Unsupervised Learning Approach to Quality Control of Proteomics Studies

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

A dissertation submitted in partial fulfilment of the requirements for the degree Master of Science.

Author: Koena Monyai (672475)

Supervisor: Prof. Terence van Zyl

Co-Supervisor: Dr. Stoyan Stoychev

School of Computer Science and Applied Mathematics,
Faculty of Science,
University of the Witwatersrand, Johannesburg.

March 2023

Abstract

Liquid chromatography and mass spectrometry-based experiments are currently used in Proteomics studies to discover novel biomarkers that can be used in clinical settings and assist in combating diseases such as cancer. Although researchers and scientists are discovering large quantities of biomarkers, only a few are applied in a clinical setting. The experimental procedure of proteomics studies is complex and prone to faults. Therefore, the inadequate number of adopted biomarkers is due to faults occurring during the experimental procedure.

Quality control solutions have been implemented to increase the number of adopted biomarkers in clinical settings. Inconsistency is one of the significant hurdles in proteomics experiments as it causes the experimental results to be unreproducible. Reproducibility is necessary for every scientific study. Therefore, the majority of the tools monitor system consistency by identifying experiments or peaks indicating evidence of high technical variability. Consistency based quality control solutions require a large number of experiments, which are not always available or require substantial time to perform. Quality control solutions monitoring consistency can, therefore, result in delayed troubleshooting.

In our study, we present an unsupervised quality control solution to classify isotopic peak pairs based on their quality, i.e. classify low-quality and high-quality peak pairs. Our solution comprises of a peak detection technique, feature engineering and unsupervised classification of the peak pairs using clustering and feature selection. Our studies focus on identifying peak pairs deviating from the expected peak shapes and show evidence of interference. Comparisons of the clustering techniques are made to determine if the different clustering techniques can classify the peak pairs based on quality. The performance of clustering techniques can be negatively affected by the presence of irrelevant and redundant features. Therefore, we also evaluate if a genetic algorithm based feature selection improves the clustering results.

Our results show that using clustering techniques is likely to results in the misclassification of isotopic peak pairs as all the clusters contain both low-quality and high-quality peak pairs. Clustering techniques resulted in the majority of the low-quality data points being misclassified. Incorporating a feature selection technique to the clustering improved the overall performance of the techniques most notably significantly reduced the number of misclassified low-quality peak pairs. Before our solution can be employed as a quality control solution and adopted in the laboratories, we need to evaluate and optimise the peak detection and feature engineering steps.

Declaration

I, Koena Monyai (student number: 672475), declare that this research project is my own work, and all information sources have been referenced. Additionally, this work has never been submitted to any other university for any other degree.

Signed: 

Date: 02 March 2023

Acknowledgements

I would like to thank my supervisor, Prof. Terence van Zyl, for his supervision and guidance through my research journey. His input on both the content of this document, as well as input in the subject matter have been of great help and are much appreciated. I would also like to thank DR. Stoyan Stoyvech who helped me navigate the field of Proteomics by sharing his knowledge and always being available to clarify things for me. To my family; my parents, brother and sisters, thank you for being my pillar of strength throughout this journey. Lastly, I would like to thank my CSIR family and friends for their words of encouragement and for being my source of motivation.

Publications

Part of this work has been presented at the following conferences:

1. Monyai K, van Zyl T, Stoychev S. Peak Detection, Feature Extraction and Clustering of Peptides Fragments Ions. In2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI) 2019 Nov 19 (pp. 144-149). IEEE.

Contents

List of Figures

List of Tables

Abbreviations

HGP : Human Genome Project

HPP : Human Proteome Project

LC : Liquid Chromatography

MS : Mass Spectrometry

DDA : Data Dependent Acquisition

MRM : Multiple Reaction Monitoring

PRM : Parallel Reaction Monitoring

DIA : Data Independent Acquisition

CRAWDAD : Chromatogram Retention Time Alignment and Warping for Diffential Analysis of Data

AUC : Area Under Curve

QC : Quality Control

PCA : Principle Component Analysis

LoOP : Local Outlier Probability

CWT : Continuous Wavelet Transformation

SPC : Statistical Process Control

MAD : Median Absolute Deviation

DBSCAN : Density-based spatial clustering of applications with noise

GMM : Gaussian Mixture Model

AIC :Akaike information criterion

Chapter 1

Introduction

1.1 Background

One of the most significant achievements in biological research is the complete sequencing of the human genome through the Human Genome Project (HGP) Calderón-Celis et al. (2018). HGP came into inception in the mid-1980s and the motivation behind the project was the need to understand diseases such as cancer. Characterising the human genome resulted in many breakthroughs in biological and medical research. In biological research, the HGP project provided a comprehensive list of genes and through inference a comprehensive list of proteins and RNAs resulting in the emergence of systems biology Hood & Rowen (2013). For medical research, HGP helped pilot research focusing on disease risk associated with human genetic variations Hood & Rowen (2013), Munoz & Heck (2014).

Genomics studies were able to use inference to compile a protein and RNA list because all the bio-molecules are linked Hood & Rowen (2013). Genes contain information about the making of specific proteins. Genes in the form of chromosomes are located in the nucleus of the cell and proteins are formed in the cytoplasm and this is where the amino acids are located. Proteins are a chain of amino acids and the amino acids sequence tells us about the structure of that particular protein which in turn tells us about the function or role of that particular protein Lodish et al. (2000). As mentioned, genes have a genetic code needed in the formation of proteins. Cells obtain the genetic code through the transcription process. During this process, the cell produces an mRNA by separating a two-stranded DNA of a gene and binding the RNA polymerase to one of the DNA strands Hernandez (2001). After the mRNA is produced it travels from the nucleus to the cytoplasm and the translation process begins Noller (1991). In the translation process, ribosome, a macro-molecule responsible for building proteins reads

a codon to code a specific amino acid. A codon is made up of three nucleotide, DNA molecules are made up of two long polynucleotide chains. Ribosome travels through the mRNA and using the codons to creates an amino acid chain that represents the protein. The order in which the amino acids are linked determines the shape and the function of that particular protein Kindler et al. (2005).

Although genomics studies proved to be useful in biological research, it was evident that studies cannot help us understand the cellular process completely as other dynamic bio-molecules such as RNA, proteins and metabolites play roles in cellular processes Calderón-Celis et al. (2018). The need to understand other bio-molecules cultivated post-genomics studies such as proteomics. Proteomics, the study of the full protein complement at any given time point, provides functional annotation to the genome by characterizing protein expression and interactions that in turn drive cellular processes. Proteomics is thus become a valuable tool in predicting course of disease, as well as monitoring disease treatment effects Munoz & Heck (2014), Zhang et al. (2013).

Similar to genomics studies, researchers developed the Human Proteome Project (HPP) to characterise the full human proteome (Legrain et al. (2011) Kim et al. (2014)). Liquid Chromatography combined with Mass Spectrometry (LC-MS) has become the most popular tools in studies involving the characterisation of the proteins in given biological samples (Lill (2003) Ioannidis & Bossuyt (2017)). The reason LC-MS gained popularity due to the ability to identify thousands of known and unknown proteins Zhang et al. (2013). Although researchers have yet to obtain sequencing of the full proteome, proteomics studies have proved to be useful in other applications such as clinical research (Kim et al. (2014) Wang et al. (2016)). Quantitative proteomics has found a way in clinical practice as a form of novel bio-marker discovery and verification technique. Quantitative proteomics is used to determine both the identities and quantities of detected proteins in a biological sample Calderón-Celis et al. (2018). A bio-marker is a reporter bio-molecule, for example a protein,that indicates a transition from "normal" to an "abnormal" state and can thus be used to detect onset and progression of a disease. Novel bio-markers are needed to assist with early detection, accurate prognosis and personalised as well as effective treatment (Wang et al. (2016) Roointan et al. (2018)).

Proteins have four levels of structure, primary, secondary, tertiary and quaternary structure Branden & Tooze (2012). The primary structure is the amino sequence obtained during the translation process Kindler et al. (2005). The secondary structure depends on hydrogen bonding, the bonding of amino hydrogen and carboxyl oxygen atoms of the peptides or polypeptides. The two common types of secondary structures are alpha-helix and b-sheet. The tertiary structure is the three-dimensional shape the protein forms to reach maximum stability at minimum energy Richardson (1981), Kindler et al. (2005).

Stability of the protein is commonly achieved by burying the non-polar amino acids side chains insides of hydrophilic acidic or basic amino acid side chains. The interactions of the polypeptides of the proteins make up the quaternary structure and can result in a larger aggregate protein complex Richardson (1981).

It is common to digest proteins into peptides when performing analytical studies such as Proteomics due to the complexity of the protein structure. Through digestion, the secondary, tertiary and quaternary structures fall away and Proteomics studies only focus on the primary structure of the proteins which the amino acid sequence of the peptide Zhang et al. (2013).

1.1.1 Proteomics Experiments Procedure

There are two strategies mainly followed in proteomics experiments: a bottom-up approach and a top-down approach. In the top-down approach, we analyse the actual proteins and in a bottom-up approach, peptides are extracted from the proteins and analysed instead (Zhang et al. (2013) Bittremieux, Tabb, Impens, Staes, Timmerman, Martens & Laukens (2017)). Bottom-up experiments are more popular since peptides ionize more efficiently than proteins and are also easier to partition chromatographically Zhang et al. (2013). There are three stages in a proteomic experiment; the pre-analytical, analytical and post-analytical stages. Sample collection and sample preparation make up the first stage; the analytical stage involves the characterisation of the proteins and post-analytical is the actual processing of raw spectra from the second stage to a list of all the detected proteins Feist & Hummon (2015). Figure 1.1 shows the bottom-up proteomics experiment procedure.

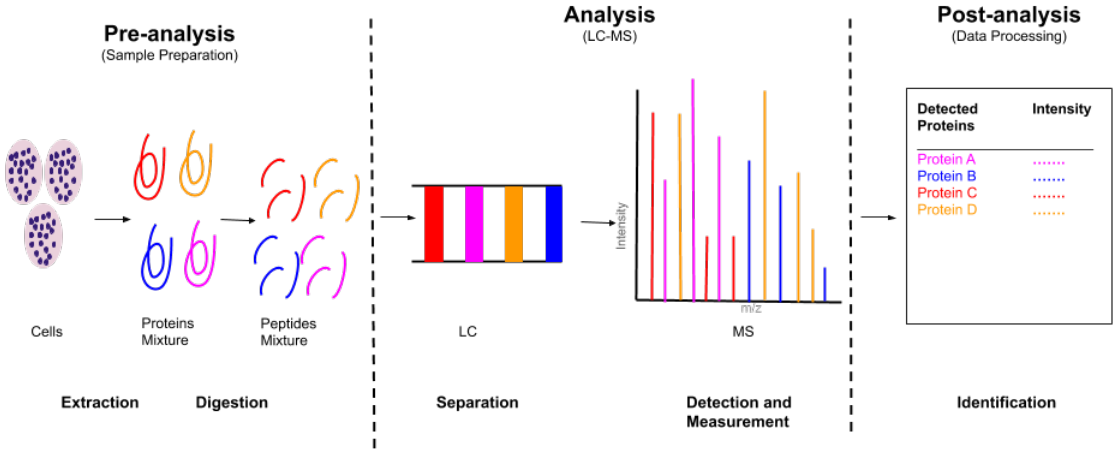


FIGURE 1.1: Bottom-up proteomics experimental procedure

1.1.1.1 Sample Collection and Preparation

Sample collection, which is the first step of any proteomics study includes harvesting the sample from subjects. The sample harvesting procedure can either be invasive like biopsy or drawing blood, or non-invasive such as collecting a urine sample Feist & Hummon (2015). The sample goes through a preparation process to make it suitable for analysing instruments. Proteins are isolated from other molecules through lysis and extracted from the biological sample Zhang et al. (2013). In bottom-up proteomics peptides are generated from the proteins via enzymatic digestion (proteolysis) and inference is used to identify the proteins in the data processing stage Zhang et al. (2013), Bittremieux, Tabb, Impens, Staes, Timmerman, Martens & Laukens (2017). The removal of contaminants during preparation is vital as contaminants are likely to interfere with chromatography, affect the ionisation efficiency and contaminate the mass spectrometry system Feist & Hummon (2015). A proteomic experiment can either be label-free or labelled and this will determine how the proteins are quantified. In the case of the labelling, a specific amount of the isotopic labelled analogue compound is added to the biological sample. Isotopic labelling results in a pair of light and heavy peptide isotopes Ong et al. (2002).

1.1.1.2 Liquid Chromatography

The output of preparation stage is a complex heterogeneous mixture of peptides which is not favourable for the analytical systems Boersema et al. (2008), Tabb et al. (2009). To ensure quality analysis of the peptides by the mass spectrometry system, liquid chromatography(LC) uses a column to separate peptides by their physio-chemical properties such as charge or more commonly hydrophobicity Tabb et al. (2009). An LC column has a stationary phase and a mobile phase. Peptides adsorb to the stationary phase and the purpose of the mobile phase is to move the peptides through the column Shi et al. (2004). In reverse phase, hydrophilic peptides are easier to move through the column, and hydrophobic peptides tend to have a stronger bond with the stationary phase and are more difficult to move or elute through the column Boersema et al. (2008). The time it takes the peptide to elute through the LC is therefore influenced by its hydrophobicity. The duration of the elution process is referred to as the peptide's retention time and is crucial information needed in the data processing stage. The separation process not only simplifies the peptides' mixture for ease of analysis but also introduces a time dimension to the results Premstaller et al. (2001). A Gaussian function can be use to estimate the shape a peptide's elution profile is likely to follow Zhang et al. (2014).

1.1.1.3 Mass Spectrometry

The purpose of the analytical stage is to analyse all the peptides by using the mass spectrometry (MS) to measure their mass-charge ratios (m/z) and intensities. The components of the MS responsible for the analysis are ion source, detector and mass analyser Karpievitch et al. (2010). The ion source charges or ionises eluting peptides from the LC. The charged peptides are called precursor ions. The peptides are ionised because the MS uses an electromagnetic field to move ions through the system Zhang et al. (2013). The m/z of the precursor ions is measured using the mass analyser and the detector keeps track of the number of the different precursors. Most proteomics studies use tandem MS where specific precursor ions are not only analysed but fragmented, resulting in fragment ions. Fragment ions go through the analytic process and are analysed and quantified using the second MS's analyser and detector. Tandem proteomics provides more analytical specificity than using a single MS Karpievitch et al. (2010).

The analysis of the peptides takes place in several cycles. A single cycle consists of the m/z and intensities of either the selected precursor ions, fragment ions or both. As mentioned, only specific precursor ions are fragmented, and the selection of these ions depends on the MS technique used. The first MS technique used in proteomics was Data Dependent Acquisition (DDA). In DDA, the N top intense ions are fragmented Michalski et al. (2011). The shortfall of DDA is the exclusion of precursors with low intensity, and the selection of precursor ions is stochastic, making it challenging to reproduce the results Doerr (2014). Targeted Selection, e.g. Multiple Reaction Monitoring (MRM), improved on the specificity of DDA experiments by allowing the technician to specify the m/z precursor ions of interest and only those ions are fragmented Bereman et al. (2012). Although targeted selection improved the specificity of the MS, it is only useful if the technicians know which peptides are of interest, thus the introduction of Data Independent Acquisition (DIA). DIA improved on DDA while still maintaining the specificity of the targeted monitoring by analysing all the precursor ions. This is possible because in DIA a wider precursor transmission window is applied in comparison to DDA. This removes the necessity to pre-select precursors one at a time as DDA does Tsou et al. (2015), Doerr (2014). The output of the analytic stage is a raw spectral of measurement obtained during each cycle; this includes the m/z, intensity and time.

1.1.1.4 Data Processing

Data processing stage converts the raw spectrum into meaningful information containing a list of all the detected peptides and their quantities. There are different mass spectrometry tools with different output files formats. Most of data processing software tools

do not have the capabilities to process all the different vendors' file formats. Therefore, raw spectra is converted into open file format (XML) before data processing Karpievitch et al. (2010), Deutsch et al. (2010), Kessner et al. (2008). Although the data processing tools are different, there are generic steps amongst the tools such as chromatogram extraction, peak detection and peptide identification. Most tools start by first extracting the chromatograms, retention time and the intensity, for each m/z. The chromatograms undergo re-sampling to ensure that the time intervals are consistent. Peaks are extracted from the chromatograms using Chromatogram Retention Time Alignment and Warping for Differential Analysis of Data (CrAWDAD) Pino et al. (2017). A peptide is represented by multiple peaks, which are a precursor ion peak, isotope ion peak; in the case the isotopic labelling; and fragments ions peaks. Before identification, we do not know which peaks belong to which peptides. Therefore, peaks with the same start, apex and end times are grouped. The groups are matched against databases or libraries containing the sequences of the peptides, i.e. all m/z and retention time (time at apex) of all the fragment ions of that particular peptide Deutsch et al. (2010), Pino et al. (2017). Tandem MS data is noisy and not all peak groups represent an actual peptide. A distinct peptide will match with several peak groupings. To determine the correct peak grouping, the top ten groups are evaluated using shapes, co-elutions score, library intensity and log intensity to determine the probability of them representing an actual peptide. The peak grouping with the highest score is selected Pino et al. (2017) as a 'true' group. In the case of the labelled proteomics data, the ratio of the labelled peptide and unlabelled peptide is used to quantify the sample. In a label-free experiment, we use area under the curve (AUC) quantification method to quantify the peptides Kessner et al. (2008). Each protein contains proteolytic peptides; these are unique peptides in each protein and are used to determine the proteins found in the sample Parker & Borchers (2014).

1.1.2 Errors in the Experiments

The bottom-up approach is a complex and an error-prone procedure. Errors affect the reliability of the results of the experiments. The process starts at sample collection; the subjects should be selected such that biological variation is reduced. This is because the variation can be mistaken to be of biological significance. Incorrect storage of the samples results in degradation Goebel-Stengel et al. (2011). Regulating temperature, pressure and utilizing the suitable type of buffers used during sample preparation is essential; inappropriate setting can alter properties of the peptides. Sample loss and contamination in the sample preparation stage can also result in poor results Bittremieux, Tabb, Impens, Staes, Timmerman, Martens & Laukens (2017), Karpievitch et al.

(2010). Protocols are defined to reduce errors associated with the sample preparation stage Kuzyk et al. (2010).

LC introduces the most variability. A good peak shape shown in Figure 1.3, indicates good separation. Overloading the column results in the tailing the peaks, i.e. skewing to the left. Elution of the stationary phase causes a sample bleed and fronting peaks, i.e. peaks tend to skew to the right Rudnick et al. (2010). Skewed peaks indicate that the column needs replacement Bittremieux, Tabb, Impens, Staes, Timmerman, Martens & Laukens (2017). Loading the LC with insufficient sample results in a low signal to noise ratio and a chromatogram noisier than the chromatogram indicated in Figure 1.2 Bielow et al. (2015). Improper connection and cleaning of the column results in dead volumes and carryovers from previous experiments interfering with the current sample causing peaks to broaden, prolonged elution times and co-elution Noga et al. (2007). The MS ion spray instability results in jagged peaks and premature fragmenting Rudnick et al. (2010). Interference in the ion channel results in dissimilarity in fragments and isotopic ions of the same peptide Percy et al. (2014) and deviate from the expected shape demonstrated in Figure 1.4 and Figure 1.5. Table 1.1.2 shows some of the peaks deviating from expected shapes.

There is still no standard protocol for data processing. The analytical stage can produce high-quality results, but if the settings of the database search are incorrect, the quality of the final results will be of affected Bell et al. (2009). The different data processing tools use databases that are different in curation, completeness and comprehensiveness and the search engine makes different assumptions. The data processing stage is easier to standardise, once the optimal settings are obtained, the maintenance of the software is easier than maintaining the experiment setup.

