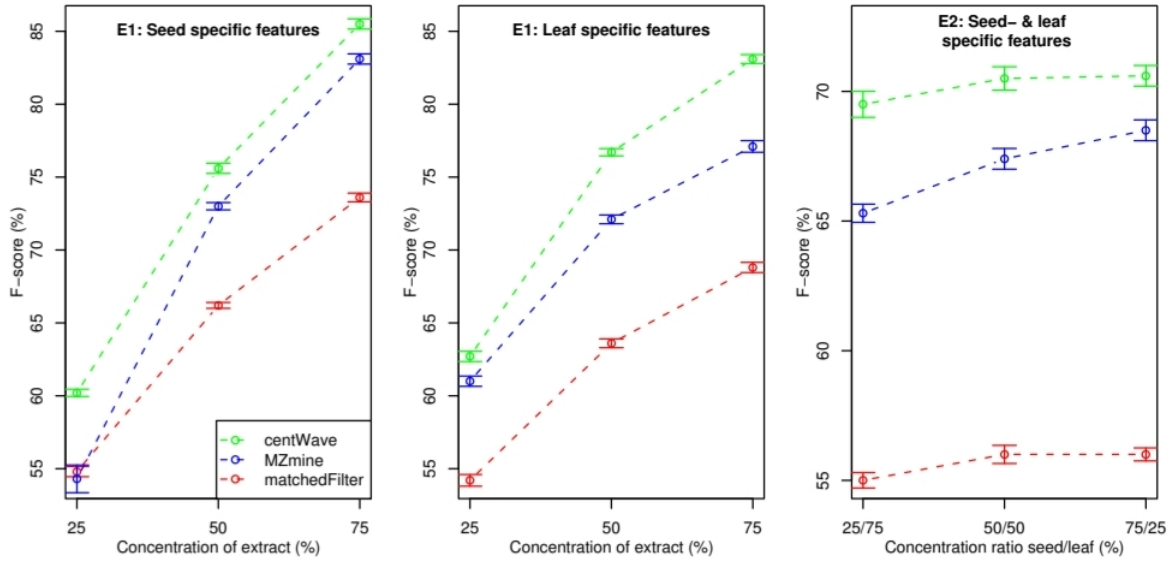


Figure 2-14: F-score values for two experiments. The first experiment(left) consisted of looking for features in dilution series of seed extract. The second experiment(middle) consisted of looking for features in dilution series of leaf extract. The third experiment consisted of looking for features in a mixture of seed and leaf extract. The F-score is the benchmark for the three feature detection algorithms in all the three experiments. Higher F-Score values represent better feature detection performance.[87]

values of the features. In the case of Self Adjusting Feature Detection Algorithm, the presence of IS compounds were needed to be confirmed. The IS are correctly detected, if the algorithm reported them within the range of $\leq$ 0.003Da in the m/z domain and a retention time of $\leq$ 10s.

2. **Computational efficiency** - The time taken by the algorithm to complete feature detection as well as the complete details of hardware(computer system) used.

3. **Parameter optimization** - The number of parameters that were needed to be optimized in order to obtain optimum results.

4. **False Discovery Rate (FDR)** - The number of false positives and false negatives detected by the algorithm.
   Various methods have been proposed for estimating the quantity of FDR. Centwave uses F-Score[90] for evaluation of the FDR. The F-Score is evaluated as :

$$F - score = \frac{2.R.P}{R + P} \tag{2-1}$$

   where $R = \frac{TP}{NP}$ and $P = \frac{TP}{N}$. TP is the number of True Positives, NP is total number of real features and N is total number of features detected by the algorithm. A perfect feature detection will achieve a F-Score of 100% and the presence of false positives and false negatives will lower its values (Figure (2-14)).
   Gridmass used rAnova[56] to estimate the FDR. The experiment comprised of using seven biological replicates of a sample specimen followed by three technical replicates injected non-consecutively. The difference between the features detected in technical replicate and the biological replicate represented the false positives.

Performance evaluation on a complex data sample

After the evaluation on a data sample with annotated features, the algorithms were tested on complex samples where the number of features are unknown. In these cases, the evaluation of the algorithms becomes difficult. One of the first criteria for performance comparison is to identify the number of features which are detected by both of the algorithms. This is then put into perspective by taking into account the total number of features detected by each algorithm individually. Venn diagram constructions are often used for visualizing the results of this procedure (Figure (2-13)). Additional information such as feature width, differences in m/z width and retention are also compared.

Chapter 3

Wavelet Transforms

3-1 Introduction to wavelet transforms

Wavelet analysis refers to a class of time-frequency (or time-scale) representation of a signal and is a standard tool used in signal and image processing applications.The technique has found its application across various fields of science such as geophysics [91], astrophysics [6], medical imaging [92].

Mathematically, wavelet transforms are characterized by: ψ which is the wavelet function, a which characterizes the frequency of the signal and b which registers the position(or "time")in the signal. All of these parameters simultaneously provide a time-frequency snapshot of the signal.

Let $s(x)$ be a finite energy, square integrable function i.e. $s(x)\,\epsilon\,\mathbf{L^2}(\mathbf{R})$. The wavelet transform of the signal is given as:

$$s(x) \mapsto W(b,a) = \int_{-\infty}^{\infty} \overline{\psi_{b,a}(x)} s(x) dx \tag{3-1}$$

where $\psi_{b,a}(x)$ is the wavelet analyzing function. For wavelet transforms, $\psi_{b,a}(x)$ is given as :

$$\psi_{b,a}(x) = \frac{1}{\sqrt{a}} \psi(\frac{x-b}{a}) \tag{3-2}$$

In Equation (3-2), a controls the effective support of ψ. If $a < 1$, the wavelet analysis function ψ gets contracted(or "squeezed") and if $a > 1$, ψ is dilated (or "stretched"). In both of these cases, the shape of the wavelet function remains unchanged and only the effective width of the wavelet function varies as a function of a. The parameter b controls the position of the wavelet function.

For wavelet analysis, we introduce a frame of reference known as the "mother wavelet" given as $\psi_{1,0}(x) = \psi(x)$. All other wavelet functions, which are derived from the mother wavelet, by

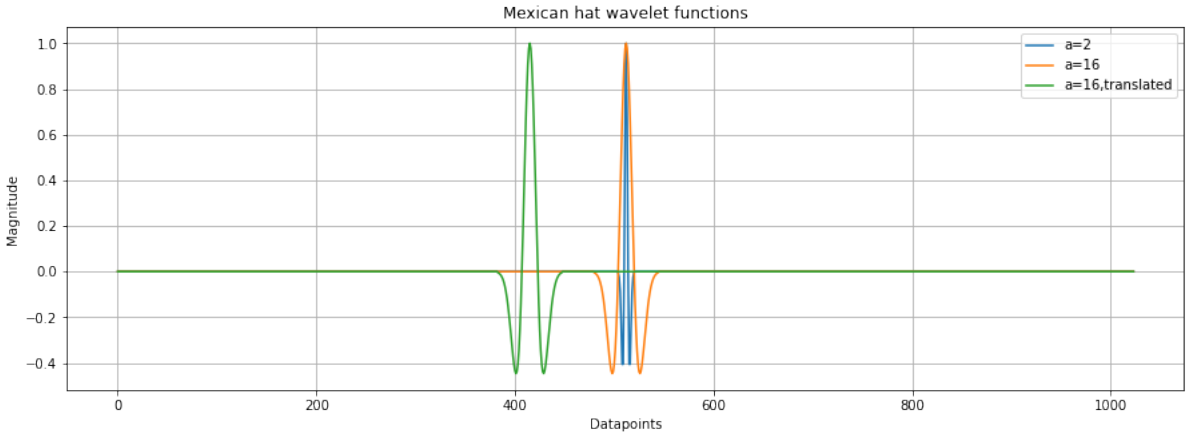


Figure 3-1: Mexican hat wavelet function for different parameters a and b. Blue: $a = 2$ and $b = 0$, Orange: $a = 16$ and $b = 0$, Green: $a = 16$ and $b = 100$

varying a and b are called "daughter wavelets". Figure (3-1) displays the mexican hat wavelet function for varying a and b.

3-1-1 Properties of wavelet functions

Existence conditions

In order to represent time-frequency space accurately, the wavelet function $\psi(x)$ must satisfy the following conditions [22] :

1. **Finite energy**: The energy of the wavelet function should be finite i.e.

$$\int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty \tag{3-3}$$

2. **Finite support**: The wavelet function should have finite support and should be well localized in frequency domain and time domain. This is given by the admissibility criteria. The admissibility criteria also guarantees invertibility of the wavelet transform. Mathematically, the criteria is expressed as:

$$c_\psi = 2\pi \int_{-\infty}^{\infty} \frac{|\psi(\hat{\xi})|^2}{|\xi|} d\xi < \infty \tag{3-4}$$

A slightly weaker condition, derived from Equation (3-4), can be given as :

$$\hat{\psi}(0) = 0 \Leftrightarrow \int_{-\infty}^{\infty} \psi(x) = 0 \tag{3-5}$$

Intuitively, Equation (3-3) and Equation (3-5) express that wavelets are oscillating functions, well localized in time and frequency domains. If we combine this information with Equation (3-1), we observe that wavelet transform is the convolution of the signal s with scaled, flipped and conjugated wavelet $\psi_a^*(x) = |a|^{-1/2}\overline{\psi(x/a)}$ i.e:

$$W(b,a) = (\psi_a^* * s)(b) = \int_{\infty}^{\infty} \psi_a^*(b-x)s(x)dx \tag{3-6}$$

This means that the wavelet transforms performs a local filtering operation with a zero mean function $\psi_{a,b}(x)$ i.e. the transform coefficient $|W(b,a)|$ is non-negligible when the wavelet function $\psi_{b,a}(x)$ matches any small part of the signal.

General Properties

Apart from the properties derived from the existence conditions, wavelet functions have other significant properties that are useful for singularity detection :

1. **Vanishing moments**: A wavelet function is said to have a N vanishing moments iff:

$$\int_{-\infty}^{\infty} x^n \psi(x)dx = 0,\ n = 0,1,2,3...N \tag{3-7}$$

   This property guarantees that the wavelet transform is insensitive to polynomials up to order N, which constitute the smooth part or the trend in the signal.

2. **Constant relative bandwidth**: If $\hat{\psi}$ has a bandwidth of $\Delta\xi$ in frequency domain, then $\hat{\psi_{b,a}}$ has a bandwidth of $\Delta\xi/|a|$. This implies that wavelet functions work like a filter with constant relative bandwidth i.e $\Delta\xi/\xi$ = constant. This property is helpful in localizing the position of singularities in the time-frequency domain.

3. **Invertibility** An important property of the wavelet transform is invertibility i.e if $s(x) \mapsto W(b,a)$, then $s(x)$ can be reconstructed from the wavelet coefficients $W(b,a)$. The reconstruction formula is given as:

$$s(x) = c_\psi \int_{-\infty}^{\infty} db \int_{-\infty}^{\infty} \frac{da}{a^2} \psi_{b,a}(x) W(b,a) \tag{3-8}$$

   where c_ψ denotes the admissibility constant in Equation (3-4). This implies that the signal $s(x)$ can be seen as a linear superposition of the wavelets $\psi_{b,a}$ with coefficients $W(b,a)$.

3-1-2 Type of wavelet transforms

The wavelet transform in equation(3-1) needs to be discretized for implementation. This is achieved by restricting the parameters a and b to a discrete set of points i.e. $\Gamma = (a_j, b_j,\ j,\ k\ \epsilon\ Z)$ in the (a,b) plane. Different discretization strategies lead to different types of wavelet transforms. The two major types of wavelet transforms are:

1. **Continuous Wavelet Transform (CWT)**: In this case, the parameter a is chosen independently based on some prior knowledge of the application and the parameter b_j varies from $j = 0,1,2...N$ where N is the length of the signal $s(x)$. This type of discretization has the advantage of making the whole analysis invariant to global translations and is very useful for feature detection. However, the discretization method will result in a redundant transform space and will require different strategies for exploitation of relevant information from the transform space.

2. **Discrete wavelet transform (DWT)**: In DWT, the parameters are chosen as:

   (a) scale parameter $a_j = a_o\lambda^{-j}$, $j \epsilon \mathbb{Z}$ for some $\lambda > 1$

   (b) translation parameter $b_k \equiv b_{k,j} = kb_oa_o\lambda^{-j} \quad j,k \epsilon \mathbb{Z}$

   Thus we get :

$$\psi_{j,k}(x) = \lambda^{j/2}\psi(a_o^{-1}\lambda^j x - kb_o), \quad j,k \epsilon \mathbb{Z} \tag{3-9}$$

   The most common choice is $\lambda = 2$ (octaves) and $a_o = b_o = 1$ which gives :

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad j,k \epsilon \mathbb{Z} \tag{3-10}$$

   DWT generates a dyadic lattice $((k2^{-j}, 2^{-j}),\ j,k \epsilon \mathbb{Z})$ where the scale parameter and translation parameter are dependent on choice of λ. The advantage of choosing the discretization strategy is that it yields fast reconstruction algorithms and generates a sparse representation of transform space. DWT is primarily used for signal denoising [29] and compression applications [19].

3-1-3 Visualization of CWT

Given a one-dimensional signal, $s(t)$ with length N, the CWT of the signal will be an M x N matrix where M represents the number of scales, and N is the length of the signal. This means that wavelet transform adds another dimension to the original signal governed by the scale parameter a.

In two dimensional signals, the number of dimensions present in the wavelet transform is also dependant on the nature of the wavelet function being used. In case of isotropic wavelets, CWT consists of three dimensions (two spatial dimensions and the scale dimension). Isotropic wavelets are useful in singularity detection. For anisotropic wavelets, the number of dimensions increase to four. The additional dimension is contributed by the rotation parameter θ which characterizes the skew of the wavelet function. Anisotropic wavelet functions are useful in identifying orientations or directional elements in the 2D signal.

In practice, CWT is visualized from an energy perspective i.e we plot $|W(a,b)|^2$ as a function of a and b. This is known as a scalogram (analogous to spectrogram used in short-time fourier transform). However, depending on the application, $|W(a,b)|$, $|W(a,b)|/c_\psi$ [4] are some of the other quantities that can be plotted as a function of a and b. For our purpose, we will use the quantity $|W(a,b)|$ for the scalogram.

As an example, we plot the scalogram for a gaussian signal $g(t)$ in Figure (3-2).

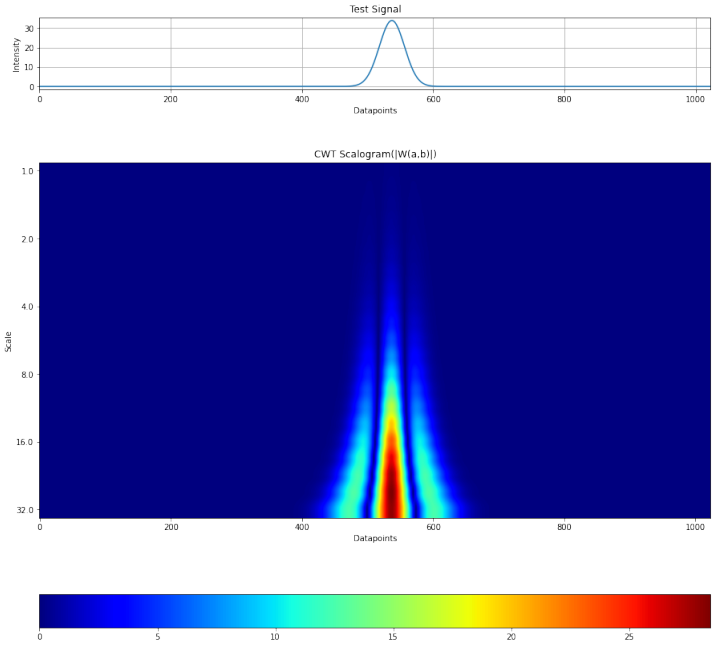


Figure 3-2: Top: Gaussian signal $g(t)$. Bottom: Scalogram of the gaussian signal

3-2 Wavelet transform maxima (WTM)

For feature (peak) detection, the discretization used in CWT is very useful. CWT preserves the dimension of the signal which helps in tracking the position and evolution (amplitude of wavelet coefficient) of the feature in the transform space. However, CWT is redundant and the results are often hard to interpret.

In order to work around the redundancy, we limit the transform space to a set of points that can characterize the signal effectively. These set of points are chosen based on the intended application. For our application, we are interested in characterizing the peaks present in the signal. Peaks can be defined as local maxima points present in the signal. Therefore, we limit the analysis to a set of local maxima points that are present in the wavelet transform space. We define this set as:

$$\Gamma_{lm} = \Big\{(b_j, a_k) \quad |W(b_{j-1}, a_k) < W(b_j, a_k) \;\&\; W(b_j, a_k) > W(b_{j+1}, a_k) \;\forall a_k \quad j, k \,\epsilon\, \mathbb{Z}\Big\} \tag{3-11}$$

Intuitively, Equation (3-11) means that for every scale a, we look for local maxima coefficients in the position parameter b. The idea is demonstrated for the previously used gaussian signal $g(t)$ in Figure (3-3).

From Figure (3-3), we see that the set of maxima points can be grouped together to form a connected structure in the transform space. These are called as chains (ridges). Roughly speaking, a chain is a collection of maxima points that are connected (or correlated via some metric) to each other in the transform space. The main idea is that these chains can be used to completely characterize the nature of the local maxima point present in the signal.

The technique was developed by Mallat et al. [61] under the name of Wavelet transform modulus maxima (WTMM) for the application of signal denoising. Arneodo et al. [8] extended

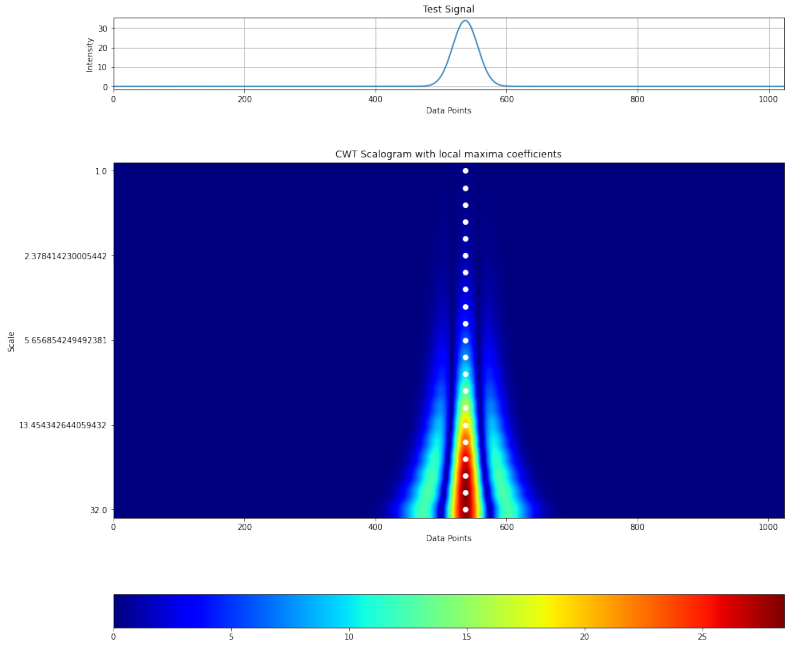


Figure 3-3: Top: Gaussian signal $g(t)$. Bottom: CWT Scalogram with maxima points. The white dots represent the location of local maxima wavelet coefficients at every scale a. These local maxima points can be grouped together to form a connected structure called chain.

the technique for the analysis of fractals. Carmona et al. [17] proposed the technique for the application of signal reconstruction.

3-2-1 Properties of wavelet transform maxima

1. **Characterization of Lipschitz exponents** [7]: Given a signal $s(t)$ which consists of a singularity $\gamma_\alpha(x - x_o)$ of order α i.e :

$$\gamma_\alpha(x - x_o) = \begin{cases} 0, & x \leq x_o \\ (x - x_o)^\alpha & x > x_o \end{cases} \tag{3-12}$$

Differentiating with respect to $x^{\alpha+1}$ we get:

$$\frac{\mathrm{d}^{\alpha+1}\gamma_\alpha}{\mathrm{d}x^{\alpha+1}}(x - x_o) = \Gamma(\alpha + 1)\delta(x - x_o) \tag{3-13}$$

where $\delta(x)$ is known as the dirac function. Now, let the wavelet be an nth derivative function of a smooth function ϕ, i.e. $\psi(x) = \frac{d^n}{dx^n}\phi(x)$ with $n \geq \alpha + 1$. Then, the wavelet transform of $\gamma_\alpha(x - x_o)$ with respect to $\psi(x)$ is given as :

$$W_{\gamma_\alpha}(b, a) = \Gamma(\alpha + 1)a^\alpha \frac{d^{n-\alpha-1}\phi}{dx^{n-\alpha-1}}(\overline{\frac{x_o - b}{a}}) \tag{3-14}$$

Assume that the modulus of $(n - \alpha - 1)$th derivative of ϕ has N maxima points at positions at $(x_l = 1, ..., N)$. Then, for each a, $|W_{\gamma_\alpha}|$ has N maxima points at locations

$(b_l = ax_l + x_o,\ l = 1, ..., N)$ which converges to x_o as a tends to 0. Therefore, maxima of $|W_{\gamma_\alpha}|$ will have N chains converging to x_o as a tends to 0.

For a particular $b_l = ax_l + x_o$, we have :

$$|W_{\gamma_\alpha}(b_l = (ax_l + x_o), a)| = \Gamma(\alpha + 1)a^\alpha \phi_l \tag{3-15}$$

Taking log(base e) on both sides, we get :

$$|W_{\gamma_\alpha}(b_l = (ax_l + x_o), a)| \sim \alpha ln(a) + ln\phi_l \tag{3-16}$$

where $\phi_l = \phi_l = \frac{d^{n-\alpha-1}}{dx^{n-\alpha-1}}(\phi(\frac{x_o - b_l}{a}))$.

Intuitively, this means that by looking at the slope of the log of the wavelet coefficients along a chain, we can determine the singularity α governing the signal.

2. **Behaviour of noise [61]**: Let $n(x)$ be a stationary white noise signal with variance σ^2. Let $\psi(x) = \frac{1}{a}\psi(\frac{x}{a})$. Then, it can be shown that:

$$E(|Wn(a,b)|^2) = \frac{\sigma^2 ||\psi||^2}{a} \tag{3-17}$$

The above relation implies that the expected energy of noise in the wavelet transform space is inversely proportional to the scale parameter a. Roughly, this means that chains that are generated due to noise (i) have short length (ii) have decreasing strength as scale a increases.

This idea can be extended to the case of additive gaussian noise. Let $y(x) = y_o(x) + n(x)$ where $y_o(x)$ is the true signal and $n(x)$ is zero mean white noise with variance σ^2. Then:

$$E(|Wy(a,b)|^2) = |Wy_o(a,b)|^2 + E(|Wn(a,b)|^2) \tag{3-18}$$

Using Equation (3-17), we get:

$$E(|Wy(a,b)|^2) = |Wy_o(a,b)|^2 + \frac{\sigma^2 ||\psi||^2}{a} \tag{3-19}$$

Equation (3-19) means that the transform coefficients of the observed signal (y_x) will be strongly impacted by noise at lower scales and the impact will be reducted at higher scales.

3. **Theorem [60]**: Let $\psi_x = (-1)^n \theta^n$ be a wavelet function, where θ is a gaussian function. For any $f \epsilon \mathbf{L^2}(\mathbb{R})$, the modulus maxima of $|Wf(b,a)|$ belong to a set of connected curves that are never interrupted as scale decreases.

This means that chains can be traced back to the finest scale for wavelets that are derived from derivatives of gaussian function.

3-2-2 Synthetic Example

In this section, we demonstrate the wavelet transform maxima technique using a test signal and study the properties discussed above.

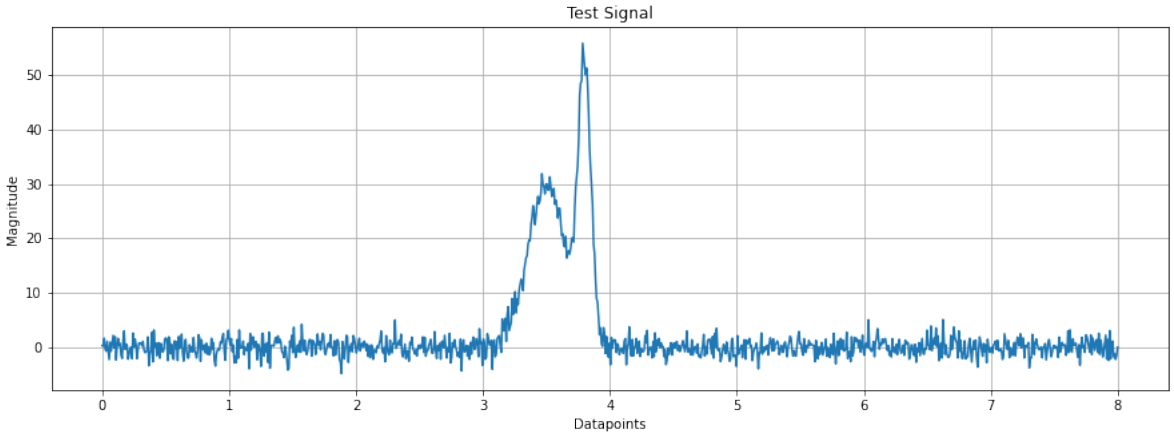


Figure 3-4: Test Signal: $s(t) = 30e^{-0.5(\frac{(3.5-t)^2}{(0.15)^2})} + 50e^{-0.5(\frac{(3.8-t)^2}{(0.05)^2})} + n(t)$

Test Signal

The modelling of the clean test signal was based on the peaks observed in the mobility domain. Mathematically, the clean test signal is given as :

$$s(t) = 30e^{-0.5(\frac{(3.5-t)^2}{(0.15)^2})} + 50e^{-0.5(\frac{(3.8-t)^2}{(0.05)^2})} \tag{3-20}$$

where $t = nT/N$, $N = 1024$, $T = 8$ and $n = 1, 2, ...1024$. On the noiseless signal, we added stationary zero mean gaussian noise $n(t)$ with standard deviation $\sigma_{noise} = 1.5$. Therefore, the final test signal consisted of two gaussian peaks with gaussian noise (Figure (3-4)). The final model can be represented as:

$$y(t) = s(t) + n(t) \tag{3-21}$$

Wavelet transform

The next step was to compute the CWT of the test signal. For this, we chose mexican hat function as our wavelet function. We used dyadic scales(octaves) with three voices per octave as the scale parameter a. Mathematically, the scale parameter can be given as :

$$a = 2^{nj/K}, \ K = 4, \ n = 1, 2, 3, \ j = 0, 1...5 \tag{3-22}$$

In equation(3-22), j controls the number of octaves and $K - 1$ controls the number of voices per octave. The minimum and maximum scale values were chosen as $a_{min} = 1$ and $a_{max} = 32$ respectively. This choice of scales is motivated by the conservative bounds governed by the size of the filter. The size of the dilated wavelet filter should not increase the size of signal.

In CWT, the normalization factor of the wavelet function is an important factor that governs the behaviour of the transform space. In our case, we use L1 normalized wavelet functions. We say that the wavelet is L1 normalized when:

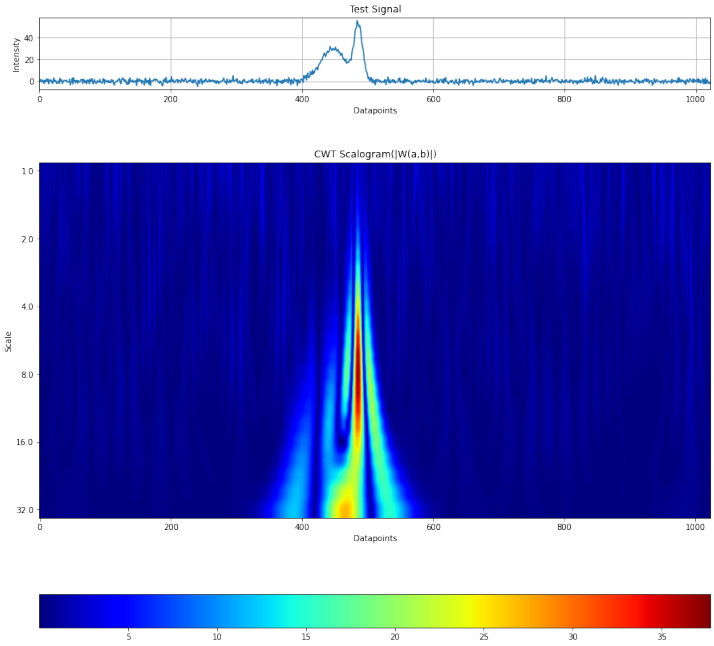


Figure 3-5: Scalogram for the test signal

$$\int_{-\infty}^{\infty} |\psi(x)|dx = \int_{-\infty}^{\infty} |\psi_{b,a}(x)|dx \quad \forall a, b \tag{3-23}$$

For L1 normalization, the wavelet function is given as:

$$\psi_{b,a}(x) = \frac{1}{a}\psi(\frac{x-b}{a}) \tag{3-24}$$

Mathematically, using L1 normalization has the advantage of penalizing the transform space at higher scales. This can be used to reduce the impact of noise at higher scales (Equation (3-17) and Equation (3-19)).

After deciding the scale parameter, the normalization factor and the number of scales, the wavelet transform was computed using Equation (3-1). We present the scalogram for the test signal in Figure (3-5).

Wavelet transform maxima

After computing the wavelet transform, we identified the local maxima coefficients at every scale a. Note: WTM is slightly different from WTMM. The former computes the local maxima over wavelet coefficients for every scale while the latter computes the local maxima over absolute wavelet coefficients for every scale. The plot for local maxima coefficients is given in Figure (3-6).

From the figure, it can be inferred that: (a) the density of local maxima points decreases as scale increases (b) local maxima points corresponding to gaussian peaks have higher amplitude than than local maxima points corresponding to noise.

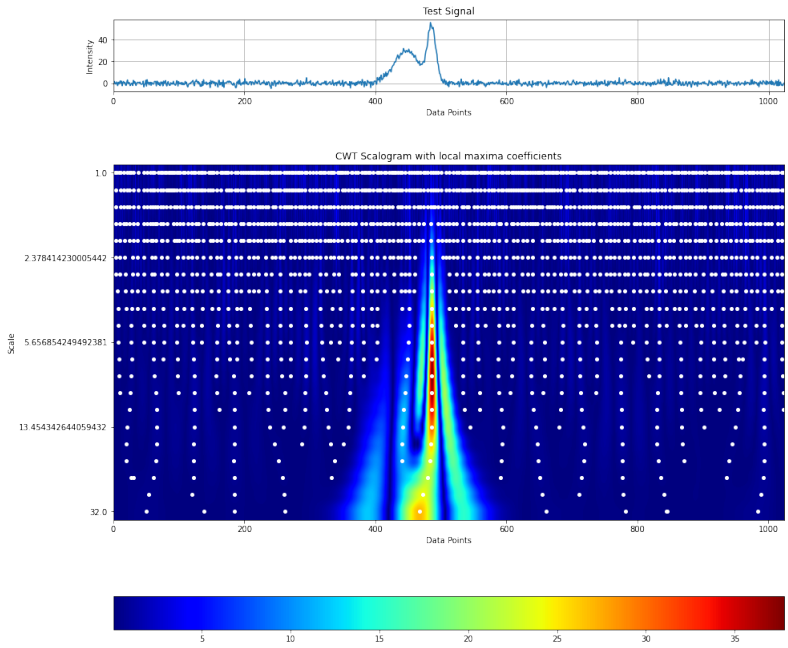


Figure 3-6: Scalogram for the test signal along with the local maxima coefficients. The white dots represent the location of local maxima coefficient at every scale a

Analysis of chains

After local maxima points across scales were identified, we grouped these points across scales to form chains. For demonstration, the chain construction was based on a naive distance based algorithm i.e maxima points that are close together in scale space were grouped together to form a chain. We present a chain generated by a gaussian peak (Figure (3-7)) as well as a chain generated by the noise in the signal (Figure (3-8)).

The analysis lead to the following findings:

1. The majority of noise-induced chains were short and only dominated smaller scales.
2. Even if noise-induced chains were long, on average, the strength of wavelet coefficients associated with noise-induced chains decreased as scale increased.
3. Chain associated with low intensity peaks behaved like noise at lower scales but gradually gained strength at higher scales.
4. Chain associated with gaussian peaks tend to have their maximum wavelet coefficient occuring at the scale which best represented the feature.
5. Along a chain, the finest scale represented the location of the peak in the test signal.

Overall, the initial analysis lead to the conclusion that wavelet transform maxima technique can be used for peak detection. However, there were a number of important quantative and qualitative criteria that were still needed to be defined in order to make the technique robust and sensitive to peaks present in the IM-IMS datasample.

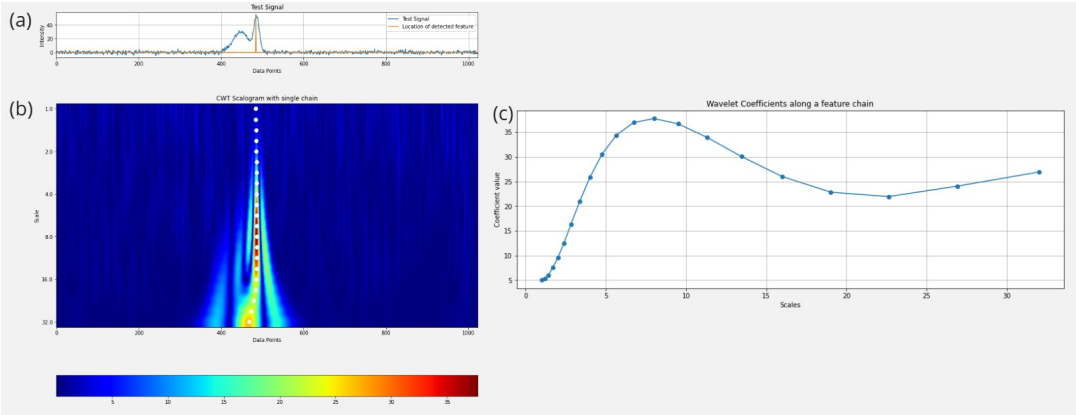


Figure 3-7: (a) Test Signal. The blue line represents the test signal and the orange line orange line corresponds to the finest scale location($a = 1$) for the given chain. (b) Scalogram of the test signal. White dots correspond to maxima points that form a chain belonging to a gaussian peak. (c) Wavelet coefficients along the chain. Maximum wavelet coefficient = 37.767 occuring at scale a=8

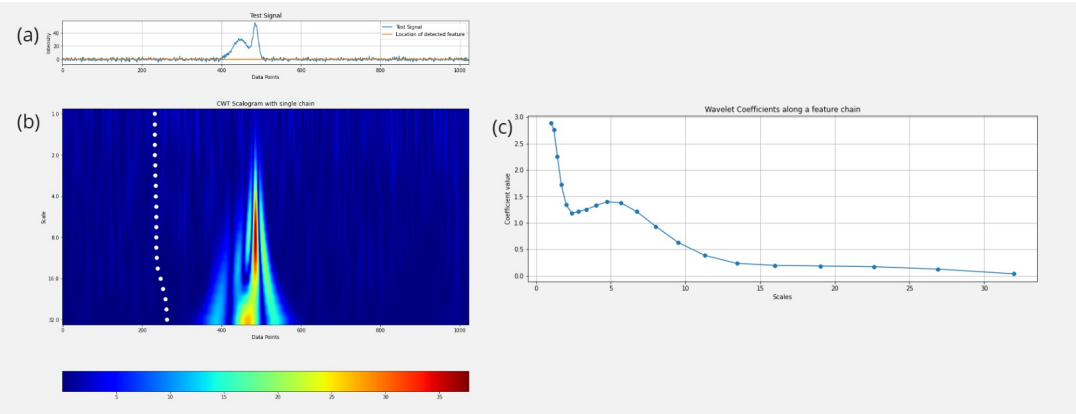


Figure 3-8: (a) Test Signal. The blue line represents the test signal and the orange line orange line corresponds to the finest scale location($a = 1$) belonging to the chain. (b) Scalogram of the test signal. White dots correspond to maxima points that form a chain belonging to noise. (c) Wavelet coefficients along the chain. Maximum wavelet coefficient = 2.88 occuring at scale a=1

Having performed some initial analysis, we now present existing literature pertaining to WTM based peak detection.

3-3 Existing literature related to peak detection using WTM

Wavelet transform maxima has previously been used by Du et al.[30] for peak detection in MS based datasample.

This short review explains the existing algorithm and emphasizes the parameters in use.

3-3-1 Wavelet parameters

1. **Choice of wavelet**: The first step would be to decide the wavelet function $\psi(x)$. In order to obtain good results, the wavelet function should match the characteristic nature of the peaks that are present in the data sample. Therefore, the authors chose mexican hat function (Figure (3-1)) to be the wavelet function for peak detection.