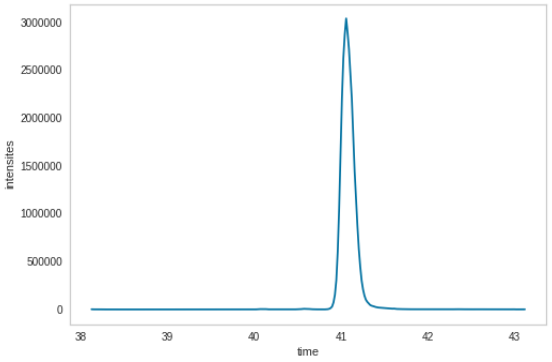


FIGURE 1.2: Good Peak Chromatogram

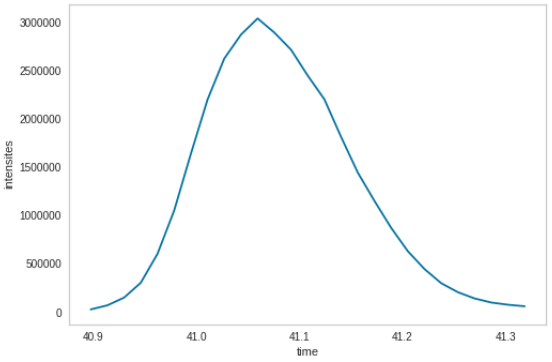


FIGURE 1.3: Good peak

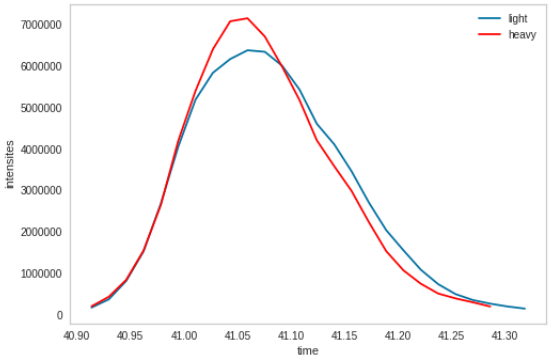


FIGURE 1.4: Isotopes Peaks

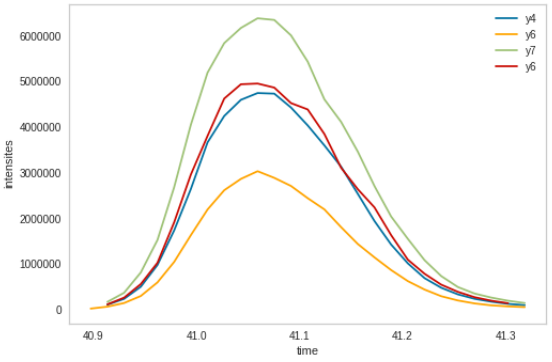
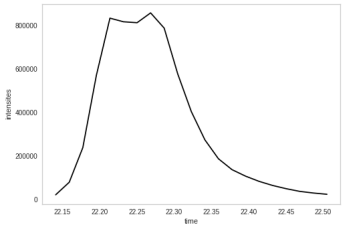
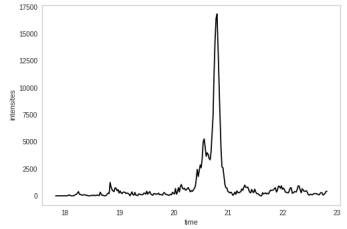
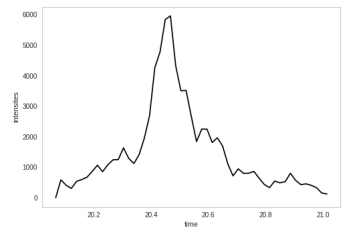


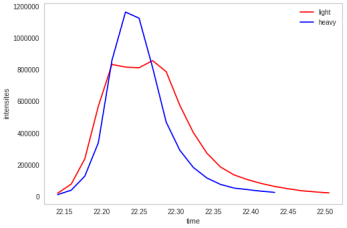
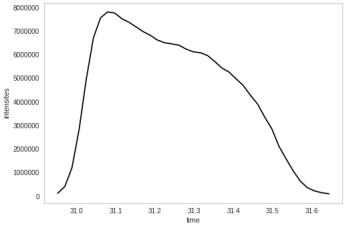
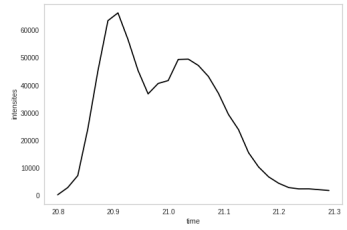
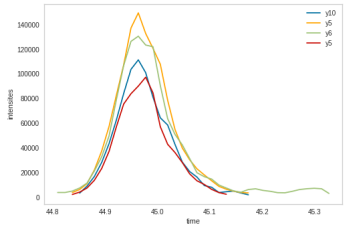
FIGURE 1.5: Fragments Peaks

TABLE 1.1: Problems related to the analysis stage

Problem	**Stage**	**Cause**	**Peak**
Tailing	LC	Overloading	
High Background	MS	Electronic Noise	
Jagged	MS	Ion Spray Instability	

Continued on next page

Table 1.1 – *Continued from previous page*

Problem	**Stage**	**Cause**	**Peak**
Isotopic Similarity	MS	Interference	
Shoulder	LC	Void in the column	
Bimodal	LC	Coelution	
Fragments Dissimilarity	MS	Interefence	

1.2 Problem Statement

Genomic and post-genomic studies aim at understanding complex diseases such as cancer Aslam et al. (2017). Proteomics studies proved to be useful in cancer research and have achieved an increase in life expectancy for the patients. Although there is progress in cancer diagnoses and treatment, there is still a need for novel bio-markers to reduce the fatality rate of cancer. The novel bio-markers are needed to help with early diagnoses, prognosis, understanding the risk of infection associated with different ancestry groups, the introduction of personalised and effective treatment and therefore, the reduction of the fatality rate of cancer Parker & Borchers (2014).

Research involving the discovery of the novel bio-markers gained popularity in the last decades and there is an increase in the number of bio-markers proposed in literature Milardi et al. (2019). Every bio-marker reported needs to be taken through the validation and verification processes for FDA approval purposes before they can be applied in clinical settings. Most of the bio-markers fail these processes and therefore do not get the necessary approval for clinical use. Therefore, one major problem researchers are trying to solve in proteomics studies is to increase the number of reported bio-markers that are FDA approved. Errors associated with the proteomics experimental procedure contribute to the low number of approved bio-markers. These errors include insufficient sample quantities and improper sample storage during the preparation stage; LC or MS malfunction and flawed data processing techniques. Errors occurring during the pre-analytical stage have been reduced through the standardisation of sample preparation processes Feist & Hummon (2015), Bodzon-Kulakowska et al. (2007), Visser et al. (2005). The standardisation protocols of the analytical stage have yet to be defined.

The need to eliminate errors occurring during the proteomics experiments has given rise to the formation of quality control (QC) forums. Incorporating the proper form of quality control will increase the number of bio-markers that are FDA approved. The driving force behind the QC forums was to improve the consistency of the experiments. Reproducibility is essential in all scientific studies. Inconsistency of proteomics experiments results in unreproducible results Bereman et al. (2014). Technical variabilities within the procedure were identified as the root cause of the inconsistencies Pichler et al. (2012). Therefore, QC solutions focused on detecting, monitoring and reducing technical variability within the experimental setup. The first problem with technical variability based QC solution is that there are other sources of variability like biological and systematic factors. The second problem is that to be able to detect variability, comparison of technical replicates should be made to detect any variations in the results and effect of technical variabilities. Completing the required replicates can take several months, delay the troubleshooting of the experiment, and results in wastage of the biological

sample Bittremieux, Tabb, Impens, Staes, Timmerman, Martens & Laukens (2017) . The last problem is that low technical variability amongst the replicates is not always an indication of an optimised experimental procedure and high-quality results Choi et al. (2017).

1.3 Research Objectives

The problem statement indicates we need QC tools that focus on monitoring the quality of the experimental results. Therefore, the key objective of our work is to develop a QC solution that focuses on quality and not technical variability. There is no defined quality measure for the raw MS spectrum; however, from Section 1.1.2, we know the representation of a high-quality peak and the quality of a peak is dependent on how the peak deviates from expected shape. Our specific aim is to develop a solution that will assess the quality of isotopic peak pairs in a given experiment using unsupervised learning techniques. Because of limited expert annotations, an unsupervised approach is more feasible than a supervised approach. The QC solution is made up of two steps, data pre-processing and the actual unsupervised classification of the peak pairs.

1.3.1 Data pre-processing

The purpose of data pre-processing is to transform the data such that it is suitable for the classification stage. The data pre-processing stage is vital to the success of the solution as incorrect pre-processing will have negative impact on the classification algorithms' performances. Firstly, the peaks need to be extracted from chromatograms. Like the full MS spectrum, there is no measure of high-quality for a chromatogram. Furthermore, peaks are transformed to feature space through feature engineering. Feature engineering reduces each peak into a numerical vector and each numerical value represents a feature or an attribute. This step is necessary as we can define features that provide us with information about the peaks' quality. In our study, a data instance is made up by merging the features of light and heavy isotopic peak pairs.

The sub-objectives of the data processing stage are:

- to develop a peak detection technique, and
- to engineer quality-based features for the peaks and merge isotopic pairs.

1.3.2 Unsupervised Classification

After data pre-processing, feature representations of the peak pairs are presented to clustering algorithms. Clustering algorithms perform unsupervised classification by determining the natural grouping of the data points indicated by maximised similarity within the clusters and maximised dissimilarity amongst the clusters. The objective of the unsupervised classification is to group the peaks pairs such that high-quality isotopic pairs and low-quality peak pairs are in separate clusters. Clustering algorithms assume that all the features are essential in the correct classification of the peak pairs; however, redundant and irrelevant features affects the performance of the algorithms. Feature selection techniques can help improve the clustering results.

The sub-objectives the of classification stage are:

- to evaluate the performance of the different clustering techniques, and
- to determine if incorporating a feature selection solution to our clustering techniques can enhance their performances

1.4 Research Questions

Clustering techniques make different assumptions and use different similarity measures to perform clustering. The performance of clustering techniques is also dependent on the data structure. In our study we use k-means, agglomerative, fuzzy-c-means (FCM), Gaussian mixture models (GMM) and Density-based spatial clustering of applications with noise (DBSCAN) to classify the Proteomics isotopic pairs based on their quality. To determine the clustering technique suitable for classifying the isotopic pairs, we compare the performance and this lead us to our first research question:

1. Which clustering technique perform best in classifying isotopic pairs?

One of the factors that can hinder the performance of the clustering techniques is the inclusion of irrelevant features. We incorporate a genetic algorithm feature selection technique to determine the relevant features and improve the clustering performance. To determine if the feature selection technique improves the performance of the clustering techniques, we use our study to answer the following question:

2. Which of the clustering results improved the most by adding feature selection?

The selected features can also assist analysts to troubleshoot the experiment system as the features are quality attributes of the isotopic pairs. Therefore, analysts can use the features to determine a step in the whole experimental procedure that is performing sub-optimally. Using the experts notes regarding the issues with every low-quality pair we can evaluate our if feature selection technique selects features that related to the specified issues. This leads us to our next research question:

3. Which are the prominent features selected by the different algorithms?

The main aim of our solution is to develop a quality control solution to be used in laboratories by analyst to determine which isotopic pairs are high-quality and can be used in further studies and which pairs need to be re-evaluated. Therefore, the overall question that our study aims to answer is as follows:

4. Is using unsupervised learning as a quality control tool for proteomics experiments a viable solution?

1.5 Delineation, Assumptions and Limitations

Below are some of the limitations, delineations and assumptions made in our work;

- The work assumes that only analytical errors are responsible for low-quality peak pairs; therefore, the attributes will be focusing on detecting analytical errors.
- The data is from MRM isotopic labelled experiments. Therefore we do not know how well our approach works with other proteomics experiments, DDA, DIA and SWATH or labelled-free data. To our knowledge, the data used in our study are the only publicly available labelled data to date.
- There is an extensive list of different clustering techniques and it is not feasible to compare all the techniques. Our work compares Gaussian Mixture Model, k-means, agglomerative Clustering, fuzzy c-means and DBSCAN.

1.6 Research Contribution

The main contribution of our work is an unsupervised quality assessment solution for isotopic peak pairs; the contributions are as follows:

- implementation of a peak detection technique that will extract the peaks from the original chromatograms,
- engineering of quality-based attributes of the peaks,
- comparison of different clustering techniques to determine which techniques suitable for classifying peak pairs, and
- implementation of a feature selection technique to improve the clustering.

1.7 Thesis Overview

The rest of the thesis is as follows; chapter 2 provides a literature review of quality control in proteomics, various clustering techniques and selection methods. Chapter 3 presents our research methodology. Chapter 4 gives the results of our studies and it also includes the analysis and discussion of our findings. Our conclusion and future works are presented in chapter 5 and chapter 6 respectively.

Chapter 2

Background and Literature Review

Researchers performing proteomics experiments realise the importance of in-depth quality control; therefore, they have implemented various quality control solutions to improve the quality of the experimental results Bittremieux, Walzer, Tenzer, Zhu, Salek, Eisenacher & Tabb (2017). In this section, we will be reviewing the different quality control approaches in proteomics and related fields found in the literature. We will also be discussing the background of clustering and unsupervised feature selection techniques we use in our work.

2.1 Literature Review

Various forums and groups have been established to standardize QC in proteomics studies Ivanov et al. (2013) Bittremieux, Walzer, Tenzer, Zhu, Salek, Eisenacher & Tabb (2017) Bell et al. (2009). QC solutions focus on assessing the quality of experiments, experiment-level QC, or the quality of distinct peaks of an experiment, peak level QC Eshghi et al. (2018). The solutions rely on the data-processing stage and integrate peptides identifications to the solution or are independent of the data processing stage and perform QC on raw files. There is no universal definition of high-quality experiment or high-quality peak. Since consistency is a major challenge in proteomics experiments, most of the tools choose to associate consistency with high-quality Rudnick et al. (2010). As a results most tools focus on monitoring if the experiment results obtained during a specified period of time are consistent Bramwell (2013) Bramwell (2013).

2.1.1 Quality Control at Experiment Level

Initially using quality control samples was the common approach in monitoring the experimental system's performance. A QC sample is a mixture of known peptides with specific quantities Bittremieux, Tabb, Impens, Staes, Timmerman, Martens & Laukens (2017). The samples are either interleaved between actual samples or spiked into an experiment sample Bereman et al. (2012). To assess the experimental performance, laboratory technician compares the expected and actual results, raw spectra, of the QC samples. A significant difference between the measured and expected results indicates a sub-optimal analytical performance. The technician is then required to identify the root cause of the error Bittremieux, Tabb, Impens, Staes, Timmerman, Martens & Laukens (2017) Rudnick et al. (2010). This form of QC requires manually monitoring of the results, which is time-consuming and results in quality control depending on the technician's discretion as the technician decides which differences of results are alarming and require a cause of action. The number of identified peptides is another direct measure of system consistency, for experiments analysing the same sample; ideally, the results of the experiments should contain the identical number of correctly identified peptides and a significant decrease in that particular number indicates lack of consistency Rudnick et al. (2010). Although a differences number of peptides indicates lack of repeatability, it does not tell us what went wrong in experiment Pichler et al. (2012).

An experiment-based QC solution identifies experiment showing evidence of high technical variability in a group of experiments analysing the same sample, i.e. an experiment with different results from the rest experiments Bramwell (2013). Proteomics experimental data is notably complex; a single experiment can contain in excess of ten thousand peaks Michalski et al. (2011). Feature engineering is a vital part of reducing the complexity of the data and capturing critical information about the consistency of the experiments. Quality control solutions are, therefore, divided into two stages: engineering features to monitor consistency and analysing the features using statistical or machine learning techniques to determine which experiment results demonstrate evidence of high technical variability.

Clinical Proteomic Tumor Analysis Consortium (CPTAC) and National Institute of Standard and Technologies (NIST) performed multi-laboratories studies with the intention of defining features that can capture both intra-laboratory and inter-laboratory variabilities. Experiments in the studies involved introducing controlled perturbations as sources of variability. Forty-six features were empirically derived to capture the differences in the results of the experiments. These features depend on the data-processing stage and can be used to detect variabilities due to poor chromatography, ionisation, dynamic sampling and differences in precursor ions spectrum; MS spectra and fragment

ion spectra; MS2 spectra Rudnick et al. (2010). In the chromatography stage, the features can be used to monitor differences in the retention time of the identified peptides, the time it takes for 50% of the identified peptides to elute, the order the peptides elute in, and the median, first and tenth deciles and the interquartile of the peaks' widths. The dynamic sampling based features are used to capture the stochastic nature of DDA, monitor oversampling and to ensure that all the identified peptides are sampled at maximum intensity. To monitor electrospray instability features that capture changes in the total ion current of the cycles and m/z of the median peptide were defined. During the studies, it was discovered that MS and MS2 spectra intensities differ significantly; however, the root cause of the differences was not determined. The following features were defined to capture the differences: the median total ion current, maximum and median intensities per cycle and the injection time, which is the time the peptides are recorded by the MS. The features defined by NIST are dependent on the identification of the peptides and proteins and the data processing stage can introduce other variabilities and the features can only be extracted for a single MS technology. Ma et al. (2012) developed an improvement to NIST solution by developing a feature extraction tool, Quameter, that caters for a wider range of mass spectrometry workflows. Quameter can be used to extract features that are independent of the identification information for any MS technology. The dimensionality of the raw files is reduced by taking the first, second and third quartiles of the total ion current, peaks' width and retention times as features . Bittremieux et al. (2015) defined a new set of features that do not rely on the experimental results but the instrument sensors. These features capture the configuration parameters and status log information. Instrument based features only monitor the performance of the MS and exclude the LC.

Although the features can assist in identifying sources of variability in the complex system, there are not easy to interpret and might require expert knowledge to identify experiments exhibiting proof of technical variability Bittremieux, Tabb, Impens, Staes, Timmerman, Martens & Laukens (2017). Taylor et al. (2013) developed a tool to visualise the NIST features and the lab technicians can monitor them over time and see if there is any sudden and unexpected deviation from the usual performance. Visualisations are a crucial tool in quality control. However, they rely on the lab technician discretion on acceptable variabilities amongst the experiments and require full-time monitoring. Pichler et al. (2012) used control charts on 19 of the 46 Quameter features to perform performance test for complex samples. To determine the experiments that are out of control, i.e. experiments showing evidence of significant technical variability, the median absolute deviation (MAD) of each feature is calculated. An experiment with a feature that is not within the bounds $\pm 2 \times MAD$ is classified as being out control, in

other words, the experiment demonstrates high technical variability. Proteomics workflow stages are related; therefore, uni-variate analysis of each feature does not capture inter-relation of the system and correlation of the features.

Wang et al. (2014) introduce using multivariate statistics to analyse Quameter features and capture the correlation of the features. They use robust Principal Component Analysis (PCA) for feature reduction. To determine out-of control experiments the L1 distance from principal components to the multivariate median is calculated. Any data point with L1 distance out of bound $\pm 3 \times IQR$ is classified as an erroneous experiment. Although the use of the PCA reduces the complexity of the data, it can result in the omission of important information Chandola et al. (2009), Latecki et al. (2007). The limitation of statistics techniques resulted in the introduction of machine learning techniques to analyse the NIST or Quameter features. Amidan et al. (2014) used logistic regression model, which they have termed Lasso Logistic Regression Classier (LLRC), on features extracted from NIST and Quameter to detect if a data instance shows proof of high-variability. An experiment x_i is considered flawed if it's probability $\pi(x_i)$ is greater than a threshold τ

$$\pi(x_i) = \frac{e^{g(x_i)}}{e^{g(x_i)} + 1}, \tag{2.1}$$

and

$$g(x_i) = \beta_0 + \beta_1 x_i \tag{2.2}$$

where $g(x_i)$ is the linear component of the logistic regression, β_0 and β_1 are the intercept and the coefficient of x_i, which is a feature representation of an experiment. β_0 and β_1 are obtained using lasso, which obtains the best coefficient by maximising the penalised log-likelihood function. Regularisation is used to avoid over-fitting. τ and regularisation are obtained by minimising loss that penalises misclassified flawed experiments. Experiments demonstrating high technical variability, i.e. erroneous experiments, have more detrimental consequences than misclassified high-quality experiments. All the experiments classified as poor quality are subjected to further evaluation by a laboratory technician for troubleshooting. Although LLRC manages to achieve high sensitivity and specificity, the model requires expert annotated data for training. Obtaining annotated data is be time-consuming. To date, the only publicly available annotated datasets for experiment-based QC are from the PNNL studies Wang et al. (2014).

Bittremieux et al. (2016) use an unsupervised anomaly detection technique to determine anomalous experiments. Anomalies are experiments showing evidence of high technical variability. The solution relies on Quameter for feature extraction and Local Outlier Probability (LoOP) to determine anomalies . LoOP uses of nearest neighbour concept to determine anomalous data points Bittremieux et al. (2016). LoOP is a local anomaly detection technique as it uses probability set distances as outlier scores instead of the

average distances Goldstein & Uchida (2016), Chandola et al. (2009). The probability sets distances are obtained by assuming a half-Gaussian distribution of the distances Bittremieux et al. (2016), Kriegel et al. (2009). The solution also uses random forest to perform feature selection by subspace separability. Feature selection allows is used to validate if an experiment is a true anomaly and identify features responsible for the classification. One of the constraints of unsupervised learning is that misclassification of data instances is unavoidable Amidan et al. (2014).

2.1.2 Quality Control at Peak Level

As mentioned a proteomic experiment can contain up to ten thousand peptides chromatograms, therefore compressing raw spectra into quartiles and 46 features result in an excessively compressed representation of the raw spectra and insufficient information about the quality of the whole experiment. There is no definition of a high-quality chromatogram. Therefore, the peaks are assessed instead. Peak-level quality control solution is made of three stages: peak detection, feature engineering and classification of good and bad quality peaks.

2.1.2.1 Peak Detection

A precursor or fragment ion chromatogram is made up of three components namely: the actual ion peak, signal offset and a baseline Du et al. (2006). The peak detection problem involves the extraction of peaks by identifying the start and endpoint of peak or peak boundaries Finney (2012). Errors in peak detection affect the quality of the peak. Majority of the tools that perform peak quality assessment rely on Skyline for peak detection Eshghi et al. (2018), Dogu et al. (2018). Skyline is a data processing platform; it performs chromatograms extraction from the raw file, re-sampling, peak detection, peak grouping and peptide identification Pino et al. (2017). Skyline uses Chromatogram Retention Time Alignment and Warping for Differential Analysis of Data (CRAWDAD) to perform peak detection. Points, where the first derivative is equal to zero, are used to obtain the maxima and minima points of the chromatogram. The second derivative is used to find inflection points. The maxima, minima and inflection points make up the peak Finney (2012), Pino et al. (2017). Manual inspection of the peaks detected using Skyline platform indicates that the might be a need for manual correction of the detected boundaries. Incorrect boundaries are the result of the unevenness of the chromatograms which is demonstrated by multiple local extrema (minima and maxima) Pino et al. (2017). Figure 2.1 shows all the detected local minima and maxima of a single chromatogram.

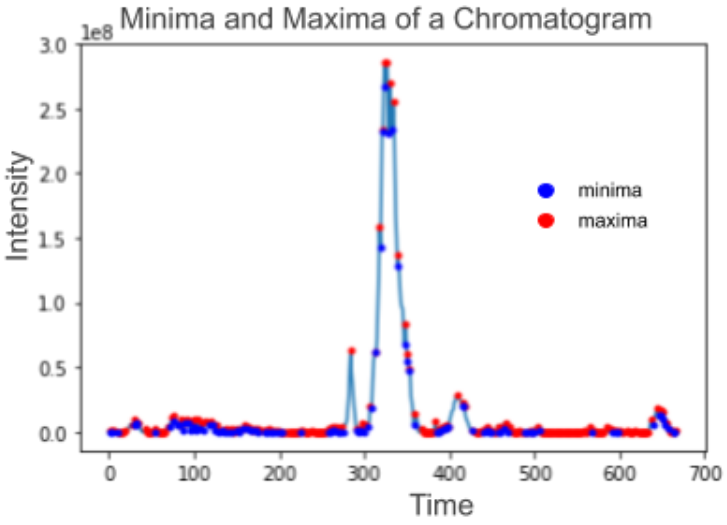


FIGURE 2.1: Local minima and maxima of a chromatogram

Aoshima et al. (2014) developed a more straightforward approach for peak detection, AD3D, to reduce the processing time as peak detection of more than ten thousand peaks can be a time-consuming task. Their approach was to identify the peak apex using global maximum. The peak apex of a real peak has to be above a specified threshold, and a horizontal line drawn at 50% of peak height should intercept the peak twice, on the left and right side of the peak apex. AB3D chooses the local minima closest to the peak apex with intensities less than a specified threshold as peak boundaries Aoshima et al. (2014).