2. **Choice of scales** a: The next step is to decide the range for the scale parameter a that will characterize the transform space. As wavelet functions $\psi_{a,b}(x)$ can be seen from a matched filtering point of view, the range is governed by prior knowledge of the peak widths present in the signal and should be given as inputs by the user. Conservative bounds for a_{min} and a_{max} are 1 and N respectively where N is the length of the signal.

   The number of scales between a_{min} and a_{max} is also assumed as a prior input given by the user. However, Carmona et al. [17] define that chains should be slow varying and smooth in nature (with respect to their amplitude $|W(b_j, a_k)|$) for better characterization of the signal. So, we assume that the number of scales between a_{min} and a_{max} should be high (in the range of 20-50). Choosing high number of scales also compensates for the lack of knowledge regarding peak widths present in the signal.

3. **Translation parameter** b: As the algorithm uses CWT, the translation parameter is same as the length of the signal.

3-3-2 Algorithm

The algorithm is divided in two parts: (i) Construction of chains from the CWT space (ii) Peak detection criteria for discriminating chains triggered due to noise from chains triggered due to actual peaks.

3-3-3 Chain construction

Suppose the 2D CWT coefficient matrix is M x N where M is the number of scales and N is the length of the data sample.

1. Identify the local wavelet maxima coefficients present at the coarsest scale a_{max}. These are the initialization points for the chains that start from the scale a_{max}. Set the initial gap number corresponding to the chain as 0.

2. For every local wavelet maxima coefficient present at a_{max}, go to scale a_{max-1}. Search for the nearest local maxima point (nearest with respect to the translation parameter b) at the scale a_{max-1}. The nearest local maxima point should be within a search radius. The search radius is taken to be proportional to the scale and the wavelet being used. Gregoire et al. [39] uses a search radius as $1.5a_j$ where a_j is the scale at which we are searching for the local maxima coefficient(in this case, it would be a_{max-1}). Remove this local maxima point. If no local maxima is located within the search radius, increase the gap number by 1 for that particular chain. Repeat this procedure till a_{min}.

3. After all the chains that were initialized from the scale a_{max} are constructed, remove the chains which have gap numbers larger than the gap threshold.

4. Now, repeat Step 1 - Step 3 for all the local maxima points present at scale a_{max-1} that were not linked in the previous step. These are the initialization points for the chains that start from the scale a_{max-1}. Set the initial gap number corresponding to the chain as 0.

5. Repeat Step 1 - Step 4 until the initialization reaches the row corresponding to smallest scale a_{min}.

Peak Detection Criteria

After all the chains are constructed, four rules are used for separating chains originating from mass peaks from chains originating from noise.

1. **Maximum wavelet chain coefficient**: The scale at which the chain has maximum wavelet coefficient should be within a range. The idea is to use this maximum wavelet coefficient in order to determine the peak width associated with the feature.

2. **Length of chains**: The length of the chain should be higher than a threshold provided by the user.

3. **SNR**: The Signal-to-Noise Ratio (SNR) value should be higher than a threshold value provided by the user. The author defines SNR as :

   $$\text{SNR} = (\text{Signal strength/local noise strength})$$

   Signal strength corresponds to maximum wavelet coefficient in a chain (within a scale range). The local noise strength of a peak is defined as the 95-percentage quantile of the absolute CWT coefficient values at scale $a = 1$ within a local window.

4. **Shoulder peak criteria**: Shoulder peaks are small peaks that surround major peaks. Chemically, these peaks are associated with the matrix molecules. In the wavelet space, these peaks tend to form short chains with relatively high wavelet coefficients. Thus, by reducing the threshold value for rule 2, these peaks can be detected. The algorithm provides an option for this by selecting a window around the major peaks and then reducing threshold value associated with rule 2.

3-3-4 Parameters

Having reviewed the algorithm, we present a list of all the parameters being used as well as their function.

1. a (range): Scales
2. Gap threshold (int): Gaps that can be tolerated in a chain.
3. Window size (int): The window that is used for searching for the local maxima while constructing chains.
4. Scale range (int): The upper threshold on which the algorithm needs to calculate the maxima wavelet coefficient value present in a chain
5. Length Threshold (int): Minimum length required to be considered as a chain belonging to the feature.
6. SNR Threshold (float): Minimum SNR value to be considered as a peak.
7. Local window (int): Size of the local window used for calculation of SNR.
8. Shoulder peak window (int): Size of the window surrounding the major peak for detection of shoulder peaks.

Analysis of the parameters

Based on the functioning of the algorithm, we present our analysis and understanding of the parameters.

1. Window size : We found this parameter is difficult to specify as a prior input given by the user. This is because the MS spectrum is noisy in nature. Having too narrow of a window will lead to maxima points not forming chains and having too wide of a window will lead to formation of wavy (incorrect) chains.
2. Gap threshold : We found the gap parameter is not justified from a theoretical point of view. Suppose, a maxima point is present at scale a_j, $j \epsilon \mathbb{Z}$. If that maxima point is not getting connected to any other maxima point at scale a_{j-1}, then it means that maxima point is characteristic to that scale a_j and it should be terminated at that scale. The use of gap in chains will lead to construction of incorrect chains. Also, the gap threshold is dependent on the window size being used. Having a large window will lead to chains with zero gaps but are incorrectly constructed. Having a small window size will lead to chains with many gaps.
3. SNR : The algorithm assumes that the scale where the wavelet coefficient is maximum in chain best represents the feature. However, this is only true for isolated features. In MS, depending on the sample being studied, the peaks can be very close to each other. This leads to bias (peaks begin to merge in transform space) in the maximum wavelet coefficient present in a chain. As a result, the wavelet coefficient gets maximized at

coarse scales which do not represent the width of the peak. In order to mitigate this effect, the author introduces another parameter called scale range which caps the scales being used for identification of the maxima coefficient in a chain. However, users with no prior information about wavelet transforms or MS will find this parameter difficult to interpret.

4. Local Window : The window size used for calculating the noise strength requires prior knowledge about the instrument and the sample.
5. Noise Strength : In the calculation of SNR, the author estimates noise strength as 95% quantile of absolute CWT coefficients at scale $a = 1$ in a local window. However, we found that CWT coefficients tend to bias this value as they are correlated to each other in the transform space. Also, CWT coefficients at scale $a = 1$ does not capture the high frequency components which usually represents the noise in the signal. In order to capture the high frequency component, the scale parameter a should be less than 1.
6. Shoulder peak window : This window size requires prior knowledge about the data sample.

3-4 Existing literature regarding to thresholding of wavelet chains

Real world signals are noisy in nature and this noise gets translated to wavelet transform space. As a result, after construction of wavelet chains, we see that there are some chains that are triggered due to noise and some chains that are triggered due to information (in our case, chemical peaks). Therefore, it becomes important to remove the chains that are triggered due to noise for feature detection.

Different authors have proposed different techniques related to thresholding of wavelet chains. We provide a brief overview of these techniques below.

3-4-1 Slope-Amplitude Histogram

Antoine et al. [7] proposed a classification method based on computing the slope and amplitude of the wavelet chain (in their work, they refer to chains as 'ridges') and used it to discriminate bright spots present on the surface of sun from cosmics in space in Extreme-ultraviolet Imaging Telescope (EIT) images. They used isotropic wavelets for the wavelet transform, since directions were irrelevant in this context. The method is as follows:

1. Define a ridge $\mathbf{R}$ as a 3D curve $(\vec{r(a)}, a)$ such that for every scale $a \, \epsilon \, \mathbb{R}^+$, $|W[f](\vec{r(a)}, a)|^2$ is a local maxima coefficient in transform space and $\vec{r}$ is a smooth and continuous function of a.Here, f is the signal and $W[f](\vec{r(a)}, a)$ refers to the wavelet transform of the signal at scale a. The condition $\vec{r}$ is continuous and smooth function means that the 3D curve (ridge) is smooth and does not show heavy variations as a function of a.
2. Compute the amplitude of the ridge using the following equation.

$$\boldsymbol{A_R} = \lim_{a \to 0} |W[f](\vec{r}(a), a)|^2 \tag{3-25}$$

Essentially, this refers to the squared wavelet coefficient of the ridge as scale $a \to 0$.

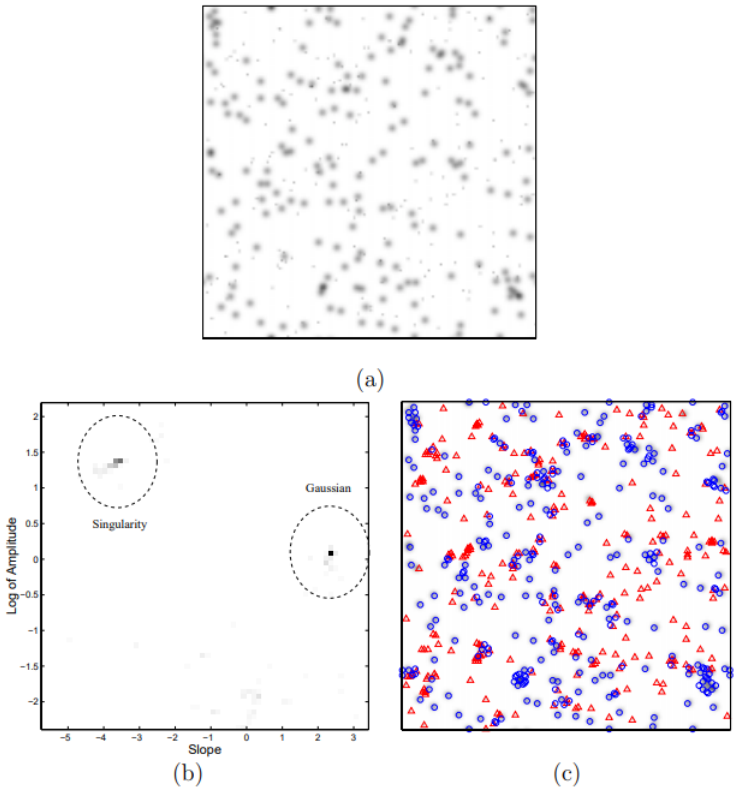


Figure 3-9: (a) Academic Signal (a mixture of singularities and gaussian functions). (b) Slope-Amplitude Histogram (Logarithm of amplitude is plotted to reduce the range). (c) Discrimination of features based on slope-amplitude histogram. Triangles point to singularities and circles point to gaussian functions.[7]

3. Compute the slope of the ridge using the following equation.

$$\boldsymbol{S_R} = \lim_{a \to 0} \frac{d\ ln(|W[f](\vec{r}(a), a)|^2)}{d\ ln(a)} \tag{3-26}$$

4. Compute these quantities for all the ridges present in the transform space.
5. Compute a 2D histogram of these two quantites to show the distribution of ridges as a function of their slope and amplitude.
6. Use the 2D histogram to discriminate between features.

Figure (3-9) presents an academic example where the 2D histogram was used to discriminate between simulations of impulse functions and gaussian functions.

3-4-2 Norm and length based thresholding

Donoho et al. [13] in their software package *WAVELAB* implement a norm and length based thresholding of wavelet chains. The implementation is as follows:

1. Define a threshold value for every scale, $T_a = D_{range}||W[f](\cdot, a)||_p$. Here,$||W[f](\cdot, a)||_p$ is the p-norm (default is infinity norm) of the wavelet transform of the signal $f(x)$ at scale a and D_{range} is referred as the "dynamic range" which is given as an input by the user (parameter value should lie between 0 and 1). If any wavelet local maxima coefficient (belonging to a chain) at scale a, is less than this threshold value T_a, then the complete wavelet chain is removed.

2. Count the number of scales for which a chain exist. This will be referred to as the length of the wavelet chain. If the wavelet chain does not persist an octave (in terms of length), then the chain is considered weak and removed.

3-4-3 Bootstrap based thresholding

Carmona et al. [17] proposed a bootstrap based thresholding of wavelet local maxima coefficients in the wavelet transform space in their software package *SWAVE+*. The procedure is based on learning the nature of noise in the signal and then developing a thresholding value T_a for every scale a in the transform space. Wavelet local maxima coefficients that are below this threshold value are then removed. The complete procedure is given below:

1. Given a signal $f(x)$, compute the mean $m(x)$ of the signal using a moving average window. The default window size used in the algorithm is 8.

2. Compute $n(x) = f(x) - m(x)$. This is a representation of noise in the signal.

3. For B = 1:128:

   (a) Perform sampling with replacement of the signal $n(x)$. Represent this signal as $p(x)$

   (b) Compute the wavelet transform of $p(x)$ for every scale a.

   (c) For every scale a, compute the 95th percentile wavelet coefficient in the transform space. Store this value.

4. Average the stored values for every scale a (as every scale will have 128 values). This will give a threshold value T_a for every scale.

5. Remove wavelet local maxima coefficients, at scale a, if it is less than this threshold value T_a.

It is to be noted that bootstrapping assumes that no parametrization of the noise in the signal is available. If we assume a distribution for the noise in the signal, say $\sim N(0, \sigma)$, then $n(x)$ can be used to compute σ. Step 3(a) is then replaced by simulation of the noise $\sim N(0, \sigma)$. The rest of the steps remain the same.

3-5 Research Objectives

After studying WTM and its application in MS spectrum, we present our research objectives:

1. **Design of 2D wavelet function**: In the previous section, we saw that the wavelet function was tailored to the peaks present in the MS spectrum. For 2D feature detection, our objective would be to design(or use) a wavelet function adapted to the peaks in the IM-IMS data samples.
2. **Minimize the hyperparameters used by the algorithm**: Current algorithm uses many window parameters for chain construction, calculation of local noise and detection of shoulder peaks. All of these parameters require knowledge about the instrumentation, data sample being studied and wavelet transform. The objective would be to completely automate the process of chain construction.
3. **Minimize the criteria for peak detection**: In the previous paper [30], the author introduces three criteria for peak detection namely, maximum wavelet chain coefficient, length of chains and SNR for detection of MS peaks. Our objective would be to minimize the criteria used for peak detection.

Chapter 4

2D Feature Detection Algorithm for Ion Mobility Imaging Mass Spectrometry

The proposed 2D feature algorithm can be divided into the following sections (Figure (4-1)):

1. **Partitioning the data sample into overlapping sections**
2. **2D CWT of each section of the data sample**
3. **Denoising of CWT coefficients**
4. **Automatic local maxima clustering and chain construction**
5. **Peak Detection Criteria: Effective length thresholding**

In order to demonstrate the working mechanics of the algorithm, we will use a small section of a real world data sample. Details regarding the data sample and the section are given in Figure (4-2). In IM-IMS data, the selected test section is considered to be on the lower side of SNR. However, this region is good for demonstration of the algorithm. Performance and evaluation for other sections will be demonstrated in Chapter 5.

4-1 Step 1 : Partitioning of the data sample

IM-IMS instruments tend to create large data files. Mathematically, the spectral information generated from the IM-IMS instrument can be represented in a 2D M x N matrix where M characterizes the number of mobility bins and N represents the number of m/z bins. Usually, the number of m/z bins is in the order of $\sim$200k to 400k and the number of mobility bins is in the order of $\sim$6k.

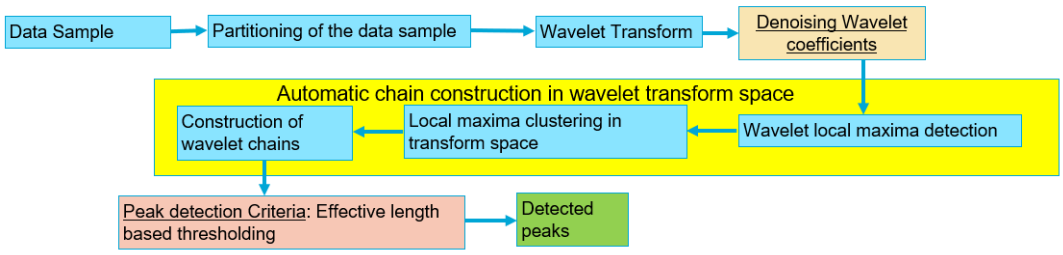


Figure 4-1: Layout of the feature detection algorithm

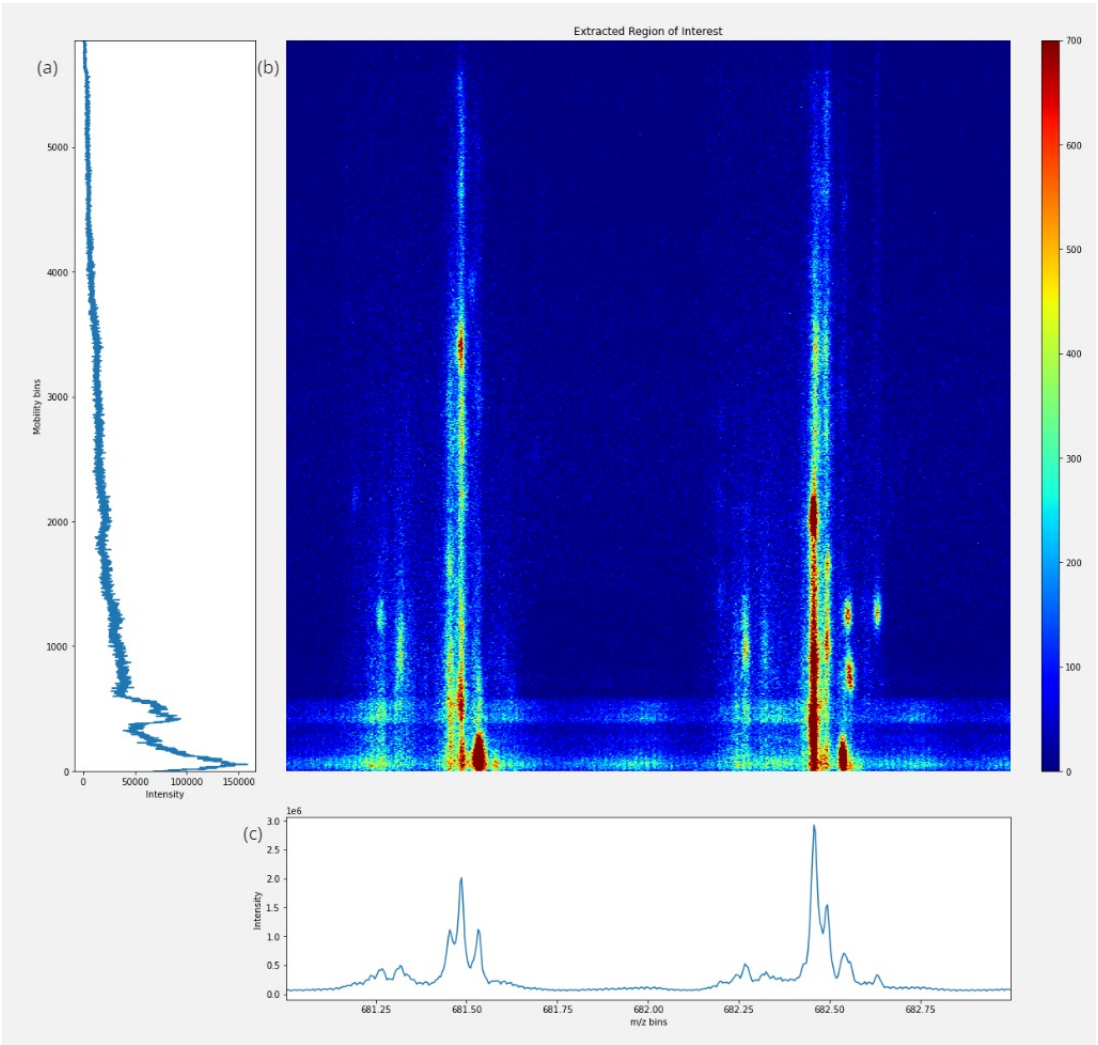


Figure 4-2: Test section extracted from real world IM-IMS data sample. (a) Complete mobility information. The intensity values in the plot is obtained by summing the 2D test section along columns. (b) 2D Test Section (c) m/z information. The intensity values in the plot is obtained by summing the 2D test section along rows. Here, m/z values are in the range of 681-683 m/z.

Based on this information, we partition the 2D matrix along the m/z bins and retain the complete mobility information in each partition. Each partition contains some overlapping portions in order to minimize the boundary effects. We run the peak detection algorithm on each section separately and the common peaks detected in overlapping sections are removed.

4-2 2D Continuous Wavelet Transform

,

4-2-1 Design of the wavelet function

In the previous chapter, we discussed that the wavelet function $\psi(x)$ should resemble the peaks present in the data sample for better characterization in terms of wavelet coefficients. A visual analysis (Figure (4-2)) of the data sample reveals that the peaks are anisotropic in nature. The anisotropic nature of the peak is because the information in the 2D matrix is generated from two different sub instruments. Information on the m/z axis is governed by the mass spectrometer which usually characterizes a peak in the range of 3-20 mass bins (rows) whereas mobility information is governed by the resolution set by the user for the ion mobility spectrometer and usually characterizes peaks in the range of 200-500 mobility bins (columns). Therefore, the 2D peaks tend to appear stretched along the mobility axis.

Based on this information, we use a generalized 2D mexican hat function as our wavelet function. The function allows us to control the width of the wavelet function along both dimensions separately. Mathematically, generalized 2D mexican hat wavelet function can be expressed as:

$$\psi(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}(2 - (\frac{x^2}{\sigma_x^2}) - (\frac{y^2}{\sigma_y^2}))e^{-(\frac{x^2}{2\sigma_x^2})-(\frac{y^2}{2\sigma_y^2})} \tag{4-1}$$

In equation (4-1), σ_x and σ_y control the width of the wavelet function along rows and columns respectively. Details regarding the numerical implementation of the wavelet function can be found in the works of Freeman et al. [36].

4-2-2 Normalization factor

As in the previous case, we use L1 Normalization for the wavelet function. For the given wavelet function $\psi_{x,y}$, the L1 normalization factor was found to be $1/(2\pi\sigma_x\sigma_y)$ and $||\psi_{x,y}||_1^1$ was approximately equal to 1.4 for different values of σ_x and σ_y. This was numerically verified by the following procedure:

1. Numerically implement equation (4-1) for different values of σ_x and σ_y
2. For a given wavelet function $\psi_{x,y}$, compute $\sum_x \sum_y |\psi(x,y)|$
3. Record this value for various wavelets functions having different σ_x and σ_y.

The plot for $\sum_x \sum_y |\psi(x,y)|$ for various $\psi_{x,y}$ with varying σ_x and σ_y is presented in Figure (4-3).

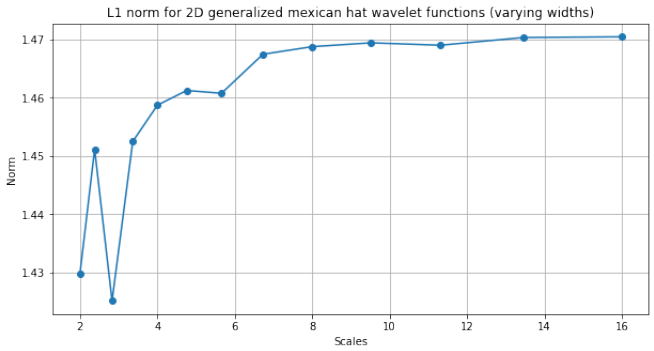


Figure 4-3: L1 Norm values for varying widths. The normalization factor is given as: $1/(2\pi\sigma_x\sigma_y)$. The widths (equivalently scales) of the wavelet functions are chosen according to Table (4-1).

4-2-3 Parameters

In equation (4-1), we introduced two parameters namely σ_x and σ_y. Both of these parameters function as the scale parameter a for different dimensions, where changing σ_x or σ_y will change the effective support of the wavelet function without affecting it's shape. In this section, we define the parameter and explain the method of choosing σ_x and σ_y for our peak detection algorithm .

1. σ_x (float64): The parameter governs the width of the wavelet function along rows. In our case, rows correspond to the mobility axis. The choice of σ_x is based on the minimum expected peak width along the mobility dimension. A working estimate of this quantity would be 1% of the total mobility bins present in the data sample.

2. σ_y (Array[Float64]): The parameter governs the width of the wavelet function along columns. In this case, we specify a range of values for σ_y based on the expected peak widths along the m/z axis. Conservative lower bound and upper bound for σ_y would be 1 and $N/10$ respectively, where N is the total number of mass bins present in each section.

Explanation of parameters

It is important to understand the reason behind the datatype used for σ_x and σ_y and it's relation with respect to 1D case scenario demonstrated in the previous chapter.

In the 1D case scenario, wavelets can also be viewed as a set of filters of increasing widths where a determines the width of the wavelet function. If we assume that the user has no information about the expected peak widths in the data sample, the user can compensate for the lack of information by selecting maximum number of widths within the conservative bounds. However, the range, which the user supplies, should be monotonically increasing for correct characterization of scalogram and chains.

Now, if we consider the 2D case, we have two parameters σ_x and σ_y which govern the width of the wavelet function and are independent of each other. If we assume that the user has no information about the expected peak widths in the data sample, then the user will need

to supply a range for σ_x and σ_y independently. This will lead to a total number of $\#\{\sigma_x\}$ x $\#\{\sigma_y\}$ wavelet functions and will make the algorithm computationally expensive. Also, the monotonic nature of peak widths will be violated. For example, we take the wavelet function $\psi(x, y)$ generated from $(\sigma_x, \sigma_y) = (2, 8)$ and $\psi'(x, y)$ generated from $(\sigma'_x, \sigma'_y) = (4, 2)$. These two wavelets are not monotonically related to each other as the width is increasing along rows but decreasing along columns. This violation of monotonic nature of widths will lead to incorrect construction of chains.

One way to work around this case, would be to map a single parameter s to (σ_x, σ_y) . For example:

$$\forall s \, \epsilon \, \mathbb{R}^+, s \mapsto (\sigma_x, \sigma_y) := \begin{cases} \sigma_x & = 4s \\ \sigma_y & = s \end{cases} \tag{4-2}$$

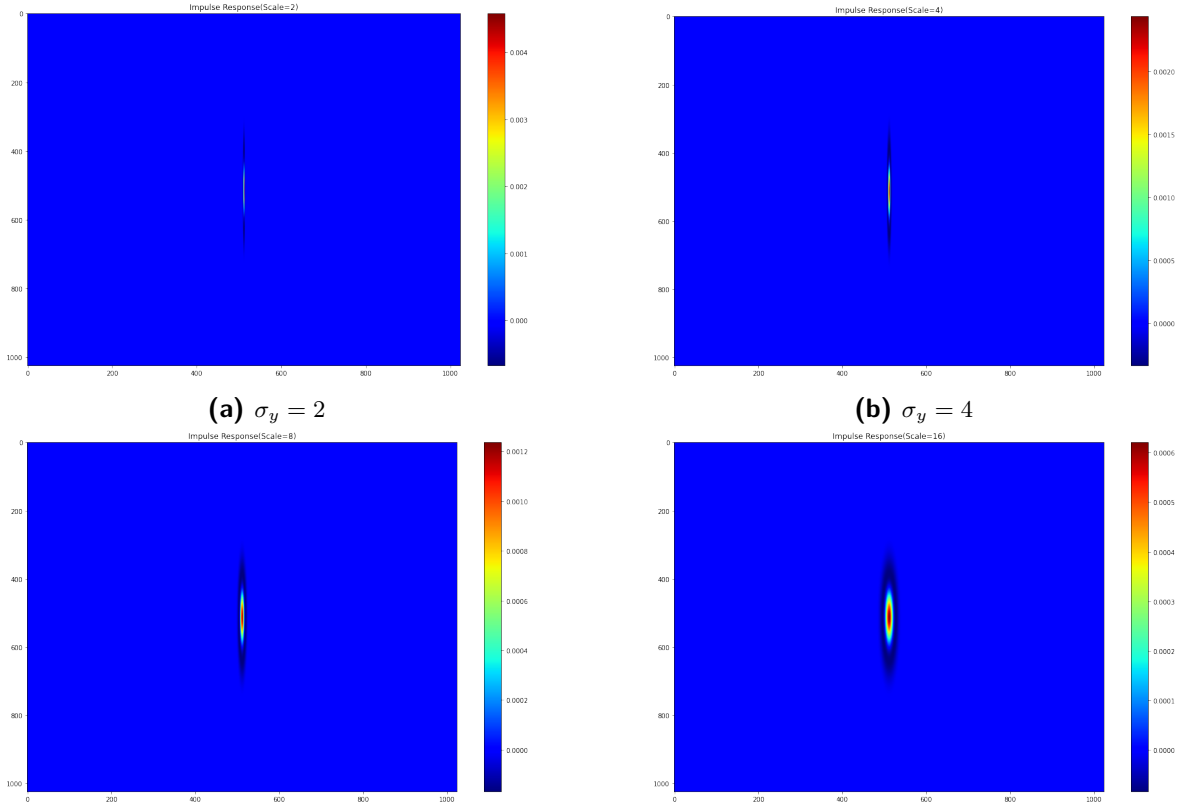
This type of mapping can be used to preserve the monotonically increasing nature of widths. However, in our case, the 2D data is generated from two different and fairly independent sub-instruments. Introducing a mapping similar to equation (4-2) would mean that there is an implicit relation between mass spectrometer and ion mobility spectrometer which is not true.

As a result, our final choice was to fix one of the widths as a single parameter and take a range of values for the other width. This will preserve the monotonically increasing nature of peak width in one dimension. As peaks along the m/z axis are characterized by less number of data points, we found that the overall 2D peak is more sensitive to m/z information. Therefore, our final design decision was to fix σ_x as a single parameter given as an input by the user and to fix σ_y as a range, given as an input given by the user. Experiments and results related to choice of σ_x will be discussed in chapter 5 and chapter 6 respectively.

For the given chapter, the choice of σ_x and σ_y is presented in Table (4-1). These choices are motivated based on the peaks observed in the real world IM-IMS data sample.

a	$\sigma_y = 0.5a$	σ_x	Size of wavelet: floor[$10\sigma_x$] X floor[$10\sigma_y$]
2.0	1.0	64	640 X 10
2.3784	1.1892	64	640 X 11
2.828	1.414	64	640 X 14
3.3635	1.6817	64	640 X 16
4.0	2.0	64	640 X 20
4.7568	2.3784	64	640 X 23
5.6568	2.828	64	640 X 28
6.7271	3.3635	64	640 X 33
8.0	4.0	64	640 X 40
9.5136	4.7568	64	640 X 47
11.3137	5.656	64	640 X 56
13.454	40	64	640 X 67
16.0	8.0	64	640 X 80

Table 4-1: Scale parameters. σ_x and σ_y correspond to width of the wavelet function along rows and columns respectively. The last column presents the size of the wavelet filter



(a) $\sigma_y = 2$ **(b)** $\sigma_y = 4$

(c) $\sigma_y = 8$ **(d)** $\sigma_y = 16$

Figure 4-4: Different mexican hat wavelet functions obtained using equations (4-1) and Table (4-1). The plots are obtained by computing the impulse response of the wavelet filter. In all of the cases, $\sigma_x = 64$.

4-2-4 Convolution

After deciding the width parameters, the next step would be convolution of the partitioned section with the wavelet function. For the given algorithm, convolution is carried out in the spatial domain. In order to minimize the boundary effects, we use edge based padding. Figure (4-5) demonstrates the convolved product of the test section with different 2D mexican hat wavelet functions.

4-3 Denoising of CWT coefficients

After computing the wavelet transform, the next step is denoising of CWT coefficients. As we want to perform peak detection in the transform space (and not perform reconstruction), we define denoising as removing (zeroing out) wavelet coefficients that originate due to the noise in the signal. Denoising CWT coefficients has the advantage of removing wavelet local maxima coefficients generated due to noise, thus improving the overall speed of the algorithm. Our approach to denoising is based on Term-by-Term Hypothesis testing (significance testing) [84] of wavelet coefficients.

4-3-1 Thresholding by hypothesis testing

In order to denoise the wavelet transform space, we need to decide which coefficients should be kept and which coefficients should be zeroed out. This can be formulated as a binary hypothesis test for every wavelet coefficient present in the the transform space.

Mathematically, the test can be formulated as :

$$H_0 : w_{a,x,y} = 0 \quad against \quad H_1 : w_{a,x,y} \neq 0 \tag{4-3}$$

Here, a is the scale parameter (equivalent to width parameter σ_y as discussed previously) and x, y are the spatial location in the 2D plane.

The main idea for the formulation of the null hypothesis H_0 is based on the observation that regions where the data matrix is locally homogeneous will yield nearly zero-valued wavelet coefficients.

Rejection of the null hypothesis depends on double sided p-value of each coefficient.

$$p = Prob(|w_{a,x,y}| > \tau | H_0) + Prob(-|w_{a,x,y}| < -\tau | H_0) \tag{4-4}$$

In this equation, τ refers to the detection threshold.

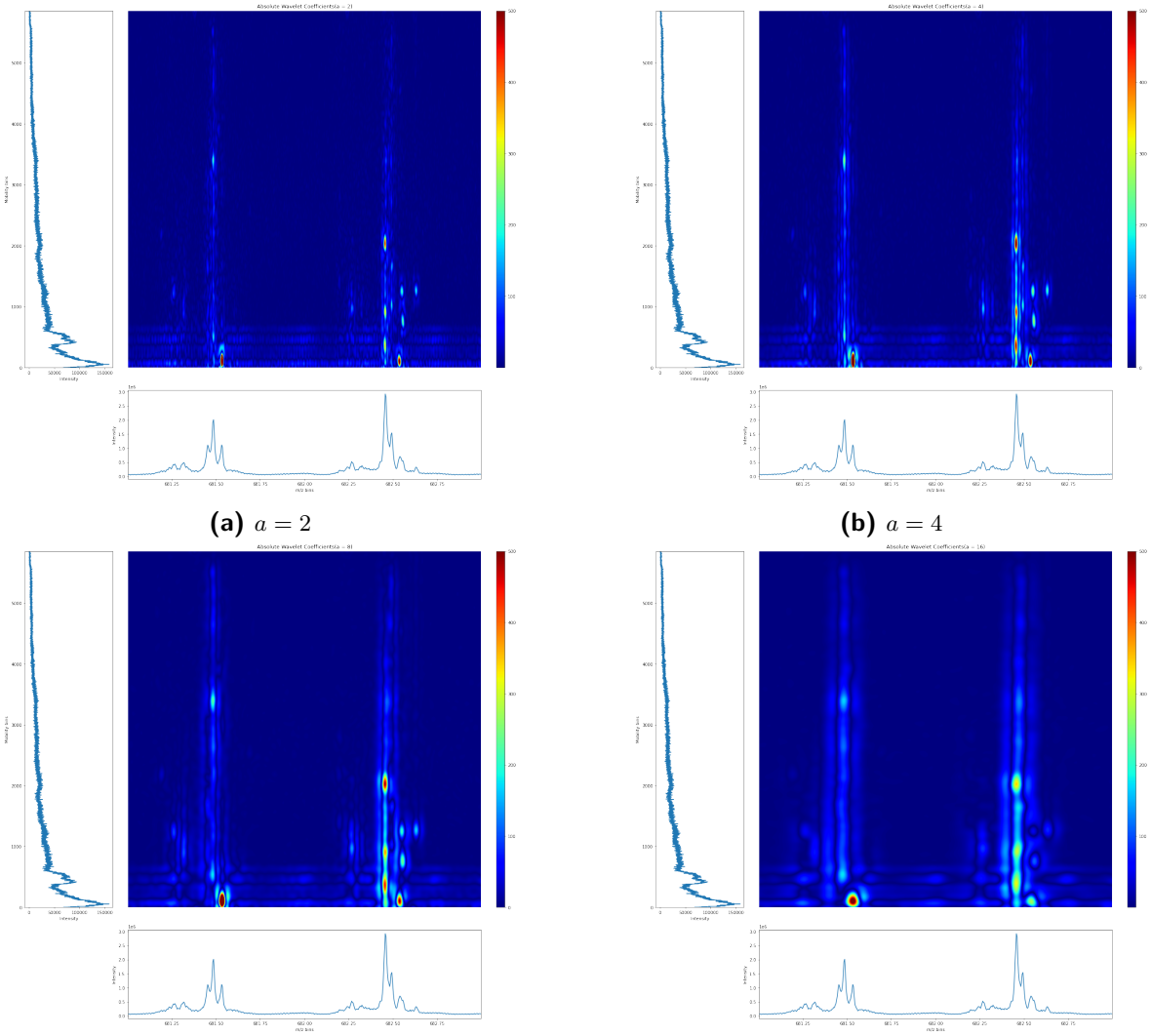
Implementing equation (4-4) would require the knowledge of distribution of $w_{a,x,y}$ under the null hypothesis H_0 and numerical value of critical threshold τ.

If we assume that the coefficients have a zero mean gaussian distribution i.e.

$$w_{a,x,y} \sim N(0, \sigma_a) \tag{4-5}$$

where σ_a is the distribution of the coefficients at scale a, then the p-value can be given as:

$$p = 2(\frac{1}{\sqrt{2\pi}\sigma_n}) \int_{|w_{\sigma_y,x,y}|}^{+\infty} e^{-t^2/2\sigma_n^2} dt = 2(1 - \Phi(|w_{\sigma_y,x,y}|/\sigma_n)) \tag{4-6}$$



(a) $a = 2$ **(b)** $a = 4$

(c) $a = 8$ **(d)** $a = 16$

Figure 4-5: Absolute CWT coefficients of the test section for different scale parameters a (equivalent to width parameter σ_y).

where Φ is the standard normal cumulative distribution function.

Now given a Type 1 error level, α. Here, type 1 error is the probability of rejecting the null hypothesis when the null hypothesis holds true and α is the risk of committing of the error. If $p > \alpha$, then the null hypothesis H_0 is not excluded, i.e. the value of the coefficient could be due to noise. However, if $p \leq \alpha$ then the coefficient is likely not generated due to noise and hence the null hypothesis H_0 is rejected.

In equation (4-6), we define critical threshold τ as:

$$\tau = \Phi^{-1}(1 - \alpha/2) \tag{4-7}$$

Using (4-6) and (4-7), we get:

$$\begin{array}{ll} |w_{\sigma_x,x,y}| \geq \tau\sigma_n & w_{\sigma_x,x,y}\ is\ significant \\ |w_{\sigma_x,x,y}| < \tau\sigma_n & w_{\sigma_x,x,y}\ is\ insignificant \end{array} \tag{4-8}$$

Choosing $\tau = 3$, corresponds to $\alpha = 0.002$.

Based on equation (4-8), the thresholding policy for wavelet coefficients at scale a is given as:

$$w(a,x,y) = \begin{cases} 0 & if\ w(a,x,y) < 3\sigma_a \\ w(a,x,y) & if\ w(a,x,y) \geq 3\sigma_a \end{cases} \tag{4-9}$$

This equation retains only positive significant coefficients. The negative coefficients correspond to locations which are out of phase with the wavelet function and are thus removed.

4-3-2 Noise level estimation

The final step before using equation (4-9) as the threshold policy would be to estimate σ_a. We estimate σ_a with the following relation:

$$\sigma_a = \sigma_{2D}\sigma_a^{0,1} \tag{4-10}$$

Here, σ_{2D} corresponds to standard deviation of the noise in the 2D data and $\sigma_a^{0,1}$ is the standard deviation of the wavelet transform of zero mean white noise with variance = 1 at scale a.

We observe that there are two terms that requires to be estimated in order to calculate σ_a. The second $\sigma_a^{0,1}$ term is obtained by the following procedure:

1. Simulate zero mean white noise with variance = 1.
2. Compute the wavelet transform using scale parameter a.
3. Compute the standard deviation of the wavelet coefficients obtained in step 2.

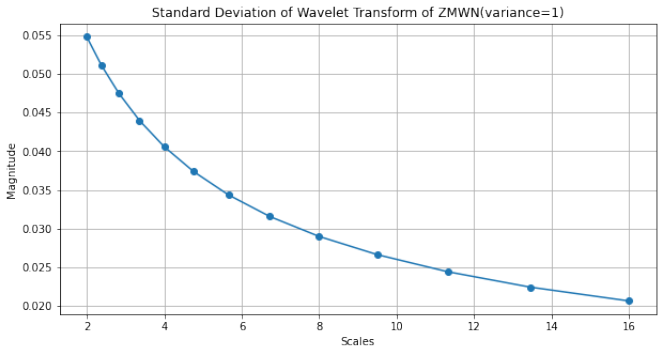


Figure 4-6: Standard Deviation of wavelet transform of Zero mean white noise with variance = 1 for different scales a.

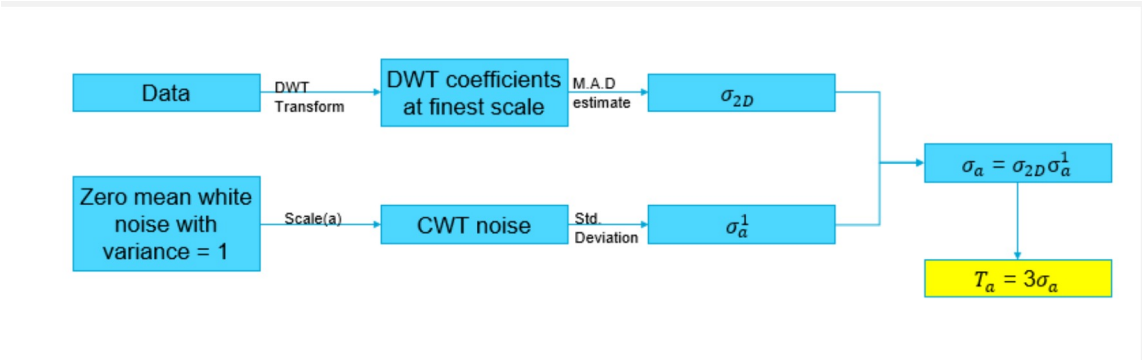


Figure 4-7: Layout of the denoising pipeline at a given scale a in the feature detection algorithm.

The results for the second term, corresponding to the scales considered in equation is displayed in Figure (4-6).

We estimate the first term σ_{2D} using DWT [28]. The estimate of σ_{2D} using DWT is given as:

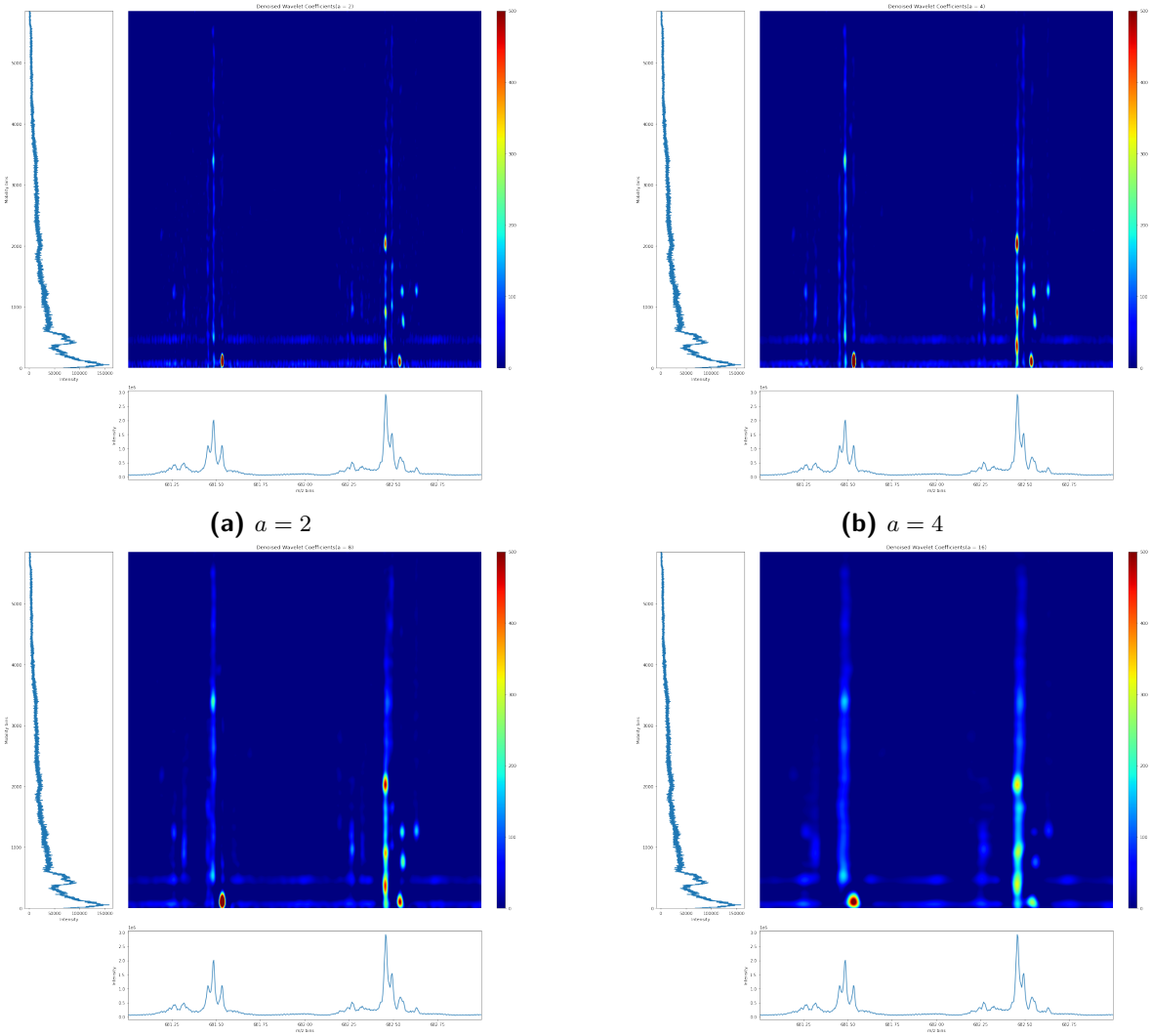
$$\hat{\sigma}_{2D} = M.A.D(w_1)/0.6745 = median(|w_1 - median(w_1)|)/0.6745 \tag{4-11}$$

where M.A.D. stands for median absolute deviation, w_1 corresponds to DWT coefficients obtained at the finest scale using orthogonal wavelet functions and 0.6745 is a correction factor specific to gaussian distribution. For 2D data, the estimator is applied at the diagonal sub-band of the finest scale.

4-3-3 Overall Pipeline

Having discussed the specifics of the denoising algorithm, we now present the complete pipeline for denoising of CWT coefficients for a single scale (a). The pipeline can be extended to all values of a independently.

The results of denoising on the test section is presented in Figure (4-8). Figure (4-9) presents the fraction of wavelet coefficients retained at each scale a after denoising.



(a) $a = 2$ **(b)** $a = 4$

(c) $a = 8$ **(d)** $a = 16$

Figure 4-8: Denoised wavelet coefficients of the test section at different scales.

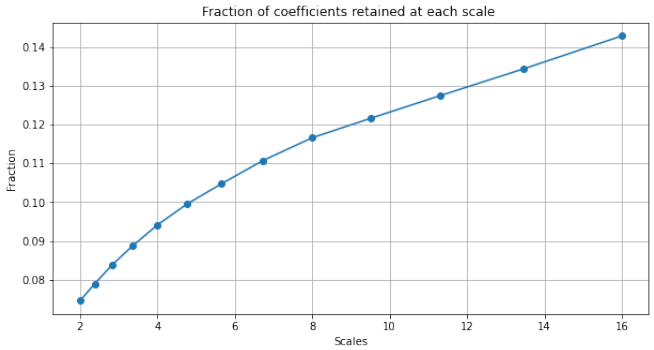


Figure 4-9: Fraction of wavelet coefficients retained after denoising for different scales. The fraction is obtained by : (No.of non-zero wavelet coefficients at scale a/ No. of absolute no-zero wavelet coefficients before denoising at scale a)

4-4 Automatic Local maxima clustering and chain construction

The implementation of this idea is adapted from the works of Bijaoui et al. [10]. After denoising, the next step is to identify and connect local maxima coefficients that are related to each other in transform space. This procedure can be broken down into three sections:

1. **Local maxima detection** - Local maxima coefficients are identified at every scale a.
2. **Automatic local maxima clustering** - This section is required to account for shoulder peaks contained in the data sample. Shoulder peaks typically exist surrounding dominant peaks, forming short chains around long chains (which corresponds to dominant peaks). Therefore, we first cluster local maxima coefficients across transform space. Every cluster will lead to formation of at least one chain.
3. **Chain construction**- After clustering the local maxima coefficients, we construct chains based on some optimization criteria.

4-4-1 Local maxima detection

We define a local maxima at any scale a as a point whose coefficient value is strictly greater than the coefficient values of it's surrounding eight point neighbourhood. In our algorithm, we ignore the local maxima coefficients that exist at the border of the partitioned data i.e points that have less than eight point neighbourhood are not candidates for local maxima coefficients.

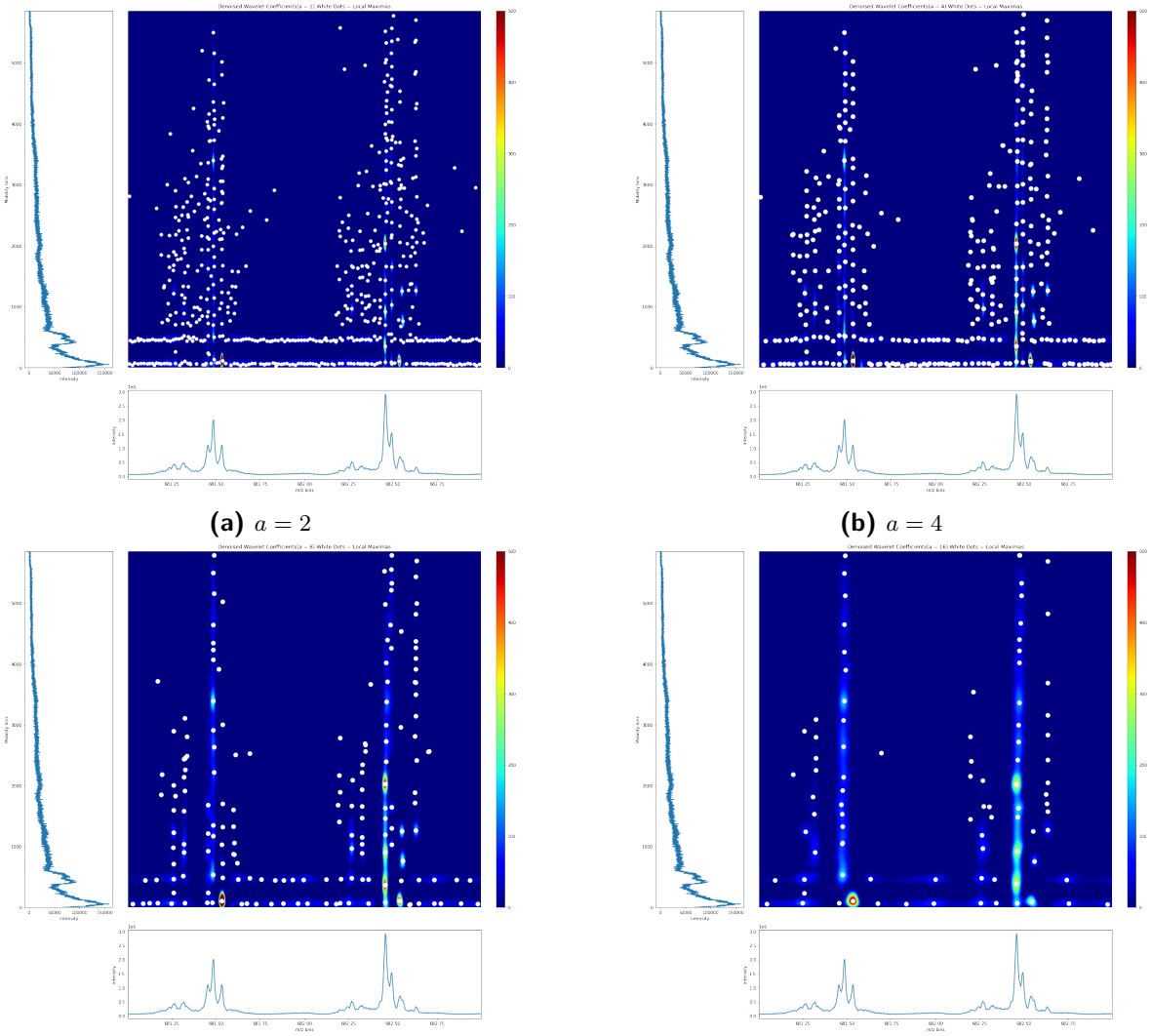
After defining the local maxima point, we detect local maxima coefficients at every scale a. Figure (4-10) displays the detected local maximas in the denoised wavelet transform of the test section at scale for different scales.

4-4-2 Automatic local maxima clustering

Next step is to cluster local maxima points that are related to each other in transform space. Mathematically, if there exists a local maxima coefficient at position b_{loc} and scale a represented as $W(b_{loc}, a)$, then the objective is to identify local maxima coefficients at scale $a-1$ that are related to the coefficient $W(b_{loc}, a)$ (Note that in this step, there could be more than one local maxima coefficient related to $W(b_{loc}, a)$).

Gregoire et al. [39] specify a window around the position b_{loc} to capture local maxima points present at scale $a-1$. This window is based on the prior knowledge of the wavelet being used and the scale being investigated. However, this window was found to be restrictive and it did not take the behaviour of the data or the behaviour of surrounding wavelet coefficients into account.

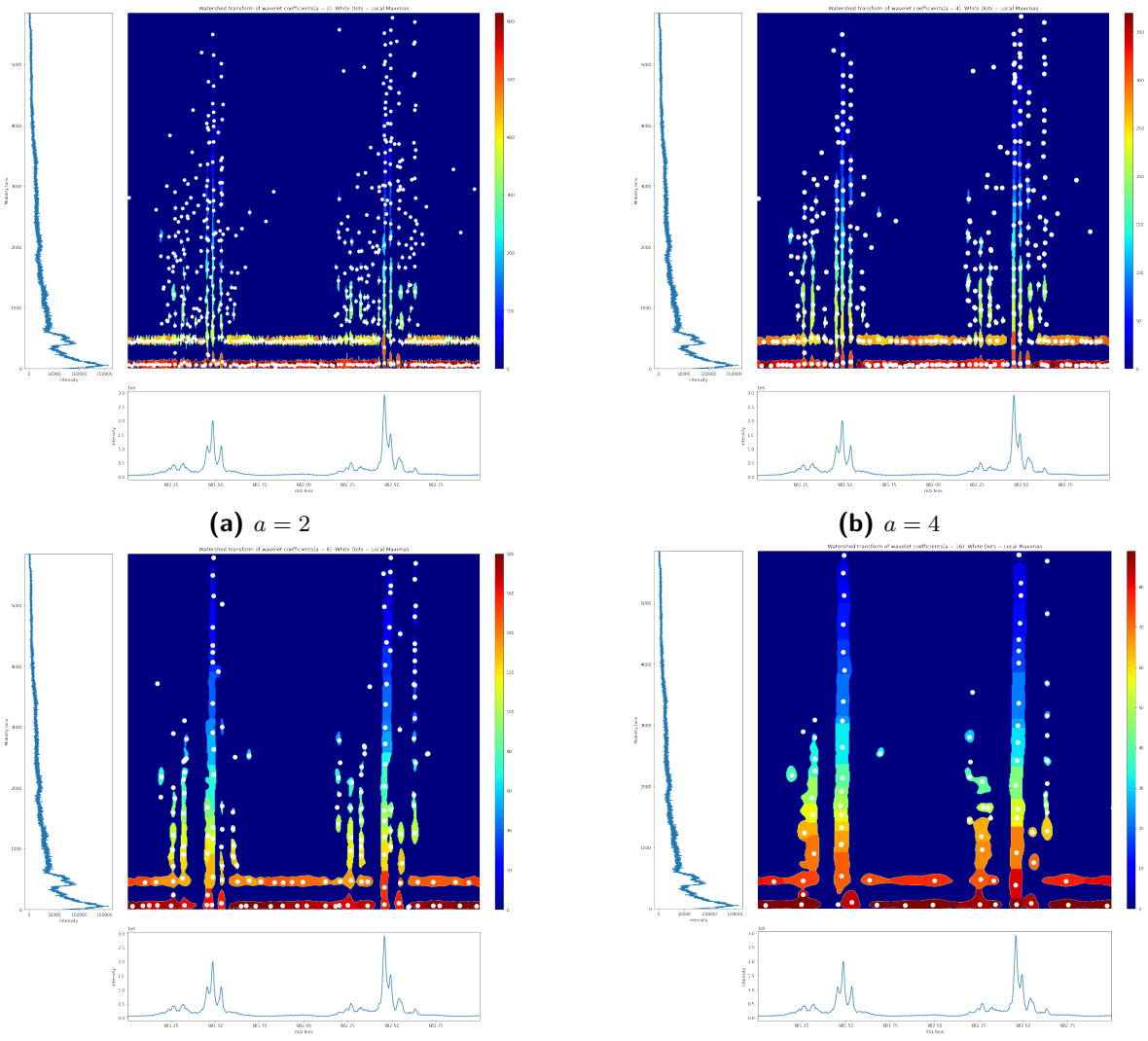
Thus, in order to make this procedure data driven, we use watershed segmentation [81] to establish a region of influence around the local maxima point. Starting from user-defined markers, the watershed algorithm treats bin values as a local topography (elevation). The algorithm floods basins from the markers until basins attributed to different markers meet on watershed lines.



(a) $a = 2$ **(b)** $a = 4$

(c) $a = 8$ **(d)** $a = 16$

Figure 4-10: Detected local maxima coefficients present in the denoised wavelet coefficients of test section at different scales a. White dots represent the local maximas at each scale.



(a) $a = 2$ **(b)** $a = 4$

(c) $a = 8$ **(d)** $a = 16$

Figure 4-11: Watershed segmentation of denoised wavelet coefficients at different scales a. White dots represent the local maximas detected at this scale. Each local maxima is associated with its region of influence.

In our case, we define markers as the local maxima coefficients that are present on a scale a. Thus, using watershed segmentation, we are able to characterize a region of influence around a given local maxima point. This region of influence will act as the search window to look for local maxima points present at scale $a-1$. Figure (4-11) demonstrates the watershed segmentation performed on the denoised wavelet transform of the test section at dyadic scales.

Implementation of local maxima clustering

1. Initialize from the coarsest scale a_{max}, set $a' = a_{max}$.
2. For every local maxima point at scale a' (Note that this will give us clusters that start from $a' = a_{max}$):
   (a) Set search radius = Region of influence corresponding to the local maxima point.
   (b) Identify all local maxima points in $a' - 1$ that belong in the search radius. These local maxima points are now part of the cluster.
   (c) Modify the search radius. New search radius = Region of influence corresponding to local maxima point at scale a' + Region of influence corresponding to local maxima point at scale $a' - 1$.
   (d) Remove the local maxima points that got detected at scale $a' - 1$.
   (e) Go to scale $a' - 2$ and repeat till $a' = a_{min}$.
3. Set a' to $a' - 1$ and repeat till a_{min} (This will give us clusters that are initialized from $a' = a_{max-1}$).

Figure (4-12) gives a pictorial representation of the algorithm.

Denoising Clusters

After the clustering operation is completed, we remove clusters that have a length of one. Clusters having a length of one means that there is only one local maxima point in that cluster. This local maxima point was not able to find other local maxima points at coarser or finer scales and therefore no chains can be constructed in this cluster. This means that the local maxima point is specific to that scale. We attribute this type of local maxima point to noise and therefore remove them.

4-4-3 Chain construction

The final step after clustering is chain construction. The process of chain construction is similar to clustering operation described above. A key difference is that a local maxima point at scale a will get linked to only one local maxima point at scale $a-1$. This local maxima point should be closest (in terms of position parameter b) to the previous local maxima point and should be within the region of influence of the previous local maxima point. Also, chain construction is done within a cluster i.e. local maxima points that are candidates for chain construction should belong to the same cluster.

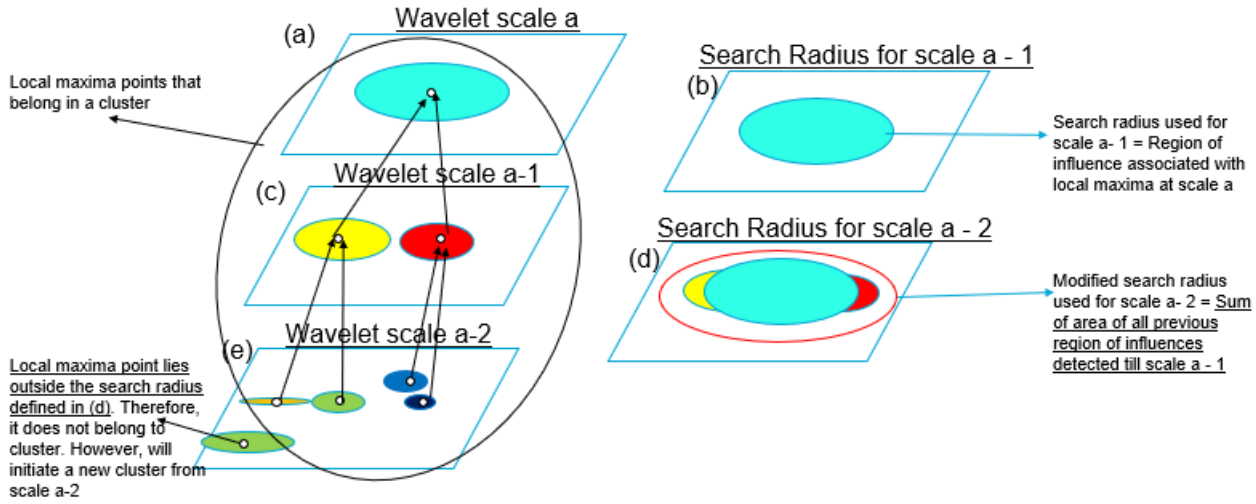


Figure 4-12: Visual demonstration of the local maxima clustering algorithm. (a) Cluster initialized by a local maxima coefficient at scale a. (b) Search radius defined for detecting local maxima coefficients at scale $a-1$. The search radius = Region of influence associated with local maxima coefficient at scale a (c) Local maxima coefficients detected at scale $a-1$. These local maxima coefficients lie within the search radius defined earlier and are linked to the initialized cluster. After linking, they are removed. (d) Modified search radius. The search radius now includes regions of influence associated with local maxima coefficients detected at scale $a-1$. (e) Local maxima detected at scale $a-2$. The local maxima coefficients that lie within the search radius belong to the initialized cluster and are linked and removed. This process continues till scale a_{min} is reached. The local maxima coefficient that lie outside the search radius will not be linked. Instead, they will initialize a new cluster from scale $a-2$ and the process from step(a) to step(e) will be repeated.

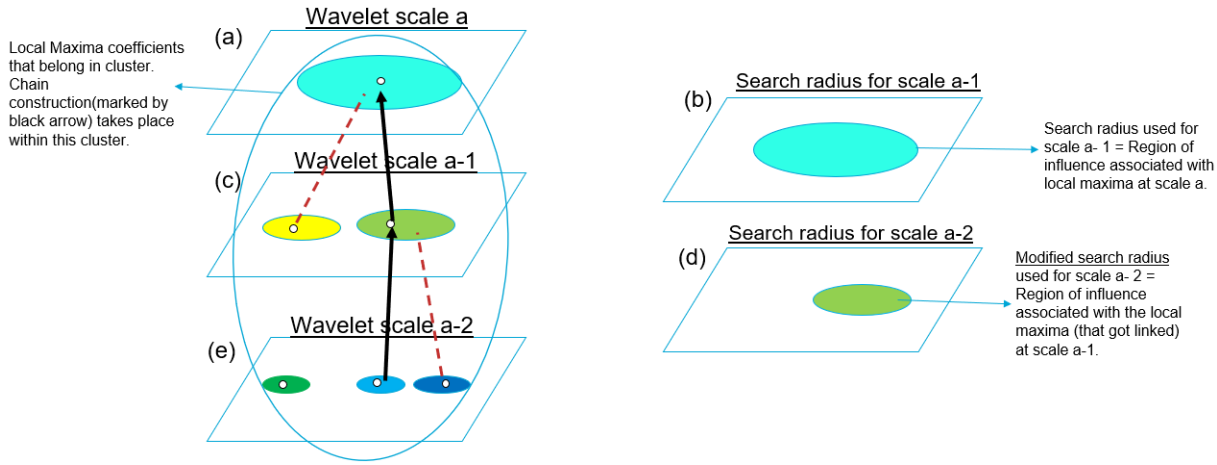


Figure 4-13: Pictorial demonstration of the chain construction algorithm. The chain construction takes place within a cluster (marked with a circle). The black arrows represent a chain. The dashed red lines represent potential local maxima coefficients that were rejected based on the distance criteria. (a) Chain initialized by a local maxima coefficient at scale a. (b) Search radius defined for detecting local maxima coefficients at scale $a-1$. The search radius = Region of influence associated with local maxima coefficient at scale a (c) Local maxima coefficients detected at scale $a-1$. These local maxima coefficients lie within the search radius defined earlier. The closest local maxima coefficient (in terms of position parameter b) gets linked and removed. This is indicated with a black arrow. The remaining local maxima coefficient will initiate a new chain from scale $a-1$ (d) Modified search radius. New search radius = region of influence associated with local maxima coefficient linked at scale $a-1$. (e) Local maxima coefficients detected at scale $a-2$. The closest maxima point that was within the search radius defined in (d) was linked and removed (marked with black arrow). This process continues till scale a_{min} is reached.

Implementation of Chain Construction

1. Initialize from the coarsest scale in a cluster a_{max}, set $a' = a_{max-cluster}$.
2. For every local maxima point at scale a'(Note that this will give us clusters that start from $a' = a_{max-cluster}$):
   (a) Set search radius = Region of influence corresponding to the local maxima point.
   (b) Identify all local maxima points in $a'-1$ that belong in the search radius. The closest local maxima point gets linked.
   (c) Modify the search radius. New search radius = Region of influence corresponding to local maxima point at scale $a'-1$.
   (d) Remove the local maxima point that got linked at scale $a'-1$.
   (e) Go to scale $a'-2$ and repeat till $a' = a_{min-cluster}$
3. Set a' to $a'-1$ and repeat till $a_{min-cluster}$ (This will give us clusters that are initialized from $a' = a_{max-cluster-1}$).

Figure (4-13) gives a pictorial representation of the algorithm.

4-5 Peak detection criteria : Effective length thresholding

After construction of chains, the final task is to separate chains that are triggered due to noise from chains that are triggered due to chemical peaks. In the literature, we found that thresholding of wavelet chain can be done based on their length, magnitude of wavelet coefficient (belonging to a chain) at the the finest scale or the wavelet chain energy (refer Section 3-4). We wanted to define a single parameter that combines all of the above three criteria. For this, we define a parameter called effective length. Effective length represents the number of scales at which the local maxima coefficients (belonging to a chain) are greater than the surrounding noise level. The algorithm takes user defined threshold length as input and if the effective length of a given chain is greater than the threshold length, the chain is likely triggered due to a chemical peak.

The procedure for calculating effective length can be broadly broken down into three steps:

1. Assuming gaussian distribution of local noise, calculate the noise parameter σ_{local} from the local surrounding of the local maxima wavelet coefficient belonging to a chain.

2. Simulate the wavelet transform of σ_{local} for all scales a. Generate a threshold value using the transform space.

3. If the local maxima coefficient of a chain generated by the data at scale a is greater than the the local maxima coefficient generated by the wavelet transform of σ_{local} at scale a, then effective length+=1.

We present the details about the parameters involved in each of the mentioned steps:

4-5-1 Calculation of σ_{local}

σ_{local} refers to the standard deviation noise parameter corresponding to the local surrounding of the wavelet local maxima point. We define the local surrounding as the column or the row corresponding to the local maxima point. Based on our definition of the 2D matrix, the column would correspond to the mobility information for a particular m/z value and the row would correspond to the m/z information (within the partitioned data sample) for a particular mobility bin value.

The steps for calculating σ_{local} are:

1. For every wavelet local maxima point in a chain, extract the column and the row, corresponding to the local maximum point, in the data sample.

2. Compute the standard deviation parameter for the row and column independently, using DWT. Label them as $\sigma_{m/z}$ and $\sigma_{mobility}$.

3. If $\sigma_{m/z} > \sigma_{mobility}$, $\sigma_{local} = 2\sigma_{m/z}$ else $\sigma_{local} = \sigma_{mobility}$.

Explanation for σ_{local}

It is important to understand the choice of σ_{local} with respect to the given data sample. The 2D IM-MS data sample is in continuous profile mode. As a result, the observed peaks have significant width along both the dimensions. In order to account for the nature of continuous profile data, we consider noise both in the horizontal direction and vertical direction of the local maxima point. Also, often it is observed that peaks with high signal to noise ratio are very broad with widths in the range of 80-120 m/z bins. If the width of the peak (along m/z direction) is greater than the width of the largest wavelet function then the algorithm will detect several false peaks that lie on the surface of the main peak. In order to minimize this effect, we put a penalty factor of 2 in the horizontal direction of noise.

4-5-2 Translation of the detection level to wavelet transform space

After the local noise level has been established, the next step would be to translate the noise level to the wavelet transform space. This translation needs to be adapted to the quantity being studied which are local maxima points. An additional requirement would be that the translation should generate a significant threshold value so that local maxima points that lie on the surface of the peaks can be removed effectively. Based on these criteria, we decided that the threshold level will be governed by the maximum of local maxima points of the wavelet transform of the local noise level σ_{local}. However, this procedure will become expensive as the behaviour of σ_{local} will vary for every local maxima point based on its location in the partitioned data sample. In order to simplify the procedure, we develop a relation similar to the one developed in the denoising procedure of wavelet coefficients. The complete procedure is as follows:

1. Generate a simulation of zero mean gaussian noise with variance = 1.
2. For every scale a
   (a) For B = 1:50
      i. Simulate zero mean white noise with variance = 1.
      ii. Compute the wavelet transform(corresponding to scale a).
      iii. Identify the maximum of the local maxima points in the transform space.
      iv. Store the value.
   (b) After the cycle is complete, we will have 50 values corresponding to the maximum value generated from the simulation.
   (c) Compute the average of the values. This will give the average maximum of the local maxima coefficients generated due to wavelet transform of the zero mean white noise with variance = 1 at scale a. This value can be denoted as M_a.
   (d) The contribution of local noise at scale a can now be given as:$T_{a,local} = \sigma_{local} M_a$.

Steps (i) - (iv) can be implemented separately. Essentially, the maximum of local maxima coefficients at scale a is not a stable statistic i.e. it varies with every simulation. As a result,

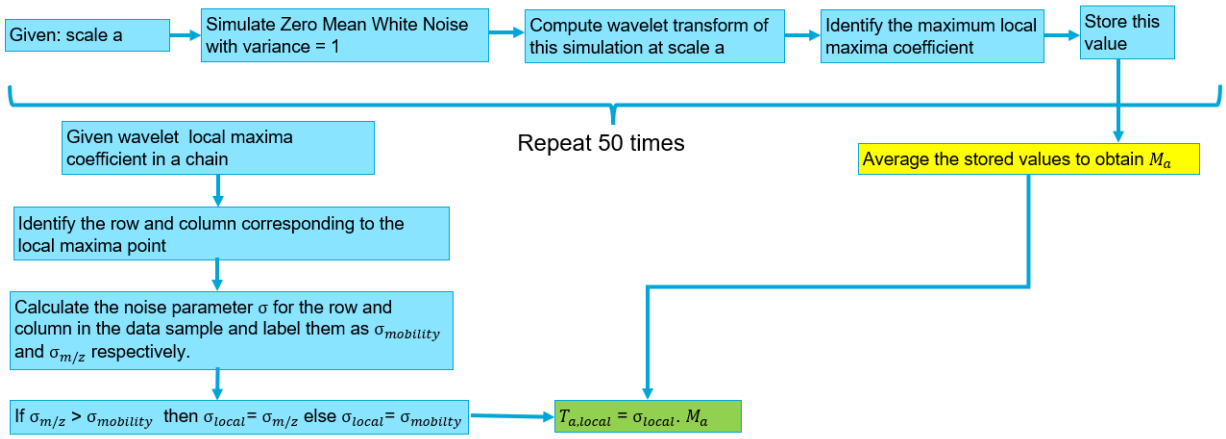


Figure 4-14: Layout of the pipeline for calculation of $T_{a,local}$. The quantity M_a can be determined independently.

we used multiple simulations followed by averaging in order obtain a stable representation of this quantity. Experiments related to the equation used in step(d) are given in Chapter 5.

The complete pipeline for deriving the threshold value $T_{a,local}$ for a wavelet local maxima coefficient is given in Figure (4-14)

4-5-3 Effective length threshold

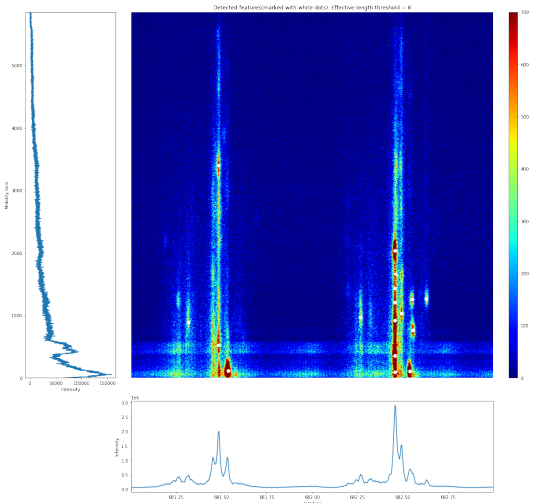
Once we have established the threshold value generated due to local noise at scale a, we compare it with the local maxima coefficient belonging to a chain at scale a. If the latter quantity is higher than the former, we say that at scale a, the local maxima point is dominant. Else, the local noise is dominant at scale a. We perform this comparison for all the scales for which the chain exists and record the number of scales at which the local maxima coefficient is dominant. If this quantity is greater than a threshold length (given by the user), we say that the chain is triggered due to a chemical peak.

Figure (4-15) presents the results with varying effective threshold length for the test section.