AD3P assumes that only peaks with high intensity represent actual precursor ions or fragments ions and low-intensity peaks are labelled as noise. Using local extrema is not ideal; the jaggedness of the peaks can result in many local minima, determining which local minima represent peak boundaries becomes a challenge. A genuine LC-MS peak possesses specific characteristics and shape. Du et al.(2006) take a different approach from AB3D and CRAWDAD to detect metabolomics peaks. Instead of relying on extrema to identify the peak boundaries and apex they used continuous wavelet transformation (CWT) to transform the chromatogram from time and intensity space into a three-dimensional space made up of the scale, time and CWT coefficient. Choosing the right wavelet, similar to the expected peak shape, is essential and transforming the chromatogram into its wavelets space, makes it easier to identify the ridgelines. The position of the ridgeline provides information about the location of the peak apex, and the length of the ridgeline inform us about the width of the peak, the width can be used to approximate the peak boundaries Du et al. (2006), Bramwell (2013).

2.1.2.2 Feature Engineering

Most of the tools or solutions that perform peak level quality control use consistency as a measure of quality. The tools compare precursor ion peaks of the same peptide amongst the different experiments and determine outlying peptides. Most of the tools monitor the peptides from the quality control samples to eliminate biological variabilities. QC solutions monitor the peak area, maximum intensity, mass error and full width at half maximum (FWHM) of a peptide across the experiments as these metrics should be unchanged Bereman et al. (2014), Pichler et al. (2012), Dogu et al. (2018).

To measure consistency, we require a sufficient number of experiments to establish the acceptable deviations amongst the peptides and in most cases, it is challenging to obtain the right number of experiments. Furthermore, it is also assumed that the system is producing good quality consistently. We need features that also measures the quality of peaks, instead of focusing only on the consistency of the system. The peaks are expected to follow a Gaussian function, or at least be triangular, be smooth and be symmetric. Deviation from the expected shape might indicate errors within the proteomics procedure. Features pertaining to the shape of the peaks have been defined and used in metabolomics quality control solutions Zhang & Zhao (2014), Zhang et al. (2014). In case of tandem MS, fragment ion peaks from the same precursor ions are expected to have the same shape and retention time, this also applies to isotope peaks in case of isotopic labelling. Interference from other ions or a cleaning detergent that co-elute with the peaks can result in time shifts and dissimilar fragments and isotope peaks shapes. Shape-based and features measuring interference are given in Table 2.1.

TABLE 2.1: Shape-based and interference-based features

Feature	Description	Type
Gaussian Similarity	Measures similarity between the peak and a Gaussian curve.	Shape Based
Triangle Area Ratio	Measure how well the peak resembles triangle.	Shape based
Symmetry	Skewness of the peak.	Shape based
Kurtosis	Sharpness of the peak	Shape based
Full width at base	Width at base of the peak.	Shape based
Full width at half maximum	Width at 50% of the peak height.	Shape based
Zig zag Index	Measures the jaggedness of the peak.	Shape based
Signal to Noise Ratio	Measures background noise.	Shape based
Modality	Measures the modality of the peak.	Shape based
Peak Significance	Measures the ratio of the intensities at the boundaries and the intensity at apex.	Shape based
Similarity	Pearson correlation of the fragment ion peaks from the same precursor.	Interference based
Shift	Indicates if the fragment ion peaks from the precursor have the same retention time.	Interference based
Isotopic Shift	Measures if light and heavy isotopic pair have the same retention time.	Interference based
Isotopic Similarity	measures if light and heavy isotopic pair have the same shape.	Interference based

2.1.2.3 Classication

To monitor if the peptides' attributes are consistent across the experiments, Bereman et al. (2014) introduced the use of Statistical Process Charts (SPC) in proteomics quality control Shewhart (1924, 1938). The SPC charts have the same principle used in Chiva et al. (2018) where a feature is an outlier or out of control if it does not fall within the three standard deviations, the charts also help visualise features over time Shewhart (1924). A precursor ion peak is considered as outlier if three of its features are out of control and an experiment with more than five outlying precursor ions is considered faulty. The mean and variation control charts manage to only capture small deviations in the features but fail to detect fluctuations in some of the features. To improve the use of control charts in proteomics QC, MSSTAT introduced using moving range and cumulative sum Dogu et al. (2017). Due to the sensitivity of the moving range and cumulative sum to outliers, the Median Absolute Deviation (MAD) is used to determine outlying features in Pichler et al. (2012), Bittremieux et al. (2015).

In metabolomics Zhang & Zhao (2014) used empirical methods to determine the appropriate cut-off threshold of each shape-based feature to classify the peaks. Their studies demonstrated that instead of establishing a threshold per feature, determining the best combination of cut-off thresholds for all features improved classification. Percy et al. (2014) used manual intervention to monitor interference of both fragment and isotopic based features. Eshghi et al. (2018) supervised classification techniques such as K Nearest Neighbour, Regularized Random Forest, Support Vector Machine and regularised logistic regression to classify peaks based on quality. Shape-based and interference-based features were used to represent the peaks. All the classifiers manage to obtain accuracies greater than 0.7.

2.2 Clustering

The lack of expert annotated proteomics data shows that unsupervised learning is a more viable quality control solution than supervised learning. Some of the available tools and solutions have used unsupervised techniques, such as outlier detection and anomaly detection. Anomaly or outlier detection techniques make two assumptions: that a dataset has a few anomalies and the anomalies are more likely to have a different generation process; therefore, vary significantly from the other data points Chandola et al. (2009). Therefore, the tools assume that low-quality data points are few which not always the case as shown in the dataset used by Eshghi et al. (2018), where bad quality peak pairs make up 40% the dataset. In case of a high number of low-quality peak pairs, the assumption made by anomaly or outlier detection techniques is invalid; therefore, other forms of unsupervised techniques should be explored. We need to use a technique that will group data points according to their quality. Classification can be used to group high quality and isotopic peak; however, the classification should not rely on experts annotated data.

Clustering is a form of unsupervised classification method that groups data instances based on a similarity measure. The techniques strive to maximise intra-cluster similarity while minimising inter-cluster similarities Saxena et al. (2017). Clustering has been used in fields such as engineering, medical research and cybersecurity for pattern detection Saxena et al. (2017), Estivill-Castro & Yang (2000). There are different clustering techniques, and it is not possible to choose an optimal technique as the performance of the techniques is dependent on the dataset's characteristics Estivill-Castro & Yang (2000). The different clustering approaches are as follows: hierarchical, partitioning, fuzzy theory, density-based, spectral based and model-based techniques. Clustering techniques can provide hard assignments; data instance can solely belong to a single cluster or soft

clustering; which provides a set of scores with a score representing the membership or probability of the data instance belonging in that particular cluster. Most of the techniques require certain parameters to be specified prior to the clustering. Choosing the optimal parameters is a non-trivial task without expert knowledge; however, evaluations indices can be used in this regard Xu & Tian (2015).

2.2.1 Dissimilarity and Similarity measures

There are different ways we can calculate the dissimilarity or similarity amongst the data points. Minkowski distances $Dist_{xy}$ are used to measure the dissimilarity for datasets containing quantitative variables (interval and ratio) Singh et al. (2013). Minkowski distance is a generalised distance measure and is made to be specific by setting the value of p, there are three popular distances, Manhattan distance $p = 1$, which measures the absolute differences between the pair of data points. Euclidean distance $p = 2$ is the root square difference between the instances, and this is the more popular measure and Chebyshev distance $p = \infty$ which is the maximum value distance Xu & Tian (2015), Singh et al. (2013). Pearson correlation is a popular measure that is usually used to determine the similarity amongst text documents Huang (2008). For qualitative variables, specifically nominal variables, distance cannot be used to measure dissimilarity because there is no natural ranking in these variables; instead, Jaccard or Hamming similarity should be used. Jaccard similarity is a ratio of intersection of a pair of data points and their union; if the similarity is one, then data points are identical. Hamming similarity measures the number of substitutions needed to make data point x_i identical to data point x_j Xu & Tian (2015), Huang (2008), Berkhin (2006), Yang & Kamel (2003).

$$Dist_{xy} = \left(\sum_{k=1}^{d} |x_{ik} - x_{jk}|^{1/p} \right)^{p} \tag{2.3}$$

2.2.2 Hierarchical Clustering

Hierarchical clustering either iteratively combine, agglomerative clustering or split clusters, divisive clustering. Agglomerative clustering starts with N single point clusters and merges the clusters until there is a single cluster containing all the data point and divisive clustering iteratively splits clusters from a single cluster to single point clusters Estivill-Castro & Yang (2000), Deb & Dey (2017). Dendrogram can be used to visualise the partitioning or merging of the clusters Berkhin (2006).

Agglomerative technique is the most popular of the two. To choose which clusters to merge or divide at each iteration, the inter-cluster distance is used. In divisive clustering, the two components that will result in the largest inter-cluster distance are parted, and in agglomerative clustering, the clusters with the smallest inter-cluster similarity are merged Berkhin (2006). The inter-cluster distances can be calculated using either graph metrics or geometric metrics Berkhin (2006), Estivill-Castro & Yang (2000). There are three graph-based metrics: single linkage, complete linkage or average linkage. In single linkage, the inter-cluster distance is the distance between the closest two points from the two clusters. Complete linkage determines the distance between the data points furthest from each other. Average linkage pairs data points, one point from each cluster, calculates the distance between each pair, the distances are then averaged to determine the inter-cluster distance Berkhin (2006), Estivill-Castro & Yang (2000). Geometric metrics use distances between the cluster centres to determine inter-cluster distances. The mean or centroid, the median of the data points in a cluster can be used as centres; ward clustering uses minimum variance as the distance measure instead. Variance is the increase in the sum squared of intra-cluster distances after merging occurs. The intra-cluster distance is the distance between a data point and its cluster centre Berkhin (2006).

Hierarchical clustering is commonly used in bioinformatics, especially in gene expression clustering. With the advancement in bioinformatics tools, raw data such as gene expression are readily available and are in large quantities making analysis of the raw data difficult with traditional methods Langfelder et al. (2008). The raw data can be simplified using clustering, allowing analysts to focus on the few clusters instead of every single gene. Clustering can be performed on the genes found in a single sample to determine genes with similar regulation or function, and infer those characteristic to genes with unknown properties. Samples can also be clusters to determine sub-types of similar clusters D'haeseleer (2005).

One of the advantages of hierarchical clustering is the ability to visualize it using a dendrogram. The graphical representation allow analysts to validate the clustering results based on prior knowledge about which objects belong in the same cluster Herrero & Dopazo (2002). A dendrogram is a network structure that is represented by a tree diagram. The tree diagram is made up root node, which represent all the objects being in single cluster and branches to nodes connected by edges until the last nodes, referred to as leaves, which represent a single object clusters or unclustered objects Reddy & Vinzamuri (2018). The length or the height of the branches is determined by the distance between the clusters. A heat-map can be used to enhance the visualization with the colours of the heat-maps representing the distance Lemenkova (2020).

2.2.3 Partition-based Clustering

In partition-based clustering, a mean or other prototypes are used to represent the whole cluster. Although the partitioning-based clustering occurs in iterations, there is no splitting or merging of the clusters like in hierarchical clustering Saxena et al. (2017). The purpose of the iterations is to re-assign the data instances to the clusters to improve clustering, i.e. to optimise the clustering solution Berkhin (2006). The prototypes eliminate the need to measure the intra-cluster similarity using pairwise distances causing the clustering problem to have a linear time complexity as the data points are only compared to a single prototype. The data points are parted into different clusters such that the clusters are homogeneous. An objective function is used to ensure homogeneity in the clusters. Most of the partitioning clustering techniques' objective function to minimise the sum of squared errors (SSE) as minimum SSE indicates minimum variance within the clusters Kutbay (2018).

A commonly used clustering technique is k-means due to its simplicity and computational efficiency Saxena et al. (2017), Estivill-Castro & Yang (2000). The clusters are represented by centroids. The euclidean distance is used to determine the distance between a data point and the centroids and data point is assigned to the closest cluster centre Kutbay (2018). The clustering is performed such that the objective function (J) provided in Equation (2.4) is minimised. K-medoid is similar to k-means, however, instead of using centroid to represent the cluster they use medoid instead. Medoids are data points in the cluster closest to each point in their particular cluster Berkhin (2006).

2.2.4 Fuzzy Theory

Artificial Intelligence was developed to mimic human intelligence. Classical logic systems are able to mimic human intelligence in instances of certainty and precision. However, these systems fail at making decisions in instances where an approximate mode of reasoning is required Nikravesh (2007), Zadeh (1988). These instances require using inference to make decisions even in cases of incomplete knowledge; inferring with partial information is an action that human beings do with ease. Fuzzy logic was introduced to allow intelligent systems to use approximate reasoning Klir & Yuan (1995). The ability of fuzzy logic to mimic the human being's way of reasoning even with vague inputs stems from the approach of borrowing from natural language and defining rules to map linguistic variables instead of numerical variables Zadeh (1988). For example, instead of classifying a person as tall or not tall based on their height as classical models do, fuzzy logic will give us a degree of tallness of that individual. In other words, instead of returning one for individuals with height above a certain threshold and zero for anyone

below the threshold, fuzzy logic will return a value in the interval [0,1] with values close to zero indicating a short person and values close to one indicating tall people. Fuzzy logic allows us to map infinite inputs to infinite outputs usually using if-then rules Klir & Yuan (1995).

A standard fuzzy logic system is made up of four components; a fuzzification unit, a knowledge database or a set of rules, a decision-making unit and a defuzzification unit. A fuzzy logic system takes in an input and converts it to a fuzzy set using the fuzzification unit and uses rules and knowledge database to determine a fuzzy set of the output which is then defuzzified into a crisp output. Unlike the crisp set of the classical logic systems where an element can either have a full membership, a fuzzy set allows for partial membership indicated in the height classification example above Klir & Yuan (1995), Zadeh (1988). In the universe U, a fuzzy set is an ordered pair of element x and the membership function μ, which tells us about the membership of x in the crisp set or the term set. The term set is usually linguistic variables, for example; tall, medium and small can be the term set if the element x is height measurement. The fuzzy set in the case of x being height will then be $fs = (x, \{\mu_{tall}, \mu_{medium}, \mu_{short}\})$ where μ_i is the membership of element x in the term set which is any value between [0,1] with $\mu = 0$ indicating no membership, $0 < \mu < 1$ indicating partial membership and $\mu = 2$ indicating full membership Klir & Yuan (1995).

Fuzzy logic theory has been used to implement feature extraction, classification and clustering techniques. An example of a fuzzy-based clustering technique is fuzzy c-means (FCM) Klir & Yuan (1995). In FCM, instead of determining the cluster an element x belongs to, we determine the membership function of that element. The membership function provides us with partial memberships of that particular data point in all the clusters, it provides with the fuzzy set $FS = (x, \{\mu_1, \mu_2, ..., mu_k\})$ where μ is the membership value and K is the number of clusters, therefore mu_i is the membership of x in cluster i. FCM uses the distance of the data point to the cluster which is represented by the cluster centre in comparison to the data point's distance to other clusters to determine the membership value. By determining the membership function and partial memberships of the elements in the clusters, FCM performs soft clustering instead of hard clustering performed by K-Means and k-medoids Kutbay (2018). This clustering technique aims to minimise the objective function which is given by Equation (2.5).

$$J = \sum_{j=1}^{K} \sum_{i=1}^{N} ||x_i^J - c_j||^2 \tag{2.4}$$

$$J.m = \sum_{j=1}^{K}\sum_{i=1}^{N} u_{ij}^{m}||x_j - v_i||^2 \tag{2.5}$$

2.2.5 Model Based Clustering

Model-based clustering assumes that the process generating the dataset can be described by a mixture of finite number of models and the general formula is given by Equation (2.6) with π_g being mixing proportion and $\sum_{g=1}^{G} \pi_g = 1$. For each model, g acts as a cluster and f_g is the distribution of the model Fraley & Raftery (2002). The parameters of the models are unknown and therefore the need to be determined Berkhin (2006), McNicholas (2016)

$$f(x) = \sum_{g=1}^{G} \pi_g f(x, \psi_g). \tag{2.6}$$

Gaussian Mixture Models (GMM) is a probabilistic model based clustering technique assuming that the process generating data points can be described using a mixture of Gaussian models. The parameters of the models are obtained using Expectation Maximization (EM) and the objective function is given by Equation (2.7) Berkhin (2006)

$$L(X|C) = \prod_{i=1:N} j = 1 : kPr(x_j|C_j). \tag{2.7}$$

2.2.6 Density Based Clustering

Density-based clustering define high density areas as clusters. The clusters are usually separated by low density areas. To identify clusters in the dataset these techniques can either define a density function or use the density-based connectivity approach Schubert et al. (2017), Berkhin (2006).

In the density-based connectivity approach, density and connectivity definition rely on the nearest neighbour to determine the local distribution of a cluster. An actual cluster is made up of core points and border points and data points in low density areas are outliers Berkhin (2006), Schubert et al. (2017). Although density-based clustering does not rely the number of clusters, there are parameters that need to specified prior to the clustering to determine the core, border and outlying points. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) requires the minimum number of neighbours, $minPoints$ and radius, ϵ to be specified before the actual clustering. To determine if a point belongs to a cluster or is an outlier, the following definitions should be considered Berkhin (2006), Schubert et al. (2017):

- Direct Reachability : A data instance d_a is a core point c_a if it has neighbours greater than minPoints in a neighbourhood n_a within a radius of ϵ.
- Density Reachability : If a data point d_b is in neighbourhood n_a of a core point c_a is also a core point, all the data point in its neighbourhood n_b, are also included in c_a's cluster.
- Border points: All the data points that are within a neighbourhood but are not considered to be core points are referred to as border points.
- Clusters: the different neighbourhood are considered as actual clusters and any data point that does not form part neighbourhood is an outlier.

Density function clustering uses a density function over the feature space. An example of such a clustering technique is DENCLUE 2.0, which uses kernel density estimation to perform clustering. The local maximum or local attractor of the estimated density is used to define the cluster and data points that have the same local maximum are assigned to the same cluster Berkhin (2006), Hinneburg & Gabriel (2007).

2.2.7 Clustering Validation Indices

It is challenging to determine the clustering techniques' parameters without expert knowledge. Choosing a low number of clusters can result in inferior data compression and using an elevated number of clusters results in small-sized clusters resulting loss of the clusters' characteristics Xu & Tian (2015). Choosing the right parameters is vital in ensuring a good quality clustering solution Berkhin (2006). External and internal validation indices can be used to assist in this regard. The external validation indices measure the homogeneity of the clusters by using the actual labels of the data. Therefore external indices cannot be used in choosing parameters of the clusters in an unsupervised manner. However, internal validation indices do not depend on the labels; therefore, internal indices are appropriate for an unsupervised setting Petrovic (2006).

Internal indices validations define and measure clustering quality in different ways. For clustering techniques relying on a dissimilarity measure to group data instances, internal indices are calculated using compactness within clusters and separation between the clusters Petrovic (2006). Factors such as noise, clusters' shapes and type of linkage used have an effect on the internal validation indices Liu et al. (2013). The compactness of the cluster can be measured using the variance of the cluster and separation is measured by the pairwise distance between the centres of the clusters or minimum pairwise distance between the objects of two clusters Liu et al. (2013). Although fuzzy c-means relies on

distances as a dissimilarity measure, it performs soft clustering; therefore, we can use the fuzziness to measure the quality of the clusters and choose the number of clusters Zhang et al. (2008). For probability models such as Gaussian Mixture Model (GMM), we can use parameters that depend on likelihood to determine number of the models suitable for our dataset Xu & Tian (2015). Density-Based techniques such as Density-based spatial clustering of applications with noise (DBSCAN) relies on two parameters: minimum points in a neighbourhood and the radius of the neighbourhood. Sander et al. (1998) implemented a heuristic method that depends on K-nearest neighbour to determine the optimal parameters Saxena et al. (2017).

The following indices rely on measuring the compactness and separation and work well with techniques that can detect convex-shaped clusters Xu & Tian (2015).

- Davies-Bouldin indicator (DBI) - Dispersion of the clusters relies on the diameters and the separation between clusters is the distance between the cluster centres. DBI is calculated using Equation (2.8) and it is a maximum ratio of sum of diameter dim of cluster c_i and c_j and the distance between the clusters, where $i \neq j$ and $j = [1, ..., K]$. A low value of DBI indicates good separation between clusters and good compactness within clusters Coelho et al. (2012), Liu et al. (2013):

$$DBI = \sum_{i=k}^{K} max_{i \neq j} \frac{dim_i + dim_j}{d(c_i, c_j)}. \tag{2.8}$$

- Calinski Harabasz (CH) - The variance of a cluster is the average distance between data points and their cluster centre. Separation of the clusters is the sum of distances between each cluster centres and global centre $\bar{C}$, where $\bar{C}$ is the mean of all the cluster centres. A high value of CH indicates low variance within the clusters and good separation between the clusters. CH is given by (2.9) Łukasik et al. (2016), Arbelaitz et al. (2013)

$$CH = \frac{N-K}{K-1} \frac{\sum_{j=1}^{K} \sum_{i=J}^{n} |cj| d(x_i^j, cj)}{\sum_{j=1}^{K} d(c_j, \bar{C})}, \tag{2.9}$$

- Silhouette coefficient (SC) - Silhouette coefficient measures the compactness by summing the distance between the points in the same cluster and separation is measured by distance between a data point and its inter-cluster nearest neighbour and is calculated using Equation (2.10). The higher SC the better the clustering quality Xu & Tian (2015), Łukasik et al. (2016), Berkhin (2006).

$$SC = \frac{(b(x) - a(x))}{max\{a(x), b(x)\}} \tag{2.10}$$

where $a(x)$ is the mean distance between data point x_i and all points in the same cluster and $b(x)$ is the smallest distance of x_i and the data points in the other clusters.