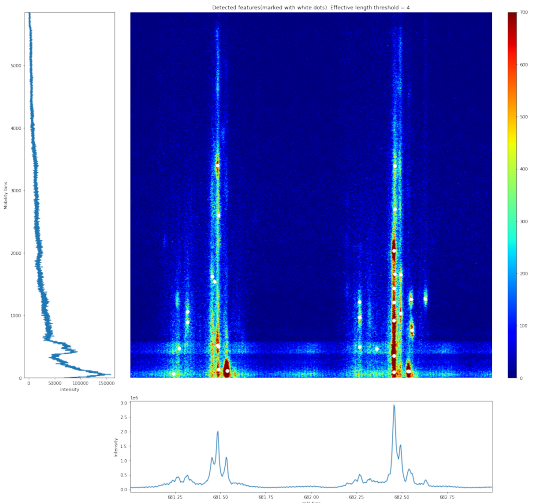
4-6 Parameters

Having presented the functioning of the algorithm, we provide a brief overview of the parameters, with their description, used in the algorithm is presented below.

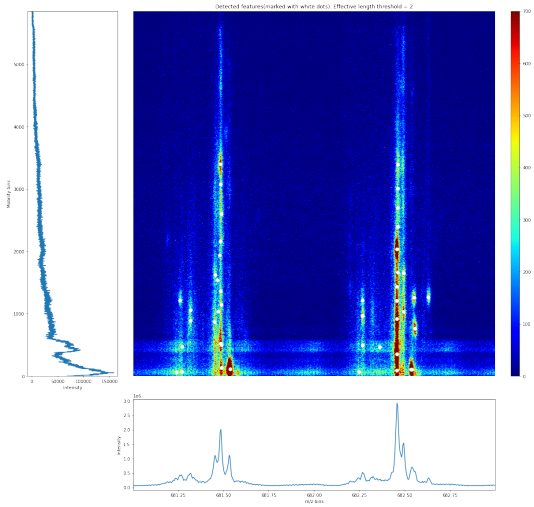
In the next chapter, we will perform experiments related to these parameters in order to understand their impact on the performance of the algorithm and to gain some working knowledge about the optimal values of these parameters.



(a) Effective length threshold = 8. Total number of detected peaks = 17



(b) Effective length threshold = 4. Total number of detected peaks = 31



(c) Effective length threshold = 2. Total number of detected peaks = 43

Figure 4-15: Detected peaks(white dots) using different effective length threshold for the given test section.

1. σ_x - Describes the width of the wavelet function (along rows). The parameter is based on the minimum width of the peak in the mobility dimension. Default value = 64 (1% of the mobility dimension).

2. $\sigma_y = 0.5a$ - Describes the widths of the wavelet function (along columns). The parameters is based on the possible widths of peaks along the m/z dimension and therefore takes a range of values as input. Default value = Dyadic values starting from 1 to 8 with three voices per octave.

3. **Effective length threshold** - Parameter that thresholds the wavelet chains constructed in transform space. Default value = 4.

4. **Number of noise simulations** - Hyperparameter that controls the number of noise simulations required for calculating $T_{local,a}$. Default value is 50.

5. **Penalty Factor** - Hyperparameter that is used in calculation of σ_{local}. Default value = 2.

Chapter 5

Performance evaluation of the algorithm

Having discussed the functioning of the algorithm in the previous chapter, we will now evaluate the performance of the algorithm by conducting various experiments. First, we will evaluate the parameters that are being estimated by the algorithm. Next, we will evaluate the performance of the algorithm based on the parameters that are given as inputs by the user. This will help us to establish some working knowledge about the input parameters and their overall impact on the algorithm. Using this knowledge, the final evaluation will be to compare the designed algorithm with an existing feature detection algorithm.

The chapter is divided into four major experiments:

1. **Experiment 1 - Denoising CWT coefficients**
   (a) Evaluate the performance of the estimator $\hat{\sigma}_{2D}$ obtained using DWT.
   (b) Evaluate the equation: $\sigma_a = \sigma_{2D}\sigma_a^{0,1}$
2. **Experiment 2 - Effective length threshold parameter**
   (a) Evaluate the equation: $T_{a,local} = \sigma_{local}M_a$
   (b) Evaluate the performance of the algorithm by varying the effective length threshold parameter for the synthetic IM-IMS data sample.
   (c) Evaluate the performance of the algorithm by varying the effective length threshold parameter for the real world IM-IMS data sample.
   (d) Evaluate the penalty factor used in estimating σ_{local} for the real world IM-IMS data sample.

3. **Experiment 3 - Scale parameter**

   (a) Evaluate the performance of the algorithm by varying σ_x on a synthetic IM-IMS data sample.

   (b) Evaluate the performance of the algorithm by varying σ_x on a real world IM-IMS data sample.

4. **Experiment 4 - Compare the performance of the designed algorithm with an existing peak detection algorithm.**

5-1 Data samples

Before we begin the performance evaluation of the algorithm, we present a brief description of the data samples being used for this evaluation.

5-1-1 Real world data sample

Real world IM-IMS data sample

The first IM-IMS data sample we will be using is obtained from a cross section of a mouse kidney tissue. In the data sample, the m/z values range from 600-950 and the number of mobility bins is 5857. The total size of the 2D matrix is 80049 x 5857. Details about the sample preparation and the instrument used can be found in [83].

For the evaluation of the algorithm, we will use small sections of this data sample partitioned along the m/z axis. Each of these partitioned sections are different from one another in a quantitative and qualitative manner.

Based on our observations, we briefly highlight some of the properties of the selected sections. It is important to note that we refer to these sections by their m/z range. This is because we retain the complete mobility information associated with the m/z range i.e no. of columns = 5857 for all partitions.

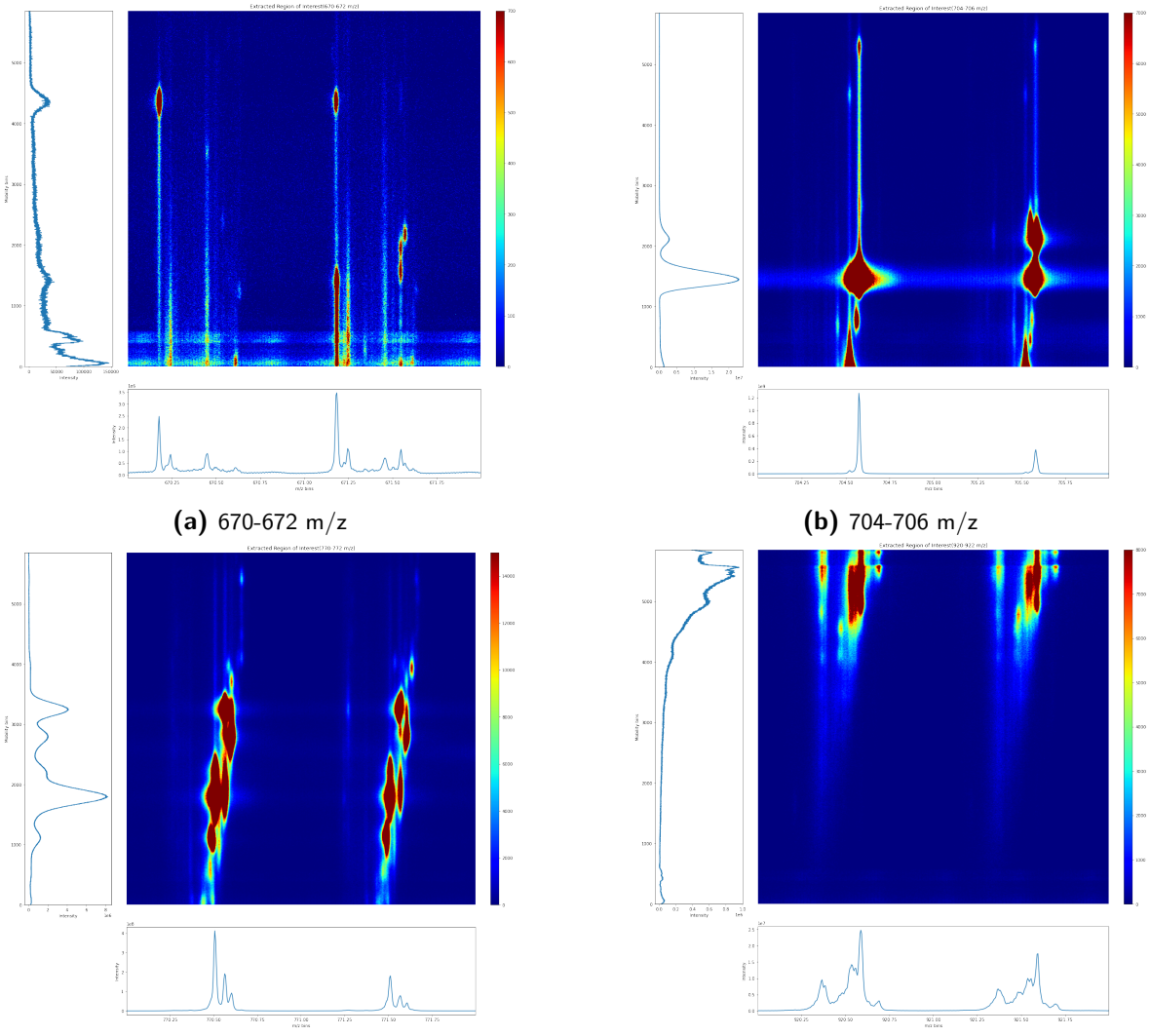
1. 670-672 m/z - Low SNR with scattered peaks

2. 704-706 m/z - High SNR with broad peaks.

3. 770 - 772 m/z - High SNR with overlapping peaks.

4. 920 - 922m/z - Moderate SNR with narrow peaks widths.

Figure (5-1) presents a visual representation of the partitioned 2D data matrices.



(a) 670-672 m/z

(b) 704-706 m/z

(c) 770-772 m/z

(d) 920-922 m/z

Figure 5-1: Different sections of the real world 2D IM-IMS data sample. The sections are obtained by partitioning the data sample along the m/z axis. The mobility dimension is completely preserved in all of these sections i.e. no. of columns in the 2D matrix = 5857 for all the partitioned sections.

5-1-2 Synthetic data samples

Synthetic IM-IMS data sample

Our synthetic data sample is modelled and developed based on our observations of the real world data sample. The procedure used for developing this synthetic data sample is briefly described below.

1. We select a small range of 207-212 m/z. We set the resolution of the m/z instrument equal to 65000. In Time of Flight (ToF) instruments, the sampling interval linearly increases as the m/z value increases. This information combined with the resolution gave a total of 1500 m/z bins (data points) for the selected range of m/z values.

2. Based on the range of the m/z values, we select chemical compounds that will be used in the data sample. These compounds were selected using the online molecular database [69].

3. We insert isotopes (with their relative intensities) related to every chemical compounds in the data sample. This is done using the python package *Molmass*. Given a chemical compound, *Molmass* computes the possible isotopes of that chemical compound as well as the relative percentage of these isotopes.

4. For every chemical compound:

   (a) We model the m/z peak. For the m/z peaks, we use pseudo-voigt function [23] as our template. Psedo-voigt function is defined as the sum of a gaussian function and a lorentzian function with a weighing parameter η which shifts the function profile towards gaussian function or lorentzian function. The equation is given as:

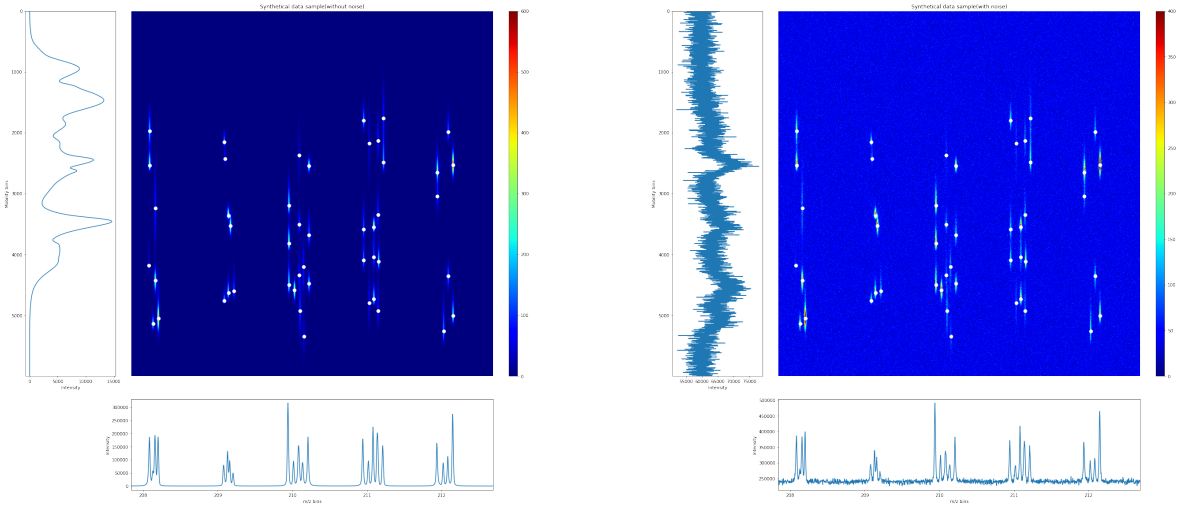
   $$pV(x) = I.[\eta(\frac{1}{\sigma\sqrt{2\pi}})e^{-\frac{(x-x_o)^2}{2\sigma^2}} + (1-\eta)\frac{1}{\pi}\frac{\frac{\Gamma}{2}}{(x-x_o)^2 + (\Gamma/2)^2}] \tag{5-1}$$

   where x_0 describes the peak location, I is the shared intensity , η is the weighing parameter controlling the behaviour of the peak function and Γ is the full width half maximum parameter shared between gausssian function and lorentzian function. We implement this function using the python package *Hyperspy*.

   (b) After modelling the m/z peak, we model the mobility peaks associated with the m/z peak. Our observation suggested that mobility peaks tend to be of gaussian shape with extended tails. As a result, the mobility peaks are modelled using skew normal distribution function. This function has a parameter α which controls the skew of the normal distribution function. Mathematically, this function can be represented as :

   $$f(x) = \phi(x)\Phi(\alpha x) \tag{5-2}$$

   where $\phi(x) = e^{-\frac{x^2}{2}}/\sqrt{2\pi}$ and $\Phi(\alpha x) = \int_{-\infty}^{\alpha x}\phi(t)dt$. Positive value of α causes the distribution function to skew towards right and negative value of α causes the distribution function to skew towards left. Magnitude of the α controls the degree of skew.



(a) Synthetic data sample (without noise)

(b) Synthetic data sample (with noise). Noise parameters (0,100)

Figure 5-2: Synthetic data sample. M/z range: 208 m/z-212 m/z. Mobility bins: 6000. White dots mark the true peaks associated with chemical compounds.

The number of mobility peaks, associated with an m/z peak, is randomized between 2 and 5. It is more important to note that this randomization does not mimic the actual mobility information associated with the molecule and is mostly used for introducing variation while modelling the synthetic data sample.

(c) After modelling the peaks in both the dimensions, we compute the outer product of the mobility peaks and m/z peak.

5. After all the 2D peaks have been modelled, we add zero mean white noise with parameters $\sim N(0, 100)$ to the 2D matrix. Adding gaussian white noise may lead to negative values in the 2D matrix. These values are zeroed out.

Figure (5-2) presents the smooth and noisy synthetic data sample with marked peaks.

Synthetic Image data sample

In the last step of design of 2D IM-IMS data sample, we zeroed out the negative values which may occur due to additive gaussian noise. This skews the noise parameter σ to a different value and therefore cannot be used for estimation of σ_{2D}. As a result, we use a different 2D data where we do not zero out the negative values. Figure (5-3) presents the sample 2D image data which will be used for our experiment of estimation of the noise parameter σ. We use this image because it has natural high frequency components (water droplets) and low frequency components (background forest, mountains,sky) which can impact the estimation of the noise parameter σ_{2D}.

Figure 5-3: 2D Image - Happy elephant

5-2 Experiment 1 - Denoising CWT coefficients

The experiments in this section are related to section (4-3). The quantities being studied are: $\hat{\sigma}_{2D}$ and σ_a from equation (4-10).

5-2-1 Experiment 1.1 - Estimation of standard deviation parameter σ_{2D} of gaussian white noise

Methodology

For this experiment, we use Figure (5-3) as our reference 2D image and add gaussian noise to this image. After this we evaluate the average performance of the estimator and compute it's bias and variance. The procedure is as follows:

1. For a given σ_{2D}
2. For B = 1:100 :
   (a) Add gaussian noise σ_{2D} to this image.
   (b) Compute the estimate $\hat{\sigma}_{2D}$ using DWT.
   (c) Store the value.
3. Average the stored values. This will give $E[\hat{\sigma}_{2D}]$.
4. Compute the bias: $\sigma_{2D} - E[\hat{\sigma}_{2D}]$.
5. Using the stored values, compute the variance of the estimator: $Var(\hat{\sigma}_{2D})$.
6. Repeat the procedure for varying σ_{2D}.

Results

σ_{2D}	$E[\hat{\sigma}_{2D}]$	$\lvert\sigma_{2D} - E[\hat{\sigma}_{2D}]\rvert$	$Var(\hat{\sigma}_{2D})$
0	0.5311	0.5311	1.232e-32
7	7.8831	0.8831	0.00037
21	21.5829	0.5829	0.0031
64	64.1515	0.1515	0.0313
120	119.960	0.0391	0.1060
254	253.640	0.3591	0.4822

Table 5-1: Results for experiment 1.1 with varying noise parameter $\sim (0, \sigma_{2D})$. $E[\hat{\sigma}_{2D}]$ is the estimated σ_{2D} averaged over 100 simulations. The last two columns represent the bias and the variance of the estimator.

Discussion

Before we begin our analysis, we highlight the following equation. Given a signal with zero mean additive gaussian noise.

$$y(x) = f(x) + e(x) \ ....e(x) \sim N(0, \sigma_{2D}) \tag{5-3}$$

The wavelet transform of $y(x)$ can be represented as:

$$W(y(x)) = W(f(x)) + W(e(x)) \tag{5-4}$$

1. When the signal is smooth, i.e. $y(x) = f(x)$, then the estimate computes the M.A.D. of the high frequency wavelet coefficients inherent to the image. This introduces a bias in the estimate.

2. This bias has an impact at lower values of σ_{2D} ($\sigma_{2D} = 0$ to 7). Mathematically, the contribution of second term in R.H.S. of equation (5-4) is negligible compared to the first term. So the estimator remains skewed towards the high frequency components inherent to the image. The lower variance values further establishes the negligible impact of noise to the image

3. As σ_{2D} ($\sigma_{2D} = 21$ to 120) increases, the contribution of the second term in the RHS of equation (5-4) starts dominating the first term i.e. more wavelet coefficients (at the finest scale) start demonstrating noisy behaviour. Now the estimator starts to capture the behaviour of noise present in the image. As a result, the bias starts to decrease.

4. At higher values of σ_{2D} the variance of the estimator starts to increase. This is because every simulation of noise has a strong (and different) impact on the image which leads to variation in the estimate.

5. Overall, the estimator is robust. We found that the bias of the estimator was always less than 1 from the true values.

5-2-2 Experiment 1.2 - Estimation of σ_a using the equation: $\sigma_a = \sigma_{2D}\sigma_a^{0,1}$

In this experiment, we evaluate the equation:

$$\sigma_a = \sigma_{2D}\sigma_a^{0,1} \tag{5-5}$$

where σ_a stands for standard deviation of wavelet coefficients at scale a, σ_{2D} refers to the standard deviation of the noise in the 2D data and $\sigma_a^{0,1}$ is the standard deviation of wavelet transform of zero mean white noise, with standard deviation equal to 1, at scale a.

The motivation of this experiment is to validate the performance of the equation so that parameter σ_{2D} (or $\hat{\sigma_{2D}}$) gets correctly translated to wavelet transform space. This translation varies depending on the wavelet function and the scale being used.

Methodology

For our experiment, we use generalized 2D mexican hat as our wavelet function. The scale parameters used in this experiment is given in Table (5-2). This choice is based on our observations regarding the size of peaks that are present in the data sample. The procedure for evaluating the equation is as follows:

1. Given: scale a, noise parameter $\sim N(0, \sigma_{2D})$
2. Simulate zero mean white noise with variance equal to σ_{2D}. The size of the 2D matrix, for the simulation, is chosen as 4096 x 4096. This is to mitigate any possible boundary effects.
3. Compute the wavelet transform of this simulation at scale a.
4. Compute the standard deviation of the transform coefficients. This will give $\tilde{\sigma_a}$ (we use a different notation for this term to differentiate between the value computed using simulation and the value computed using the equation).
5. Store the value
6. Compute σ_a by using the equation (5-5).
7. Compare the two values obtained in Step 6 and Step 4.
8. Repeat the steps for different values of a.
9. Repeat the experiment for different values of σ_{2D}.

a	$\sigma_y = 0.5a$	σ_x	Size of wavelet: floor$[10\sigma_x]$ X floor$[10\sigma_y]$
2.0	1.0	64	640 X 10
2.3784	1.1892	64	640 X 11
2.828	1.414	64	640 X 14
3.3635	1.6817	64	640 X 16
4.0	2.0	64	640 X 20
4.7568	2.3784	64	640 X 23
5.6568	2.828	64	640 X 28
6.7271	3.3635	64	640 X 33
8.0	4.0	64	640 X 40
9.5136	4.7568	64	640 X 47
11.3137	5.656	64	640 X 56
13.454	40	64	640 X 67
16.0	8.0	64	640 X 80

Table 5-2: Scale parameters. σ_x and σ_y correspond to width of the wavelet function along rows and columns respectively. The last column presents the size of the wavelet functions

Results

a	$\tilde{\sigma}_a$	$\sigma_a = \sigma_{2D}\sigma_a^{0,1}$
2.0	0.5485	0.5450
2.3784	0.5110	0.5084
2.828	0.4747	0.4728
3.3635	0.4402	0.4377
4.0	0.4077	0.4036
4.7568	0.3772	0.3714
5.6568	0.3483	0.3414
6.7271	0.3210	0.3139
8.0	0.2955	0.2890
9.5136	0.2718	0.2664
11.3137	0.2500	0.2458
13.454	0.2298	0.2264
16.0	0.2112	0.2081

Table 5-3: Results for experiment 1.2 with $\sigma_{2D} = 10$. a correspond to scale parameter of the wavelet function. $\tilde{\sigma}_a$ is the standard deviation of the wavelet coefficients corresponding to scale a computed using simulation. The last column is the standard deviation of the wavelet coefficients at scale a using equation (5-5).

a	$\tilde{\sigma}_a$	$\sigma_a = \sigma_{2D}\sigma_a^{0,1}$
2.0	2.711	2.7225
2.3784	2.527	2.538
2.828	2.3491	2.3590
3.3635	2.1758	2.1850
4.0	2.0102	2.0179
4.7568	1.8538	1.8586
5.6568	1.7075	1.7083
6.7271	1.5717	1.5679
8.0	1.4458	1.4390
9.5136	1.3290	1.3226
11.3137	1.2206	1.2176
13.454	1.1219	1.1207
16.0	1.0342	1.0293

Table 5-4: Results for experiment 1.2 with $\sigma_{2D} = 50$. a correspond to scale parameter of the wavelet function. $\tilde{\sigma}_a$ is the standard deviation of the wavelet coefficients corresponding to scale a computed using simulation. The last column is the standard deviation of the wavelet coefficients at scale a using the equation (5-5).

a	$\tilde{\sigma}_a$	$\sigma_a = \sigma_{2D}\sigma_a^{0,1}$
2.0	8.2318	8.1633
2.3784	7.6695	7.6039
2.828	7.1152	7.0619
3.3635	6.5737	6.5403
4.0	6.0568	6.0467
4.7568	5.5712	5.5848
5.6568	5.1167	5.1505
6.7271	4.6910	4.7383
8.0	4.2975	4.3512
9.5136	3.9411	3.9958
11.3137	3.6198	3.6717
13.454	3.3267	3.3747
16.0	3.0552	3.1035

Table 5-5: Results for experiment 1.2 with $\sigma_{2D} = 150$. a corresponds to scale parameter of the wavelet function. $\tilde{\sigma}_a$ is the standard deviation of the wavelet transform of noise corresponding to scale a computed using simulation. The last column is the standard deviation of the wavelet coefficients at scale a using the equation (5-5).

Discussion

1. The equation produces a robust estimate of σ_a. The values obtained by the equation were able to approximate the values obtained by the simulation for all scales a and for varying σ_{2D}.

5-3 Experiment 2 - Effective length based thresholding

The experiments in this section are related to effective length based thresholding. We evaluate the performance of the parameters (both estimated and user-driven) related to this quantity. See Section 4-5 for further details.

5-3-1 Experiment 2.1 : Evaluation of the equation $T_{a,local} = \sigma_{local} M_a$

For this experiment, we will evaluate the performance of the equation:

$$T_{a,local} = \sigma_{local} M_a \tag{5-6}$$

Here, $T_{a,local}$ is the threshold level at scale a. This threshold level is defined as the maximum local maxima that is generated due to wavelet transform of local noise (with the assumption that the local noise has a gaussian distribution with parameters $\sim N(0, \sigma_{local})$ at scale a. The parameter M_a is the average of maximum local maxima that is generated by the wavelet transform of zero mean white noise with variance = 1 at scale a (number of simulations used for averaging = 50). Equation (5-6) is similar to equation (5-5) except that in this case, M_a is obtained by averaging a number of simulations to get the desired quantity. This averaging is motivated as the maximum local maxima generated due to wavelet transform of noise is not a stable statistic i.e. different simulations of wavelet transform of noise (with same distributions) can result in different values of maximum local maxima.

Methodology

In order to evaluate the performance of the given equation, we simulate noise with different noise parameters and assess how the equation performs in these scenarios.

For the given experiment, we use generalized 2D mexican hat function as our wavelet function. The scale parameters used in this experiment is given in Table (5-2). The procedure for evaluation is as follows:

1. Given: scale a, local noise parameter $\sim N(0, \sigma_{local})$
2. For B = 1:100 :
   (a) Simulate zero mean white noise with variance equal to σ_{local}. The size of the 2D matrix, for the simulation, is chosen as 2048 x 2048. This is to mitigate the boundary effects. Note that this size is different from 4096 x 4096 which was used in experiment 1.2. This is to reduce the computation cost of the experiment (as we are running it 100 times). A rule of thumb is that as long as the size of simulation is greater than twice the size of the largest wavelet function (in this it will be 640 x 80), it can be used for simulation.
   (b) Compute the wavelet transform of this simulation for scale a.
   (c) Identify all the local maxima wavelet coefficients and select the one with the maximum value.

(d) Store the selected coefficient.

3. Average the selected coefficients. This will give the average maximum local maxima generated due to wavelet transform of noise at scale a. We will refer to this quantity as $\tilde{T}_{a,local}$ (we use a different notation for this term to differentiate between the value computed using simulation and the value computed using the equation).

4. Compute $T_{a,local}$ using the equation (5-6).

5. Compare the values obtained in Step 3 and Step 4.

6. Repeat the steps for different values of a.

7. Repeat the experiment for different values of σ_{local}.

Results

a	$\tilde{T}_{a,local}$	$T_{a,local} = \sigma_{local} M_a$
2.0	7.1775	7.1864
2.3784	6.7035	6.5958
2.828	6.2405	6.2222
3.3635	5.7437	5.7354
4.0	5.2829	5.3081
4.7568	4.8117	4.9500
5.6568	4.4353	4.4292
6.7271	4.0562	3.9378
8.0	3.6405	3.6492
9.5136	3.2777	3.2271
11.3137	3.0061	2.8673
13.454	2.6731	2.7235
16.0	2.4091	2.3211

Table 5-6: Results for experiment 2.1 with noise parameter $\sim \sigma_{local} = 10$. a corresponds to scale parameter of the wavelet function. $\tilde{T}_{a,local}$ is the threshold level which is defined average maximum local maxima of the wavelet transform of noise corresponding to scale a (obtained using simulation). $T_{a,local}$ is the threshold level obtained using equation (5-6).

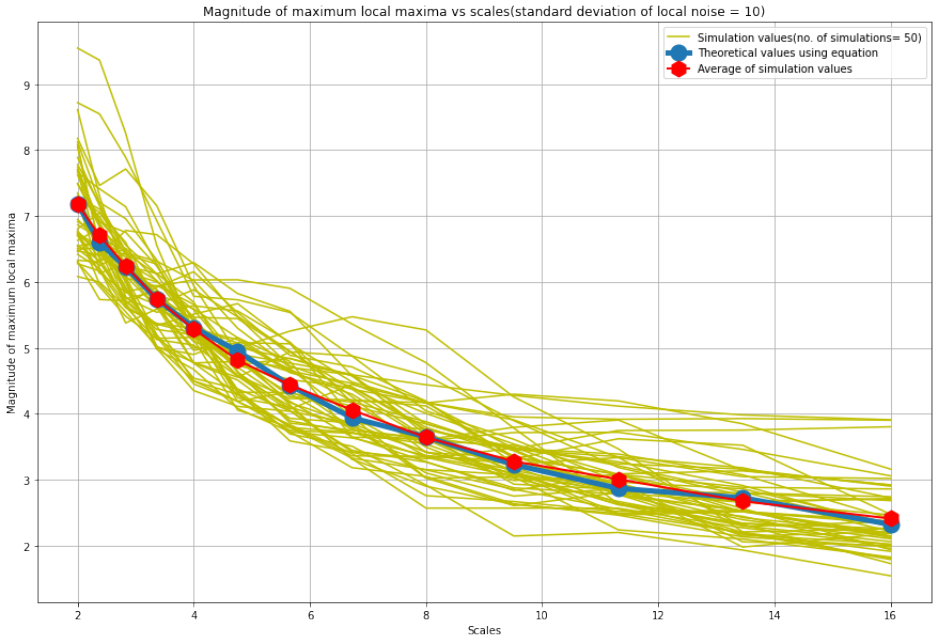


Figure 5-4: Results for experiment 2.1: Magnitude of maximum local maxima coefficients at different scales. Noise parameters ∼(0,10). The yellow lines show the maximum local maxima obtained at different scale a for 50 simulations. The red line is the average of the values obtained using simulation for different scales a ($\tilde{T}_{a,local}$). The blue line represents the theoretical value $T_{a,local}$ obtained using equation (5-6).

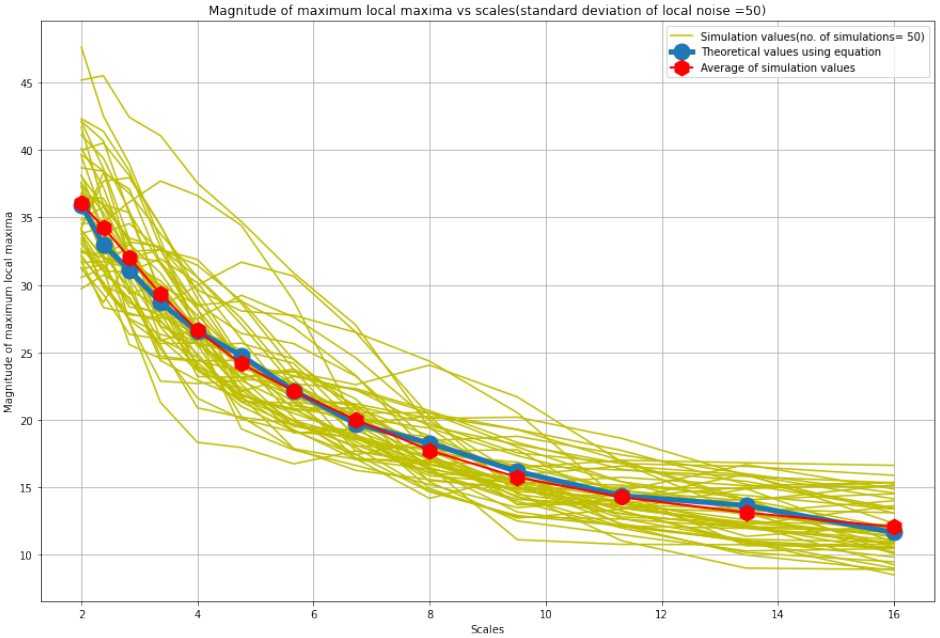


Figure 5-5: Results for experiment 2.1: Magnitude of maximum local maxima coefficients at different scales. Noise parameters ∼(0,50). The yellow lines show the maximum local maxima obtained at different scale a for 50 simulations. The red line is the average of the values obtained using simulation for different scales a ($\tilde{T}_{a,local}$). The blue line represents the theoretical value $T_{a,local}$ obtained using equation (5-6).

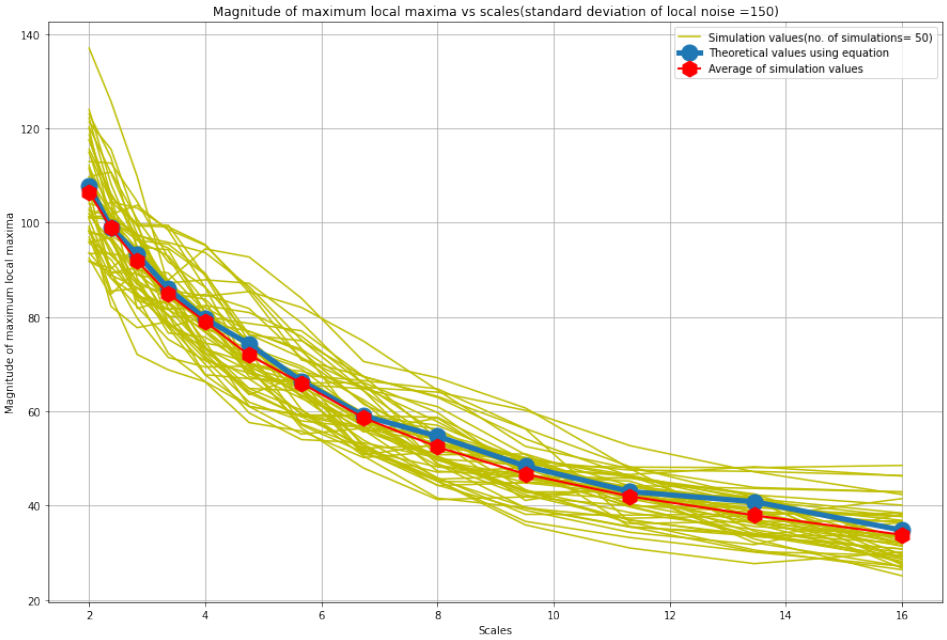


Figure 5-6: Results for experiment 2.1: Magnitude of maximum local maxima coefficients at different scales. Noise parameters ∼(0,150). The yellow lines show the maximum local maxima obtained at different scale a for 50 simulations. The red line is the average of the values obtained using simulation for different scales a ($\tilde{T}_{a,local}$). The blue line represents the theoretical value $T_{a,local}$ obtained using equation (5-6).

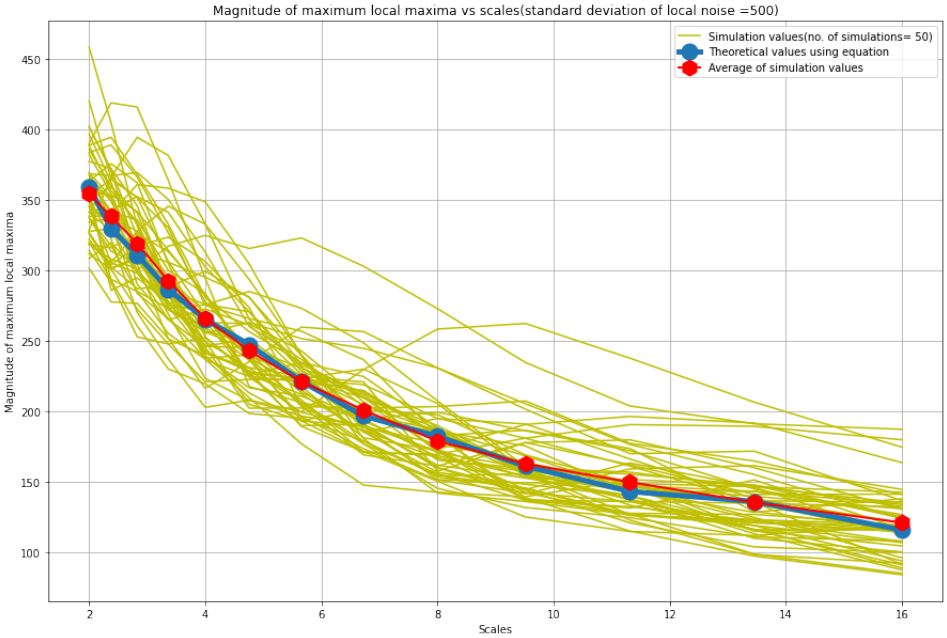


Figure 5-7: Experiment 2.1: Magnitude of maximum local maxima coefficients at different scales. Noise parameters ∼(0,500). The yellow lines show the maximum local maxima obtained at different scale a for 50 simulations. The red line is the average of the values obtained using simulation for different scales a ($\tilde{T}_{a,local}$). The blue line represents the theoretical value $T_{a,local}$ obtained using equation (5-6).

a	$\tilde{T}_{a,local}$	$T_{a,local} = \sigma_{local} M_a$
2.0	36.0283	35.9323
2.3784	34.2775	32.9790
2.828	32.0075	31.1113
3.3635	29.3288	28.6773
4.0	26.6104	26.5408
4.7568	24.1344	24.7503
5.6568	22.1072	22.1463
6.7271	19.9878	19.6893
8.0	17.7061	18.2461
9.5136	15.7023	16.1355
11.3137	14.2217	14.3368
13.454	13.0898	13.6175
16.0	12.0285	11.6055

Table 5-7: Results for experiment 2.1 with noise parameter $\sim \sigma_{local} = 50$. a corresponds to scale parameter of the wavelet function. $\tilde{T}_{a,local}$ is the threshold level which is defined average maximum local maxima of the wavelet transform of noise corresponding to scale a (obtained using simulation). $T_{a,local}$ is the threshold level obtained using equation (5-6).

a	$\tilde{T}_{a,local}$	$T_{a,local} = \sigma_{local} M_a$
2.0	106.3778	107.7971
2.3784	98.8475	98.9372
2.828	91.8654	93.3341
3.3635	84.9134	86.0321
4.0	79.0019	79.6225
4.7568	71.9171	74.2510
5.6568	65.9728	66.4390
6.7271	58.6215	59.0680
8.0	52.4662	54.7384
9.5136	46.6948	48.4065
11.3137	42.0021	43.0104
13.454	37.9815	40.8526
16.0	33.8542	34.8165

Table 5-8: Results for experiment 2.1 with noise parameter $\sim \sigma_{local} = 150$. a corresponds to scale parameter of the wavelet function. $\tilde{T}_{a,local}$ is the threshold level which is defined average maximum local maxima of the wavelet transform of noise corresponding to scale a (obtained using simulation). $T_{a,local}$ is the threshold level obtained using equation (5-6).