For clustering techniques that rely on fuzzy logic, we can use indices that measure the fuzziness of the clustering solution such as Dunn's fuzzy partition coefficient (FPC) Trauwaert (1988). The fuzziness is determined by how easy it is to transform soft clustering solution; data point having a membership in all the clusters to hard clustering solution where a data point assigned to a distinct cluster. For example if the memberships m_{ij} of the point x_a is $m_{aj} = [0.33, 0.33, 0.33]$ and x_b is $m_{bj} = [0.8.0, 15, 0.05]$ and where $j = [1, 2, 3]$ it is easier to assign x_b to single cluster c_1, whereas it will be difficult for x_a. If the memberships of the data points are similar to that of x_a, then the fuzziness will be high and will be low for a dataset where the instances have memberships similar to x_b. Partition coefficient vpc is given by Equation (2.11) Zhang et al. (2008). The partition coefficient can be normalised vpc such that $FPC = 0$ indicates complete fuzziness and $FPC = 1$ indicates that soft clustering solution is the same hard clustering solution. The normalised index is given by Equation (2.12).

$$vpc = \frac{1}{n}\sum_{j=1}^{c}\sum_{i=1}^{n} u_{ij}^2 \tag{2.11}$$

$$fpc = \frac{vcp - (1/K)}{1 - (1/K)} \tag{2.12}$$

Using the probabilistic foundation of mixture models, instead of focusing on the quality of the clusters to choose K, we can view the problem of choosing the number of the models as the problem of finding the best fit. The way to evaluate or determine the best fit is using indices that uses the combination of log-likelihood L, the number of clusters and the total number of estimated parameters. Examples of such indices are Akaike information criterion (AIC), given by Equation (2.14) and Bayesian information criterion (BIC), given by Equation (2.13). These indices aim at penalising complex models. AIC tends to overestimate because it is order consistent. The difference between AIC and BIC is penalising term Xu & Tian (2015)

$$BIC(K) = -2L(K) + v(k)\ln n, \tag{2.13}$$

and

$$AIC(K) = -2L(K) + 2v(K), \tag{2.14}$$

where $v(K)$ is the number of free parameters in a mixture model with K components and for a GMM fitted on dataset with d dimensions is determined using Equation (2.15)

Xu & Tian (2015)

$$v(K) = (K-1) + dK + \frac{dK(K+1)}{2}. \tag{2.15}$$

2.3 Feature Selection

Irrelevant and redundant features can negatively affect the quality of clustering and increase computational complexity of the clustering techniques. To improve clustering techniques results, feature selection techniques can be used to obtain a subset of the features that are relevant in separating the different classes. A feature selection technique has two objectives: to reduce the number of features and to improve the clustering solution. The feature selection techniques can either be filter and wrapper selection techniques Zhu et al. (2019).

2.3.1 Filter

Filter feature selection techniques do not rely on any clustering technique to select the relevant features. It ranks each feature according to specified criteria Alelyani et al. (2018) Jović et al. (2015). Univariate feature evaluates and ranks each feature and ignores interaction amongst the different features, and multivariate features selection takes into account the relationships of the features. Filter features selection criteria depend on measures such as entropy-based distance and laplacian scores Jović et al. (2015) Zhu et al. (2019).

2.3.2 Wrapper

Wrapper feature selection techniques depend on classification or clustering models to select as a subset that will improve the objective function of that particular model. Wrapper techniques treat the classification and clustering techniques as black boxes. In a supervised settings the objective of a feature selection technique is to maximise the accuracy of the classification model Jović et al. (2015). However, in an unsupervised approach, we cannot use accuracy as the approach does not rely on labels. Clustering validation indices discussed in section 2.2.7 can be used as the objective function Xue et al. (2015).

Wrapper methods chooses a subset of features, then clustering is performed on the subset of features and the performance of the subset is evaluated using the CVI and this is repeated until the CVI converges Alelyani et al. (2018). Feature selection can be complex because of all the possible feature subsets which are 2^D, therefore an exhaustive search

is not always feasible, especially in high dimensional spaces as this is computationally expensive Xue et al. (2015), Alelyani et al. (2018). Heuristic search techniques are used in selection methods such as sequential forward selection (SFS) and selection backward selection (SBS). However, the problem with these selection techniques is the nested effect, meaning if a feature is selected or removed, it cannot be removed or selected later Xue et al. (2015). Evolutionary computation such as Genetic Algorithm (GA), Genetic Programming and particle swarm optimisation (PSO) proved to be valid feature selection because of the global search algorithm. Another advantage of evolutionary computation is the ability to have multiple objective functions Xue et al. (2015).

2.4 Summary

The are multiple quality control solutions used for monitoring the quality of the results of the Proteomics experiments. Most of the solutions look at the system's ability to produce consistent results by monitoring technical variability. These solutions require multiple technical replicates and can delay troubleshooting. For peak-based solutions, there are features defined in the literature that focuses on monitoring peaks deviating expected shape and these eliminate the need to compare the peaks of the same peptides. Our solution we will build on existing peak-based solutions by implementing a peak detection using CWT, engineering shape-based and interference-based features and using clustering and feature selection to classify the isotopic peak pairs.

Chapter 3

Research Methodology

The purpose of our research is to develop an unsupervised quality control (QC) solution that can classify isotopic peak pairs based on quality. In this section we discuss the pre-processing stage which includes using continuous wavelet transformation (CWT) for peak extraction and engineering peak shape and interference-based features. In addition, we compare different clustering techniques to classify the peaks pairs and implement a feature selection solution using a Genetic Algorithm to improve the clustering results.

Figure 3.1 shows the functional block diagram of our study. Our solution comprises of a peak detection algorithm for extracting ion peaks from their chromatograms; the peak pairs are transformed into their feature space by our feature engineering component. There are two approaches followed in classifying peaks; the first approach involves comparing different clustering techniques to group the isotopic pairs and in the second approach, each of the clustering technique is embedded with a feature selection technique. We use different evaluation metrics to determine if the approaches are successful in grouping isotopic pairs based on quality.

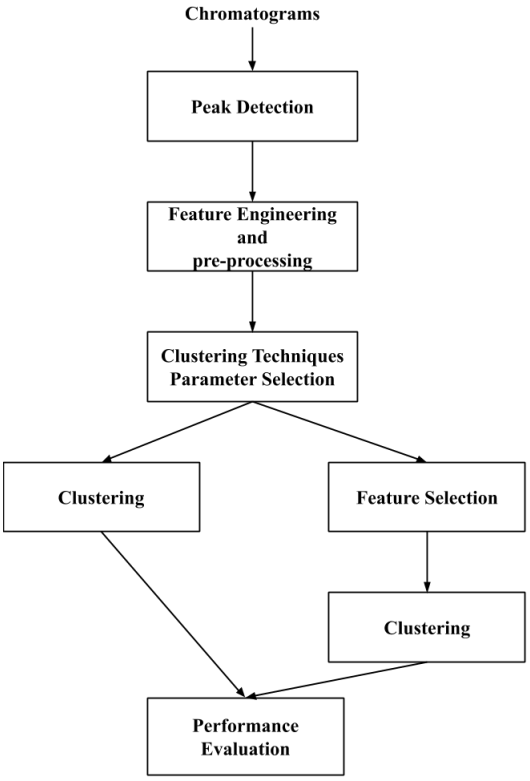


FIGURE 3.1: Functional Block Diagram for classifying the peaks

All analysis was performed using Python (3.6.15) on Jupyter Notebook. All the libraries or modules used in our study are in Table 3.1.

TABLE 3.1: Python Libraries and modules used in our study

Library	**Version**	**Algorithms**
scikit-learn	0.22.1	KMeans, Agglomerative Clustering, Gaussion Mixture Models and DBSCAN
scikit-fuzzy	0.4.2	Fuzzy
Deap	1.3.0	Genetic Algorithm

3.1 Data

The data used for our studies is from a study analysing the cerebrospinal fluid. In the study, Multiple Reaction Monitoring (MRM) was used to discover novel biomarkers for Alzheimer's disease Wildsmith et al. (2014). A stable-isotope labelled internal standard is used in the experiment and this will result in pairs of light and heavy fragment ions. The raw files from the analysis have been used in a supervised quality assessment tool, TargetedMSQC. TargetedMSQC used Skyline to perform the data analysis and extracted 1152 fragments chromatograms, 576 light isotopes and 576 heavy isotopes; i.e. 576 isotopic pairs which are publicly available. TargetedMSQC performed a quality

assessment on the isotopic pairs instead of individual peaks. TargetedMSQC also publicised expert annotation of each isotopic pair. The dataset contains 377 isotopic pairs annotated as 'ok' and 199 flagged peak pairs. The 377 pairs make up the good class and represent the high-quality isotopic pairs and 199 pairs are of low quality and belong to the bad class Eshghi et al. (2018). Moreover, they also have notes with reasons for every bad classification shown in Table 3.2. We use the 1152 fragment ion chromatograms in our study and the experts' labels to evaluate the performance of clustering techniques and feature selection techniques.

TABLE 3.2: Reasons of why experts flagged the isotopic peak pairs

Expert Notes / Reasons	**No. of bad pairs**
bimodal / interference	1
interference/ high background	1
bimodal / tailing	2
heavy transition below threshold / tailing	2
bimodal/transition ratios or orders are inconsistent between light and heavy	4
tailing/bimodal/transition ratios or orders are inconsistent between light and heavy/high background	4
tailing	7
tailing / high background	7
tailing/transition ratios or orders are inconsistent between light and heavy	8
tailing/transition ratios or orders are inconsistent between light and heavy	8
bimodal	13
jagged	31
transition ratios or orders are inconsistent between light and heavy	34
high background	37
shoulder	40

3.2 Peak Detection Methodology

An extracted ion chromatogram (XIC) $Y(t)$, is the intensity signal; $I = I_1, I_2, ..., I_N$ of an ion measured over time $T = t_1, t_2...t_N$. $Y(t)$ is made up of a baseline function $B(t)$, offset constant C and actual peak $p(t)$ as shown in Equation (3.1) Du et al. (2006). Multiples peaks can be present in a single chromatogram, therefore, the peak detection problem starts from identifying the peak apex of the actual fragment ion. The next step in peak detection is to identify the boundaries of the peaks, the start point and the endpoint of the peak. The length M of peak is less than the length N of the chromatogram. The peak detection algorithm is given in Figure 3.2

$$Y(t) = p(t) + B(t) + C. \tag{3.1}$$

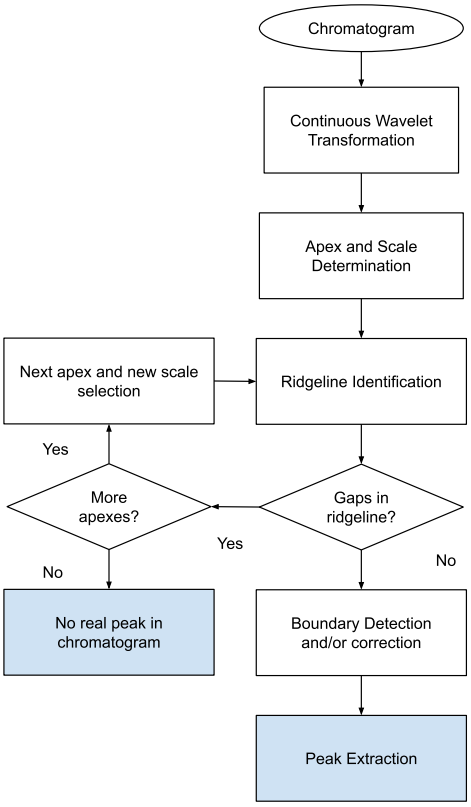


FIGURE 3.2: Peak detection process

3.2.1 Signal Transformation

A genuine precursor or fragment ion peak has specific shape characteristics. We can use analysing signals to transform the peaks into another space that will reveal these characteristics and simplify the extraction of the peaks from their chromatograms Du et al. (2006), Grossmann et al. (1990), Addison (2018). Although Fourier transformation is a popular signal transformation technique, it uses infinite analysing signals. Seeing that peaks are finite, Continuous Wavelet Transform (CWT) is a more suitable transformation technique as it uses localised waveform function as the analysing signal Grossmann et al. (1990). There are different types of wavelets or analysing signals; choosing the correct wavelet with similar properties to the peak is essential. CWT transform the chromatogram from the Intensity I and time t space into the scale s, time t and coefficients C space, the scale conveys information about the width and the time tells us more about the location of the peak. Coefficient C at scale s is determined using Equation (3.2):

$$C(s,\tau) = \int_{-\infty}^{\infty} Y(t)\psi_{s,\tau}(t)dt, \tag{3.2}$$

and

$$\psi_{s,t}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right), \tag{3.3}$$

where $Y(t)$ is the chromatogram, ψ given by Equation (3.4) is the mother wavelet or the analysing signal, $s = [1, 100]$ and $\tau = \Delta t$ of the chromatogram. A large C indicate a good match between analysing wavelet $\psi(t)$ and the peak $p(t)$. For the transformation, we use the Ricker wavelet as the analysing signal. Ricker wavelet is suitable because of its proportionality to the second derivative of a Gaussian curve, and the peak shape follows a Gaussian function. In the transformed space, we expect larger coefficients, local maxima, around the peak area and the maximum coefficient at the peak apex. Using the wrong wavelet that has a different shape to the peak might result in larger coefficients located outside the peak bounds

$$\psi(t) = \frac{2}{\sqrt{3}}\pi^{-1/4}\left(1-t^2\right)e^{-t^2/2}. \tag{3.4}$$

By combining Equation (3.1) and Equation (3.2) the CWT coefficients of $Y(t)$ can be calculated using Equation (3.5)

$$C(s,\tau) = \int_{-\infty}^{\infty} p(t)\psi_{s,\tau}(t)dt + \int_{-\infty}^{\infty} B(t)\psi_{s,\tau}(t)dt + \int_{-\infty}^{\infty} C\psi_{s,\tau}(t)dt. \tag{3.5}$$

The third term of Equation (3.1) is a constant and falls away during integration in Equation (3.5) because the mean of the analysing signal is zero. The integration of the base function $B(t)$ shown by the second term of Equation (3.5) is zero because of the symmetry of the mother wavelet $\psi(t)$ given that $B(t)$ is monotonic and gradually changing Du et al. (2006), Tautenhahn et al. (2008). Therefore, Equation (3.5) becomes Equation (3.6):

$$C(s,\tau) = \int_{-\infty}^{\infty} p(t)\psi_{s,\tau}(t)dt. \tag{3.6}$$

Figure 3.3 shows the transformation from the original space to the Continuous Wavelets space.

3.2.2 Maximum Scale

At each scale s we expect the largest coefficient $Cmax$ to be located at the peak apex, the centre of the largest and broadest peak. As the scales increases from $s = 1$ to $s = 100$, the maximum coefficient at the actual peak apex will increase until the scale matches the width of peak, i.e. $(s = w/\tau)$ and decrease for $s > w/\tau$ as shown Figure

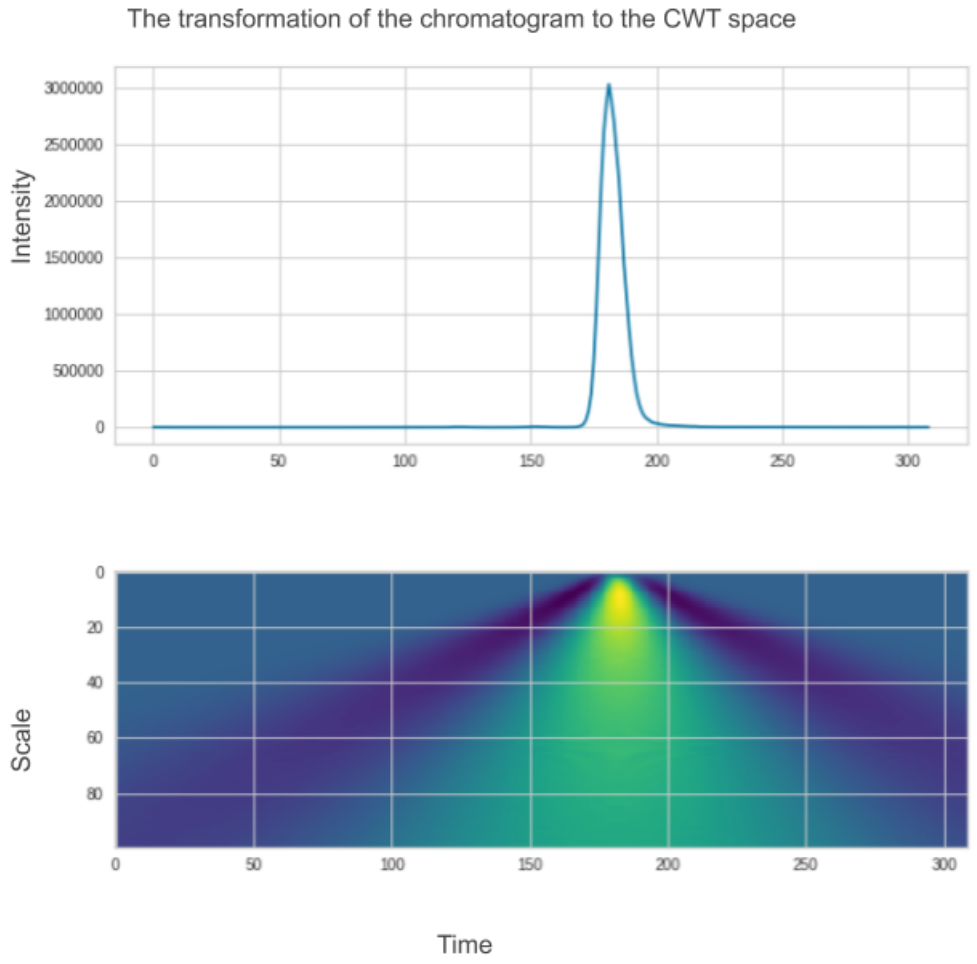


FIGURE 3.3: Normal signal and transformed signal

3.4 Du et al. (2006) . An extracted chromatogram may have multiple peaks, only one peak is an actual fragment ion peak, and the other peaks can be from noise particles that co-elute with the actual peak. Therefore, we assume that the broadest peak with the largest amplitude is the actual ion peak in the chromatogram. To determine the scale s_{max} that matches the width of the largest and broadest of the peak, we track the maximum coefficient C of each scale and if $C_s > C_{s+1}$ the width is of the peak can be approximated using s.

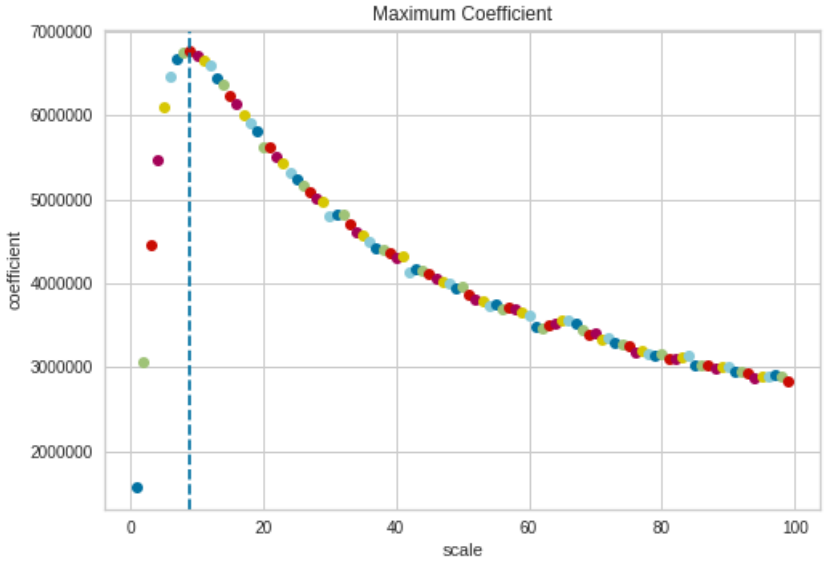


FIGURE 3.4: Highest Coefficient at each scale

3.2.3 Apex Detection

As mentioned, the maximum coefficient C_{max} at each scale s will be around the centre of the peak. Therefore, a real peak will have local maxima C_{max} at each scale. To locate the actual ion peak on the chromatogram, we use ridgelines obtained by connecting local maxima at the different scales from maximum scale to the smallest scale $s = 1$. The maximum coefficients at adjacent scales $s + 1$ and s are connected if the difference in their times is less than the translation value $t_{s+1} - t_s \leq \tau$. If the difference between t_{s+1} and t_s is large, the coefficient at s is excluded from the ridgeline, and we move to the next scale $s - 1$. The exclusion of coefficients produces gaps in the ridgelines. If a ridgeline has more than two gaps, the peak is considered as being a non-ion peak, and the next broadest peak is evaluated, i.e. we check if the peak's ridgeline contains two gaps at most. If a peak's ridgeline does not have gaps, the location of the peak apex t_p is the location of the maximum coefficient at s_{max}. Therefore, the peak apex of an actual peak is the intensity I_p at time t_p. If ridgelines of all peaks of a chromatogram have more than two gaps, then that particular chromatogram contains no actual ion peaks.

3.2.4 Peak Boundaries Detection

After identifying the peak apex of a true peak, we need to determine the starting point and the endpoint, i.e. boundaries of that particular peak. Since we know the width of the peak, which is approximated by scale, we can use it to determine the boundaries.

Therefore, if t_p is the location of the peak apex, i.e. the time in which the peak apex occurs and peak width is w, then the location of the left boundary, starting point t_l and right boundary, endpoint t_r are given by Equation (3.7) and Equation (3.8) respectively.

$$t_l = t_p - \frac{1}{2} \times w \tag{3.7}$$

$$t_r = t_p + \frac{1}{2} \times w \tag{3.8}$$

The above equations assume that all peaks are symmetrical. In the case of asymmetrical peaks, boundary corrections are necessary. The detected boundaries are considered to be actual peak boundary if the intensity I_b at the location of the boundary t_b is:

$$I_b \leq 3\% \times I_p. \tag{3.9}$$

In case Equation (3.9) is not met, local minima is used to determine new boundaries. We find all the minima that are within the compact support region of our peak ($2.5 \times s$). If all minima do not meet the criteria, the last minima within the compact support are taken as the boundaries. Therefore, a peak is made up all the intensities $I = I_l, I_{l+1}, ..., I_r$ measured at time points $T = t_l, t_{l+1}, ..., t_r$. Figure 3.5 shows the peak detection process, the apex selection, ridgeline construction and boundaries detection.

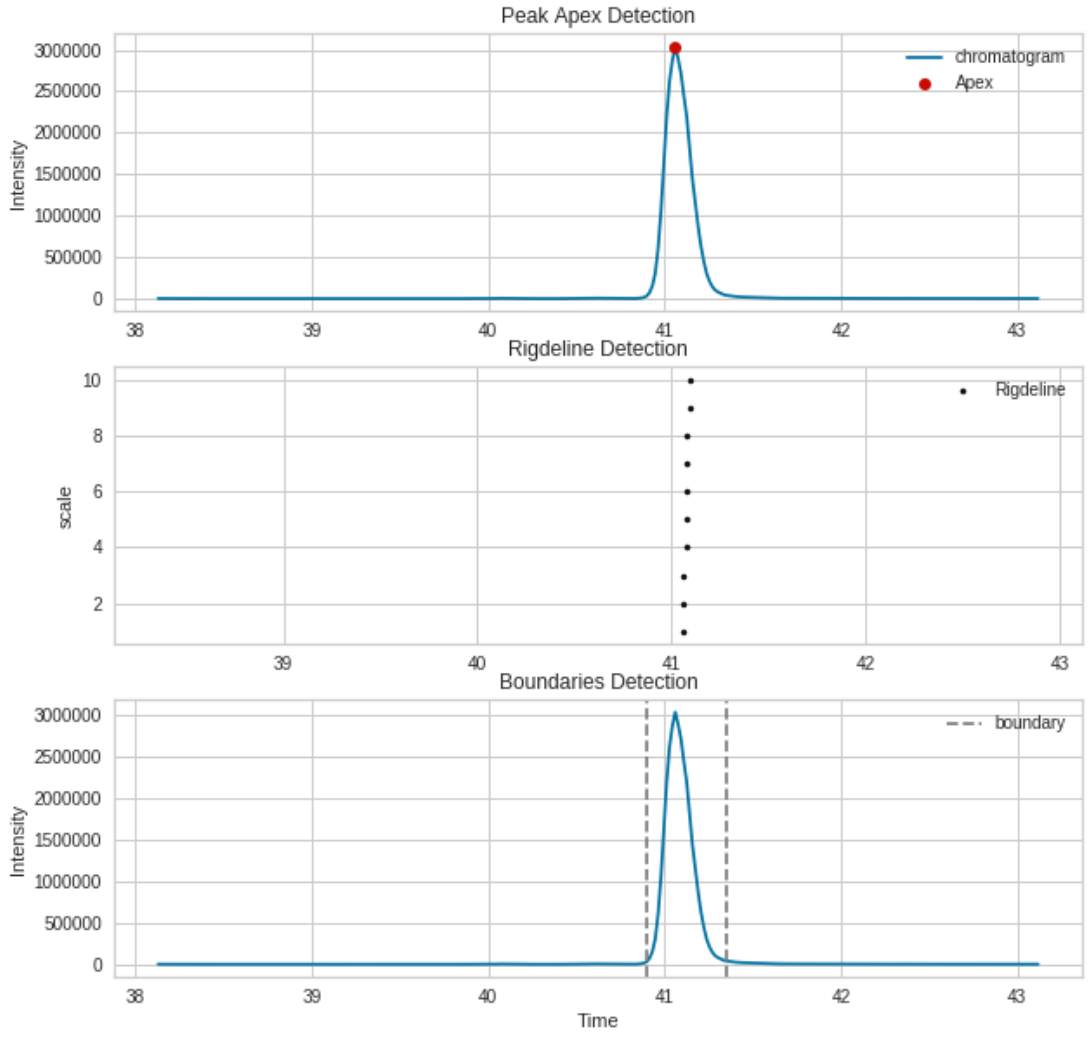


FIGURE 3.5: Peak Detection Visualization

3.3 Feature Engineering

Low-quality peaks are peaks deviating from the expected shape or have different a shape and retention time from their isotopic ion or fragment ions from the same precursor. Before clustering, we transform each of the 1172 peaks; 576 light isotopic peak and 576 heavy isotopic peaks; into the feature space. We then merge the heavy and light peaks to make up a single data point x_i. Therefore our dataset X is 576 isotopic pairs in their feature space. In our study, we extract the following shape and interference-based features:

3.3.1 Gaussian Similarity

An ideal LC-MS precursor or fragment ion peak can be estimated by an exponential modified Gaussian function (EMGF) Kalambet et al. (2011), Golubev (2017). The Gaussian function is modified through convolution with an exponential decay function and this results in EMGF given by Equation (3.10).

$$G(t) = \frac{h \cdot \sigma}{\tau}\sqrt{\frac{\pi}{2}} \cdot e^{\left(\frac{\sigma^2}{2\tau^2} - \frac{t-\mu}{\tau}\right)} \cdot \left(1 - erf\left(\frac{1}{\sqrt{2}}\left(\frac{\mu - t}{\sigma} + \frac{\sigma}{\tau}\right)\right)\right), \tag{3.10}$$

and

$$erf(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}, \tag{3.11}$$

where t is the time, h is the height of the Gaussian, μ is the mean, and σ is the sigma of the unmodified Gaussian. The Gaussian sigma is depended on the Liquid Chromatography (LC) column and determines the width of peaks, τ is the relaxation factor used to modify Gaussian. τ captures the presence of dead volumes in the column that modifies the peak and causes peak tailing Foley & Dorsey (1983).

Before fitting Equation (3.10) to the peak we need to determine the parameters; μ and σ of the unmodified function and the relaxation factor τ of the exponential decay function. Foley et al. (1983) developed a solution that uses location of the peak apex t_p, peak width at 10%, $W_{0.1}$, of the maximum intensity I_p and the asymmetrical factor B/A. $W_{0.1} = t_B - t_A$, $B = t_B - t_P$ and $A = t_P - t_A$ Foley & Dorsey (1983) which are shown in Figure 3.6. Before we can calculate the parameters, we need to determine the observed efficiency N_{sys} and the second central moment $\bar{M}_2$ as the parameters are dependent on these values.

$$N_{SYS} = \frac{41.7(t_p/W_{0.1})^2}{B/A + 1.25} \tag{3.12}$$

$$\bar{M}_2 = \frac{W_{0.1}^2}{1.764(B/A)^2 - 11.15(B/A) + 28} \tag{3.13}$$

$$\sigma = \frac{W_{0.1}^2}{3.72(B/A) + 1.2} \tag{3.14}$$

$$\tau = \sqrt{\bar{M}_2 - \sigma_G^2} \tag{3.15}$$

$$t_G = t_P - \sigma_G[-0.193(B/A)^2 + 1.162(B/A) - 0.545] \tag{3.16}$$

$$\mu = t_G + \tau \tag{3.17}$$

A Gaussian curve is fitted on $t = t_l.t_{l+1}, ..., t_p, ..., t_r$. Gaussian Similarity (GS) is the dot product of the actual peak and the fitted Gaussian.

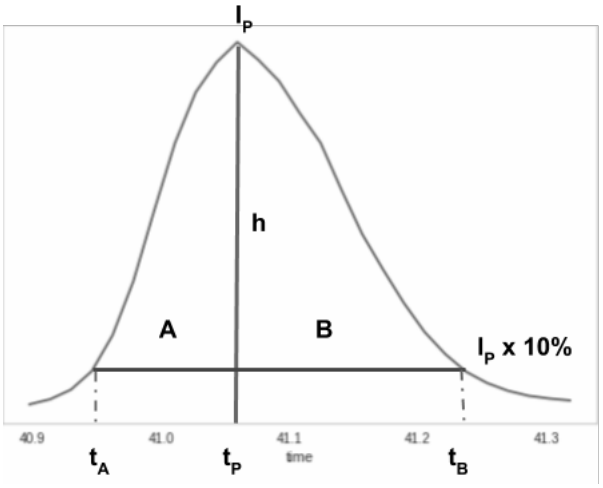


FIGURE 3.6: Exponential modified Gaussian function parameters

3.3.2 Triangle Peak Area Similarity Ratio

Triangle Peak Area Similarity Ratio (TPASR) is used to determine if the peak is triangular. LC-MS peaks are noisier than the predecessors gas chromatographic based peaks and therefore, tends to deviate from the expected shape and EMGF Zhang et al. (2014) and follow a more triangular shape. Equation (3.18) and Equation (3.19) are the areas of the fitted triangular function and the peak respectively. $TPASR$ is given by Equation (3.20).

$$TA = 0.5 \times P_w \times I_p, \tag{3.18}$$

where P_w is the width at the base of the peak,

$$PA = \sum_{t=l}^{r} I_t, , \tag{3.19}$$

where I_t is the intensity at time t, l is the time at the left boundary and r is the time at the right boundary.

$$TPASR = \frac{|TA - PA|}{TA}. \tag{3.20}$$

3.3.3 Modality

An ideal LC-MS peak is unimodal; however, due to co-elution, the detected peak may be multimodal. We use Hartigan's dip test to evaluate the modality of the peaks Eshghi et al. (2018). Hartigan et al. (1985) defined the dip test method to determine the modality of a peak. A peak $p(t)$ with mode m is unimodal if $(-\inf, m]$ is convex; increasing in density and if $[m, -\inf)$, is concave, i.e. decreasing in density. The dip test is the maximum difference between unimodal normal Gaussian distribution and the peak function $p(t)$. The modality of peak $p(t)$, $D(p)$is defined below:

$$D(p) = \rho(p, \Upsilon) \tag{3.21}$$

where, Υ is a class containing unimodal functions υ, therefore:

$$\rho(p, \Upsilon) =_{\upsilon \in \Upsilon} inf \rho(p, \upsilon) \tag{3.22}$$

and

$$\rho(p, u) =_t sup(p(t) - u(t)) \tag{3.23}$$

A peak $p(t)$ is considered to be unimodal if $D(p) = 0$ and that can only happen when $p(t) \in \Upsilon$ and $p(t)$ is multimodal if $D(p) > 0$ and thus $p \notin \Upsilon$.

3.3.4 Jaggedness

Instability of the ion-spray causes excessively jagged peaks. To determine the degree of jaggedness of each peak we define a feature Zigzag Index (ZI) Eshghi et al. (2018), Zhang & Zhao (2014), Zhang et al. (2014). The calculation of ZI is as follows Zhang & Zhao (2014):

1. Calculate Effective Peak Intensity EPI, which is the intensity at apex I_p subtracted by the baseline b at peak apex:

$$EPI = I_p - b. \tag{3.24}$$

2. Calculate the variance between two consecutive points $v(d_m, d_{m+1})$:

$$v(d_m, d_{m+1}) = 0.5 \times (2I_m - I_{m-1} - I_{m+1})^2. \tag{3.25}$$

$(2I_m - I_{m-1} - I_{m+1})^2$ indicate jaggedness of points I_{m-1}, I_m and I_{m+1},

3. Calculate the total jaggedness TJ of the peak by summing all local jaggedness TJ

$$TJ = \sum_{m=2}^{M-1} (2I_m - I_{m-1} - I_{m+1})^2, \tag{3.26}$$

which is then averaged and normalised to obtain ZI given by Equation (3.27) :

$$ZI = \frac{\sum_{m=2}^{M-1} (2I_m - I_{m-1} - I_{m+1})^2}{M \times EPI^2}. \tag{3.27}$$

3.3.5 Signal to Noise Ratio

The presence of background or foreign substances,i.e. noise can alter the actual spectrum and affect validity of the analysis. To determine the signal to noise ratio (SNR), we first need to determine the signal strength and local noise level. Local noise level (LNL) is the 95 percentage quantile CWT coefficient at the smallest scale($s = 1$), and the signal strength (SS) is the area under the curve (AUC) of the peak Zhang & Zhao (2014), Du et al. (2006). Therefore SNR is determined as:

$$SNR = \frac{SS}{LNL}. \tag{3.28}$$

3.3.6 Peak Significance

The intensity at the peak walls, i.e the boundaries should be zero. Peak significance is the ratio of intensities near the apex and the walls of the peaks Zhang & Zhao (2014), Zhang et al. (2014). Normal peaks are narrow, therefore, a low peak significance value might also indicate broadening.

3.3.7 Symmetry and Sharpness

An arbitrary distribution can be described using statistical moments. The odd moments inform us about the position of the peaks while the even moments provide information about the spread Martin & Honarvar (1995), Zhang et al. (2008), Choi & Sweetman (2010). In most cases, the first four moments are adequate in describing the distribution. The first and second moments are the mean and the variance of the distribution,

respectively. The third is related to the symmetry or skewness of the distribution. The fourth moment informs us about the sharpness, also known as kurtosis, of the distribution. For a normal Gaussian curve, the first and second moments are sufficient in describing the distribution due to its symmetry Martin & Honarvar (1995), Zhang et al. (2008), Choi & Sweetman (2010).

However, seeing that a peptide peak tends to be sharper than the Gaussian curve and malfunction during separation can result in an asymmetrical peak, the third and the fourth moment can be used to demonstrate how the peptides peaks deviate from the expected shape.

A peak $p(t)$ can be treated as grouped data with time t_i as x and intensity as the frequency f_i, therefore a peak will be analysed using grouped data statistics. The mean of grouped data is calculated as:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i \times f_i}{N}, \tag{3.29}$$

with

$$N = \sum_{i=1}^{n} f_i. \tag{3.30}$$

The rth moment is given below: Martin & Honarvar (1995), Zhang et al. (2008), Choi & Sweetman (2010):

$$m_r = \sum_{i=1}^{n} \frac{f_i(x_i - \bar{x})^r}{N}$$

.

The symmetry of a peak γ_1 is given by Equation (3.31). γ_1 is dependent on the third statistical moment m_3 and used monitor undesired peak skewness due to LC column malfunction Foley & Dorsey (1983)

$$\gamma_1 = \frac{m_3}{\sigma^3}, \tag{3.31}$$

where σ is the standard deviation of the peak.

An ideal chromatographic peak is sharper than a normal Gaussian function, i.e. it is Leptokurtic. The coefficient γ_2 used to determine the sharpness of the peaks and is given by Equation (3.32)

$$\gamma_2 = \frac{{m_4}^4}{\sigma} - 3, \tag{3.32}$$

where m_4 is the fourth statistical moment Foley & Dorsey (1983).

3.3.8 Full Width at Base and Half Maximum

High-quality peaks are narrow, however due to carryover in the column the ions peaks may be broader than expected. To monitor if there is any excessive broadening two features are used, full width at base (FWB) and full width at half maximum (FWHM). FWB is the width at the base of the peak and FWHM is the width at half of the height of the peak.

3.3.9 Fragments Shift and Similarity

Fragments ions from the same precursor ion should have the same peak shape and retention time. However, due to interference in the ion chamber, the peak shapes and retention times might differ. To measure inconsistencies of the fragment ions' attributes, two feature are used, namely Shift and Similarity. Shift is the difference between the retention times of two fragment ions whereas Similarity is the Pearson correlation of the two peaks Eshghi et al. (2018).

3.3.10 Isotopic Shift and Similarity

In the case of isotopic labelling, the pair of heavy and light ions peaks should also be similar in shape and have the same retention time. Similar to fragments features, Isotopic Shift is the time difference between isotopic pairs and Isotopic Similarity captures the similarity in the peak shapes of the isotopic pairs determined using Pearson correlation Eshghi et al. (2018).

3.4 Clustering

In this section, we discuss how we use different clustering techniques to classify the isotopic pairs based on quality. A data point can assigned to the good or the bad class; the good class represent all high-quality peak pairs and the bad class contains the low-quality pairs. There are a variety of clustering techniques and each technique uses a specific approach to group the data instances. In our study, we focus on clustering techniques which are based on hierarchical, partition, model and density approaches. The evaluation of the clustering approaches will help us determine the technique that is suitable to classify the isotopic pairs and answer our first research question. The clustering techniques require parameters to be specified before the actual clustering. Therefore, the parameter selection method is also discussed in this section.

3.4.1 Agglomerative Clustering

Agglomerative clustering is a hierarchical-based clustering technique. This technique starts with each single point clusters, i.e every data point x_i is a cluster and iteratively merge clusters until all the data points are assigned to a single cluster or until we have a specified number of clusters. In our study, we follow the latter approach of stopping the iterations at a specific number of clusters instead of merging until we have a single cluster. Therefore, before we can perform the actual clustering, we need to specify the optimal number of clusters that will result in the best clustering results. Seeing that agglomerative clustering technique works with spherical clusters, we compare CVI metrics such as Davies Bouldin Indicator (DBI), Silhouette Coefficient (SC) and Calinski Harabasz (CH) scores to select the number of clusters resulting in the best intra-clusters compactness and inter-clusters separations. After selecting the number of clusters, we then cluster the data points by merging clusters with minimum variance, given in Equation (3.33), iteratively until we reach the number of clusters specified by the metrics Xu & Tian (2015), Murtagh & Legendre (2011).

$$D(c_i and c_j) = \frac{|c_i||c_j|}{|c_i| + |c_j|}||c_i - c_j||^2 \tag{3.33}$$

3.4.2 K-Means

3.4.2.1 Random Initializing

K-means is a partition-based clustering technique that groups the data points into K clusters. Cluster centres are used to represent each cluster. K-means requires the number of cluster K to be specified before clustering. Similar to agglomerative clustering, k-means works with spherical clusters, and therefore, we use the same CVI metrics to determine the optimal K. K-means start by randomly choosing K data points as cluster centres and the euclidean distance between each data point and the centres is determined. A data point is assigned to its nearest cluster centre. The initial clustering does not represent the final clustering solution as the first centres are chosen at random. For the final clustering solution, after assigning each data point to a cluster, the new centres for those particular clusters are calculated by averaging the data points in the clusters. Each data point is reassigned to its nearest cluster. The iterations continue until the solution converges, i.e. the cluster centres remain unchanged Velmurugan (2012), Xue et al. (2015).

3.4.2.2 D2 Weighting

Randomly selecting the centres of clusters can result in clustering converging into local optimum instead of the global minimum Khan & Ahmad (2004). To prevent the clustering solution from converging to a local minimum, Arthur et al. (2007) defined an alternative randomised seeding technique, D2 weighting Arthur & Vassilvitskii (2007). Similar to random initialisation, we start by choosing random K points as centres. For each data point x in the dataset X, we determine $D(x)$ which is the distance between x and its closest centre. For each cluster, the new centre $c' = x_c \in X$ if the probability of x_c $P(x_c)$ is Equation (3.34) Yoder & Priebe (2017). After the initial cluster centres are selected, the clustering procedure is the same as in random initialisation, where we calculate the mean and reassign the data points to closest clusters at each iteration Khan & Ahmad (2004). K is the same for both random initialisation and D2 weighting.

$$P(x_c) = \frac{D(x_c)^2}{\sum_{x \in X} D(x)^2} \tag{3.34}$$

3.4.3 Fuzzy C-Means

Like k-means fuzzy c-means uses centres to represent the clusters. However, it performs soft assignments; instead of assigning the data points to single clusters, it determines a membership vector, $U_i = [u_{i1}, u_{i2}, ..., u_{ik}]$, of each data point, where u_{ik} is the membership of data point x_i in the k^{th} cluster. The membership is proportional to the proximity of the data point to the centre of the cluster. Fuzzy c-means needs the number of clusters to be specified. For agglomerative and k-means clustering, we use CVI metrics that focus on separation between clusters and cohesion within the clusters. In a fuzzy-based approach,good quality clustering can be determined using the ease of the transformation from soft assignments to a hard clustering solution. We use fuzzy partition coefficient (FPC) to choose the suitable number of clusters. The clustering of the Fuzzy-C Means procedure is as follows Zhang et al. (2008):

1. Use FPC to determine the number of optimal clusters.
2. Initialise fuzzy partition matrix $Ui = [uij]$ for $j = 1, 2, ..., K$ such that:

$$\sum_{j=1}^{K} u_{ij} = 1 \tag{3.35}$$

3. Use the partition matrix to determine the fuzzy cluster centroid $V = v_1, v_2, ..., vk$ which is the mean of all the data point weighted by memberships.

$$v_j = \frac{\sum_{i=1}^{n}(u_{ij}^{m} x_i)}{\sum_{i=1}^{n}(u_{ij}^{m}} \tag{3.36}$$

   where m us the fuzziness factor and is set at $m = 2$

4. Calculate the new memberships.

5. Repeat step 3 and 4, until the membership matrix remains the same.

The procedure above is the soft-clustering solution, to change the solution into hard clustering, the data point is assigned to cluster with the highest degree of membership Zhang et al. (2008).

3.4.4 Gaussian Mixture Model

Gaussian Mixture Model (GMM) is a linear combination of K Gaussian models as a single model is not always sufficient to capture multimodality of real-life datasets. GMM assumes that data points are sampled from a probabilistic distribution with density indicated below Li et al. (2018):

$$p(x) = \sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k), \tag{3.37}$$

where K is the number of models and will be determined using AIC and BIC. $N(.|\mu_k, \Sigma_k)$,μ_k and Σ_k is probability density, mean and covariance matrix of the kth matrix, and π is the mixing proportion Li et al. (2018).

The parameters θ, $\mu = \mu_1, \mu_2, ..., \mu_k$, $\Sigma = \Sigma_1, \Sigma_k$ and $\pi = \pi_1, \pi_2, ..., \pi_k$, of the components are not known but can be estimated using Maximum Likelihood or Expectation Maximization. Maximum Likelihood function finds parameters that maximize the likelihood of the model Bishop (2006). The log likelihood function of a GMM model is defined as:

$$\ln(X|\theta) = \sum_{i=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(x_i \mu_k, \Sigma_k) \right\}. \tag{3.38}$$

Maximum Likelihood problem becomes more complicated when there is more than one component involved. In most instances, Expectation Maximization is the method of

choice in determining the parameters for models with latent variables such as GMM. The latent variable z_ik gives us information about x_i belonging to k_th model, i.e. if $z_ij = 1$ then x_i was sampled from the jth model Li et al. (2018), Maugis et al. (2009) and $z_ik = 0, k \neq j$. We start by initialising parameters with random values, in the expectation stage, we then determine the responsibility γ_i of each component using the posterior probability defined as:

$$\gamma_i \equiv p(z_{ik} = 1|x_i) = \frac{N(x_i|u_k, \Sigma_k)}{\sum_{k=1}^{K}, N(x_i|u_k, \Sigma_k)}. \tag{3.39}$$

Maximization step uses the responsibilities to estimate new parameters. The purpose of the step is to determine the parameters that maximize the likelihood. By equating differentiation Equation (3.38) with respect to μ_k to zero we can estimate the new μ and mixing proportion using the following equations:

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \psi(z_{ik}) x_i, \tag{3.40}$$

and

$$N_k = \sum_{i=1}^{N} \psi(z_{ik}), \tag{3.41}$$

where N_k is the effective number of data points assigned to k_th component Bishop (2006). Σ is obtained by log-likelihood by Σ instead and is estimated using the following equation:

$$\Sigma_K = \frac{1}{N_k} \psi(z_{ik})(x_i - \mu_k)(x_i - \mu_k)^T. \tag{3.42}$$

The EM process iterates until the likelihood converges. GMM clustering also performs soft assignments by providing responsibility each component had in generating each data point. A component with the largest maximum responsibility is assigned to the data point to convert the clustering into hard assignments Bishop (2006).