Discussion

1. The overall decreasing nature of $T_{a,local}$ is due to the normalization factor (L1 normalization) being used.

2. The deviation between $\tilde{T}_{local,a}$ and $T_{a,local}$ starts to increase for higher values of σ_{local}.

a	$\tilde{T}_{a,local}$	$T_{a,local} = \sigma_{local} M_a$
2.0	354.3785	359.3238
2.3784	338.4224	329.7908
2.828	319.3189	311.1138
3.3635	292.8059	286.7736
4.0	266.1218	265.4085
4.7568	243.1156	247.5034
5.6568	221.2568	221.4633
6.7271	201.1383	196.8933
8.0	179.0033	182.4616
9.5136	163.0599	161.3552
11.3137	149.9761	143.3682
13.454	135.6627	136.1754
16.0	121.2517	116.0551

Table 5-9: Results for experiment 2.1 with noise parameter $\sim \sigma_{local} = 500$. a corresponds to scale parameter of the wavelet function. $\tilde{T}_{a,local}$ is the threshold level which is defined average maximum local maxima of the wavelet transform of noise corresponding to scale a (obtained using simulation). $T_{a,local}$ is the threshold level obtained using equation (5-6).

This might be due to the strong variations introduced by higher values of σ_{local}.

3. Overall, the equation produces a fair estimate for the threshold level parameter $_{local,a}$. However, the deviations start to increase with increase in the value σ_{local}. Further testing is required to assess the validity of the equation.

5-3-2 Experiment 2.2: Evaluating the performance of the algorithm by varying the effective threshold length on the synthetic IM-IMS data sample

Effective length represents the number of scales at which the wavelet local maxima coefficients (belonging to a chain) is greater than the threshold level. In our implementation of the algorithm, we take user defined input as the threshold for effective length. Chains lower than this threshold are considered to originate due to noise and are rejected.

In this experiment, we will vary the effective length threshold parameter and analyze the peaks that get detected in the noisy synthetic IM-IMS data sample (Figure (5-2b)).

Methodology

For the given experiment, we use generalized 2D mexican hat function as the wavelet function. The scale parameters used in this experiment is given in (5-2). The various effective length threshold parameters that will be used in this experiment are: 2,4,6,8,10,12.

Chains which have effective length greater than the threshold effective length are considered to originate due to chemical peaks. The spatial location of the finest scale wavelet coefficient (belonging to a chain) determines the position of the peak. We say that a chemical peak is detected if this location is at a distance of ≤ 0.01 Da in the m/z dimension and less than

100 mobility bins (rows) from the true chemical peak. Else, it will be considered as a false positive.

The performance of the parameter (and the algorithm) will be evaluated using the F-Score metric. F-score is defined as:

$$F - score = \frac{2.R.P}{R + P} \tag{5-7}$$

where $R = \frac{TP}{NP}$ and $P = \frac{TP}{N}$. TP is the number of true positives, NP represents the total number of true chemical peaks in the data sample and N is total number of peaks detected by the algorithm. A perfect feature detection will achieve a F-Score of 100% and the presence of false positives and false negatives will lower its values. In order to analyse the results, the confusion matrix associated with the detected peaks will also be studied.

Results

Results for varying effective length threshold is presented in Figure (5-8) and Figure (5-9). The confusion matrix associated with the various effective length threshold is presented in Table (5-10). The F-score plot is shown in Figure (5-10).

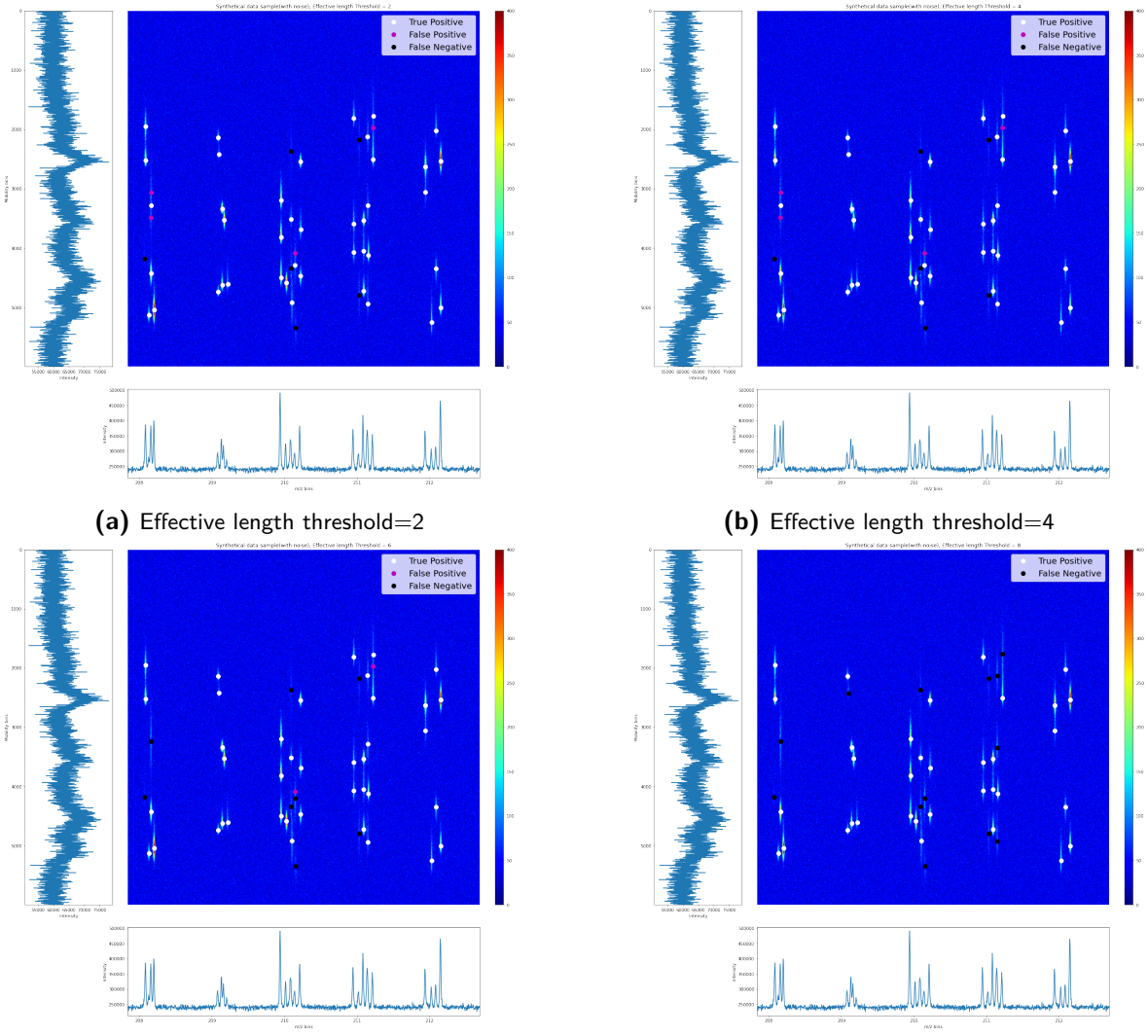
Effective length	T.P.	F.P.	F.N.	T.N.
2	42	4	6	0
4	42	4	6	0
6	40	2	8	0
8	35	0	13	0
10	33	0	15	0
12	29	0	19	0

Table 5-10: Results for Experiment 2.2: Peak detection in the noisy synthetic IM-IMS data sample with varying effective length threshold parameter (2 to 12). The table present the confusion matrix associated with the different effective length threshold parameter. T.P, F.P, T.N. and F.N. stand for True Positive, False Positive, True Negative and False Negative respectively.

Observations

Before we begin the discussion regarding the performance of the algorithm, we present our definition of SNR for characterization of peaks. As we are only concerned with detection of peaks (and not peak widths), we define SNR as:

$$S.N.R = \frac{I_{max.peak}}{\hat{\sigma}_{2D}} \tag{5-8}$$



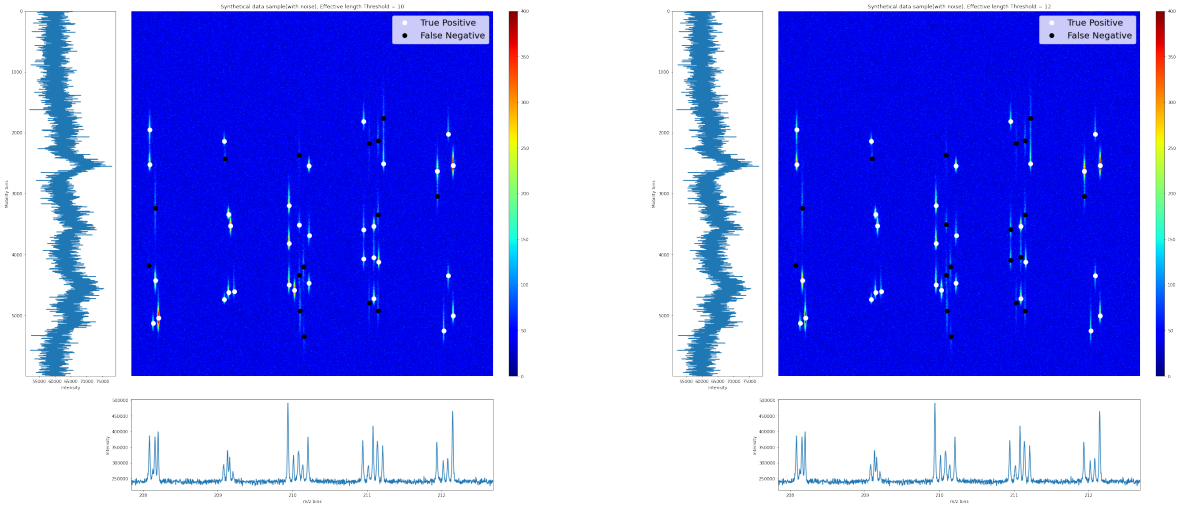
(a) Effective length threshold=2

(b) Effective length threshold=4

(c) Effective length threshold =6

(d) Effective length threshold=8

Figure 5-8: Results for Experiment 2.2: Peak detection in the noisy synthetic IM-IMS data sample with varying effective length threshold parameter (2 to 8). Total number of true peaks in the data sample = 48. In every sub-figure, the white, black and magenta dots represent the true positives, false negatives and false positives detected by the algorithm respectively.



(a) Effective length threshold=10 **(b)** Effective length threshold=12

Figure 5-9: Results for Experiment 2.2: Peak detection in the noisy synthetic IM-IMS data sample with varying effective length threshold parameter (10 to 12). Total number of true peaks in the data sample = 48. In every sub-figure, the white, black and magenta dots represent the true positives, false negatives and false positives detected by the algorithm respectively.

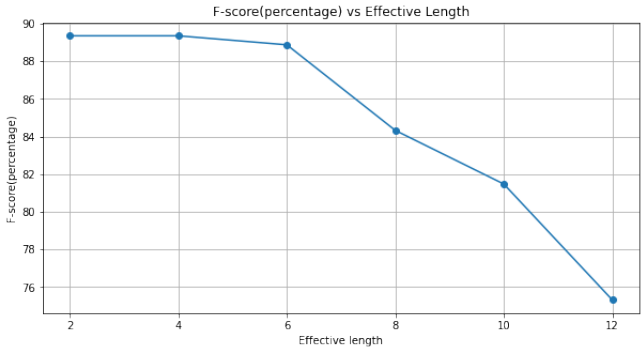


Figure 5-10: Results for Experiment 2.2: Peak detection in the noisy synthetic IM-IMS data sample with varying effective length threshold parameter. F-score(%) plot.

Here, $I_{max.peak}$ refers to maximum intensity value of a peak in the smooth data sample and $\hat{\sigma}_{2D}$ is the estimated noise in the data sample. Here we use $\hat{\sigma}_{2D}$ because the behaviour of added noise is known and uniform across the sample. However, in real world IM-IMS data sample $\hat{\sigma}_{local}$ will be a better parameter for computing SNR.

1. The general behaviour of the parameter is as follows:
   (a) Having a higher threshold value (8 to 12) implies that more number of local maxima coefficients (belonging to a chain) should be greater than the surrounding local noise level (in the transform space). This leads to discovery of only strong prominent peaks. Also, we observe that having a higher effective length threshold value decreases the number of false positives detected by the algorithm (in this case, it is zero).
   (b) Lowering the threshold value leads to discovery of more true positives but it also increases the number of false positives detected (Table (5-10)).
   (c) At threshold value = 2, we see that there are multiple false positive peaks getting detected near the true peak.
2. Despite setting the threshold parameter to it's lowest value, 6 true peaks were still not detected.
3. The F-score plot (Figure (5-10)) shows that, for a given set of scales, effective length of 4 is the optimal threshold value for detection of peaks in the synthetic IM-IMS data sample.

Discussion

1. If we look at the histograms given in Figure (5-11a) and Figure (5-11b), we see that while almost all of the true positive peaks that are detected by the algorithm form chains with length 13, their effective length is randomly distributed varying from 4 to 13.
2. Figure (5-12a) shows the peaks that are detected by the algorithm whose wavelet chains have an effective length of less than 6. We see that there are 3 false positives and 2 true positives that are detected with below this threshold effective length.The chains corresponding to these true positives usually gain their strength at higher scales (Figure (5-12b)).
3. Considering the redundancy of the scales (three voices per octave), the wavelet coefficients (in a chain) belonging to true positives tend to behave smoothly. This means that if they have gained their strength at a particular scale, their coefficients tend to remain stronger than the threshold level for a couple of scales. This behaviour can be used to evaluate if the constructed chain is correct. This is also useful in selecting the threshold value for effective length.
4. Using these observations, we found that an effective length of 4 is the optimal threshold for the given set of scales. This threshold length guarantees that chains have minimum strength to be considered as a potential true positive peak.

(a) Length **(b)** Effective length

Figure 5-11: Experiment 2.2: Evaluating the performance of the algorithm by varying the effective threshold length on a synthetic IM-IMS data sample. Histogram plots of (a) length (b) effective length corresponding to wavelet chains associated with true positives.

5. The false positives that are detected have a pattern. They usually lie on the surface of the true positive. This makes their wavelet coefficients greater than the threshold level at higher scales (Figure (5-12a)).

6. The reason for detection of these false positives is based on the size of σ_x which is fixed. Having a fixed σ_x, will lead to generation of false positives that lie on the surface of the true positives if σ_x is significantly less than the width of the true peak (along rows).

7. A brief characterization of the undetected peaks using the definition of SNR is presented in Figure (5-13). The main reasons for these undetected peaks would be (i) their maximum intensity values are comparable to the noise in the data sample and (ii) fixed value of σ_x which leads to suboptimal representation (in terms of wavelet coefficients' magnitude) in the transform space.

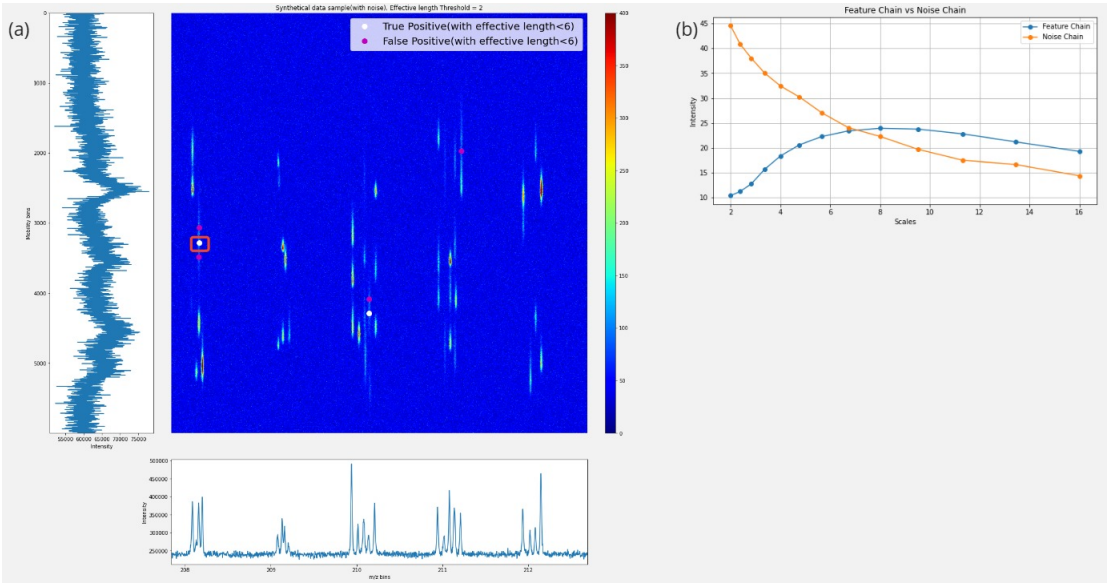


Figure 5-12: Experiment 2.2: Evaluating the performance of the algorithm by varying the effective threshold length on a synthetic IM-IMS data sample. (a) Peaks corresponding to chains that have an effective length of less than 6 and greater than 2. (b) Wavelet chain corresponding to the true positive marked by a red square. The blue line represents the wavelet coefficients of the local maxima connected in a chain and the orange line represents the threshold value generated by the local noise at every scale.

5-3-3 Experiment 2.3: Evaluating the performance of the algorithm by varying the effective threshold length on a real world IM-IMS data sample

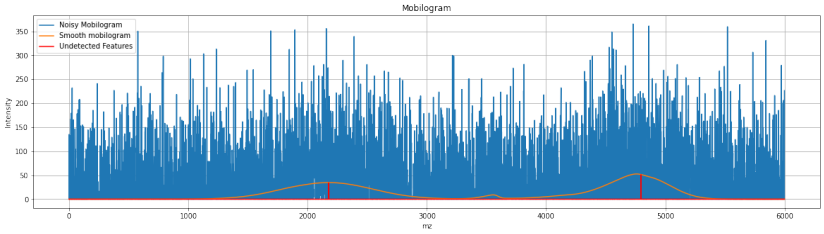
In this experiment, we will vary the effective length threshold parameter and analyse the peaks that get detected in the real world IM-IMS data sample.

Methodology

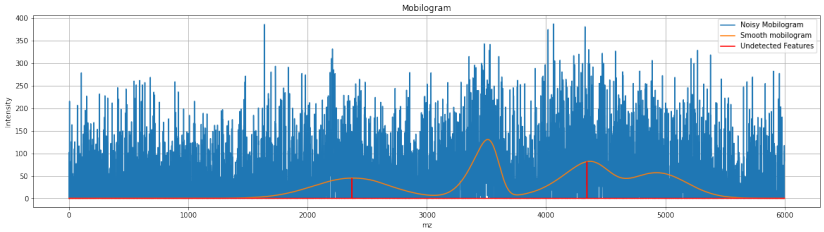
For the given experiment, we use generalized 2D mexican hat function as the wavelet function. The scale parameters used in this experiment is given in Table (5-2). The various effective length threshold parameters used in this experiment are: 2,4,6,8,10,12.

The partitioned data sections used in this experiment is briefly discussed in Section (5-1-1). Figure (5-1) presents the visual overview of the 2D data sections that will be used in this experiment.

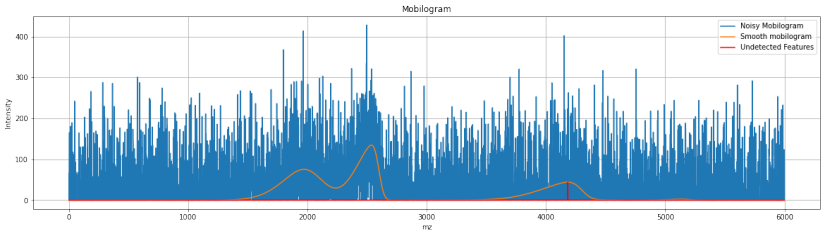
As we do not know the ground truth for this data sample, we will assess the performance of the parameter based on the peaks being detected.



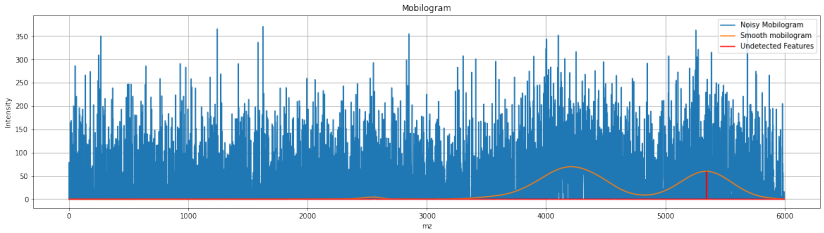
(a) Maximum peak intensity(from left to right) = 34 and 51 $\hat{\sigma}_{2D} = 52.6635$ SNR (from left to right) = 0.6538 and 0.9807



(b) Maximum peak intensity(from left to right) = 44 and 81 $\hat{\sigma}_{2D} = 52.6635$ SNR (from left to right) = 0.846 and 1.556



(c) Maximum peak intensity = 44 $\hat{\sigma}_{2D} = 52.6635$ SNR = 0.846



(d) Max peak intensity= 59 $\hat{\sigma}_{2D} = 52.6635$ SNR = 1.1346

Figure 5-13: Experiment 2.2: Evaluating the performance of the algorithm by varying the effective threshold length on a synthetic IM-IMS data sample. Mobilograms (individual columns of the data matrix) of undetected peaks. These peaks were not detected at the lowest possible value for effective length threshold. In every sub-figure, the blue line represents the noisy mobilogram, the orange line represents the smooth mobilogram and the red line marks the location of the undetected peaks. We calculate the SNR value for undetected peaks using equation (5-8).

Results

Results for varying effective length threshold for different partitioned sections of the data sample are presented in:

1. 670-672 m/z - Figure (5-14)
2. 704-706 m/z - Figure (5-15)
3. 770-772 m/z - Figure (5-16)
4. 920-922 m/z - Figure (5-17)

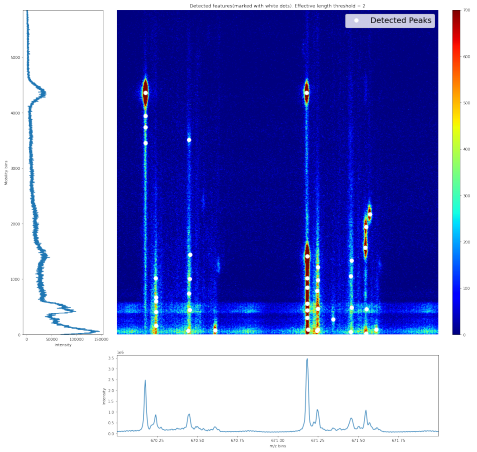
Table (5-11) presents the total number of peaks detected in the various test sections for varying effective length threshold.

Test Section	E.L = 2	E.L. = 4	E.L = 6	E.L.=8	E.L = 10	E.L = 12
670-672 m/z	40	32	29	25	18	11
704-706 m/z	79	61	48	40	29	25
770-772 m/z	97	69	54	41	36	27
920-922 m/z	83	58	37	30	28	24

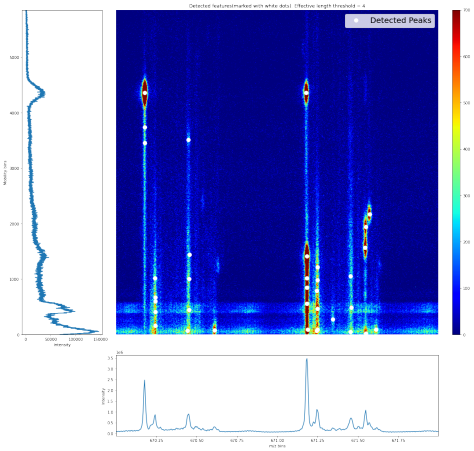
Table 5-11: Results for Experiment 2.3: Peak detection in the real world IM-IMS data sample with varying effective length threshold parameter (2 to 12). The table presents the total number of peaks detected for varying effective length threshold.

Observations

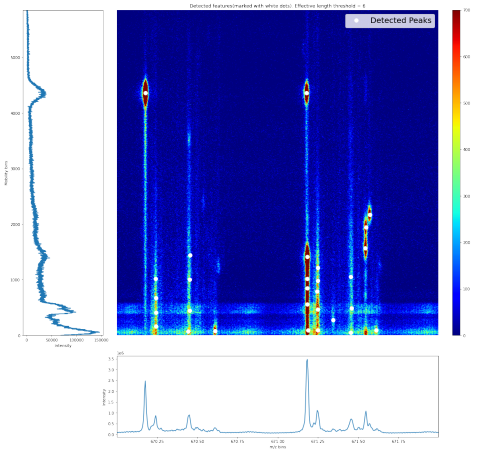
1. The general performance of the algorithm is fair across all test sections.
2. Having a high threshold value leads to detection of less number of prominent peaks and having a low threshold value leads to detection of high number of false positives.
3. In region 704-706 m/z and 770-772 m/z, that the algorithm is potentially detecting a high number of potential false positives are detected horizontally (along rows) at lower threshold levels (Figure (5-15b)).
4. In all the test regions, the algorithm detects high number of potential false positives vertically along mobility dimension.
5. In region 920-922 m/z, most of the potential true positives are only detected at lower threshold values.



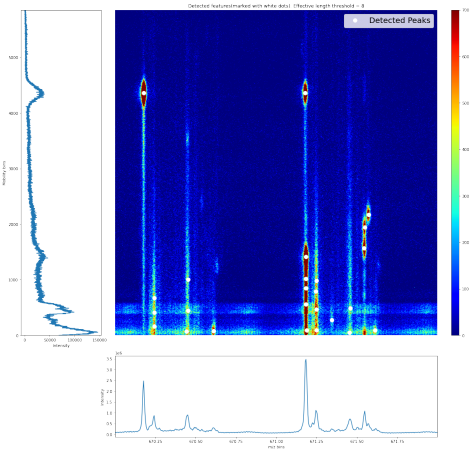
(a) Effective length threshold=2.



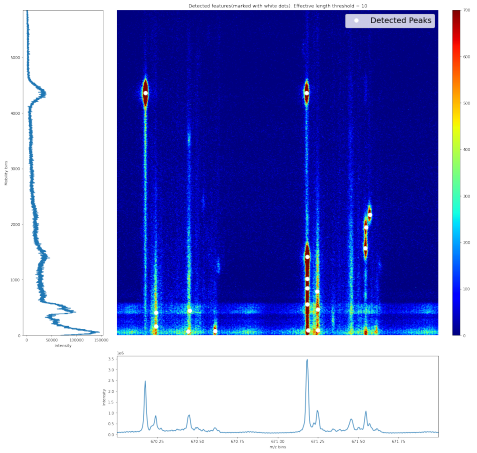
(b) Effective length threshold=4.



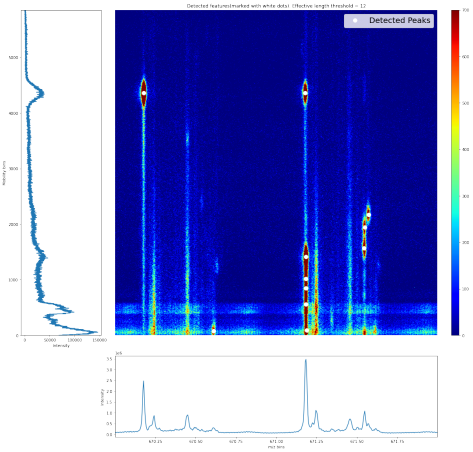
(c) Effective length threshold = 6.



(d) Effective length threshold=8.

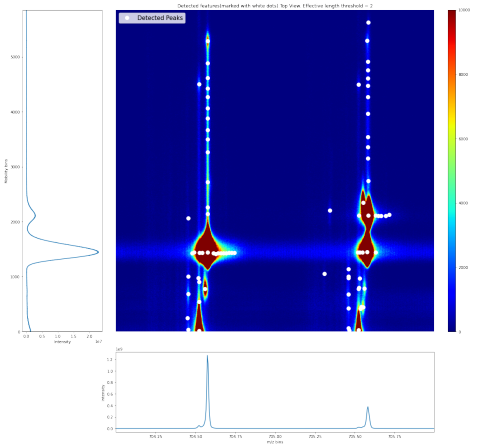


(e) Effective length threshold=10.

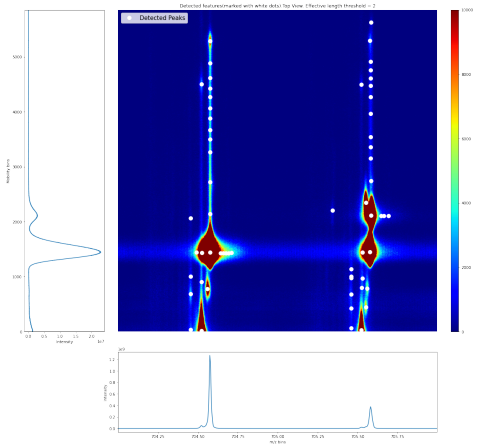


(f) Effective length threshold=12.

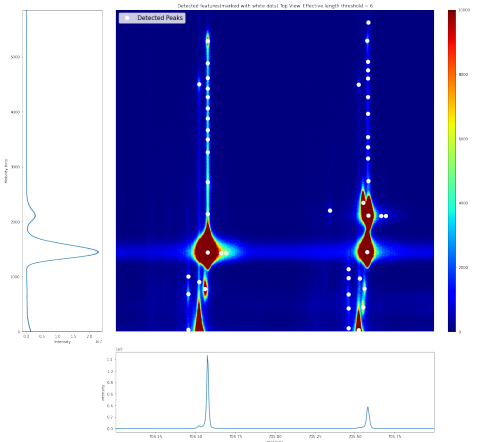
Figure 5-14: Results for experiment 2.3: Evaluating the performance of the algorithm by varying the effective threshold length on a real world IM-IMS data sample. Test section : 670-672 m/z with complete mobility information. In every sub-figure, the white dots mark the peaks detected by the algorithm.



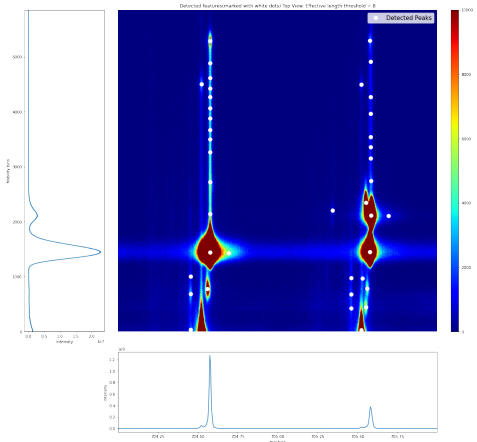
(a) Effective length threshold=2.



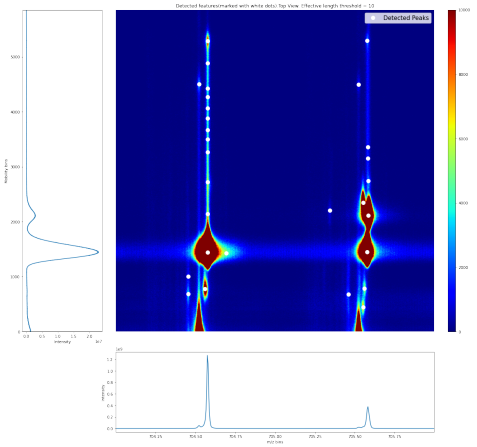
(b) Effective length threshold=4.



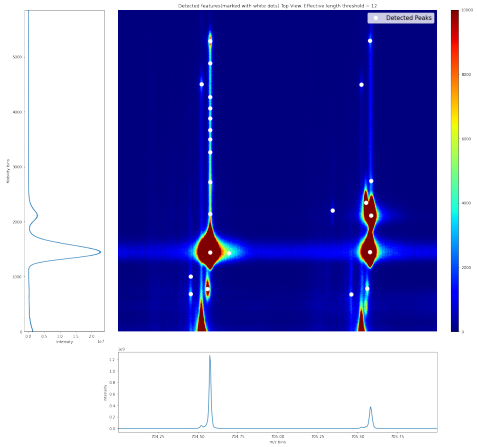
(c) Effective length threshold = 6.



(d) Effective length threshold=8.

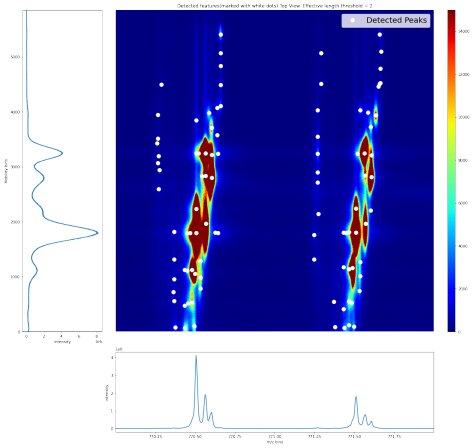


(e) Effective length threshold=10.

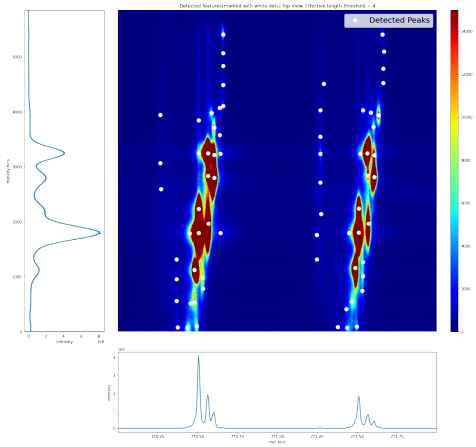


(f) Effective length threshold=12.

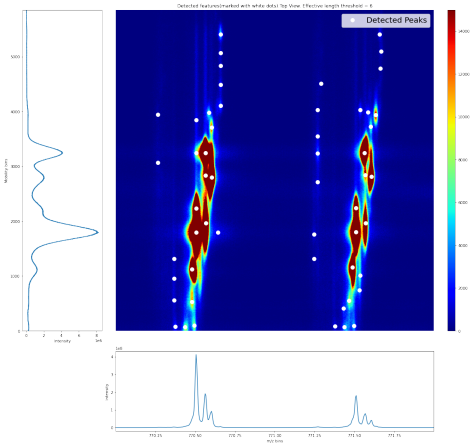
Figure 5-15: Results for Experiment 2.3: Evaluating the performance of the algorithm by varying the effective threshold length on a real world IM-IMS data sample. Test section : 704-706 m/z with complete mobility information. In every sub-figure, the white dots mark the peaks detected by the algorithm.



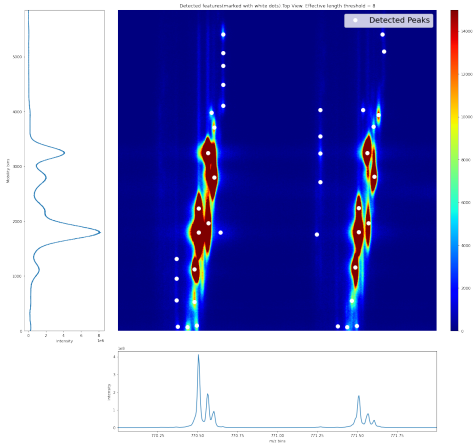
(a) Effective length threshold=2.