3.4.5 Density-based Spatial Clustering of Applications with Noise

DBSCAN is a density-based clustering technique where high-density areas represent clusters and are separated by low-dense areas, which are considered as outliers. DBSCAN is the only clustering technique in our study that can handle outliers Schubert et al. (2017). DBSCAN require the neighbourhood size, indicated by radius ϵ and density, which is

the minimum number of data points within the neighbourhood *min_points* to be specified. We use the heuristic method that relies on k-nearest neighbours to determine the parameters Sander et al. (1998), Schubert et al. (2017).

Using k-nearest neighbour adds a third parameter K. The dimensionality D of our feature space is used to determine K, $K = 2 \times D - 1$. Studies performed by Sander et al. (1998) indicates that choice K does not have a significant influence of the final clustering results Sander et al. (1998). The density of a neighbourhood is $min\_points = K + 1$. To determine ϵ, we plot the k_th distance of each data point, and the knee point of the graph is chosen as ϵ Sander et al. (1998).

To cluster the data points in the dataset, each data point x is visited until all the data points are classified as either a core point, boundary point or an outlier. A data point x_a is classified as a core point if the density of its ϵ-neighbourhood n_a is greater or equal to *min_points*; otherwise, x_a is classified as an outlier. All the data points within n_a are assigned to cluster a. If data point x_b within n_a is classified as a core point, all the data points in n_b are assigned to cluster a, and if x_b is not a core point then is classified as a boundary point of cluster a. If data point x_c classified outlier is within a neighbourhood of a core point, then x_c is reclassified as a boundary point of that particular neighbourhood Schubert et al. (2017).

3.5 Genetic Algorithm

Clustering techniques assume that all the features are relevant in separating the two classes. In this section, we discuss how the Genetic Algorithm is used to select a subset of features that are relevant to the separation of classes and therefore improve the clustering solutions. Genetic Algorithm is a randomised search and optimisation technique Maulik & Bandyopadhyay (2000). It relies on biological operators such as cross-over and mutation to generate new solutions that need to be evaluated until the optimisation is reached Whitley (1994).

3.5.1 Population

Genetic Algorithm uses evolutionary processes to generate an optimised solution to a problem, in our case selecting an optimal subset of features that will improve the results of the clustering techniques. The evolutionary processes are used to generate new populations such that the solutions in the new generations are more aligned with the objective function, i.e. the new generation contains better candidates than the

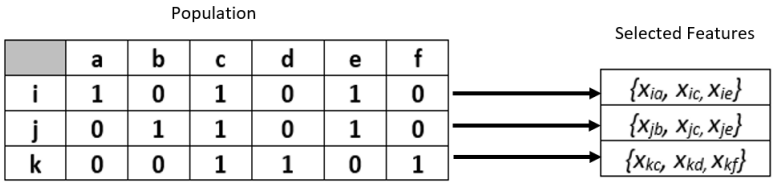


FIGURE 3.7: Selected features for each chromosome in the population. Selected features are presented to the clustering technique

previous generation. A population is made up of chromosomes, also referred to as individuals. A single chromosome is a binary string with the length D, which is the dimension of our dataset Whitley (1994). Each bit of chromosome represents the status of a feature; if the i_th bit b_i is one it means that i_th feature is selected, and if b_i is zero then the feature will be excluded as indicated in Figure 3.7. Only selected features are presented to the clustering technique. The first step in the genetic algorithm is to generate a population randomly.There is no right number of individuals for a population; therefore, in our study, we initialised the population to have $2 \times D$ individuals Whitley (1994), Abualigah et al. (2016).

3.5.2 Selection

Optimisation in the genetic algorithm is reached by selecting the best individuals from a population to generate the new population through recombination operators Srinivas & Patnaik (1994). An objective function is used evaluate the fitness of each individual. The objective of the feature selection is to improve the clustering results, i.e. the separation of low-quality and high-quality peak pairs. Improvements in the clustering results can be measured using external validation indices. External indices depend on actual labels of the data and cannot be used in an unsupervised setting. Internal indices use similarities or dissimilarities to measure the quality of clustering results. Therefore, the objective function of our genetic algorithm is the optimisation of an internal validation index, specifically minimising DBI index Whitley (1994). The fitness score of each individual in a population is the DBI index.

Post the calculation of fitness score a selection algorithm is used to select the fittest individual for the generations of new individuals Srinivas & Patnaik (1994). Popular selection methods include Roulette Wheel Selection, Tournament Selection and Rank Selection. The roulette wheel technique assigns the fitness probability score to which is proportionate to the fitness scores. Individual with high probability scores are highly likely to be selected, but due to the random nature of the selection process, an individual with a low score might still be selected. The randomness is introduced to ensure

that there is still diversity in the population. Whitley (1994), Razali et al. (2011). In tournament, n individuals are selected from the population to compete, the individual with the highest fitness score wins and is selected Mirjalili (2019). The ranking selection method proves to be useful in cases where the fitness score are quite similar and using the roulette wheel selection method will results in all individuals to have the same chance of being selected. The ranking selection methods first rank the individuals based on their fitness score, with the fittest being ranked N and the lowest being ranked 1, the chances of an individual being selected is depends on the ranks instead of the score Jebari & Madiafi (2013). In our study, we use the roulette wheel selection method due to ease of implementation.

One way of insuring that the genetic algorithm performance does not deteriorate with the generations of new populations, elitism can be used. Elitism is the selection of the best individuals to participate in the generation of the next population without going through the crossover and mutation process. This is done to ensure that average fitness value increases or at least stays the same in the new generations. This approach is not included in our solution and this can results in an unoptimized solution Jebari & Madiafi (2013).

The individual with the highest fitness score in the last generation is the solution to our feature selection problem, i.e. indicates the features that will be presented to the cluster technique and answer our third research question. The second research question is answered by the evaluation of the clustering techniques performed on the selected features.

3.5.3 Cross-Over and Mutation

Cross-over is used to combine some of the individuals to generate new individuals in the new generation. There are three common crossover methods, the single-point crossover and two-point crossover and uniform crossover also known as a multi-point crossover Mirjalili (2019). In a single-point crossover approach, new individuals are generated by swapping the parents chromosome beyond the selected single point Poli & Langdon (1998). Two points are selected in two-point crossover and the chromosomes between those two points are swapped Mirjalili (2019). Uniform crossover uses a mixing ratio Spears & De Jong (1991). For example if the mixing ratio is 0.3 this means that the first offspring will have 30% chromosomes of the first parent and 70% of the second parent and it will be the opposite for second offspring. In our study we use two-point crossover method because it is simpler than uniform crossover and is more likely to results in offspring that are not too similar to their parents like in the case of single-point crossover.

Before performing crossover we generate crossover factor cf s generated and crossover is performed only if cf is greater than a specified factor p_c Whitley (1994), Maulik & Bandyopadhyay (2000). Although crossover generates new individuals, it can result in a new generation that is similar to the previous one, and this is where mutation comes in. For every bit in a candidate we generate a mutation factor v and if $v < p_m$ then that the particular bit is then flipped Maulik & Bandyopadhyay (2000). Choosing the correct parameters p_c and p_m is a non-trivial task and there have been studies performed to determine the optimal parameters. The studies included performing multiple experiments with different values of the parameters and monitoring the performance of the genetic algorithm. These studies provided the following values are $p_c = 0.6$ and $p_m = 0,0001$ as guidelines. The p_c is selected to encourage crossover to occur, if the value was too high the new generations will be made up mostly of individuals from the previous generation and most bits of the chromosome would be flipped if p_m was too high and this will results in the offspring being too different to the parent Hassanat et al. (2019). In our study we use the values as per the guidelines.

3.6 Evaluation

To evaluate how well the techniques performed in separating good and bad peak pairs and answer our last question, we first need to label the pairs as belonging to either the good or bad class depending on their clusters. Recall, precision and f-score are used to compare the predicted labels, and the expert annotated data. Our evaluation approach is discussed in this section.

3.6.1 Labelling

Clustering is an unsupervised learning technique, and the results of the clustering techniques are labels representing the clusters data points belong to and not if the data point is a low-quality isotopic pair or a high-quality pair. To label the peak pairs, we are using expert annotations. We determine the total number of bad pairs and good pairs in each cluster and every pair in a particular cluster is labelled as a class (good or bad) that is mostly represented in it. In other words, if a cluster has 50 bad isotopic pairs and ten good pairs, all the 60 pairs in the cluster will be labelled as bad.

3.6.2 Evaluation Metrics

External validation indices are used to evaluate the performance of the clustering techniques and feature selection technique. External validation indices compare experts annotations, true labels and predicted labels, i.e. labels obtained from our study. A good performance is indicated by the true labels and predicted labels matching. However, in most cases, notably in unsupervised learning, misclassification is common. In binary classification problems, one class is considered as positive and other is negative. From the comparison of real labels and predicted labels, a data point can either be a true positive (TP), true negative (TN), false positive (FP) or false negative (FN). TP are data points classified as belonging to the positive class by both the real labels and predicted labels, FP are classified as positive by model, but according to the real labels are negative. TN are data points belonging to negative class based on both the real labels and predicted labels and FN is a data point wrongly classified as being negative by the model Hossin & Sulaiman (2015).

Confusion matrix can be used to visualise the above-defined metrics. However, it will be challenging to compare the performance of the different clustering techniques. In most cases, accuracy is used as a performance evaluation metric, $A = TP/(N)$, where N is the total number of data points. Accuracy can be misleading if there is a class imbalance as the dominant class influence the metric. In the case of class imbalances precision (P), given by Equation (3.43) and recall (R), given by Equation (3.44), are used as performance metrics. Recall (R) also known sensitivity is the ratio of TP, and the actual positive, real positive and precision is the ratio between TP and all the predicted positives, TP and FP Powers (2011). The two metrics are summarised by a harmonic mean, and this score is referred to as F1-score and is given by Equation (3.45) Hossin & Sulaiman (2015).

$$P = \frac{TP}{TP + FP} \tag{3.43}$$

$$R = \frac{TP}{TP + FN} \tag{3.44}$$

$$F1 - score = \frac{2PR}{P + R} \tag{3.45}$$

Chapter 4

Results and Discussions

In this section, we discuss the results obtained from our studies. We provide the parameters selected for each technique. We also provide performance and discussion of the clustering techniques and feature selection technique.

4.1 Parameter Selection

All the clustering techniques used in the study except for DBSCAN require the number of clusters to be specified prior clustering. Choosing a small number of clusters can result in excessive data compression and choosing a big number results in good separation but also the loss of characterisation of the clusters as this might result in single point clusters. Therefore parameter selection aims at finding optimal parameters to ensure good quality. For a valid comparison of the different clustering techniques, it is crucial to choose optimal parameters.

For Agglomerative and k-means, clustering techniques that capture with spherical clusters, Davies-Bouldin Indicator (DBI) , Calinski Harabasz (CH), Silhouette coefficient (SC) are calculated to determine the optimal number of clusters. Unlike SC and DBI that have a maximum scale of one, CH does not have a specified of maximum value. Therefore CH is scaled down using maximum CH for plotting purposes. For DBI, the lower the score means better clustering, and for CH and SC score, the higher the score, the better the clustering. We calculated SC, CH and DBI from $K = 2$ to $K = 10$ to determine the optimal number of clusters based on minimised DBI and maximised CH and SC. From Figure 4.1 and Figure 4.2 the optimal parameter according to DBI and CH is five, and according to SC, the optimal number of clusters is three. Seemingly SC tends to underestimate the number of the clusters. The number of clusters for both techniques is chosen as being five ($K = 5$).

Although fuzzy c-means also performs spherical clustering, unlike k-means and agglomerative, a data instance can belong to different clusters in varying degrees indicated by its membership vector. Therefore fuzzy partition coefficient (FPC) which measures the degree of fuzziness is used to choose the suitable number of clusters. We calculate the FPC for the same range as above. Although the highest FPC is two $K = 2$, the selected number of clusters is four $k = 4$ as the FPC of two clusters, and four clusters are not significantly different. In this case, the larger number is chosen to avoid excessive data compression, i.e. $k = 4$. The FPC plot is given in Figure 4.3.

AIC and BIC are used to determine the optimal number of Gaussian models for the GMM technique. AIC and BIC are plotted for $k = 2$ to $k = 10$ shown in Figure 4.4. A minimum AIC and BIC indicate an optimised GMM model. We obtain the minimum BIC at six models, and there is a significant dip at six models for AIC as well. However, unlike BIC which increases as the models are increased from six models, AIC decreases, and this is due to the difference in the penalty term that results in AIC usually overestimating the number of models Xu & Tian (2015). Therefore six models are used in GMM, $k = 6$.

DBSCAN requires the $minPoints$ and ϵ to be specified prior to clustering. The minimum points needed in a neighbourhood are relatively easy to determine using the dimension of the data and $minPoints = 56$. However, ϵ is not as straight forward to determine and is obtained by plotting distance between each data point and its 56th nearest neighbour shown in Figure 4.5. *epsilon* is the elbow of the plot which is 19.

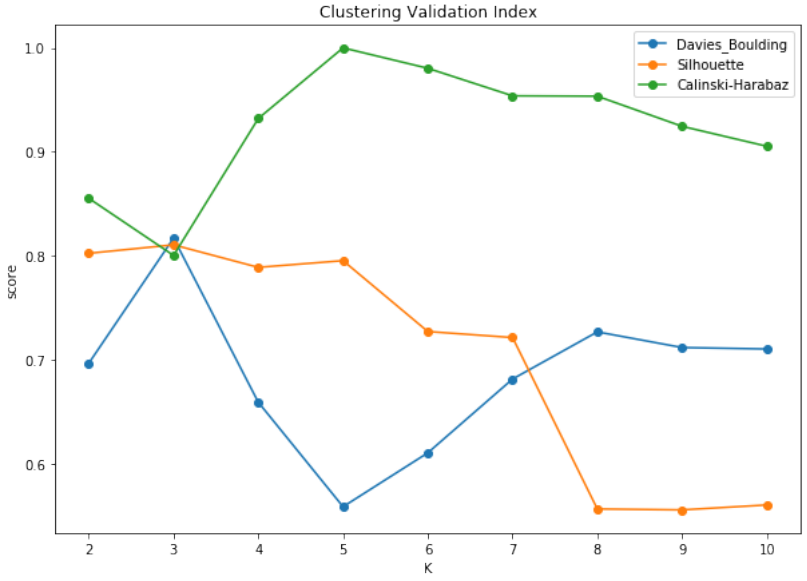


FIGURE 4.1: Clustering Validation Indices for Agglomerative Clustering

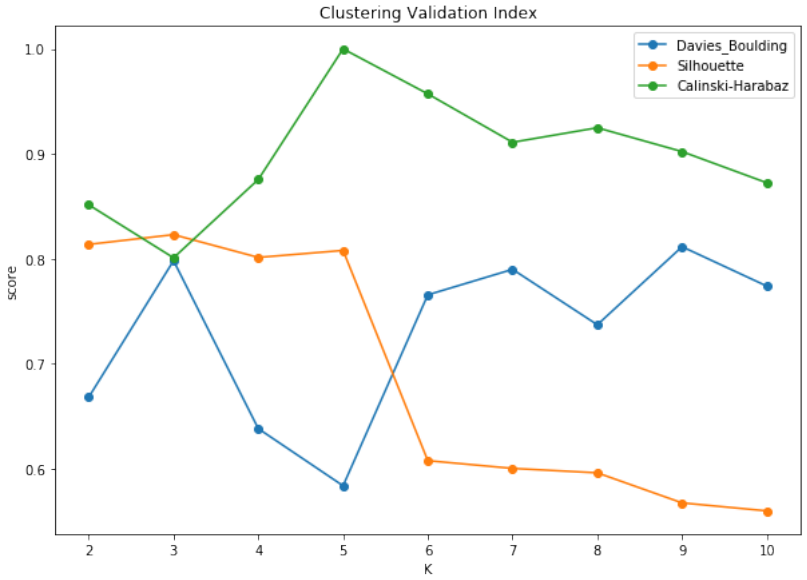


FIGURE 4.2: Clustering Validation Indices for K-means Clustering

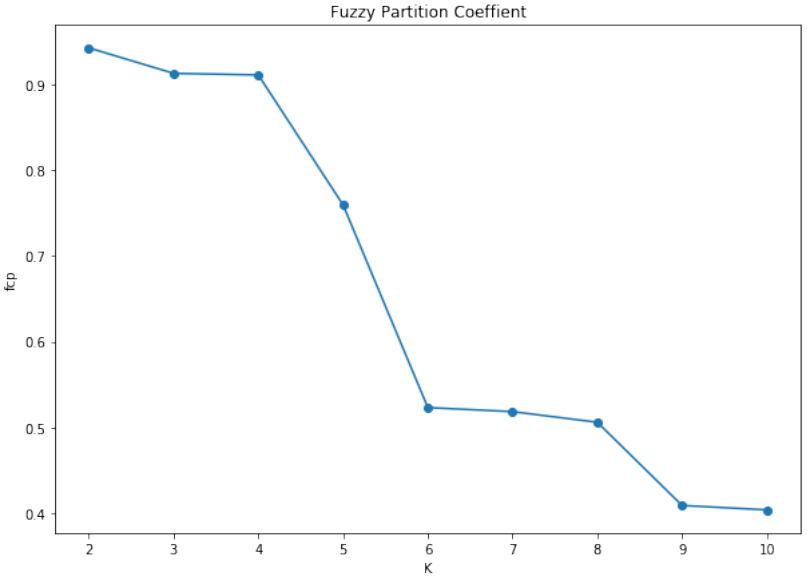


FIGURE 4.3: Fuzzy Partitioning Coefficient

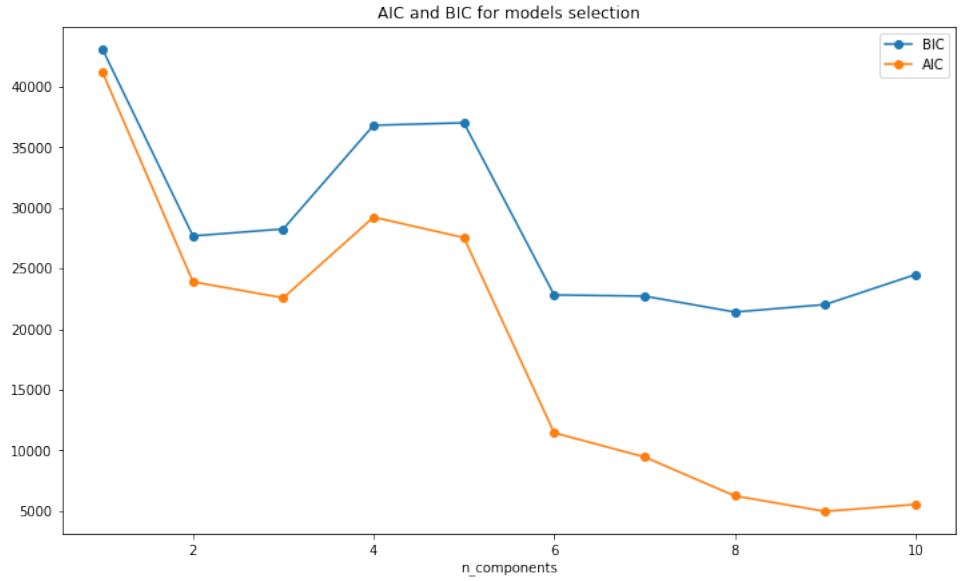


FIGURE 4.4: AIC and BIC indices

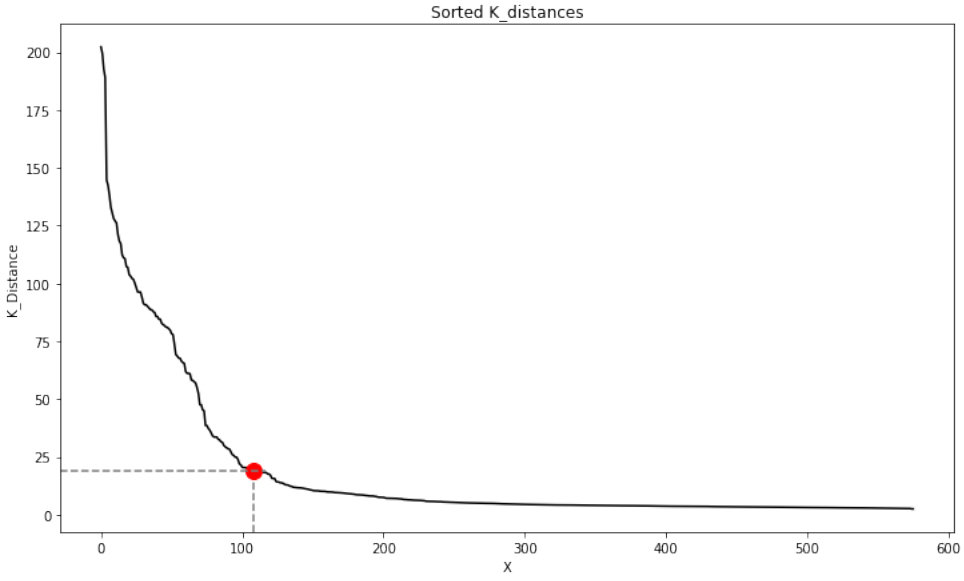


FIGURE 4.5: Determining the ϵ

4.2 Clustering Results

The different clustering techniques make an assumption when clustering, and therefore, it is not always possible to predict which technique will perform best. However, the assumptions made by each technique can be used to infer information about the characteristics of the clusters of the classes,such as shape and size. Although the main aim of this step is to determine the cluster technique that results in the correct classification of the peak pairs, we also evaluate if the different characteristics and assumptions made

by of the cluster techniques prove to be advantageous and results in better clustering solutions. We are looking at the characteristics such as type of clustering, i.e. hard or soft assignments, the ability to detect arbitrarily shapes clusters, sensitivity to outliers and ease of interpretation. Techniques such as k-means, fuzzy c-means and GMM rely on random initialisation of parameters and final clustering solution is affected by initialisation. Therefore, clustering with these techniques is performed five times, and the evaluation metrics are averaged. The resulting clusters, number of peak pairs in each cluster, of the best runs are included.