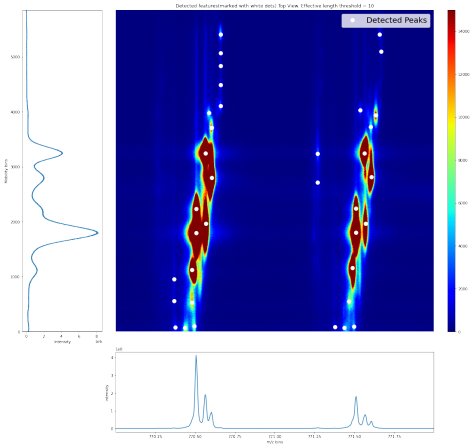
(b) Effective length threshold=4.



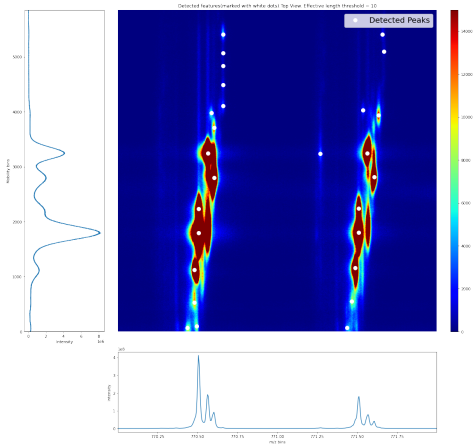
(c) Effective length threshold = 6.



(d) Effective length threshold=8.

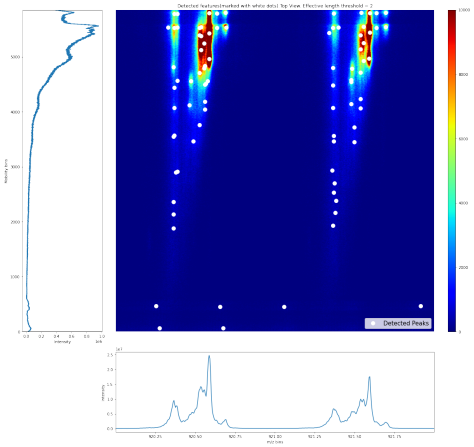


(e) Effective length threshold=10.

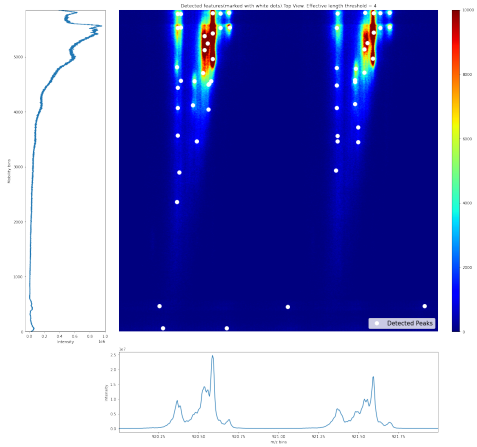


(f) Effective length threshold=12.

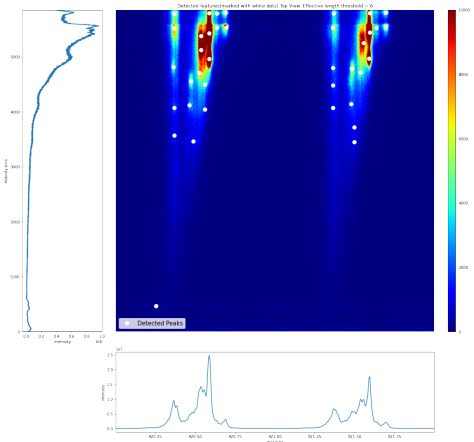
Figure 5-16: Results for Experiment 2.3: Evaluating the performance of the algorithm by varying the effective threshold length on a real world IM-IMS data sample. Test section : 770-772 m/z with complete mobility information. In every sub-figure, the white dots mark the peaks detected by the algorithm.



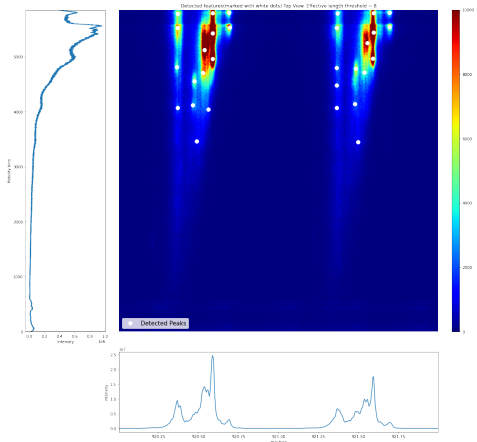
(a) Effective length threshold=2.



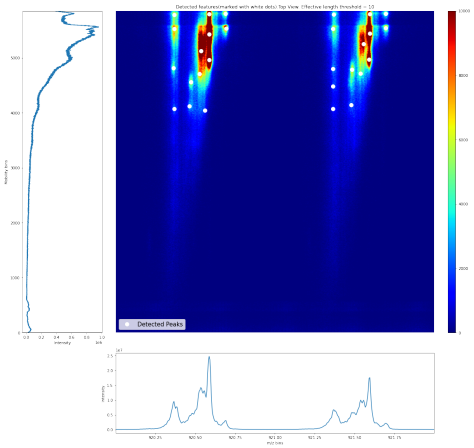
(b) Effective length threshold=4.



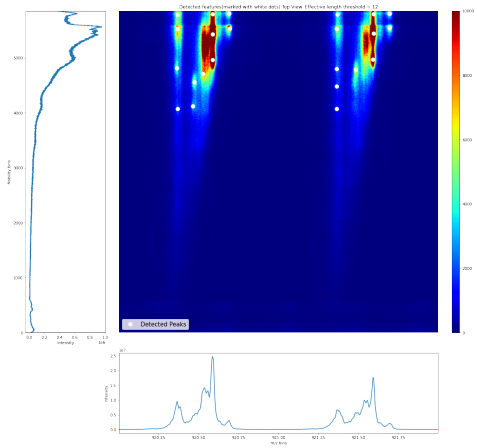
(c) Effective length threshold = 6.



(d) Effective length threshold=8.



(e) Effective length threshold=10.



(f) Effective length threshold=12.

Figure 5-17: Results for experiment 2.3: Evaluating the performance of the algorithm by varying the effective threshold length on a real world IM-IMS data sample. Test section : 920-922 m/z with complete mobility information. In every sub-figure, the white dots mark the peaks detected by the algorithm.

Discussion

1. High number of potential false positives are getting detected across various dominant mobility spectra. As the mobility peaks are noisy in nature and have extended tails, there are regions where the observed SNR of the region is high, but there are no prominent peaks. Having a fixed value of σ_x increases (or decreases) the number of peaks detected along a single mobility spectrum. Figure (5-19a) and Figure (5-18) provides an example of this problem. Analysis of wavelet chains (Figure (5-20)) corresponding to these peaks show that their chains have high wavelet coefficient values than the corresponding threshold values (noise chains) at different scales. In order to remove these peaks, we would need to vary σ_x (wavelet width along rows) as a scale parameter a'. However, our current design decision does not allow this. As a result, for the choice of σ_x in this experiment, we detect high number of potential false positives along dominant mobility spectra.

2. In the test section 704-706 m/z, there are multiple peaks that are detected horizontally (along rows) near the main dominant peak (Figure (5-19b)). This is because the peaks in this region are broad and have high SNR. As the data is noisy, this generates multiple wavelet local maxima coefficients with large coefficient values in the the transform space. As a result, we see multiple potential false positives getting detected near the main dominant peak. However, an interesting pattern to note is that wavelet chains corresponding to these type of local maxima points have a negative slope. (Figure (5-21)). This is because as scale a increases the wavelet function expands along columns and so the wavelet coefficients introduced due to noise loses its strength to the dominant peak around it. As a result, their chains have a decreasing slope.

3. In the test section 920-922 m/z, most of the potential true positives are detected at lower threshold values. This is because the region has narrow m/z peaks. As the size of wavelet function(σ_y or a) increases, the wavelet local maxima coefficients associated with these narrow peaks starts to merge with dominant peaks in the transform space. As a result, the chains corresponding to these narrow peaks have short length. Figure (5-22) presents a magnified region of the test section 920-922 m/z and manually annotated true positives. Figure (5-23) presents the peaks detected by the algorithm for different effective length thresholds. We see that at threshold length = 6, one of the manually marked peak is not detected by the algorithm.

   In order to detect these peaks, we need to keep a low value for effective length threshold. But lower effective length leads to detection of high number of false positives. As a result, there are high number of peaks getting detected in this region.

4. Based on all of the above observations, for the given set of scales, we found that effective length of 4 would be an optimal choice for peak detection. However, more tests are required to check whether should this parameter be treated as a global parameter for all clusters (group of wavelet local maxima points in wavelet transform space) or this parameter should vary for different clusters.

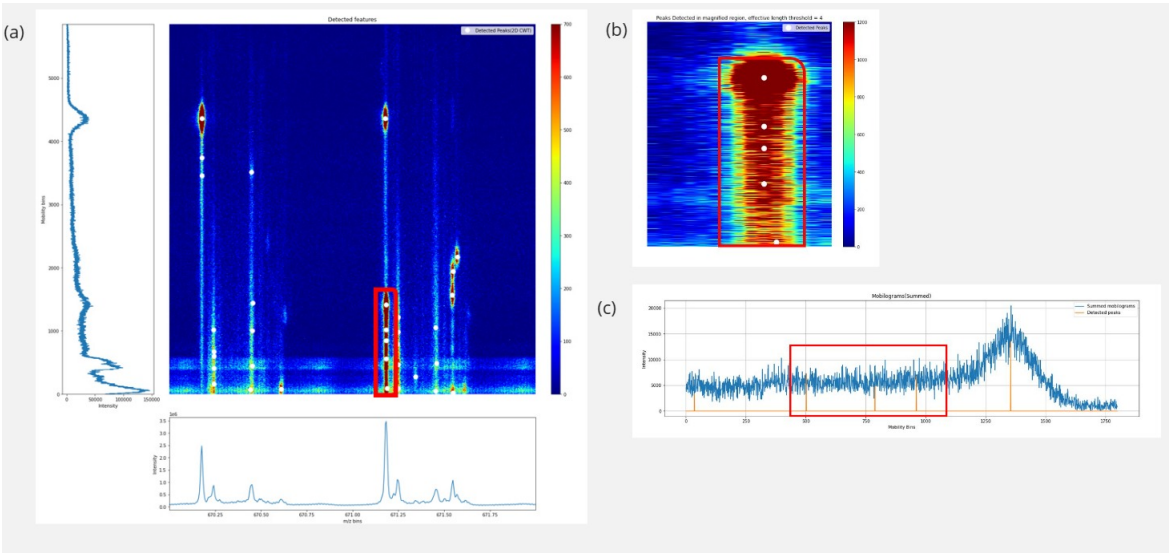
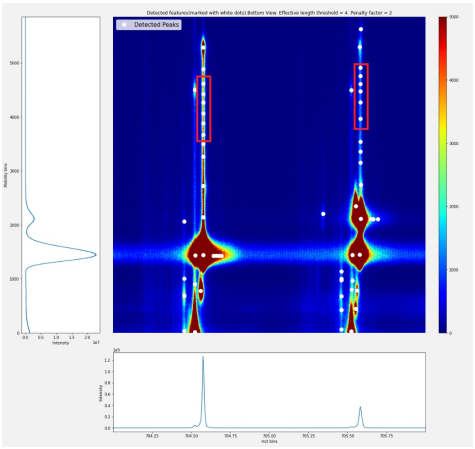
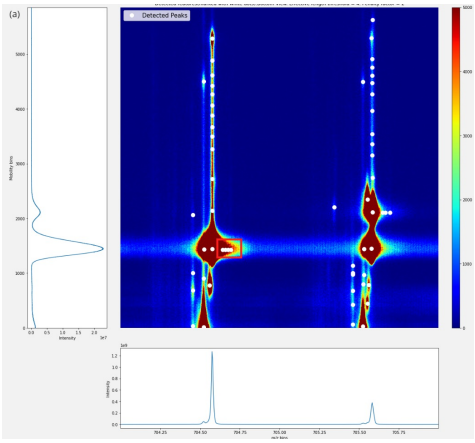


Figure 5-18: Experiment 2.3: Evaluating the performance of the algorithm by varying the effective threshold length on a real world IM-IMS data sample. Test section : 670-672 m/z with complete mobility information (a) Visual representation of the text section. White dots mark the peaks detected by the algorithm. The red box highlights the region that will be magnified for analysis (b) Magnified region. The red box marks the columns that will be summed to obtain a 1D representation of the 2D signal. (c) Summed mobilograms. Orange lines represent the peaks detected by the algorithm. The red box indicates a region where the SNR is relatively high but there is no peak like structure. As the width of the wavelet function is less than the width of the region, multiple potential false positives start getting detected.



(a) Analysis Region 1 - Potential false positives in the mobility spectrum.



(b) Analysis Region 2 - Potential false positives

Figure 5-19: Experiment 2.3: Evaluating the performance of the algorithm by varying the effective threshold length on a real world IM-IMS data sample. Test section : 704-706 m/z with complete mobility information. In every sub-figure, the red boxes mark some of the potential false positives (exhibiting a pattern) that are detected throughout the test sections.

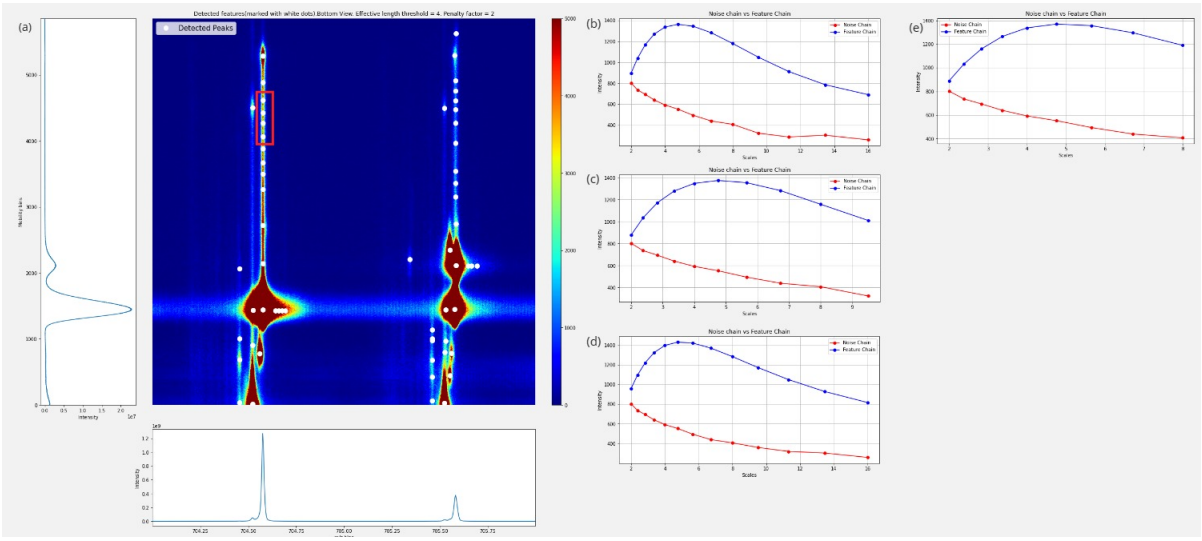


Figure 5-20: Experiment 2.3: Evaluating the performance of the algorithm by varying the effective threshold length on a real world IM-IMS data sample. Test section : 704-706 m/z with complete mobility information. (a) Visual representation of the test section. White dots mark the peaks detected by the algorithm. Red box correspond to some of the potential false positives that are detected by the algorithm. (b)-(e) Wavelet chain corresponding to the potential false positives marked by the red box. In every plot, the blue line represents the wavelet local maxima coefficients connected in a chain and the orange line represents the threshold value generated by the local noise at every scale.

5-3-4 Experiment 2.4: Evaluating the performance of the algorithm by varying the penalty factor on a real world IM-IMS data sample

In this experiment, we vary the penalty factor and analyze the peaks that are detected in the real world IM-IMS data sample.

The penalty factor controls the local noise level around the wavelet local maxima point. This level will be used for generating the threshold value for every scale a. In the current implementation of the algorithm, we use this penalty factor specifically for the row corresponding to the wavelet local maxima point. The rule is given as:

$$\text{If } \sigma_{m/z} > \sigma_{mobility},\ \sigma_{local} = 2\sigma_{m/z} \text{ else } \sigma_{local} = \sigma_{mobility}.$$

where $\sigma_{m/z}$ is the noise parameter corresponding to the row of local wavelet maxima point and $\sigma_{mobility}$ is the noise parameter corresponding to the column of the local wavelet maxima point.

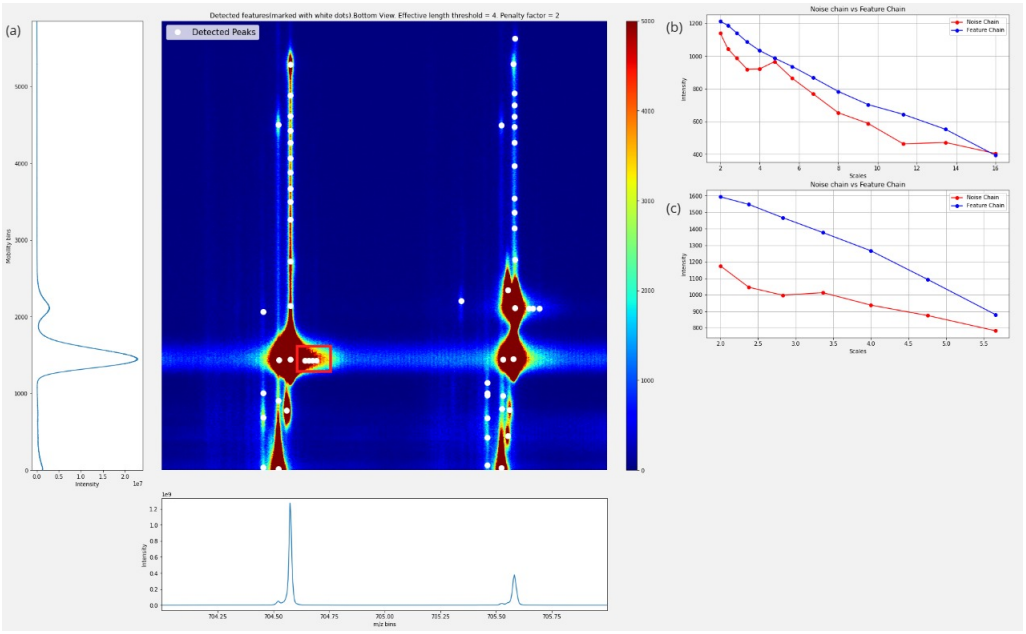


Figure 5-21: Experiment 2.3: Evaluating the performance of the algorithm by varying the effective threshold length on a real world IM-IMS data sample. Test section : 704-706 m/z with complete mobility information. (a) Visual representation of the test section. White dots mark the peaks detected by the algorithm. Red box correspond to potential false positives that are detected by the algorithm. (b)-(c) Wavelet chains corresponding to the potential false positives marked by the red box. In every plot, the blue line represents the wavelet local maxima coefficients connected in a chain and the orange line represents the threshold value generated by the local noise at every scale. These wavelet chains have a decreasing slope.

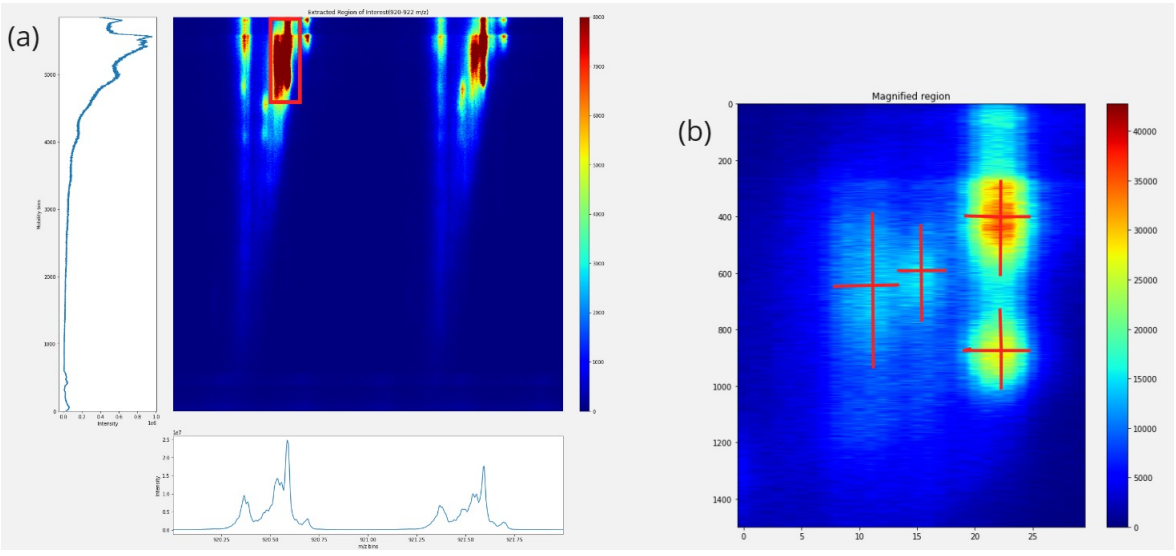
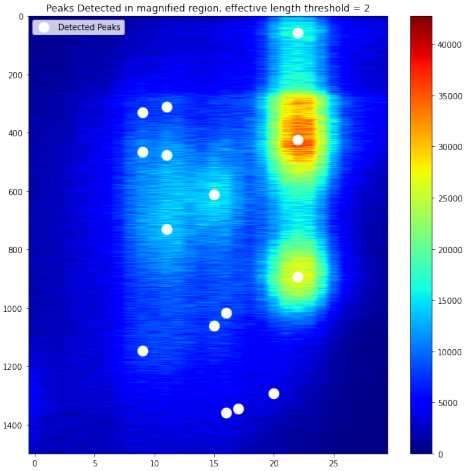
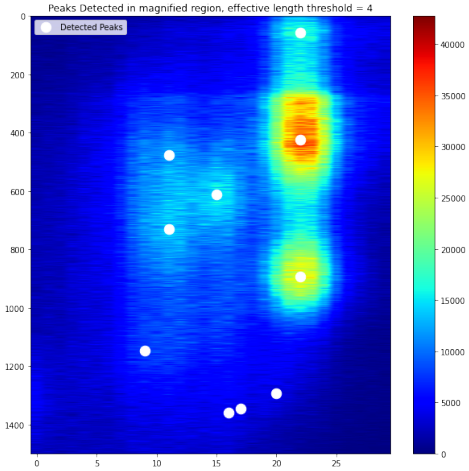


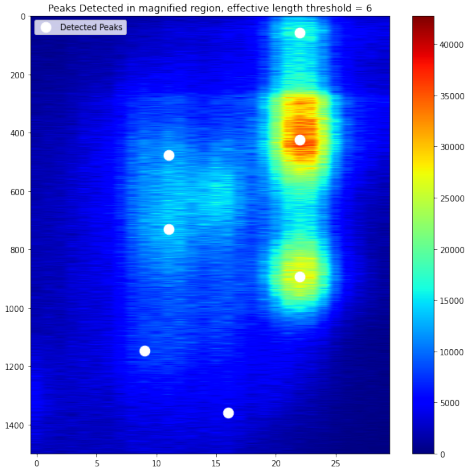
Figure 5-22: Experiment 2.3: Evaluating the performance of the algorithm by varying the effective threshold length on a real world IM-IMS data sample. Test section : 920-922 m/z with complete mobility information. (a) Visual representation of the test section.Red box indicates the region that will be magnified (b) Magnified region: The "+" sign marks manually annotated peaks. These peaks have widths in the range of 3-8 data points (along columns).



(a) Effective length threshold=2. Total no. of peaks detected = 15



(b) Effective length threshold=4. Total no. of peaks detected = 10



(c) Effective length threshold = 6. Total no. of peaks detected = 7

Figure 5-23: Experiment 2.3: Evaluating the performance of the algorithm by varying the effective threshold length on a real world IM-IMS data sample. Peaks detected by the algorithm in the magnified region (Figure (5-22)). At effective length threshold value = 6, the manually marked true positive does not get detected.

Methodology

For the given experiment, we use generalized 2D mexican hat function as the wavelet function. The scale parameters used in this experiment is given in Table (5-2). The effective length threshold parameter is set to 4. The penalty factors being studied in this experiment are: 1,2,3.

The partitioned data section used in this experiment is given in Figure (5-1b). This section represents broad and prominent peaks which are noisy in nature.

As we do not know the ground truth for this data sample, we will assess the performance of the parameter based on the peaks being detected.

Results

Results for varying penalty factors on a partitioned section of the data sample is presented in Figure (5-24). Table (5-12) presents the total number of peaks detected for varying penalty factor.

Test Section	P.f = 1	P.f = 2	P.f = 3
704-706 m/z	87	61	58

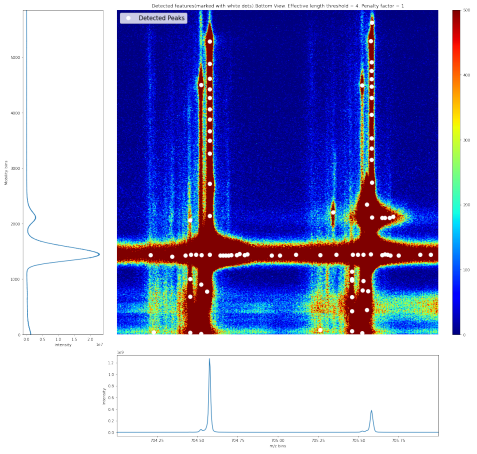
Table 5-12: Results for experiment 2.4: Evaluating the performance of the algorithm by varying the penalty factor on a real world IM-IMS data sample. Test section: 704-706 m/z with complete mobility information.The table presents the total no. of peaks detected for varying penalty factor(P.f).

Discussion

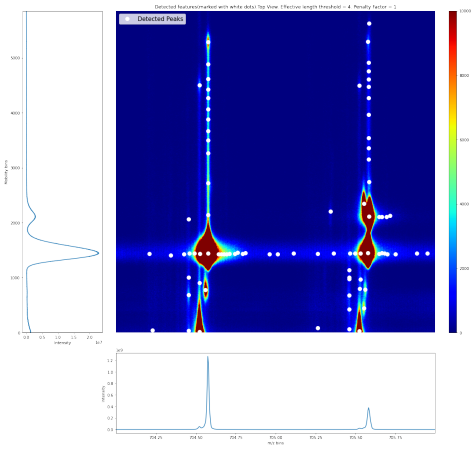
1. Not penalizing the horizontal direction (i.e. setting penalty factor = 1) leads to detection of false positives along the surface of the peak in horizontal direction. The broad nature of peak along with its noisy nature leads to local maxima points which have wavelet coefficients with large amplitudes. These coefficients have higher values than the threshold level determined by the local noise level for a number of scales. As a result, they get detected by the algorithm.

2. Increasing the penalty factor from 2 to 3 did not produce much difference in the number of potential false positives detected.

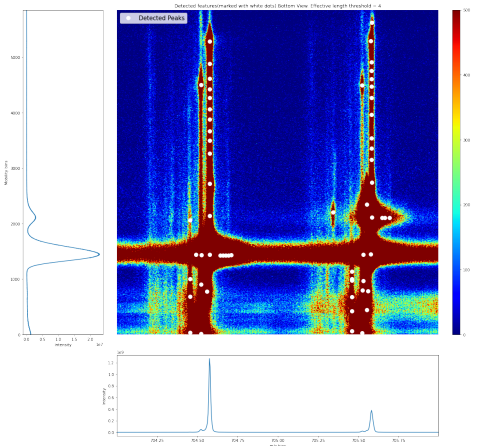
3. Having a penalty factor removes some peaks but those peaks are usually present in the bottom region of the data sample which has instrument artefacts and are not used for peak detection.



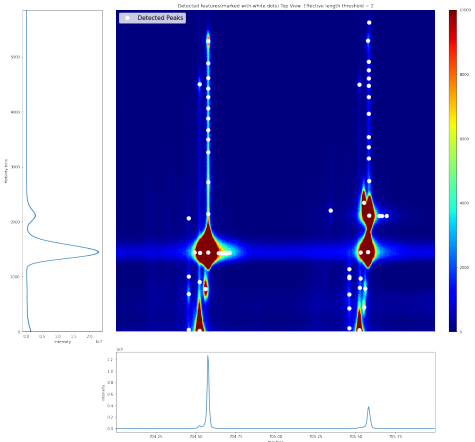
(a) Penalty factor = 1, bottom view.



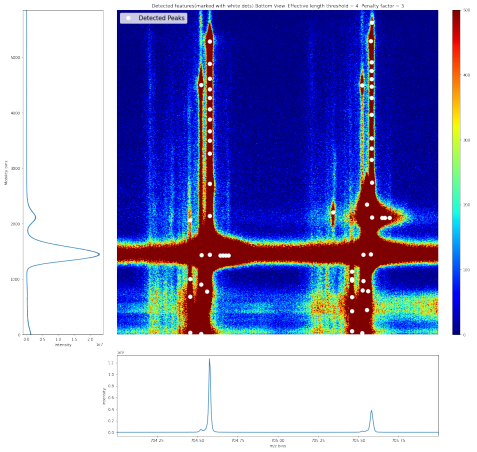
(b) Penalty factor = 1, top view.



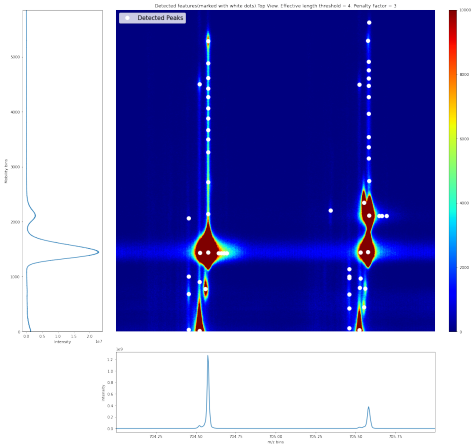
(c) Penalty factor = 2, bottom view.



(d) Penalty factor = 2, top view.



(e) Penalty factor = 3, bottom view.



(f) Penalty factor = 3, top view.

Figure 5-24: Results for experiment 2.4: Evaluating the performance of the algorithm by varying the penalty factor on a real world IM-IMS data sample. Test section: 704-706 m/z with complete mobility information. The results of every penalty factor is presented both from a top view and bottom view. In every sub-figure, the white dots mark the peaks detected by the algorithm.

5-3-5 Experiment 2.5: Evaluating the number of noise simulations required for obtaining a stable value for M_a.

For this experiment, we will evaluate the hyperparameter M_a used in equation (5-6). Here, M_a refers to the average of maximum local maxima that is generated by the wavelet transform of zero mean white noise with variance = 1 at scale a. As previously discussed in experiment 2.1, we know that maximum local maxima is not a stable statistic and varies with simulation. In order to minimize the variability, we average a certain number of simulations. Based on this idea, the objective of this experiment is to determine the optimal number of simulations N required for averaging.

Methodology

For the given experiment, we use a generalized 2D mexican hat function as the wavelet function. In order to keep the computation cost of the experiment, we demonstrate this experiment on a single scale. The scale parameter being used is $a = 2$ (which corresponds to $\sigma_y = 0.5a = 1$ and $\sigma_x = 64$).

We introduce two parameters :

1. N - Number of simulations required for averaging to obtain M_a. This is the test parameter being studied in this experiment. The values that will be considered for n are : 1,10,20,50.
2. B - The number of samples generated using the test statistic with fixed n. The idea is to generate B samples of the parameter M_a with fixed n and calculate the variation within these samples. In order to keep the computation cost low, B is taken as 10.

As we do not know the true (or theoretical value) of the parameter M_a, the experiment will be focused towards obtaining a stable representation of the value M_a. The procedure for evaluation is as follows :

1. For B = 1:10
   (a) For i = 1:N
      i. Simulate zero mean gaussian noise with variance = 1. The size of the noise matrix is taken to be 2048 X 2048. This is to reduce the computation cost and to mitigate any boundary effects.
      ii. Compute the wavelet transform of the noise matrix for scale $a = 2$.
      iii. Zero out the negative coefficients.
      iv. Identify the maximum local maxima in the wavelet transform of the noise matrix.
      v. Store the value.
   (b) Average the obtained N values to obtain M_a. This will be considered a single sample.
   (c) Store this value.

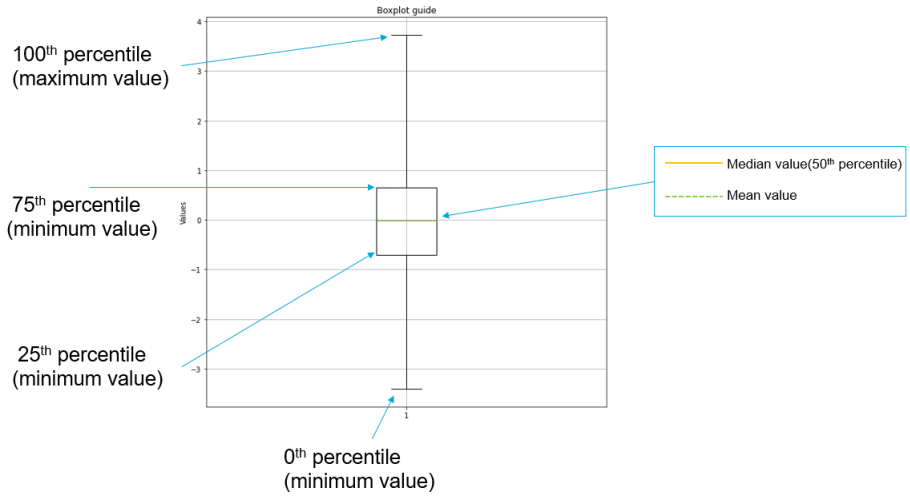


Figure 5-25: Experiment 2.5: Evaluating the number of noise simulations required for obtaining a stable value for M_a. Box plot guide. The samples used in this plot were obtained using 1024 randomly sampled points from the distribution gaussian distribution $\sim N(0, 1)$.

2. After obtaining B samples, plot the box plot to highlight the variability in the samples. Figure (5-25) presents a guide to read the box plot
3. Repeat the experiment for different values of N.

Results

The minimum , maximum, and the range (maximum - minimum) obtained for different values of N are presented in Table (5-13). The box plot representation for different values of N is presented in Figure (5-26).

N	Maximum Value	Minimum Value	Mean	Median	Range
1	0.868	0.690	0.743	0.725	0.177
10	0.795	0.716	0.757	0.752	0.079
20	0.782	0.726	0.749	0.748	0.0560
50	0.769	0.745	0.757	0.757	0.0241

Table 5-13: Results for experiment 2.5: Evaluating the number of noise simulations required for obtaining a stable value for M_a. Statistics obtained for different values of N used in estimation of M_a. B is fixed as 10 for each case.

Discussion

1. The general inference that can be drawn form Table (5-13) and Figure (5-26) is that lower values of N leads to a higher variation (range) in values of M_a. Higher variation is

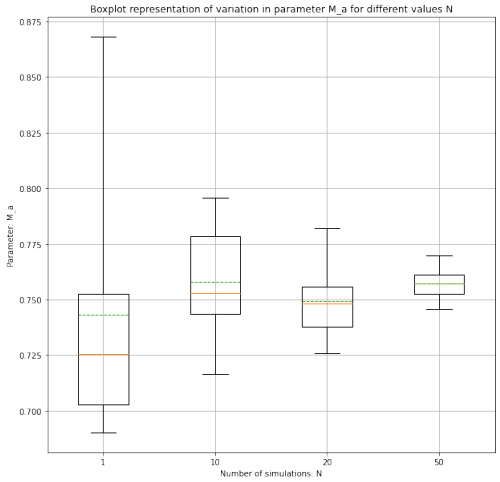


Figure 5-26: Results for experiment 2.5: Evaluating the number of noise simulations required for obtaining a stable value for M_a. Box plot for different values of N. In every box, the yellow line marks the median of the sample and the dashed green lines mark the mean of the samples.

not useful because if the experiment is repeated, there is a possibility that value $T_{a,local}$ will not remain the same and as a result it can impact the effective length of a chain.

2. As N increases, this variation starts to decrease and the mean and median obtained for the samples of M_a shift to higher value. This can be seen for $N = 50$ and is a direct result for averaging a higher number of simulations for calculation of M_a. We can infer that for that taking a higher number of simulations N will increase the stability of the parameter $T_{a,local}$.

3. For the given, experiment we would say that taking $N = 50$ ensures stability in the estimation of parameter M_a. However, this comes with a large computation cost and therefore, more research is required in reducing the computation cost of this parameter.

5-4 Experiment 3: Wavelet width parameters

The experiments in this section are related to the parameter σ_x. This parameter is associated with the width of the wavelet function (along rows) and governs the overall performance of the algorithm. We conduct experiments to understand the impact of this parameter on the performance of the algorithm.

5-4-1 Experiment 3.1: Evaluating the performance of the algorithm by varying σ_x on a synthetic IM-IMS data sample.

Methodology

For the given experiment, we use generalized 2D mexican hat function as our wavelet function. The penalty factor used in this experiment is 2. The effective length threshold parameter used is 4. The choice of σ_x and $\sigma_y(a)$ used in this experiment are presented in Tables (5-14) - (5-17).

We say that a chemical peak is detected if this location is at a distance of less than ≤ 0.01 Da in the m/z dimension and less than 100 mobility bins (rows) from the true chemical peak. Else, it will be considered as a false positive.

The performance of the parameter (and the algorithm) will be evaluated using the F-Score metric.

a	$\sigma_y = 0.5a$	σ_x	Size of wavelet: floor[$10\sigma_x$] X floor[$10\sigma_x$]
2.0	1.0	16	160 X 10
2.3784	1.1892	16	160 X 11
2.828	1.414	16	160 X 14
3.3635	1.6817	16	160 X 16
4.0	2.0	16	160 X 20
4.7568	2.3784	16	160 X 23
5.6568	2.828	16	160 X 28
6.7271	3.3635	16	160 X 33
8.0	4.0	16	160 X 40
9.5136	4.7568	16	160 X 47
11.3137	5.656	16	160 X 56
13.454	40	16	160 X 67
16.0	8.0	16	160 X 80

Table 5-14: Experiment 3.1: Evaluating the performance of the algorithm by varying σ_x on a synthetic IM-IMS data sample. Wavelet function : Generalized 2D mexican hat function. The table presents the scale parameter a, the widths (σ_x and σ_y) and the size of the wavelet function (in terms of rows and columns).

a	$\sigma_y = 0.5a$	σ_x	Size of wavelet along: floor[$10\sigma_x$] X floor[$10\sigma_y$]
2.0	1.0	32	320 X 10
2.3784	1.1892	32	320 X 11
2.828	1.414	32	320 X 14
3.3635	1.6817	32	320 X 16
4.0	2.0	32	320 X 20
4.7568	2.3784	32	320 X 23
5.6568	2.828	32	320 X 28
6.7271	3.3635	32	320 X 33
8.0	4.0	32	320 X 40
9.5136	4.7568	32	320 X 47
11.3137	5.656	32	320 X 56
13.454	40	32	320 X 67
16.0	8.0	32	320 X 80

Table 5-15: Experiment 3.1: Evaluating the performance of the algorithm by varying σ_x on a synthetic IM-IMS data sample. Wavelet function : Generalized 2D mexican hat function. The table presents the scale parameter a, the widths (σ_x and σ_y) and the size of the wavelet function (in terms of rows and columns).

a	$\sigma_y = 0.5a$	σ_x	Size of wavelet: floor$[10\sigma_x]$ X floor$[10\sigma_y]$
2.0	1.0	64	640 X 10
2.3784	1.1892	64	640 X 11
2.828	1.414	64	640 X 14
3.3635	1.6817	64	640 X 16
4.0	2.0	64	640 X 20
4.7568	2.3784	64	640 X 23
5.6568	2.828	64	640 X 28
6.7271	3.3635	64	640 X 33
8.0	4.0	64	640 X 40
9.5136	4.7568	64	640 X 47
11.3137	5.656	64	640 X 56
13.454	40	64	640 X 67
16.0	8.0	64	640 X 80

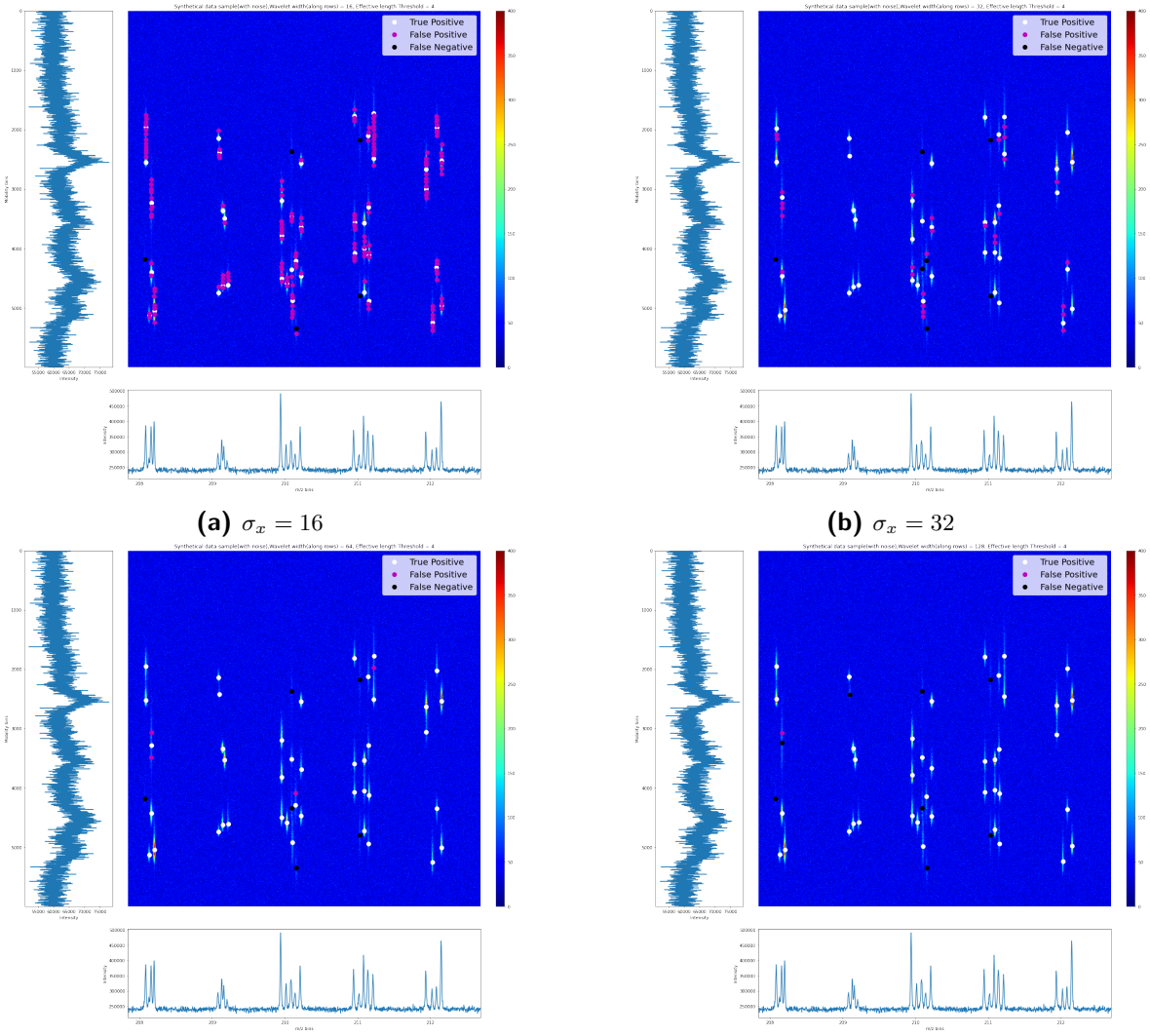
Table 5-16: Experiment 3.1: Evaluating the performance of the algorithm by varying σ_x on a synthetic IM-IMS data sample. Wavelet function : Generalized 2D mexican hat function. The table presents the scale parameter a, the widths (σ_x and σ_y) and the size of the wavelet function (in terms of rows and columns).

a	$\sigma_y = 0.5a$	σ_x	Size of wavelet: floor$[10\sigma_x]$ X floor$[10\sigma_y]$
2.0	1.0	128	1280 X 10
2.3784	1.1892	128	1280 X 11
2.828	1.414	128	1280 X 14
3.3635	1.6817	128	1280 X 16
4.0	2.0	128	1280 X 20
4.7568	2.3784	128	1280 X 23
5.6568	2.828	128	1280 X 28
6.7271	3.3635	128	1280 X 33
8.0	4.0	128	1280 X 40
9.5136	4.7568	128	1280 X 47
11.3137	5.656	128	1280 X 56
13.454	40	128	1280 X 67
16.0	8.0	128	1280 X 80

Table 5-17: Experiment 3.1: Evaluating the performance of the algorithm by varying σ_x on a synthetic IM-IMS data sample. Wavelet function : Generalized 2D mexican hat function. The table presents the scale parameter a, the widths (σ_x and σ_y) and the size of the wavelet function (in terms of rows and columns).

Results

Results for varying σ_x is presented in Figure (5-27). The confusion matrix associated with the various values of σ_x used in this experiment is presented in Table (5-18). The F-score plot is shown in Figure (5-28).



(a) $\sigma_x = 16$ **(b)** $\sigma_x = 32$

(c) $\sigma_x = 64$ **(d)** $\sigma_x = 128$

Figure 5-27: Results for experiment 3.1: Evaluating the performance of the algorithm by varying σ_x on a synthetic IM-IMS data sample. In every sub-figure, the white, black and magenta dots represent the true positives, false negatives and false positives detected by the algorithm respectively.

σ_x	T.P.	F.P.	F.N.	T.N.
16	43	171	5	0
32	41	30	7	0
64	42	4	6	0
128	40	1	8	0

Table 5-18: Results for experiment 3.1: Evaluating the performance of the algorithm by varying σ_x on a synthetic IM-IMS data sample. The table present the confusion matrix associated with the different effective length threshold parameter. T.P, F.P, T.N. and F.N. stands for True Positive, False Positive, True Negative and False Negative respectively.

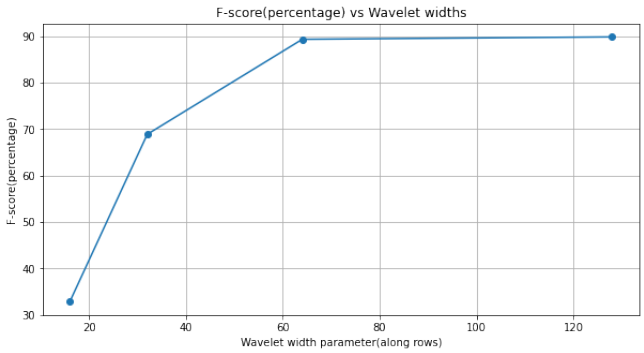


Figure 5-28: Results for experiment 3.1: Evaluating the performance of the algorithm by varying σ_x on a synthetic IM-IMS data sample. F-score(%) plot.

Discussion

1. For $\sigma_x = 16$ and $\sigma_x = 32$, we see high number of false positives being detected. This is because if the size of σ_x is lower than the size of peaks present in the noisy data sample, then multiple peaks will get detected on the surface of the true peaks.

2. For $\sigma_x = 128$, the higher value of wavelet width resulted in reduced the number of false positives. But, it also resulted in detection of less number of features (Figure (5-27d)). Most of the peaks that were not detected had short peak widths or were close to another dominant peak in the data matrix. For these peaks, the higher value of σ_x lead to reduced strength in terms of magnitude of wavelet coefficients.

3. The F-score (Figure 5-28) for chosen values of σ_x shows that the wavelets obtained using $\sigma_x = 64$ and $\sigma_x = 128$ have similar performance. But, we believe that $\sigma_x = 64$ yields better results. This is because, we prioritize detecting more features than reducing the number of false positives. Choosing $\sigma_x = 64$ resulted in detection of more peaks with minimal increase in false positives. The same cannot be said for $\sigma_x = 16$.

4. Ideally, σ_x should also vary along with σ_y for optimal detection of features. But, for the designed algorithm, choosing σ_x equal to 1% of the mobility dimension (rows) can yield satisfactory results.

5-4-2 Experiment 3.2 : Evaluating the performance of the algorithm by varying σ_x on a real world IM-IMS data sample

In this experiment, we will vary σ_x threshold parameter and analyse the peaks that get detected in the real world IM-IMS data sample.

Methodology

For the given experiment, we use generalized 2D mexican hat function as our wavelet function.The penalty factor used in this experiment is 2. The effective length threshold is taken as 4. The choice of σ_x and $\sigma_y(a)$ used in this experiment is presented in tables (5-14) - (5-17).

The partitioned data sections used in this experiment is briefly discussed in Section (5-1-1). Figure (5-1) presents the visual overview of the 2D data sections that will be used in this experiment.

As we do not know the ground truth for this data sample, we will assess the performance of the parameter based on the peaks being detected.

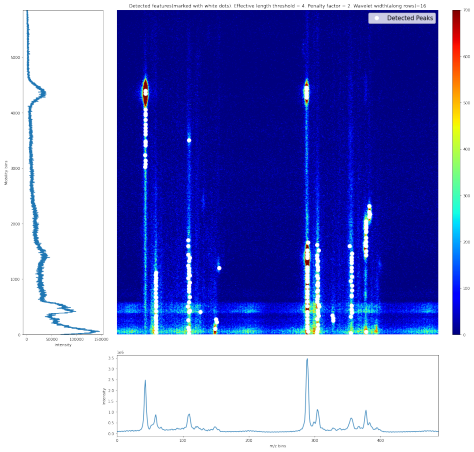
Results

Results for varying σ_x for different partitioned sections of the data sample are presented in:

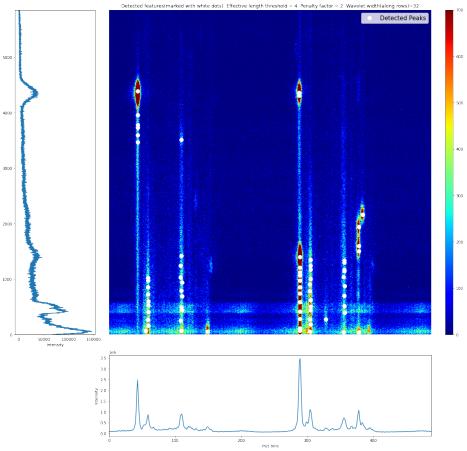
1. 670-672 m/z - Figure (5-29)
2. 704-706 m/z - Figure (5-30)
3. 770-772 m/z - Figure (5-31)
4. 920-922 m/z - Figure (5-32)

Discussion

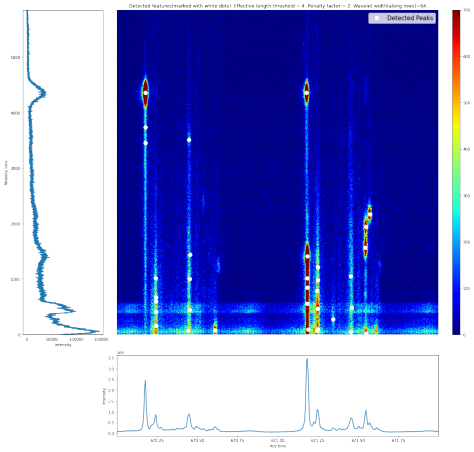
1. In general, the observations of this experiment are similar to the case of synthetic data sample. Choosing lower values of σ_x (16 or 32) results in high number of false positives getting detected whereas choosing $\sigma_x = 128$ results in less number of potential false positives but also less number of potential true positives getting detected.
2. The horizontal false positives that are detected in the case of sections with high SNR reduce for $\sigma_x = 16$ (Figure (5-31a) and Figure (5-30a)). This can be observed by looking at the transform space related to these potential false positives. In the case of $\sigma_x = 16$ we detect multiple local maximas in the region of potential false positives whereas in the case of $\sigma_x = 128$ we observe one local maxima in that region. The smaller size of $\sigma_x = 16$ leads to generation of multiple local maximas but it also reduces the coefficients strength. Choosing a large value for σ_x will make the wavelet function large (especially the positive part of the wavelet function) and will yield high coefficient values.
3. In the case of real world data sample, choosing $\sigma_x \leq 1\%$ of the mobility bins will lead to detection of high number of potential false positives. However, further testing is required for the ideal choice of σ_x.



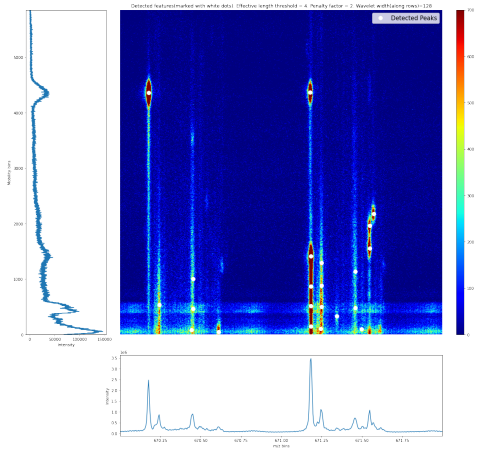
(a) $\sigma_x = 16$



(b) $\sigma_x = 32$

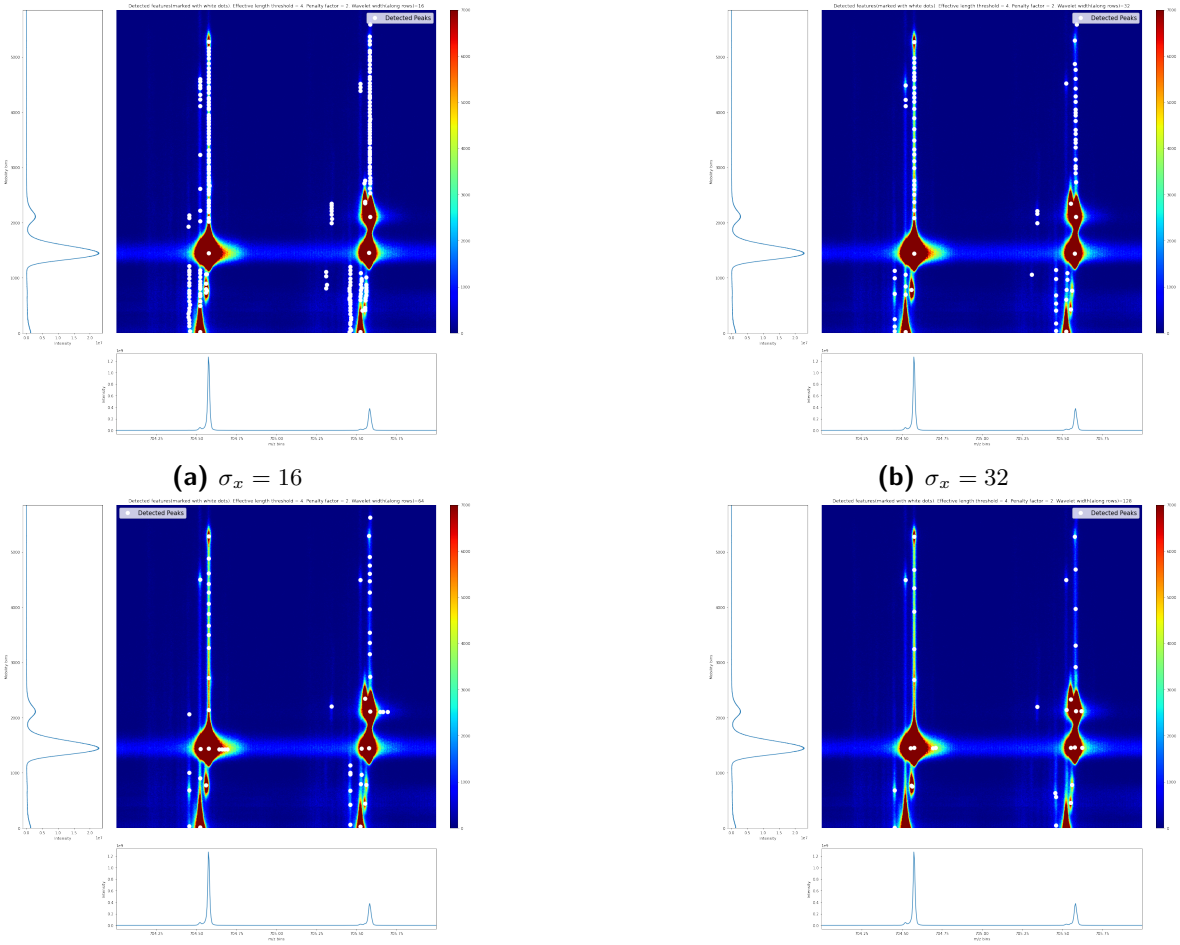


(c) $\sigma_x = 64$



(d) $\sigma_x = 128$

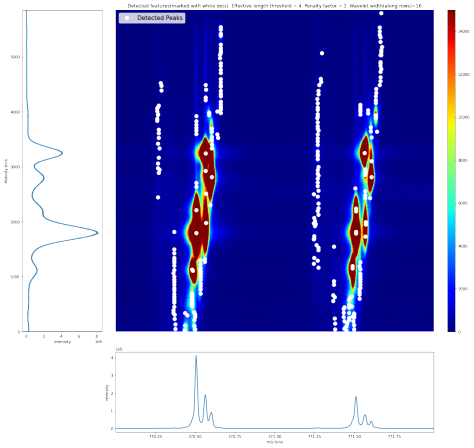
Figure 5-29: Results for experiment 3.2: Evaluating the performance of the algorithm by varying the width of the wavelet function on real world IM-IMS data sample. Test section : 670- 672 m/z with complete mobility information. In every sub-figure, the white dots mark the peaks detected by the algorithm.



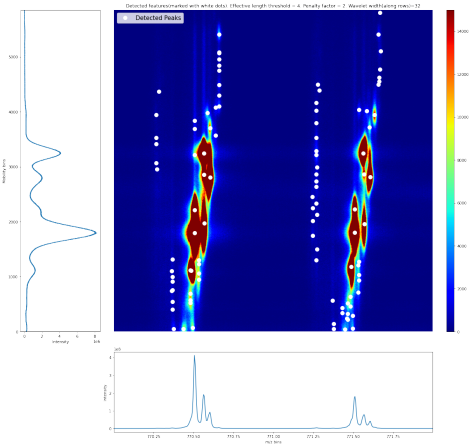
(a) $\sigma_x = 16$ **(b)** $\sigma_x = 32$

(c) $\sigma_x = 64$ **(d)** $\sigma_x = 128$

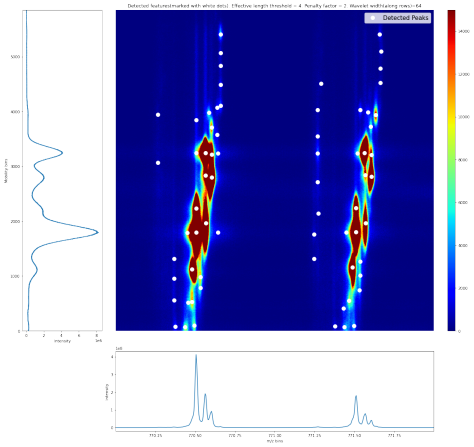
Figure 5-30: Results for experiment 3.2: Evaluating the performance of the algorithm by varying the width of the wavelet function on real world IM-IMS data sample. Test section : 704- 706 m/z with complete mobility information. In every sub-figure, the white dots mark the peaks detected by the algorithm.



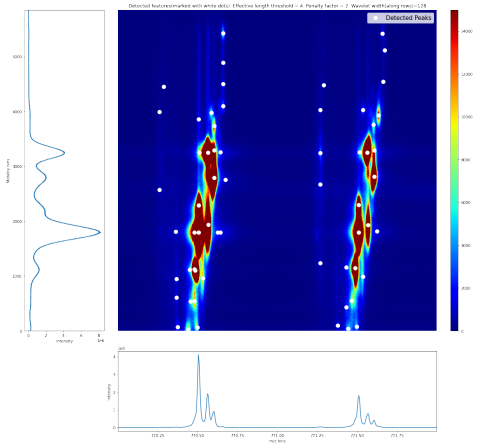
(a) $\sigma_x = 16$



(b) $\sigma_x = 32$

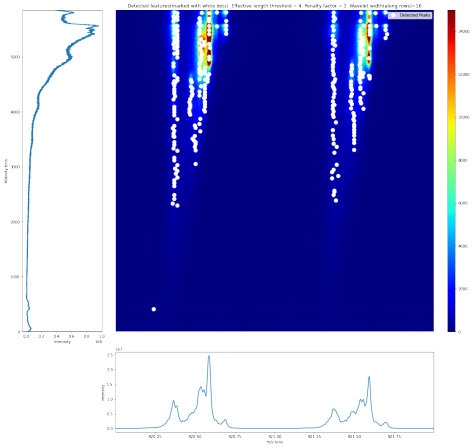


(c) $\sigma_x = 64$

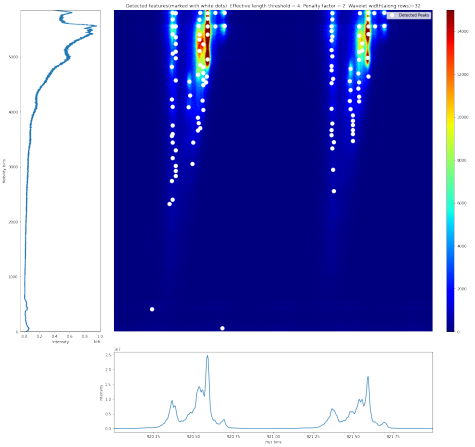


(d) $\sigma_x = 128$

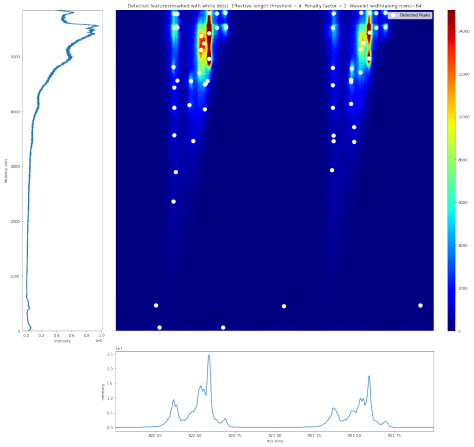
Figure 5-31: Results for experiment 3.2: Evaluating the performance of the algorithm by varying the width of the wavelet function on real world IM-IMS data sample. Test section : 770 - 772 m/z with complete mobility information. In every sub-figure, the white dots mark the peaks detected by the algorithm.



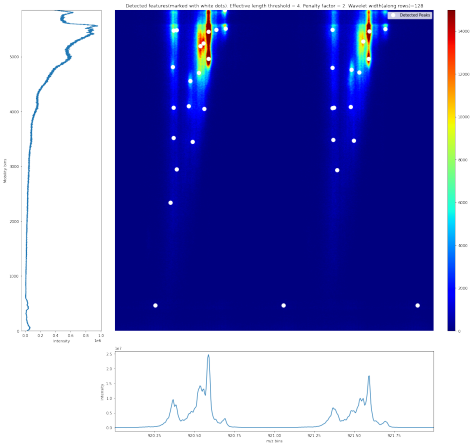
(a) $\sigma_x = 16$



(b) $\sigma_x = 32$



(c) $\sigma_x = 64$



(d) $\sigma_x = 128$

Figure 5-32: Results for experiment 3.2: Evaluating the performance of the algorithm by varying the width of the wavelet function on real world IM-IMS data sample. Test section : 920-922 m/z with complete mobility information. In every sub-figure, the white dots mark the peaks detected by the algorithm.

5-5 Experiment 4 - Comparison with an existing peak detection algorithm

The final experiment in this chapter is to compare the performance of our algorithm with an existing peak detection algorithm on a real world IM-IMS data sample. For this, we use Self Adjusting Feature Detection algorithm (SAFD) [78] as our companion algorithm. The choice for SAFD is motivated by the following criteria : (i) freely available (ii) works on continuous profile data (iii) uses python/julia as the programming language. See section 2-2-4 for more details about the functioning of the algorithm.

Methodology

2D WTM algorithm parameters

The parameters used in 2D WTM are given in Table (5-20). The parameter σ_y is adapted based on our understanding of the maximum and minimum peak width present in the IM-IMS data sample (along m/z dimension). Rest of the parameters were optimized based on the results of the previous experiments.

SAFD algorithm parameters

The parameters used in SAFD are given in Table (5-21). These parameters were optimized based on our understanding of the algorithm and the real world IM-IMS data sample.

Hardware and Software specifications

Table (5-19) presents the hardware specifications of the computer on which the experiments were performed.

Specification	Version
Processor	Intel(R) Core(TM) i7-9750HF CPU @ 2.60GHz
RAM	8 GB RAM
SSD	Yes
OS	Windows 11
Software platform for 2D WTM algorithm	Jupyter Notebook 6.3.0
Software platform for SAFD	Jupyter Notebook 6.3.0

Table 5-19: Experiment 4: Comparison of the designed algorithm with an existing peak detection algorithms. Hardware spectifications

Test Sections

The partitioned data sections used in this experiment is briefly discussed in Section (5-1-1). Figure (5-1) presents the visual overview of the 2D data sections that will be used in this experiment.

We say that a common chemical peak is detected if the peaks detected by both the algorithms are at a distance of less than 5 m/z bins (columns) and less 100 mobility bins (rows). Else, they will be considered as peaks specific to the algorithm.

As we do not know the ground truth for this data sample, the performance of the algorithms will be based on the peaks detected and the computation time of the algorithms.

2D WTM parameters	Value
σ_x (mobility dimension)	64
$\sigma_y = 0.5a$ (m/z dimension)	Dyadic scales from 1 to 8 with three voices per octave
Effective length threshold	4
Penalty factor	2
Number of noise simulation	50

Table 5-20: Experiment 4: Comparison of the designed algorithm with an existing peak detection algorithms. Optimized parameters for 2D WTM algorithm.

SAFD parameters	Value
Number of iterations	200
Maximum peak width (mobility dimension)	1200
Minimum peak width (mobility dimension)	200
Minimum peak width (m/z dimension)	4
Resolution	50000
Minimum intensity	10
Overlapping features threshold	5 (Default value)
SNR Threshold	2

Table 5-21: Experiment 4: Comparison of the designed algorithm with an existing peak detection algorithms. Optimized parameters for SAFD.

Results

Results for different test sections are presented in Figure (5-33) - Figure (5-36). Table (5-22) presents the total number of peaks detected by both the algorithms. Table (5-23) presents the computation time of both the algorithms.

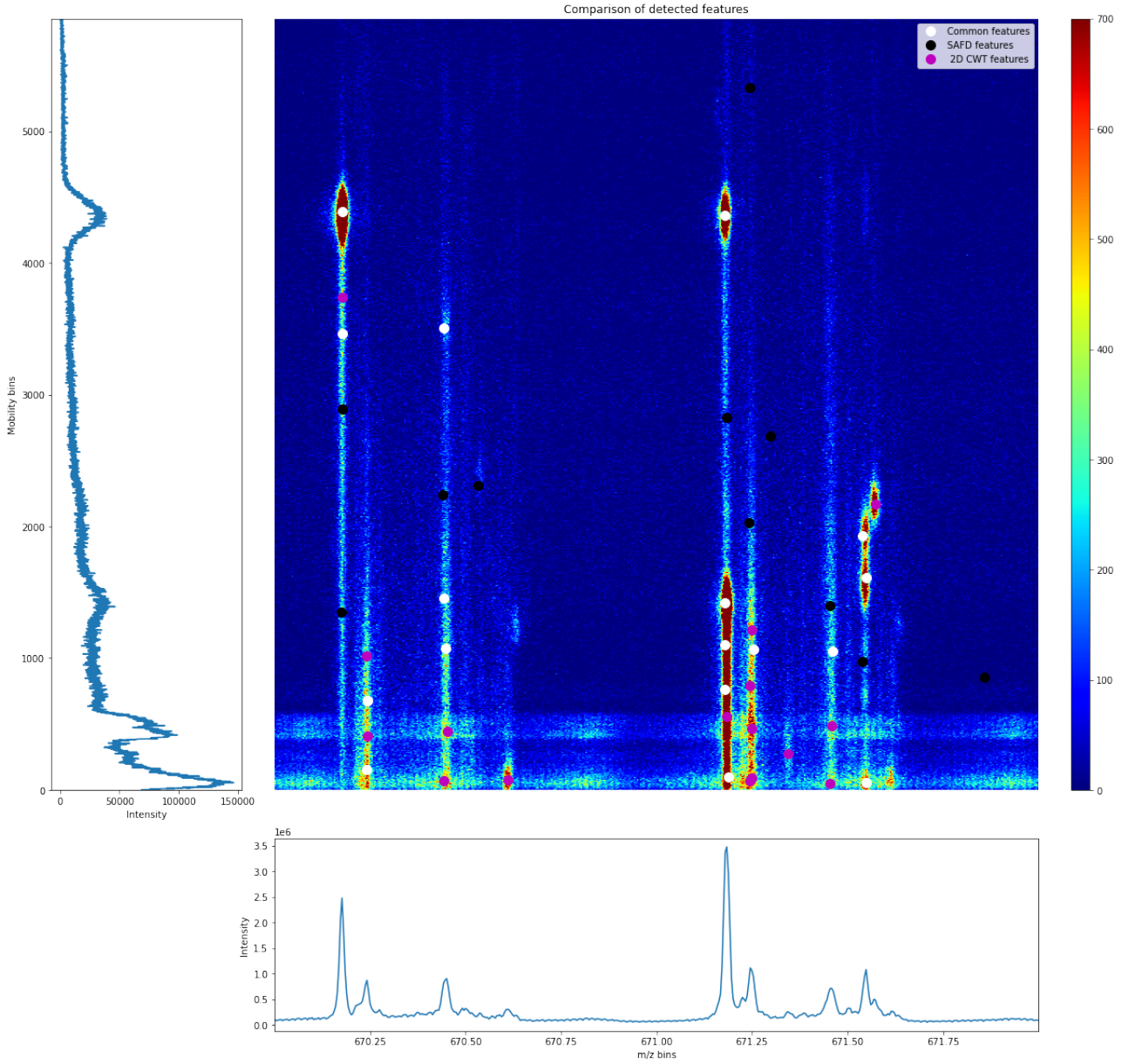


Figure 5-33: Results for experiment 4: Comparison of the designed algorithm with an existing peak detection algorithm on a real world IM-IMS data sample. Test section : 670- 672 m/z with complete mobility information. White dots mark the common peaks detected by both of the algorithms, black dots represent the peaks detected by the SAFD algorithm and the magenta dots represent the peaks detected by our 2D CWT algorithm.

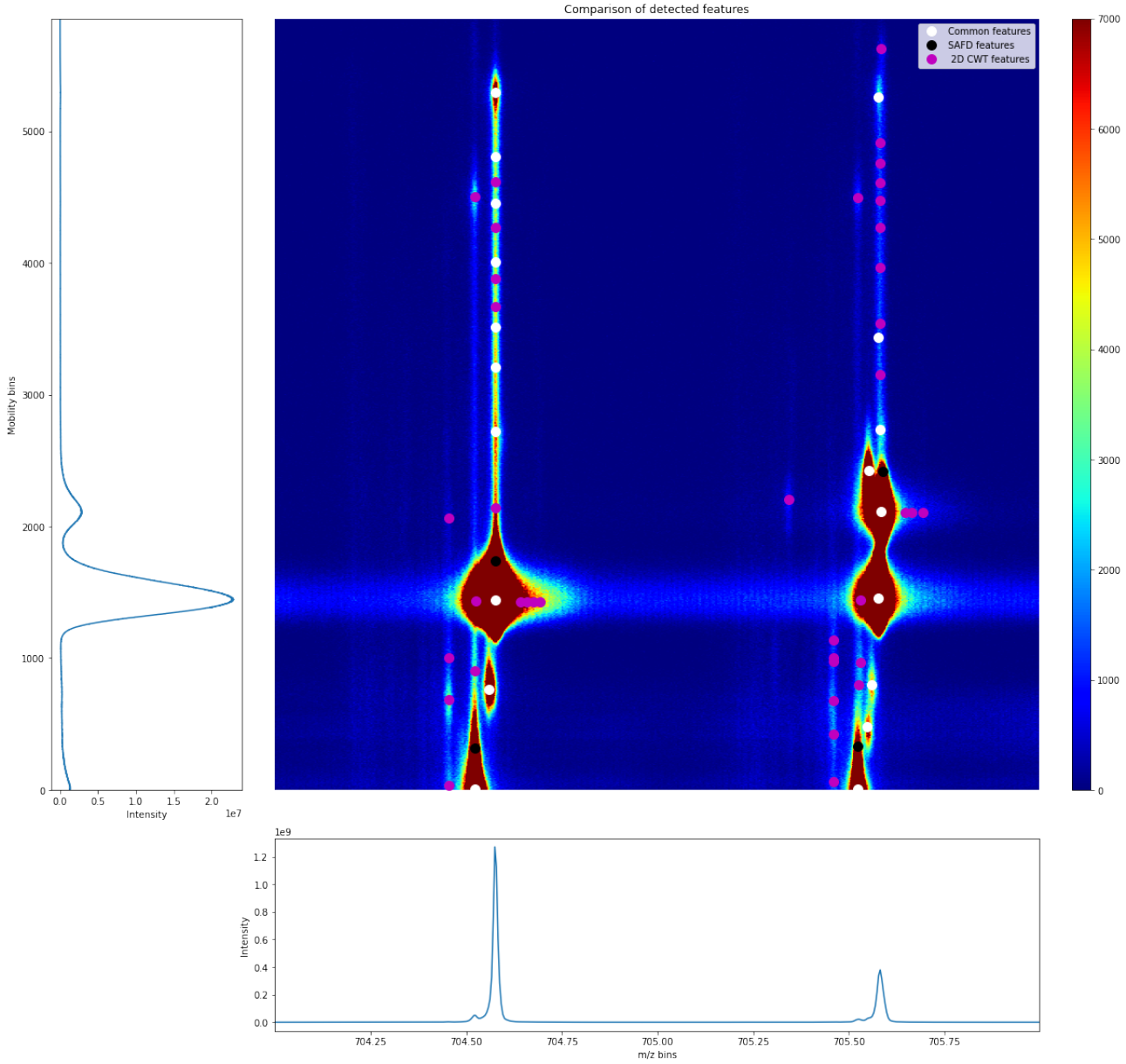


Figure 5-34: Results for experiment 4: Comparison with an existing peak detection algorithm on a real world IM-IMS data sample. Test section : 704-706 m/z with complete mobility information. White dots mark the common peaks detected by both of the algorithms, black dots represent the peaks detected by the SAFD algorithm and the magenta dots represent the peaks detected by our 2D WTM algorithm.