An ideal classification solution is indicated by homogeneous clusters, however, one of the downsides of using unsupervised learning techniques is that in most cases, false positives are unavoidable. Therefore, clusters are likely to contain both bad and good peak pairs. A misclassified low-quality pair has far worse consequences than misclassified high-quality pairs. Misclassified bad pairs do not undergo further evaluation. However, peak pairs classified as being bad are evaluated by the expert for further troubleshooting; therefore the misclassified high-quality pairs only delay troubleshooting but do not affect the quality of the final experimental results presented for validation and verification phases. Using global metrics, obtained using total TP, FN and FP of the dataset can be misleading, especially if there is a class imbalance. For example, the global f1-score is 0.65 if all the bad peak pairs are misclassified due to the good class being the the dominant class. In our study, the metrics are calculated for each class.

GMM and fuzzy c-means perform soft clustering by assigning posterior probabilities and memberships vectors to each data point. The soft clustering solution is transformed into hard assignments by assigning data points to clusters with maximum probabilities and memberships. The technique of using maximum probabilities and memberships to assign data point to clusters ignores variations in the maximum values. For example, a data point with a maximum membership of 0.4 and a data point with a maximum membership of 0.9 are both assigned clusters even though the maximum memberships are different. Variations of the maximum probabilities and memberships are plotted in Figure 4.6 and Figure 4.7 respectively. The variation of the probabilities is not significant, with the majority of the data points having a maximum probability greater than 0.98. However, there is variation in the maximum memberships. Therefore, we compared the standard transformation, assigning all data points to maximum memberships ($fuzzy_{max}$) and using thresholds, 0.5 ($fuzzy_{0.5}$) and 0.9 ($fuzzy_{0.9}$). In the threshold approach, only data points with maximum memberships greater than the threshold are assigned to clusters, and the rest data points are classified as outliers.

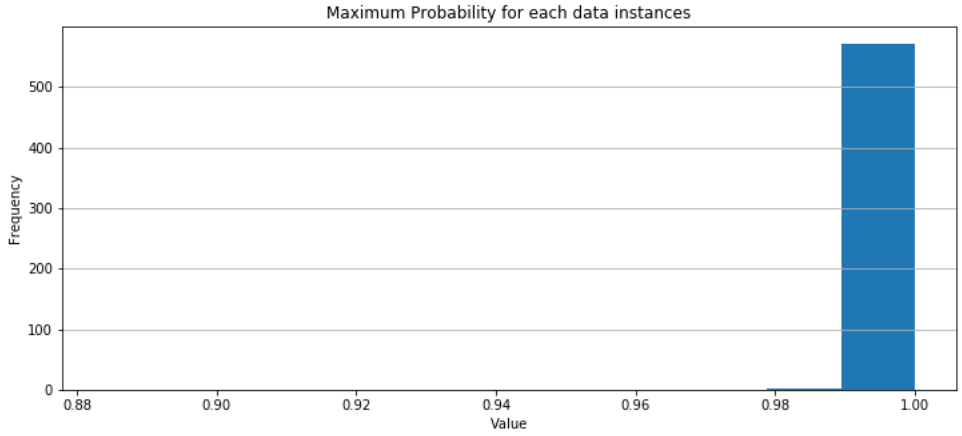


FIGURE 4.6: GMM Probabilities Histogram

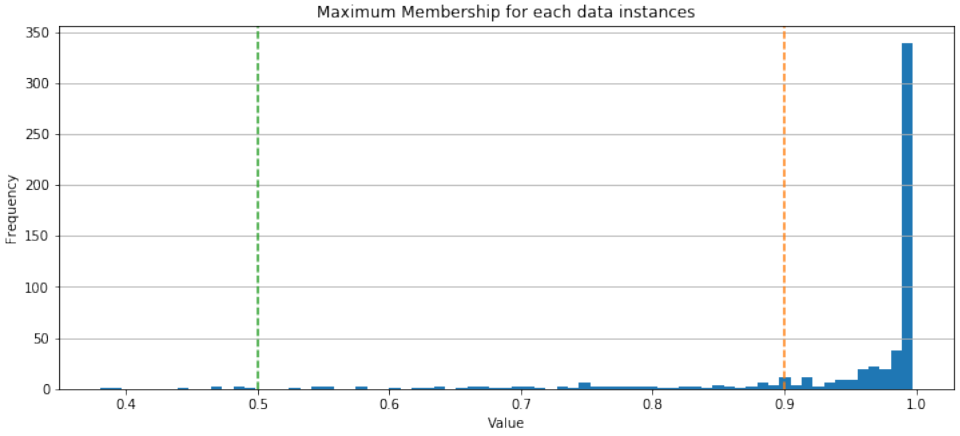


FIGURE 4.7: Fuzzy C Means Membership Histogram

Tables 4.1 - 4.8 indicates how the peak pairs are distributed amongst the clusters, the total pairs, number of bad and good pairs per cluster and outliers. The sizes of the clusters and the representation of the two classes are similar for agglomerative, $fuzzy_{max}$, $fuzzy_{0.5}$, k-means with no seeding, k-means with D2-weighting and GMM. For these techniques, we have a single dominant cluster that contains the majority of high-quality and low-quality pairs, classified as the good class and small clusters with one classified as good and the rest of the clusters are classified as bad. DBSCAN clustering resulting in 84% of the peak pairs being grouped in one cluster and the rest classified as outliers. $fuzzy_{0.9}$ classified 42% of the pairs as outliers, and the remaining pairs are grouped into two clusters of significantly different sizes; however, both of the clusters are assigned to the good class. All outliers are classified as bad. The techniques, except for the $fuzzy_{0.9}$, resulted in more than half of the bad pairs being misclassified, with agglomerative, GMM, k-means with D2-weighting and $Fuzzy_{max}$ misclassifying more 80% of low quality peak pairs. $Fuzzy_{max}$ has the highest percentage of misclassified low-quality

pairs which is 90%. The high number of misclassified pairs is indicated by metrics of the bad class given in Table 4.9. Moreover, the recall metric of the bad class as it depends on the false negatives, with the highest recall being 0.42 and that is for $fuzzy_{0.9}$ and the majority of the techniques having a recall score of less than 0.2. Although the precision of the good class also captures the misclassified bad peak pairs (FP), the lowest precision score is 0.67, and this is due to class imbalance. The low recall values of the bad class are of concern because of the consequences of misclassification of low-quality data points.

The techniques considered in studies, except for DBSCAN, are sensitive to outliers. For fuzzy c-means introducing thresholds, fuzz_0.9 and $fuzzy_{0.5}$ resulted in some of the data points being classified as outliers. $fuzzy_{0.5}$ resulted in the recall of the bad class increasing by 0.02, whereas $fuzzy_{0.9}$ resulted in an increased by 0.2. Although $fuzzy_{0.9}$ has the least number of misclassified peaks, 42% of the data points are detected as outliers, and 45% of the outliers belong to the good class. Taking into consideration that the sum of memberships of a data point should be one, real outliers are likely to have a high maximum membership depending on how far the clusters are from each other. Therefore, introducing thresholds to fuzzy c-means can result in non-outliers being incorrectly classified as outliers as in indicated by 110 high-quality isotopic pairs classified as outliers in Table 4.5. Although the DBSCAN Table 4.8 is capable of detecting outlier, it is dependent on the defined density, ϵ and minPoints. Finding the optimal combination of ϵ and minPoints is vital because if minPoints is overestimated and ϵ is underestimated, then the majority of the data points will be classified as outliers. Underestimating the minPoints and overestimated ϵ can result in all the data points being assigned to a single cluster. DBSCAN resulted in the second-highest recall score of 0.21, which is not significantly from the other techniques.

K-means, GMM and fuzzy c-means perform clustering through iterations and at each iteration, an objective function is optimised. Although the iterations can result in better clustering, the solution can converge to local optima depending on the randomly selected initial parameters. The final clustering solution is dependent on the initial parameters; this is indicated by random initialisation k-means, where one out of the five runs is different to other runs and the standard deviation in Table 4.9. When a form of seeding is used, such as D2 weighting all the five runs are the same.

K-means, fuzzy c-means and agglomerative clustering techniques work well with spherical clusters, and this is a limitation for real datasets. GMM and DBSCAN can detect arbitrarily shapes clusters. Comparing the results of the GMM and DBSCAN and other techniques, it is not easy to deduce the shape of clusters as the results are similar. If GMM and DBSCAN outperformed the other techniques, we could infer that the clusters

are non-spherical. The similarity in the results might suggest that the clusters detected by the techniques are spherical, however it also possible that the results of the techniques are similar even though the techniques detected different shaped clusters. The class imbalance implies the clusters are likely to vary in size. Therefore, the ability of a technique to detect clusters of varying sizes is advantageous in our study. For K-means, GMM, fuzzy c-means the large clusters tend to dominate the smaller clusters hence we have a single dominating cluster that has the highest number of both the bad and good data points in all the techniques except for DBSCAN and $fuzzy_{0.9}$. DBSCAN can detect a different-sized cluster of the same density; if the smaller clusters are sparser than large clusters, then they are likely to be classified as outliers instead of actual clusters.

The recall score of the bad class is a clear indication of the high number of misclassified low-quality pairs. The primary purpose of quality control is to classify isotopic pairs based quality and eliminate the need for manual inspection. Therefore, a reliable quality control solution should be able to classify both good and bad data points correctly making the other metrics also important. The clustering techniques are ranked using the unweighted mean of the f1-scores of both classes. $fuzzy_{0.9}$ has the highest mean followed by k-means with the random initialisation, DBSCAN, GMM, k-means with D2 weighting, $fuzzy_{0.5}$, agglomerative clustering and lastly standard fuzzy c-means. Although $fuzzy_{0.9}$ resulted in the highest number of correctly classified peak pairs, introducing threshold can results high-quality data points being misclassified as outliers.

The results in Table 4.9 and the distribution of data points amongst the clusters indicate there is not a clear separability between the two classes. The different techniques resulted in similar results and distribution of peak pairs, although they make different assumptions about the data and different definition of the separability of clusters and similarity within a cluster. Experts included notes on categories of bad pairs that are in our data set and Figure 4.8 shows the total distribution of the different categories and the number of low-quality pairs per category correctly classified by different techniques. Pairs indicating evidence of jaggedness and dissimilarity in shape or retention times are the most correctly classified. Only a few of the bad pairs are represented by multiple category, and our study misclassified these particular data points. Table 4.9 and Figure 4.8, particularly the results of the bad class, indicates that for clustering to form part of the quality control solution, improvements should be made.

TABLE 4.1: Number of peaks in clusters obtained using Agglomerative Clustering

Cluster	**Total Peaks**	**No. Good Peaks**	**No. Bad Peaks**	**Cluster Labels**
1	503	341	162	Good
2	15	4	11	Bad
3	37	25	12	Good
4	17	7	10	Bad
5	4	0	4	Bad

TABLE 4.2: No of peaks per cluster K-means with random centre intialization

Cluster	**Total Peaks**	**No. Good Peaks**	**No. Bad Peaks**	**Cluster Labels**
1	77	23	54	Bad
2	35	24	11	Good
3	428	319	109	Good
4	16	7	9	Bad
5	20	4	16	Bad

TABLE 4.3: KMeans Clustering -D2 weighting initialization

Cluster	**Total Peaks**	**No. Good Peaks**	**No. Bad Peaks**	**Cluster Labels**
1	505	343	162	Good
2	16	7	9	Bad
3	15	3	12	Bad
4	34	23	11	Good
5	6	1	5	Bad

TABLE 4.4: Fuzzy C Means - Membership Threshold = 0.5

Cluster	**Total Peaks**	**No. Good Peaks**	**No. Bad Peaks**	**Cluster Labels**
Outlier	34	12	22	Bad
1	442	307	135	Good
2	16	7	9	Bad
3	33	23	10	Good
4	51	28	23	Bad

TABLE 4.5: Fuzzy C Means - Membership Threshold = 0.9

Cluster	**Total Peaks**	**No. Good Peaks**	**No. Bad Peaks**	**Cluster Labels**
Outlier	244	110	134	Bad
1	329	265	64	Good
2	3	2	1	Good

TABLE 4.6: Fuzzy C Means Hard Clustering

Cluster	**Total Peaks**	**No. Good Peaks**	**No. Bad Peaks**	**Cluster Labels**
1	20	4	16	Bad
2	508	341	167	Good
3	35	25	10	Good
4	17	7	10	Bad

TABLE 4.7: GMM Clusters

Cluster	**Total Peaks**	**No. Good Peaks**	**No. Bad Peaks**	**Cluster Labels**
1	503	341	162	Good
2	15	7	8	Bad
3	21	18	3	Good
4	6	1	5	Bad
5	15	3	12	Bad
6	16	7	9	Bad

TABLE 4.8: DBSCAN clusters

Cluster	**Total Peaks**	**No. Good Peaks**	**No. Bad Peaks**	**Cluster Labels**
Outlier	93	50	43	Bad
1	483	327	156	Good

TABLE 4.9: Overall Clustering Results

Algorithim	**Bad Peaks**			**Good Peaks**		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Agglomerative	0.69	0.12	0.21	0.68	0.97	0.80
K-means+ random initialization	0.70 ± 0.01	0.18 ± 0.12	0.28 ± 0.13	0.69 ± 0.03	0.96 ± 0.03	0.80 ± 0.01
K-means+ D2 weighting	0.70	0.13	0.23	0.68	0.97	0.80
Fuzzy (max)	0.69 ± 0.03	0.12 ± 0.02	0.20 ± 0.04	0.68	0.97	0.80
Fuzzy (0.5)	0.68 ± 0.03	0.14 ± 0.01	0.23	0.68	0.97 ± 0.01	0.80
Fuzzy (0.9)	0.53 ± **0.02**	**0.41 ± 0.24**	**0.43 ± 0.15**	**0.73 ± 0.06**	**0.81 ± 0.09**	**0.76 ± 0.01**
GMM	0.67 ± 0.03	0.15 ± 0.02	0.25 ± 0.03	0.69 ± 0.01	0.96 ± 0.01	0.80
DBSCAN	0.46	0.21	0.29	0.67	0.86	0.76

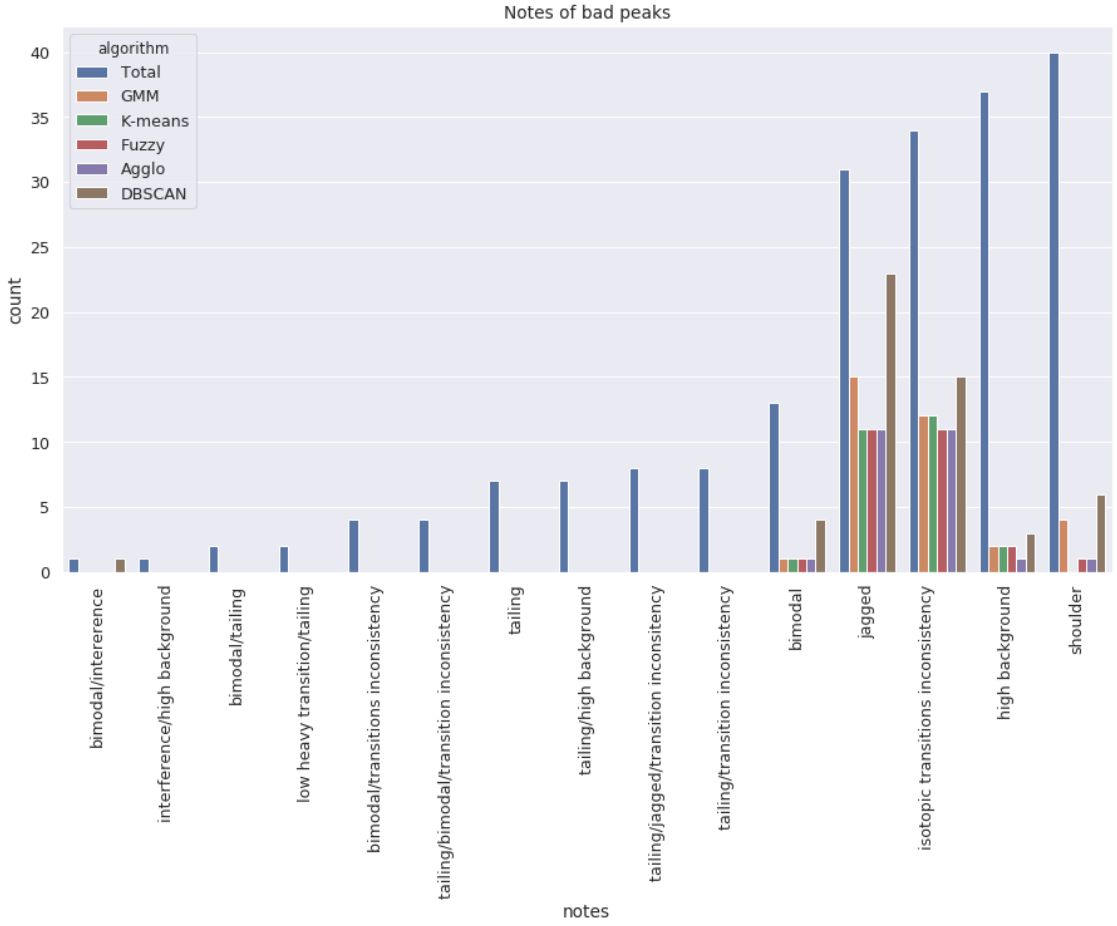


FIGURE 4.8: Notes on Bad Peaks provided by experts

4.3 Genetic Algorithm and Clustering Results

Redundant and irrelevant features could be the cause of the unsatisfactory performance of the clustering techniques. In this section, we evaluate if using genetic algorithm as a feature selection technique will improve the clustering results, precisely the number of correctly classified of low-quality peak pairs. GA performs the feature selection by the minimisation or maximization of an objective function. In our study, our objective function is to minimise the DBI index. As discussed in Chapter 2, DBI index is a CVI for techniques that can detect convex-shaped clusters; however, our feature selection technique used the index concurrently with all clustering techniques. The selection of the number of models for GMM relies on AIC and BIC instead of DBI. However, the two indices are used to ensure that the number of models is not overestimated and therefore, cannot be used for feature selection. The fuzziness of the clusters was used as a measure of clustering quality for fuzzy c-means, FPC was large in parameter selection. Therefore, no significant improvements can be made on the value during the

feature selection process, and improvements are the basis of genetic algorithm. $fuzzy_{0.9}$, $fuzzy_{0.5}$ and k-means randomised initialisation are excluded in this section. Our genetic algorithm performs feature selection through 100 generations and the individual with the lowest DBI at the 100th population is selected as the final solution. Due to the randomisation of the first population, for each clustering technique, genetic algorithm is performed in five runs. The results of the five runs are averaged.

Table 4.10 - 4.14 we observe that the distribution of the data points amongst the clusters differ from the distribution in Section 4.2. K-means and agglomerative clustering have similar distribution, with the largest two clusters classified as good and all the pairs in the remaining three clusters are classified as bad, the cluster sizes are not significantly different. Out of the four clusters obtained using fuzzy c-means, only one cluster is classified as a bad class, which is the second largest cluster. GMM resulted in two of clusters being classified as good, and the remaining four clusters are classified as bad classes. DBSCAN resulted in all the data points being grouped in a single cluster and all the low-quality peak pairs being misclassified as good. The number of misclassified bad data points reduced for all the algorithms except for DBSCAN, with the fuzzy c-means having the highest percentage of 45% misclassified low-quality pairs. K-means has the lowest percentage of misclassified bad pairs which is 39% and this is a significant improvement from the 87% obtained in the clustering section. Incorporating feature selection improved the classification of bad points; it has also increased the number of misclassified high-quality pairs.

From Table 4.15 we can see that for all the algorithms except DBSCAN, have increased the f1-score for the bad class and the f1-score of the good class remained the same. However, the recall of good class has decreased while the recall score of the bad class has increased. Similar to the clustering results, we use the unweighted mean of the f1-scores of the bad and good class to rank the techniques. K-means outperformed all the technique followed by agglomerative, GMM, fuzzy and lastly DBSCAN. Minimising DBI, in our feature selection, proved to be advantageous for k-means and agglomerative clustering. DBI index is sensitive to outliers and proved to be detrimental to DBSCAN. The randomness of the genetic algorithm is indicated by the high standard deviation of the five runs.

The increase in the number of correctly classified low-quality pairs is indicated in Figure 4.9; the majority of the types of bad pairs are correctly classified by the different clustering techniques, including the least represented categories. Table 4.16 shows the features that were selected for the best run of each cluster techniques. At most only half of the features were selected with DBSCAN having the least amount of selected features and GMM and fuzzy c-means with the highest number of selected features which is 14.

The light peaks features that are selected by four out of the five clustering techniques are sharpness, GS, modality and the heavy peak features are sharpness, FWB, z-index, modality. An interesting observation is that although the jagged pairs are one of the more prominent low-quality isotopic pair types and k-means managed to correctly classify the most jagged peaks, the z-index of both the light peak and heavy peak is not included in selected features. Features such as GS and TPASR can be attributed to k-means' ability to classify jagged peak pairs correctly. Feature selection has proved to be useful in separating high-quality and low-quality isotopic pairs, however, 39% misclassified bad data points, obtained by the best performing technique, indicates further work is needed before clustering can be used to form part of the quality control solution.

The approach of using the maximum class to label the clusters has it downfalls, especially in cases where the number of low-quality isotopic pairs and high-quality pairs are close to each other. These approach aims at imitating how an analyst in the lab will perform the classification by selecting a few data points in each clusters and decide the class labels based on the on type of isotopic pairs appearing the most. In instances where analyst cannot decide the class label due to the examples picked, the analyst will most likely investigate the whole cluster. This provides us with alternative approach of not classifying clusters with a similar number of bad and good peaks and therefore ensuring that they are manually analysed in the lab to determine the class label of each isotopic pairs. Our results, indicates that clusters that have a similar number of low-quality and high-quality pairs are classified as bad. Every instances in the bad class goes further analysed manually, therefore our approach will have a similar outcome to that of the alternative approach.