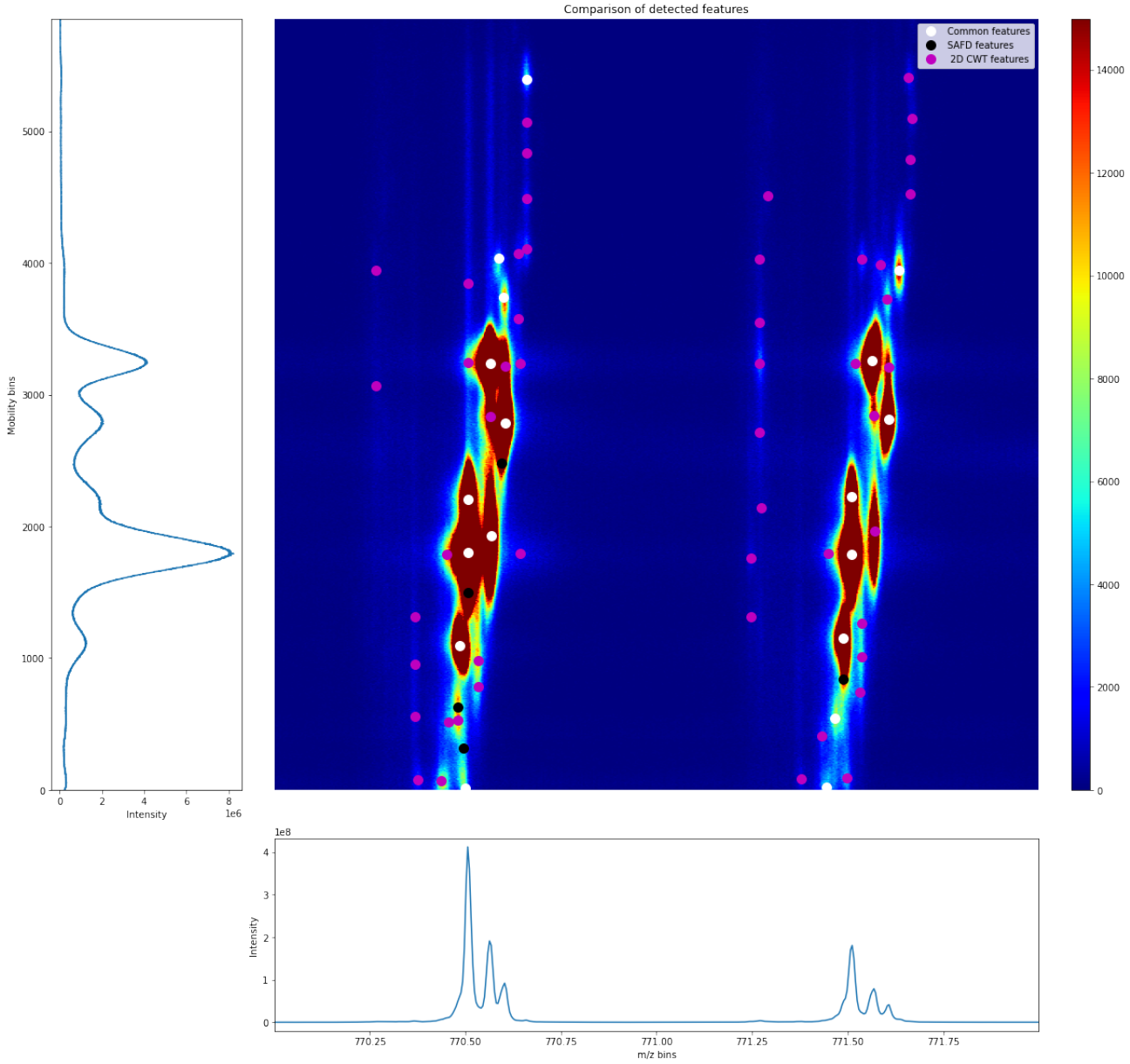


Figure 5-35: Results for experiment 4: Comparison with an existing peak detection algorithm on a real world IM-IMS data sample. Test section : 770-772 m/z with complete mobility information. White dots mark the common peaks detected by both of the algorithms, black dots represent the peaks detected by the SAFD algorithm and the magenta dots represent the peaks detected by our 2D WTM algorithm.

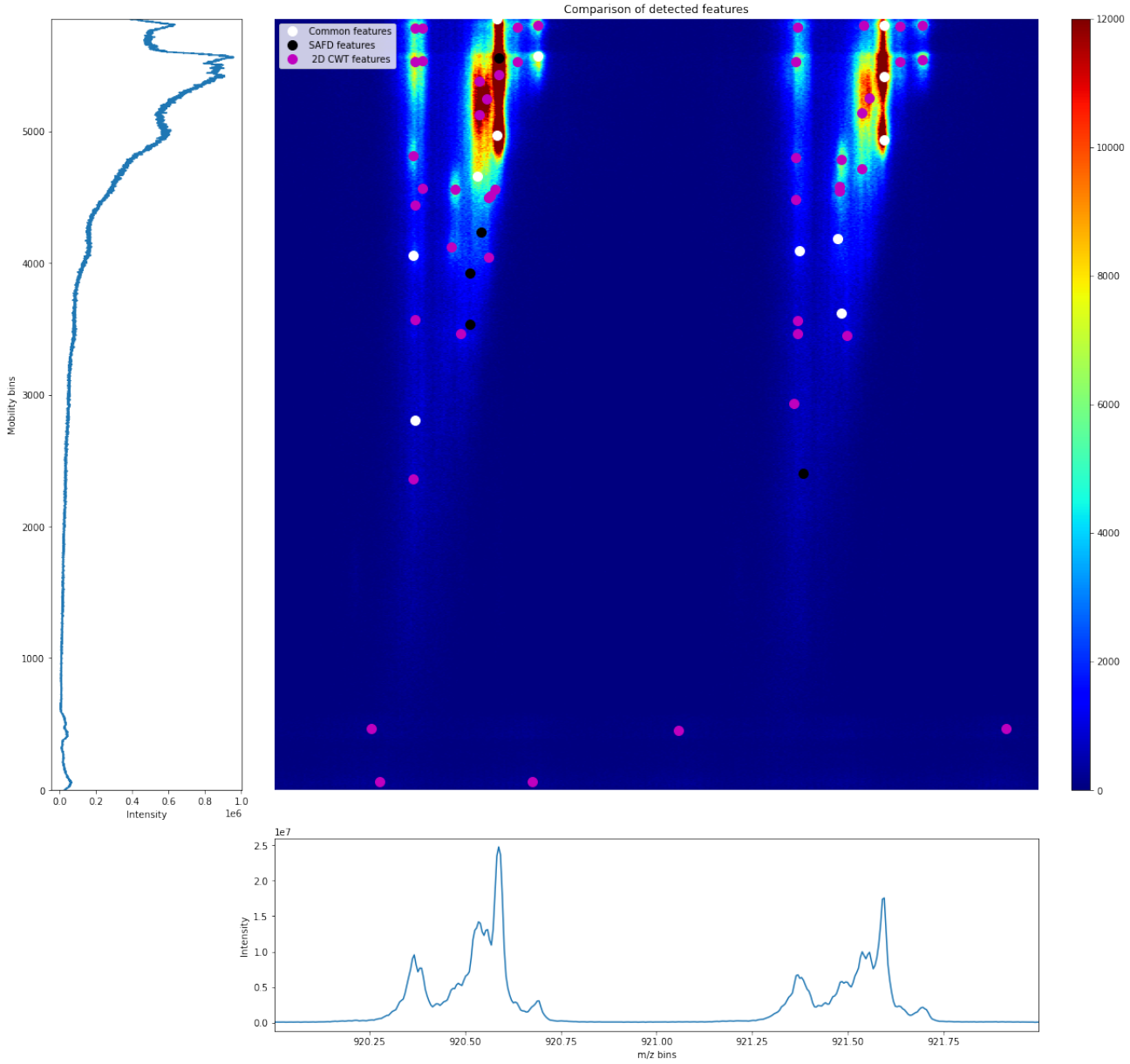


Figure 5-36: Results for experiment 4: Comparison with an existing peak detection algorithm on a real world IM-IMS data sample. Test section : 920-922 m/z with complete mobility information. White dots mark the common peaks detected by both of the algorithms, black dots represent the peaks detected by the SAFD algorithm and the magenta dots represent the peaks detected by our 2D WTM algorithm.

Region	SAFD peaks	2D WTM peaks	Common peaks(detected by both methods)
670-672 m/z	11	16	16
704-706 m/z	4	39	19
770-772 m/z	5	50	17
3.3635	5	42	12

Table 5-22: Results for experiment 4: Comparison with an existing peak detection algorithm. The total number of peaks detected by both the algorithms. The column SAFD peaks and 2D WTM peaks refer to peaks detected by the SAFD and 2D WTM algorithms exclusively. The last columns refer to peaks detected by both the algorithms.

Algorithm	Computation time
SAFD	$\sim$ 40 seconds
2D WTM (Breakdown presented below)	$\sim$ 30 minutes
(i) Simulation (for noise parameters)	$\sim$ 1500 seconds
(ii) 2D CWT + denoising + watershed transform	$\sim$ 30 seconds
(iii) Local maxima clustering + cluster denoising	$\sim$ 45 seconds
(iv) Chain construction + Effective length thresholding	$\sim$ 108 seconds

Table 5-23: Results for experiment 4: Comparison with an existing peak detection algorithm. Computation time required for processing the test section by different algorithms. The computation time is recorded for each test section used in benchmarking the algorithm. For 2D WTM based algorithm, we present the breakdown for the time taken by different sections of the algorithm.

Discussion

1. Overall, we found that the 2D WTM based algorithm is more sensitive and detects more peaks than SAFD for all the given test sections.

2. In the test section 704-706 m/z (Figure (5-34)), SAFD detects less number of potential false positives than the 2D WTM based algorithm

3. The 2D WTM based algorithm outperforms the SAFD algorithm in the 920-922 m/z test section (Figure (5-36)).

4. SAFD outperform the designed 2D WTM based algorithm with respect to computation time. The major portion of the time consumed by the latter goes in estimating the noise parameter which requires 51 (50+1) simulations of wavelet transform of zero mean white noise with variance = 1 for all scales a.

5. The computation time for chain construction depends on the number of scales (a) being used. Using higher number of scales will lead to increased computation time, using lower number of scales will reduce the resolution of the transform space which may result in construction of incorrect chains and lower number of features being detected (for a fixed effective length threshold value). Further investigation is required regarding the optimal number of scales for peak detection.

Chapter 6

Conclusion

The research was conducted for the design of a feature (peak) detection algorithm for IM-IMS data. Traditional algorithms depended on instrument and data specific parameters which cannot be optimized without prior knowledge. Therefore, the objective of this thesis was to design a robust feature detection algorithm that uses minimal information from the user.

Based on the objectives of the research, we proposed a 2D WTM algorithm for feature detection. 2D WTM has the advantage of reducing the redundant CWT space into a discrete subset of points that matches the definition of the feature (in this case peaks or local maxima points). These subsets of points can then be grouped together in the transform space to form 'chains'. Based on the various properties of these chains, we can deduce that the local maxima point in the original data space is due to noise or due to a true chemical peak.

By carefully examining the IM-IMS data sample, we decided to use 2D generalized mexican hat function as our wavelet function. This choice was motivated based on the shape of the peaks observed in the data sample. The wavelet function has two parameters σ_x and σ_y which govern the width of the wavelet function along rows and columns respectively. Based on the increasing scale condition, we decided to construct the transform space by fixing the parameter σ_x and varying σ_y.

For thresholding of constructed chains, we introduced a parameter called as "effective length". Effective length is the total number of scales for which the wavelet local maxima coefficient (belonging to a chain) is greater than its surrounding noise level. This noise level is governed by the maximum local maxima coefficient generated by the wavelet transform of local noise (assuming gaussian additive noise). We say that the a peak is detected if the effective length of the given chain is greater than the threshold effective length.

We evaluated the performance of the designed algorithm on a synthetic IM-IMS data sample and a real world IM-IMS data sample. While the performance of the algorithm was fair on the synthetic data sample, the performance of the algorithm was fair on the real world IM-IMS data sample.

6-1 On scale parameters σ_x and σ_y

The algorithm developed is constrained by the choice of σ_x which takes a single value as an input. By performing tests on synthetic and real world IM-IMS data sample, we found that this value should be approximately equal to 1% of the mobility dimension to generate fair results. Choosing a value $\leq$1% for σ_x lead to detection of high number of false positives while choosing a large value for σ_x lead to detection less number of true positives. However, there was no research conducted where σ_x considered a range of values as an input. Theoretically and through experiments, we found that considering a range of values for σ_x should be the next step for optimizing the algorithm.

For σ_y, the research was performed based on dyadic scales (from $a = 2$ to 16) with three voices per octaves. This choice had a vague motivation that peak widths along the m/z dimension belonged within this range was not adapted to the data sample being studied. While the choice of dyadic scales with three voices per octave along with an threshold effective length of 4 yielded decent results, the range of σ_y should be computed from the data itself.

We provide an outline for the future work for determining the scale parameters:

1. Determine a range for σ_x by studying the data along the mobility dimension - In this thesis, we managed to establish that 1% of the mobility dimension is fair choice for the width of the wavelet function (along rows). Choosing 2% lead to detection of less number of true positives. Therefore, the range of values that $sigma_x$ takes should be between 1% and 2% of the mobility dimension.

2. Determine a range for σ_y by studying the data along the m/z dimension - This can be done if the resolution of the instrument can be derived from the data itself. The idea would be to determine the expected peak width (in terms of data points or m/z bins) from the resolution and select a set of scales around the expected peak width. Another possible direction would be to identify the most prominent peak in the m/z dimension and the peak width associated with it. This peak width can then be used to determine a set of scales for the wavelet functions (in the m/z dimension).

3. Investigate the optimal choice for (σ_x, σ_y) as pairs and study its impact on the construction of chains - After determining the range for σ_x and σ_y, the idea would be selecting (σ_x, σ_y) such that the increasing nature of the widths is maintained. Research should be focused on whether this increasing necessary is a necessary condition for construction of chains and if not, how will the choice of (σ_x, σ_y) impact the construction of chains.

6-2 On construction of chains

In our algorithm, we used watershed segmentation in order to define a search space for the local maxima point thereby automating the process of chain construction. However, the problem with this method is that it requires the whole 3D matrix (scale parameter adds another dimension to the transform space making it a 3D matrix) in memory. We achieved this by using *Numpy* package command *np.memmap* which writes the 3D matrix into disk. However, this is not an optimal solution when working with huge data matrices.

Future work should be focused on construction of chains using only local maxima points. Carmona et al. [18] proposed a random walk algorithm for automatic construction of wavelet chains (ridges) in the transform space but the algorithm was focused for time-series based signals (speech signals). Future research should focus on developing algorithms for 2D MS based signals. One direction would be to construct a cost function that takes the wavelet coefficient value into account.

6-3 On local noise level

For determining the local noise level (for effective length based thresholding), we used a quantity "maximum local maxima" coefficient generated by wavelet transform of local noise (assuming additive gaussian noise). The local noise level determined using this quantity yielded fair results on the synthetic IM-IMS data sample. However, it was found that peaks whose maximum intensity values are comparable to the noise level will not be detected by the algorithm.

The estimation of the local noise level also required multiple simulations of zero mean white noise with unit variance as "maximum local maxima" coefficient as it is not a stable statistic. This drastically increased the computation time of the algorithm rendering the algorithm not useful for big data matrices.

Future work should be focused on generating local noise level that requires minimum number of simulations and whose values can be determined analytically. A possible direction would be to the use the equation $T_{a,local} = k\sigma_a$ where k is chosen such that it matches the threshold level as implemented in the current version of the algorithm. We believe that a correct estimation of local noise level in the transform space will yield better results.

6-4 On effective length based thresholding

In our algorithm, we used effective length based threshold as an input to be provided by the user. We found that the effective length of 4 was the optimal threshold length for the given set of scales. However, the relation between scale parameter and the effective length was not explored. It could happen that if we use a different set of scales, the optimal effective length might change. So, future work should be focused on generalizing the effective length parameter for different scales.

References

[1] Daniil A Abdrakhimov et al. "Biosaur: An open-source Python software for liquid chromatography–mass spectrometry peptide feature detection with ion mobility support". In: *Rapid Communications in Mass Spectrometry* (2021), e9045.

[2] Mark D Adams et al. "Complementary DNA sequencing: expressed sequence tags and human genome project". In: *Science* 252.5013 (1991), pp. 1651–1656.

[3] Joanne E Adaway and Brian G Keevil. "Therapeutic drug monitoring and LC–MS/MS". In: *Journal of Chromatography B* 883 (2012), pp. 33–49.

[4] Paul S Addison. *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance*. CRC press, 2017.

[5] N Leigh Anderson and Norman G Anderson. "Proteome and proteomics: new technologies, new concepts, and new words". In: *Electrophoresis* 19.11 (1998), pp. 1853–1861.

[6] Jean-Pierre Antoine et al. "Application of the 2-D wavelet transform to astrophysical images". In: *Physicalia magazine* 24 (2002), pp. 93–116.

[7] Jean-Pierre Antoine et al. *Two-dimensional wavelets and their relatives*. Cambridge University Press, 2008.

[8] A Arneodo, NGRS Decoster, and SG 2 Roux. "A wavelet-based method for multifractal image analysis. I. Methodology and test applications on isotropic and anisotropic random rough surfaces". In: *The European Physical Journal B-Condensed Matter and Complex Systems* 15 (2000), pp. 567–600.

[9] Aıcha Bagag et al. "Separation of peptides from detergents using ion mobility spectrometry". In: *Rapid Communications in Mass Spectrometry* 25.22 (2011), pp. 3436–3440.

[10] Albert Bijaoui and Frédéric Rué. "A multiscale vision model adapted to the astronomical images". In: *Signal processing* 46.3 (1995), pp. 345–362.

[11] Mark T Bokhart et al. "Quantitative mass spectrometry imaging of emtricitabine in cervical tissue model using infrared matrix-assisted laser desorption electrospray ionization". In: *Analytical and bioanalytical chemistry* 407.8 (2015), pp. 2073–2084.

[12] Amanda Rae Buchberger et al. “Mass spectrometry imaging: a review of emerging advancements and future insights”. In: *Analytical chemistry* 90.1 (2018), p. 240.

[13] Jonathan Buckheit et al. “About wavelab”. In: *Handbook of WaveLab Version* 850 (1995), pp. 1–37.

[14] Kristin E Burnum-Johnson et al. “Ion mobility spectrometry and the omics: Distinguishing isomers, molecular classes and contaminant ions in complex samples”. In: *TrAC Trends in Analytical Chemistry* 116 (2019), pp. 292–299.

[15] Bruno Campos et al. “Identification of metabolic pathways in Daphnia magna explaining hormetic effects of selective serotonin reuptake inhibitors and 4-nonylphenol using transcriptomic and phenotypic responses”. In: *Environmental science & technology* 47.16 (2013), pp. 9434–9443.

[16] Francesco Capozzi and Alessandra Bordoni. “Foodomics: a new comprehensive approach to food and nutrition”. In: *Genes & nutrition* 8.1 (2013), pp. 1–4.

[17] René Carmona, Wen-Liang Hwang, and Bruno Torresani. *Practical Time-Frequency Analysis: Gabor and wavelet transforms, with an implementation in S*. Academic Press, 1998.

[18] René A Carmona, Wen L Hwang, and Bruno Torrésani. “Multiridge detection and time-frequency reconstruction”. In: *IEEE transactions on signal processing* 47.2 (1999), pp. 480–492.

[19] M Mozammel Hoque Chowdhury and Amina Khatun. “Image compression using discrete wavelet transform”. In: *International Journal of Computer Science Issues (IJCSI)* 9.4 (2012), p. 327.

[20] Jürgen Cox and Matthias Mann. “MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification”. In: *Nature biotechnology* 26.12 (2008), pp. 1367–1372.

[21] Darren J Creek et al. “Metabolite identification: are you sure? And how do your peers gauge your confidence?” In: *Metabolomics* 10.3 (2014), pp. 350–353.

[22] Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[23] WIF David. “Powder diffraction peak shapes. Parameterization of the pseudo-Voigt as a Voigt function”. In: *Journal of applied crystallography* 19.1 (1986), pp. 63–64.

[24] Sören-Oliver Deininger et al. “Normalization in MALDI-TOF imaging datasets of proteins: practical considerations”. In: *Analytical and bioanalytical chemistry* 401.1 (2011), pp. 167–181.

[25] Nicholas J DelRaso et al. “Air force research laboratory integrated omics research”. In: *Military Medicine* 180.suppl_10 (2015), pp. 67–75.

[26] James N Dodds and Erin S Baker. “Ion mobility spectrometry: fundamental concepts, instrumentation, applications, and the road ahead”. In: *Journal of the American Society for Mass Spectrometry* 30.11 (2019), pp. 2185–2195.

[27] Bruno Domon and Ruedi Aebersold. “Mass spectrometry and protein analysis”. In: *science* 312.5771 (2006), pp. 212–217.

[28] David L Donoho and Iain M Johnstone. "Adapting to unknown smoothness via wavelet shrinkage". In: *Journal of the american statistical association* 90.432 (1995), pp. 1200–1224.

[29] David L Donoho and Iain M Johnstone. "Threshold selection for wavelet shrinkage of noisy data". In: *Proceedings of 16th annual international conference of the IEEE engineering in medicine and biology society*. Vol. 1. IEEE. 1994, A24–A25.

[30] Pan Du, Warren A Kibbe, and Simon M Lin. "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching". In: *bioinformatics* 22.17 (2006), pp. 2059–2065.

[31] Warwick B Dunn et al. "Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics". In: *Metabolomics* 9.1 (2013), pp. 44–66.

[32] John B Fenn et al. "Electrospray ionization–principles and practice". In: *Mass Spectrometry Reviews* 9.1 (1990), pp. 37–70.

[33] Oliver Fiehn et al. "Metabolite profiling for plant functional genomics". In: *Nature biotechnology* 18.11 (2000), pp. 1157–1161.

[34] Oliver Fiehn et al. "The metabolomics standards initiative (MSI)". In: *Metabolomics* 3.3 (2007), pp. 175–178.

[35] Anna Floegel et al. "Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach". In: *Diabetes* 62.2 (2013), pp. 639–648.

[36] Peter E Freeman et al. "A wavelet-based algorithm for the spatial analysis of Poisson data". In: *The Astrophysical Journal Supplement Series* 138.1 (2002), p. 185.

[37] Erin Gemperline, Stephanie Rawson, and Lingjun Li. "Optimization and comparison of multiple MALDI matrix application methods for small molecule mass spectrometric imaging". In: *Analytical chemistry* 86.20 (2014), pp. 10030–10035.

[38] Eva Gorrochategui et al. "Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow". In: *TrAC Trends in Analytical Chemistry* 82 (2016), pp. 425–442.

[39] John M Gregoire, Darren Dale, and R Bruce Van Dover. "A wavelet transform algorithm for peak detection and application to powder x-ray diffraction data". In: *Review of Scientific Instruments* 82.1 (2011).

[40] Matthew P Greving, Gary J Patti, and Gary Siuzdak. "Nanostructure-initiator mass spectrometry metabolite analysis and imaging". In: *Analytical chemistry* 83.1 (2011), pp. 2–7.

[41] Jean-Luc Guerquin-Kern et al. "Ultra-structural cell distribution of the melanoma marker iodobenzamide: improved potentiality of SIMS imaging in life sciences". In: *Biomedical engineering online* 3.1 (2004), pp. 1–7.

[42] Xianlin Han and Richard W Gross. "Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: a bridge to lipidomics". In: *Journal of lipid research* 44.6 (2003), pp. 1071–1079.

[43] Kenneth Haug et al. "MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data". In: *Nucleic acids research* 41.D1 (2013), pp. D781–D786.

[44] Jiuming He et al. “MassImager: A software for interactive and in-depth analysis of mass spectrometry imaging data”. In: *Analytica chimica acta* 1015 (2018), pp. 50–57.

[45] Franz Hillenkamp et al. “Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers”. In: *Analytical chemistry* 63.24 (1991), 1193A–1203A.

[46] Demian R Ifa et al. “Desorption electrospray ionization and other ambient ionization methods: current progress and preview”. In: *Analyst* 135.4 (2010), pp. 669–681.

[47] Minoru Kanehisa et al. “KEGG for integration and interpretation of large-scale molecular data sets”. In: *Nucleic acids research* 40.D1 (2012), pp. D109–D114.

[48] Abu B Kanu et al. “Ion mobility–mass spectrometry”. In: *Journal of mass spectrometry* 43.1 (2008), pp. 1–22.

[49] Peter D Karp et al. “Expansion of the BioCyc collection of pathway/genome databases to 160 genomes”. In: *Nucleic acids research* 33.19 (2005), pp. 6083–6089.

[50] Hye Kyong Kim, Young Hae Choi, and Robert Verpoorte. “NMR-based plant metabolomics: where do we stand, where do we go?” In: *Trends in biotechnology* 29.6 (2011), pp. 267–275.

[51] Takashi Kobayashi et al. “A novel serum metabolomics-based diagnostic approach to pancreatic cancer”. In: *Cancer Epidemiology and Prevention Biomarkers* 22.4 (2013), pp. 571–579.

[52] Naomi L Kuehnbaum and Philip Britz-McKibbin. “New advances in separation science for metabolomics: resolving chemical diversity in a post-genomic era”. In: *Chemical reviews* 113.4 (2013), pp. 2437–2468.

[53] Jennifer E Kyle et al. “Uncovering biologically significant lipid isomers with liquid chromatography, ion mobility spectrometry and mass spectrometry”. In: *Analyst* 141.5 (2016), pp. 1649–1659.

[54] James N Kyranos et al. “High-throughput high-performance liquid chromatography/-mass spectrometry for modern drug discovery”. In: *Current opinion in biotechnology* 12.1 (2001), pp. 105–111.

[55] Eva Lange et al. “High-accuracy peak picking of proteomics data using wavelet techniques”. In: *Biocomputing 2006*. World Scientific, 2006, pp. 243–254.

[56] Michael A Lawrence and Maintainer Michael A Lawrence. “Package ‘ez’”. In: *R package version* 4.0 (2016).

[57] Rohan Lowe et al. “Transcriptomics technologies”. In: *PLoS computational biology* 13.5 (2017), e1005457.

[58] Roger S Macomber. “A complete introduction to modern NMR spectroscopy”. In: *Nova York* (1998).

[59] Marko A Makinen, Osmo A Anttalainen, and Mika ET Sillanpää. *Ion mobility spectrometry and its applications in detection of chemical warfare agents*. 2010.

[60] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

[61] Stephane Mallat and Wen Liang Hwang. “Singularity detection and processing with wavelets”. In: *IEEE transactions on information theory* 38.2 (1992), pp. 617–643.

[62] Jody C May et al. “Conformational ordering of biomolecules in the gas phase: nitrogen collision cross sections measured on a prototype high resolution drift tube ion mobility-mass spectrometer”. In: *Analytical chemistry* 86.4 (2014), pp. 2107–2116.

[63] Liam A McDonnell and Ron MA Heeren. “Imaging mass spectrometry”. In: *Mass spectrometry reviews* 26.4 (2007), pp. 606–643.

[64] Anna Meiliana, Nurrani Mustika Dewi, and Andi Wijaya. “Metabolomics: An Emerging Tool for Precision Medicine”. In: *The Indonesian Biomedical Journal* 13.1 (2021), pp. 1–18.

[65] Raanan A Miller et al. “A MEMS radio-frequency ion mobility spectrometer for chemical vapor detection”. In: *Sensors and Actuators A: Physical* 91.3 (2001), pp. 301–312.

[66] Owen D Myers et al. “Detailed investigation and comparison of the XCMS and MZmine 2 chromatogram construction and chromatographic peak detection methods for preprocessing mass spectrometry metabolomics data”. In: *Analytical Chemistry* 89.17 (2017), pp. 8689–8695.

[67] Gabe Nagy et al. “Distinguishing enantiomeric amino acids with chiral cyclodextrin adducts and structures for lossless ion manipulations”. In: *Electrophoresis* 39.24 (2018), pp. 3148–3155.

[68] Meritxell Navarro-Reig et al. “Untargeted comprehensive two-dimensional liquid chromatography coupled with high-resolution mass spectrometry analysis of rice metabolome using multivariate curve resolution”. In: *Analytical chemistry* 89.14 (2017), pp. 7675–7683.

[69] *NIST Chemistry WebBook,* https://webbook.nist.gov/chemistry/mw-ser/. Accessed: June,2022.

[70] Matthew B O’Rourke and Matthew P Padula. “A new standard of visual data representation for imaging mass spectrometry”. In: *PROTEOMICS–Clinical Applications* 11.3-4 (2017), p. 1600098.

[71] Nathan H Patterson et al. “Three-dimensional imaging MS of lipids in atherosclerotic plaques: Open-source methods for reconstruction and analysis”. In: *Proteomics* 16.11-12 (2016), pp. 1642–1651.

[72] Dominic S Peterson. “Matrix-free methods for laser desorption/ionization mass spectrometry”. In: *Mass spectrometry reviews* 26.1 (2007), pp. 19–34.

[73] Steven D Pringle et al. “An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument”. In: *International Journal of Mass Spectrometry* 261.1 (2007), pp. 1–12.

[74] JD Rabinowitz et al. “Metabolomics in drug target discovery”. In: *Cold Spring Harbor symposia on quantitative biology*. Vol. 76. Cold Spring Harbor Laboratory Press. 2011, pp. 235–246.

[75] Manfred Rauh. “LC–MS/MS for protein and peptide quantification in clinical chemistry”. In: *Journal of Chromatography B* 883 (2012), pp. 59–67.

[76] Sebastian Rauschert et al. “Metabolomic biomarkers for obesity in humans: a short review”. In: *Annals of Nutrition and Metabolism* 64.3-4 (2014), pp. 314–324.

[77] Guillaume Robichaud et al. “MSiReader: an open-source interface to view and analyze high resolving power MS imaging files on Matlab platform”. In: *Journal of the American Society for Mass Spectrometry* 24.5 (2013), pp. 718–721.

[78] Saer Samanipour et al. “Self adjusting algorithm for the nontargeted feature detection of high resolution mass spectrometry coupled with liquid chromatography profile data”. In: *Analytical chemistry* 91.16 (2019), pp. 10800–10807.

[79] Arun Kumar Shanker, Maduraimuthu Djanaguiraman, and Bandi Venkateswarlu. “Chromium interactions in plants: current status and future strategies”. In: *Metallomics* 1.5 (2009), pp. 375–383.

[80] Melanie M Sinanian et al. “Multivariate curve resolution-alternating least squares analysis of high-resolution liquid chromatography–mass spectrometry data”. In: *Analytical chemistry* 88.22 (2016), pp. 11092–11099.

[81] Pierre Soille et al. “Morphological image analysis: principles and applications”. In: 2.3 (1999).

[82] Bernhard Spengler and Martin Hubert. “Scanning microprobe matrix-assisted laser desorption ionization (SMALDI) mass spectrometry: instrumentation for sub-micrometer resolved LDI and MALDI surface analysis”. In: *Journal of the American Society for Mass Spectrometry* 13.6 (2002), pp. 735–748.

[83] Jeffrey M Spraggins et al. “High-performance molecular imaging with MALDI trapped ion-mobility time-of-flight (timsTOF) mass spectrometry”. In: *Analytical chemistry* 91.22 (2019), pp. 14552–14560.

[84] J-L Starck and Fionn Murtagh. “Astronomical image and data analysis”. In: (2007).

[85] Ragnar Stolt et al. “Second-order peak detection for multicomponent high-resolution LC/MS data”. In: *Analytical chemistry* 78.4 (2006), pp. 975–983.

[86] Lloyd W Sumner et al. “Proposed minimum reporting standards for chemical analysis”. In: *Metabolomics* 3.3 (2007), pp. 211–221.

[87] Ralf Tautenhahn, Christoph Böttcher, and Steffen Neumann. “Highly sensitive feature detection for high resolution LC/MS”. In: *BMC bioinformatics* 9.1 (2008), pp. 1–16.

[88] Sara Tortorella et al. “LipostarMSI: comprehensive, vendor-neutral software for visualization, data analysis, and automated molecular identification in mass spectrometry imaging”. In: *Journal of the American Society for Mass Spectrometry* 31.1 (2019), pp. 155–163.

[89] Victor Treviño et al. “GridMass: a fast two-dimensional feature detection method for LC/MS”. In: *Journal of Mass Spectrometry* 50.1 (2015), pp. 165–174.

[90] C Van Rijsbergen. *Information Retrieval (Book 2nd ed)*. 1979.

[91] Andrew T Walden and Roy E White. “Seismic wavelet estimation: A frequency domain solution to a geophysical noisy input-output problem”. In: *IEEE transactions on Geoscience and Remote Sensing* 36.1 (1998), pp. 287–297.

[92] Barbara Weyn et al. “Automated breast tumor diagnosis and grading based on wavelet chromatin texture description”. In: *Cytometry: The Journal of the International Society for Analytical Cytology* 33.1 (1998), pp. 32–40.

[93] Chalini D Wijetunge et al. "A new peak detection algorithm for MALDI mass spectrometry data based on a modified Asymmetric Pseudo-Voigt model". In: *BMC genomics* 16.12 (2015), pp. 1–12.

[94] D Willingham, A Kucher, and N Winograd. "Molecular depth profiling and imaging using cluster ion beams with femtosecond laser postionization". In: *Applied Surface Science* 255.4 (2008), pp. 831–833.

[95] Matthias Wilm. "Principles of electrospray ionization". In: *Molecular & cellular proteomics* 10.7 (2011).

[96] Alan Hb Wu et al. "Role of liquid chromatography–high-resolution mass spectrometry (LC-HR/MS) in clinical toxicology". In: *Clinical toxicology* 50.8 (2012), pp. 733–742.

[97] Xiaowei Zhang et al. "Omics Advances in Ecotoxicology". In: *Environmental Science & Technology* 52.7 (2018). PMID: 29481739, pp. 3842–3851. DOI: 10.1021/acs.est.7b06494. eprint: https://doi.org/10.1021/acs.est.7b06494. URL: https://doi.org/10.1021/acs.est.7b06494.

[98] Dao-Xiu Zhou, Yongfeng Hu, and Yu Zhao. "Epigenomics". In: *Genetics and Genomics of Rice*. Springer, 2013, pp. 129–143.

Glossary

List of Acronyms

LC-MS	Liquid Chromatography Mass Spectrometry
NMR	Nuclear Magnetic Resonance spectroscopy
HR-NMR	High Resolution Nuclear Magnetic Resonance spectroscopy
MS	Mass Spectrometry
SIMS	Secondary Ion Mass Spectrometry
DESI	Desorption Spray Ionization
MALDI	Matrix Assisted Laser Desorption Ionization
NIMS	Nanostructure Initiator Mass Spectrometry
ESI	Electrospray Ionization Technique
ESI	Electrospray Ionization
IM	Ion Mobility Spectrometry
IM-MS	Ion Mobility Spectrometry-Mass Spectrometry
IM-IMS	Ion Mobility Spectrometry-Imaging Mass Spectrometry
TIMS	Trapped Ion Mobility Spectrometry
CCS	Collision Cross Section
DTIMS	Drift Tube Ion Mobility Spectrometry
TWIMS	Travelling Wave Ion Mobility Spectrometry
DMA	Differential Mobility Analyzer
SNR	Signal-to-Noise Ratio
LDI	Laser Desorption Ionization
ESI	Electrospray Ionization
IMS	Imaging Mass Spectrometry
TIC	Total Intensity Count
IS	Internal Standards

CWT Continuous Wavelet Transform

ROI Region Of Interest

RT Retention Time

HPLC/MS High Performance Liquid Chromatography coupled to Mass Spectrometry

LC-HRMS High Resolution Mass Spectrometry coupled with Liquid Chromatography

UPLC/ESI-MS Electrospray Ionization based Mass Spectrometry coupled with Ultra High Performance Liquid Chromatography

FDR False Discovery Rate

DWT Discrete wavelet transform

WTM Wavelet transform maxima

WTMM Wavelet transform modulus maxima

SAFD Self Adjusting Feature Detection algorithm

ToF Time of Flight

EIT Extreme-ultraviolet Imaging Telescope