Table 4.10: Agglomerative and GA

Cluster	**Total Peaks**	**No. Good Peaks**	**No. Bad Peaks**	**Cluster Labels**
1	32	13	19	Bad
2	163	114	49	Good
3	211	182	29	Good
4	74	33	41	Bad
5	96	35	61	Bad

Table 4.11: K-means and GA

Cluster	Total Peaks	No. Good Peaks	No. Bad Peaks	Cluster Labels
1	39	15	24	Bad
2	144	114	30	Good
3	138	60	78	Bad
4	216	177	39	Good
5	39	11	28	Bad

Table 4.12: Fuzzy and GA (Davies Bouldin)

Cluster	Total Peaks	No. Good Peaks	No. Bad Peaks	Cluster Labels
1	30	22	8	Good
2	209	99	110	Bad
3	322	246	76	Good
4	15	10	5	Good

Table 4.13: GMM and GA

Cluster	Total Peaks	No. Good Peaks	No. Bad Peaks	Cluster Labels
1	101	20	81	Bad
2	11	4	7	Bad
3	31	21	10	Good
4	12	4	8	Bad
5	26	7	19	Bad
6	395	321	74	Good

Table 4.14: DBSCAN and GA

Cluster	Total Peaks	No. Good Peaks	No. Bad Peaks	Cluster Labels
1	576	377	199	Good

Table 4.15: Clusters and GA evaluation metrics

Algorithim	**Bad Peaks**			**Good Peaks**		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Agglomerative	0.69 ± 0.09	0.39 ± 0.14	0.48 ± 0.09	0.74 ± 0.03	0.90 ± 0.07	0.81 ± 0.01
K-means+ D2 weighting	**0.64 ± 0.04**	**0.49 ± 0.15**	**0.54 ± 0.10**	**0.76 ± 0.04**	**0.86 ± 0.06**	**0.80 ± 0.01**
Fuzzy (max)	0.46 ± 0.3	0.27 ± 0.30	0.28 ± 0.30	0.71 ± 0.06	0.89 ± 0.14	0.79 ± 0.02
GMM	0.63 ± 0.08	0.39 ± 0.23	0.45 ± 0.19	0.74 ± 0.06	0.88 ± 0.07	0.80
DBSCAN	0	0	0	0.65	1	0.79

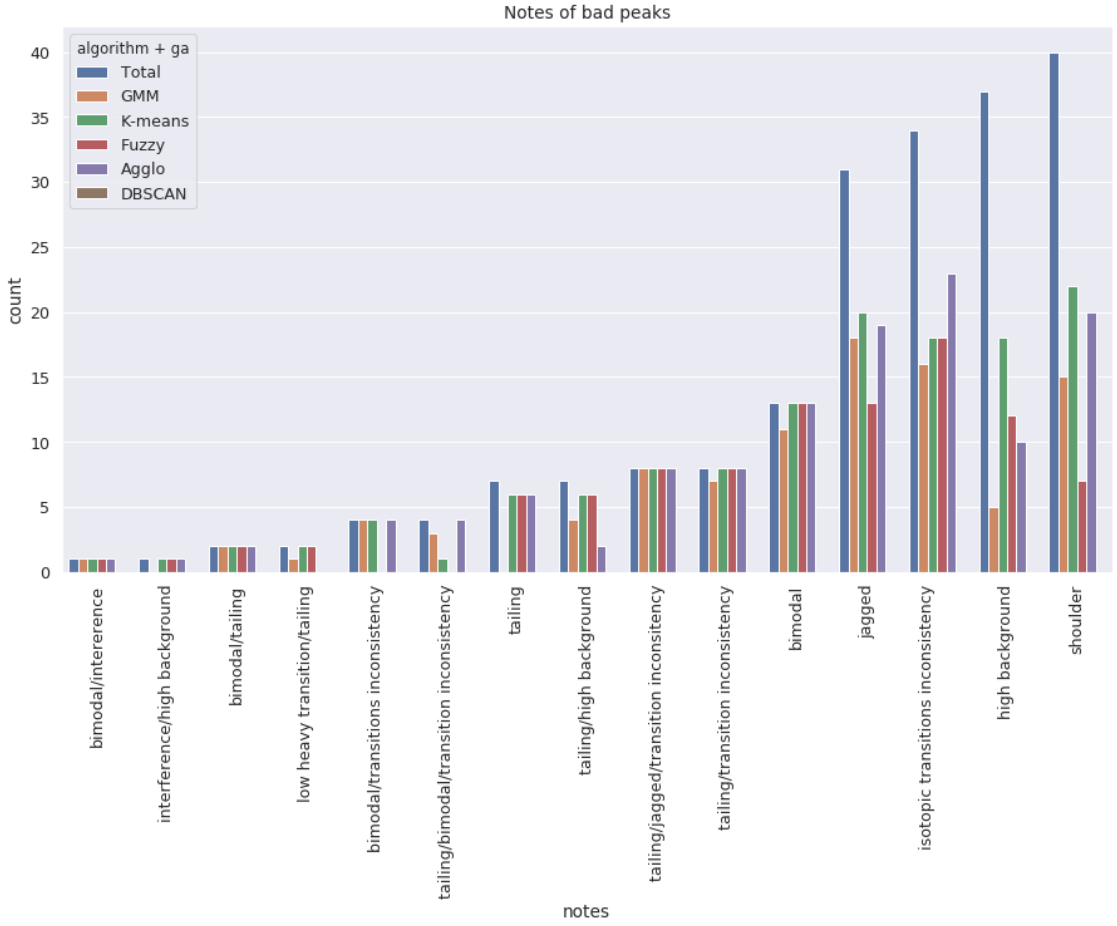


FIGURE 4.9: Notes on Bad Peaks provided by experts after feature selection

4.4 Discussions

Proteomic quality control solutions can either be depended or independent of the data processing stage. Solutions independent of the data processing stage were implemented to eliminate the errors introduced by the stage. Our study uses data processed by Skyline; however, instead of using peaks detected using Skyline, we use the chromatograms of all the identified fragments. The peak detection method of Skyline is prone to errors; therefore, by using chromatograms, this error is eliminated. However, errors occurring during chromatogram extraction and resampling are unavoidable. The advantage of using data processed using Skyline instead of the raw data is that chromatogram extraction procedure and distinguishing actual fragments from noise is a non-trivial process.

Data pre-processing, peak detection method and feature engineering, are vital to the success of our study. Therefore, there is a need to evaluate the performance of data - pre-processing independent of the clustering techniques. For experimental-based quality

TABLE 4.16: Selected Features

Algorithms	**Features**													
	symmetry_h	sharpness_h	fwhm_h	fwb_h	snr_h	GS_h	tpasr_h	zz_index_h	peak_significance_h	modality_h	fragment_shift_h	fragment_similarity_h	isotope_shift_h	isotope_similarity_h
Agglomerative		x	x	x	x	x	x	x	x	x				
K-means	x	x		x		x				x			x	
Fuzzy	x		x	x	x	x	x	x	x	x				
GMM		x	x	x		x	x	x						
DBSCAN		x				x	x	x	x	x				

Algorithms	**Features**													
	symmetry_l	sharpness_l	fwhm_l	fwb_l	snr_l	GS_l	tpasr_l	zz_index_l	peak_significance_l	modality_l	fragment_shift_l	fragment_similarity_l	isotope_shift_l	isotope_similarity_l
Agglomerative		x				x		x	x			x		
K-means		x				x	x		x	x		x		
Fuzzy	x			x		x	x			x				
GMM	x	x				x				x				x
DBSCAN	x	x					x			x				

control, extensive inter-lab and intra-lab studies were performed to define metrics that capture any sub-optimal performance. Comprehensive studies are also needed to define new features and evaluate if existing features capture the differences between known high-quality and low-quality data peaks.

Our study used clustering techniques that rely on concepts such as partitioning, hierarchy, density, fuzziness and statistical models to classify to separate low-quality and high-quality peak pairs. Although the various clustering techniques are based on different concepts and perform clustering differently, all the techniques resulted in a significant high number of low-quality isotopic pairs misclassified. The unsatisfactory performance of the clustering techniques can be due to errors in the pre-processing stage, i.e. errors in peaks detection and the engineered features. The other reasons might include the class imbalance between good and bad peak pairs and the inclusion of irrelevant features, and this is indicated by the improvement of the results with the incorporation of feature selection. Although the number of misclassified low-quality isotopic pairs decreased due to feature selection, there is also an increase in the number of the misclassified high-quality pairs. The objective function, precisely minimising the DBI index, resulted in DBSCAN misclassifying of all the bad data points because the index is sensitive to outliers. DBSCAN's performance can be improved by using a density-based index capable of handling outliers. However, the objective function resulted in major improvements in k-means and agglomerative clustering results. The disadvantage of using a genetic algorithm as a feature selection technique affects the ability to reproduce the results due to the randomisation of the initial population. The recall scores the bad class even after feature selection indicates that further studies are still needed to improve the classification of the peaks pairs based on quality.

4.5 Research Questions Answers

1. **Which clustering technique performs best in classifying isotopic pairs?**

   Introducing thresholds based on data points' maximum memberships distribution to fuzzy c-means improved the results of the standard fuzzy c-means and fuzzy c-means with a threshold of 0.9 outranked all the techniques. The unweighted mean of the f1-scores is used to rank the techniques. Excluding deviants of fuzzy c-means and randomised k-means, then DBSCAN outperforms all the other techniques.

2. **Does the integration of the feature selection technique improve the clustering results?**

The majority of clustering techniques improved with the incorporation of a feature selection technique. DBSCAN is the only technique with a decline in performance. The performance of DBSCAN can be improved by using density-based CVI.

3. **Which of the clustering results improved the most by adding feature selection?**

   By minimising DBI index proved to be advantageous to clustering technique that works with convex-shaped clusters because of how the index define compactness and separability. K-means is the most improved technique and also outranks other techniques.

4. **Which are the prominent features selected by the different algorithms?**

   Gaussian similarity, modality, full width at base and zig-zag index are features selected by four out of five features. Although the feature selection technique was introduced to improve the clustering techniques, it can also assist in troubleshooting.

5. **Is using unsupervised learning as a quality control tool for proteomics experiments a viable solution?** Our study indicates that a quality control solution made up of pre-processing and clustering is likely to results in a significant number of low-quality peak pairs being incorrectly classified. Our unsupervised feature selection technique improved the clustering techniques results. However, further studies are needed to improve the results before our unsupervised learning solution can be used as a quality control solution.

Chapter 5

Conclusion

Our work presents an unsupervised quality control solution for proteomics isotopic pairs that relies on clustering and genetic algorithms. The purpose of the quality control solution is to monitor the quality of peak pairs, thus monitoring the performance of experimental procedure. We used various clustering techniques to classify the isotopic pairs based on quality. Redundant and irrelevant features hinder the performance of the clustering techniques. Therefore, we implemented a feature selection technique to improve on the results.

Although clustering techniques are based on different concepts such as hierarchical clustering, partitioning, statistical models and density, to group the data points, the performance of the methods was similar. One of the disadvantages of the agglomerative clustering, k-means and fuzzy c-mean is the inability to handle non-spherical clusters. The performance of the GMM is not significantly different from the result of convex-shaped techniques despite its ability to detect arbitrarily shaped clusters. The techniques with exception to the DBSCAN had similar data points distribution, one dominant cluster containing the majority of good and bad pairs and smaller clusters. DBSCAN resulted in a single cluster and other data points classified as outliers. The data points classified as outliers can either be true outliers or belong in clusters less dense than expected based on ϵ and minPoints.

The soft assignments obtained using fuzzy c-means is transformed into hard assignments by assigning data points to clusters with the highest degree of membership for fuzzy c-means and probabilities for GMM. Due to variation of maximum memberships, We introduced thresholds, of 0.5 and 0.9, and data points with maximum memberships less than thresholds are classified as outliers. Although using 0.9 as a threshold resulted in the highest recall score for the bad class, 45% of the high-quality peak pairs are classified as outliers.

Introducing feature selection improved the performance and the highest percentage of misclassified bad peaks reduced from 90% to 45% when excluding DBSCAN feature selection results. The objective function proved to be advantageous for k-means and agglomerative as they are significant improvements in the results of the techniques, with k-means outperforming all the other techniques. The combination of the feature selection and DBSCAN resulted in poor performance; all the low quality peak pairs were misclassified. This poor performance indicates a more suitable CVI is needed. Although feature selection proved to be useful further studies such as evaluating and optimising the pre-processing stage of the solution.

Our studies show that unsupervised learning, specifically clustering and feature selection, has the potential to be used in quality control of proteomics studies. However, before it is employed as a quality control solution, further studies such as optimising pre-processing stage, are needed to improve the results. The availability of larger experts-annotated data will also be beneficial in improving the performances of the techniques.

Chapter 6

Future Works

Currently the solution is only evaluated after clustering, this makes is difficult to determine the cause of the unsatisfactory performance. Therefore, each component of the solution needs to be evaluated individually, and this will enable us to identify components that need to improving. For feature engineering, although there is scientific theory substantiating the engineered features, evaluations are needed to make sure that a feature such similarity can detect an actual tailing peak from a normal peak as defined by an expert. If the evaluation indicates that the feature does not actually detect the problem, a new approach of engineering that particular feature is needed. For run based quality control, there have been extensive studies performed to come up with a list of metrics. The same is needed for peak based QC to make sure that all the bad peak categories are covered and to improve the availability of expert annotated data. The dataset used in this study has 1172 peaks, and a single DIA experiment can have up to ten thousands peaks per run, indicating that the categories present do not represent every possible type of low-quality peak. For the scope of the project data processing was excluded, and we relied on the data processed through Skyline, and this also introduces errors in the work. Therefore we need to develop a data processing algorithm.

The clustering results can be improved by using enhanced versions of the techniques used in this study that can handle outliers, different sized clusters such as Chameleon. For feature selection Davies Bouldin Index was used as the objective function; however, this metric works with spherical clusters detecting techniques such as k-means and agglomerative, suitable validation indices are needed for the other techniques such as Density Based Clustering Validation for DBSCAN.

In proteomics, visualisations and ease of use are important in ensuring that the solution is adopted in the laboratories. Therefore, after optimising all the stages: peak detection, feature engineering and clustering with feature selection, we will design a graphic user

interface (GUI). The GUI will allow the user to load the raw data, perform the QC, view all the classified low-quality data points and evaluate them at ease.

Bibliography

Abualigah, L. M., Khader, A. T. & Al-Betar, M. A. (2016), Unsupervised feature selection technique based on genetic algorithm for improving the text clustering, *in* '2016 7th international conference on computer science and information technology (CSIT)', IEEE, pp. 1–6.

Addison, P. S. (2018), 'Introduction to redundancy rules: the continuous wavelet transform comes of age'.

Alelyani, S., Tang, J. & Liu, H. (2018), Feature selection for clustering: A review, *in* 'Data Clustering', Chapman and Hall/CRC, pp. 29–60.

Amidan, B. G., Orton, D. J., LaMarche, B. L., Monroe, M. E., Moore, R. J., Venzin, A. M., Smith, R. D., Sego, L. H., Tardiff, M. F. & Payne, S. H. (2014), 'Signatures for mass spectrometry data quality', *Journal of proteome research* **13**(4), 2215–2222.

Aoshima, K., Takahashi, K., Ikawa, M., Kimura, T., Fukuda, M., Tanaka, S., Parry, H. E., Fujita, Y., Yoshizawa, A. C., Utsunomiya, S.-i. et al. (2014), 'A simple peak detection and label-free quantitation algorithm for chromatography-mass spectrometry', *BMC bioinformatics* **15**(1), 376.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., PéRez, J. M. & Perona, I. (2013), 'An extensive comparative study of cluster validity indices', *Pattern Recognition* **46**(1), 243–256.

Arthur, D. & Vassilvitskii, S. (2007), k-means++: The advantages of careful seeding, *in* 'Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms', Society for Industrial and Applied Mathematics, pp. 1027–1035.

Aslam, B., Basit, M., Nisar, M. A., Khurshid, M. & Rasool, M. H. (2017), 'Proteomics: technologies and their applications', *Journal of chromatographic science* **55**(2), 182–196.

Bell, A. W., Deutsch, E. W., Au, C. E., Kearney, R. E., Beavis, R., Sechi, S., Nilsson, T., Bergeron, J. J., Beardslee, T. A., Chappell, T. et al. (2009), 'A hupo test sample study reveals common problems in mass spectrometry–based proteomics', *Nature methods* **6**(6), 423.

Bereman, M. S., Johnson, R., Bollinger, J., Boss, Y., Shulman, N., MacLean, B., Hoofnagle, A. N. & MacCoss, M. J. (2014), 'Implementation of statistical process control for proteomic experiments via lc ms/ms', *Journal of the American Society for Mass Spectrometry* **25**(4), 581–587.

Bereman, M. S., MacLean, B., Tomazela, D. M., Liebler, D. C. & MacCoss, M. J. (2012), 'The development of selected reaction monitoring methods for targeted proteomics via empirical refinement', *Proteomics* **12**(8), 1134–1141.

Berkhin, P. (2006), A survey of clustering data mining techniques, *in* 'Grouping multidimensional data', Springer, pp. 25–71.

Bielow, C., Mastrobuoni, G. & Kempa, S. (2015), 'Proteomics quality control: quality control software for maxquant results', *Journal of proteome research* **15**(3), 777–787.

Bishop, C. M. (2006), *Pattern recognition and machine learning*, springer.

Bittremieux, W., Meysman, P., Martens, L., Valkenborg, D. & Laukens, K. (2016), 'Unsupervised quality assessment of mass spectrometry proteomics experiments by multivariate quality control metrics', *Journal of proteome research* **15**(4), 1300–1307.

Bittremieux, W., Tabb, D. L., Impens, F., Staes, A., Timmerman, E., Martens, L. & Laukens, K. (2017), 'Quality control in mass spectrometry-based proteomics', *Mass spectrometry reviews* .

Bittremieux, W., Walzer, M., Tenzer, S., Zhu, W., Salek, R. M., Eisenacher, M. & Tabb, D. L. (2017), 'The hupo-psi quality control working group: Making quality control more accessible for biological mass spectrometry', *Analytical chemistry.-Washington, DC, 1948, currens* **89**(8), 4474–4479.

Bittremieux, W., Willems, H., Kelchtermans, P., Martens, L., Laukens, K. & Valkenborg, D. (2015), 'imondb: mass spectrometry quality control through instrument monitoring', *Journal of proteome research* **14**(5), 2360–2366.

Bodzon-Kulakowska, A., Bierczynska-Krzysik, A., Dylag, T., Drabik, A., Suder, P., Noga, M., Jarzebinska, J. & Silberring, J. (2007), 'Methods for samples preparation in proteomic research', *Journal of Chromatography B* **849**(1-2), 1–31.

Boersema, P. J., Mohammed, S. & Heck, A. J. (2008), 'Hydrophilic interaction liquid chromatography (hilic) in proteomics', *Analytical and bioanalytical chemistry* **391**(1), 151–159.

Bramwell, D. (2013), 'An introduction to statistical process control in research proteomics', *Journal of proteomics* **95**, 3–21.

Branden, C. I. & Tooze, J. (2012), *Introduction to protein structure*, Garland Science.

Calderón-Celis, F., Encinar, J. R. & Sanz-Medel, A. (2018), 'Standardization approaches in absolute quantitative proteomics with mass spectrometry', *Mass spectrometry reviews* **37**(6), 715–737.

Chandola, V., Banerjee, A. & Kumar, V. (2009), 'Anomaly detection: A survey', *ACM computing surveys (CSUR)* **41**(3), 15.

Chiva, C., Olivella, R., Borràs, E., Espadas, G., Pastor, O., Solé, A. & Sabidó, E. (2018), 'Qcloud: A cloud-based quality control system for mass spectrometry-based proteomics laboratories', *PloS one* **13**(1), e0189209.

Choi, M., Eren-Dogu, Z. F., Colangelo, C., Cottrell, J., Hoopmann, M. R., Kapp, E. A., Kim, S., Lam, H., Neubert, T. A., Palmblad, M. et al. (2017), 'Study: detection of differentially abundant proteins in label-free quantitative lc-ms/ms experiments', *J Proteome Res* **16**(2), 945–957.

Choi, M. & Sweetman, B. (2010), 'Efficient calculation of statistical moments for structural health monitoring', *Structural Health Monitoring* **9**(1), 13–24.

Coelho, G. P., Barbante, C. C., Boccato, L., Attux, R. R., Oliveira, J. R. & Von Zuben, F. J. (2012), Automatic feature selection for bci: an analysis using the davies-bouldin index and extreme learning machines, *in* 'The 2012 international joint conference on neural networks (IJCNN)', IEEE, pp. 1–8.

Deb, A. B. & Dey, L. (2017), 'Outlier detection and removal algorithm in k-means and hierarchical clustering', *World Journal of Computer Application and Technology* **5**(2), 24–29.

Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B. et al. (2010), 'A guided tour of the trans-proteomic pipeline', *Proteomics* **10**(6), 1150–1159.

D'haeseleer, P. (2005), 'How does gene expression clustering work?', *Nature biotechnology* **23**(12), 1499–1501.

Doerr, A. (2014), 'Dia mass spectrometry', *Nature methods* **12**(1), 35.

Dogu, E., Mohammad-Taheri, S., Abbatiello, S. E., Bereman, M. S., MacLean, B., Schilling, B. & Vitek, O. (2017), 'Msstatsqc: Longitudinal system suitability monitoring and quality control for targeted proteomic experiments', *Molecular & Cellular Proteomics* **16**(7), 1335–1347.

Dogu, E., Taheri, S. M., Olivella, R., Marty, F., Lienert, I., Reiter, L., Sabido, E. & Vitek, O. (2018), 'Msstatsqc 2.0: R/bioconductor package for statistical quality control of mass spectrometry-based proteomics experiments', *Journal of proteome research* **18**(2), 678–686.

Du, P., Kibbe, W. A. & Lin, S. M. (2006), 'Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching', *Bioinformatics* **22**(17), 2059–2065.

Eshghi, S. T., Auger, P. & Mathews, W. R. (2018), 'Quality assessment and interference detection in targeted mass spectrometry data using machine learning', *Clinical proteomics* **15**(1), 33.

Estivill-Castro, V. & Yang, J. (2000), Fast and robust general purpose clustering algorithms, *in* 'Pacific Rim International Conference on Artificial Intelligence', Springer, pp. 208–218.

Feist, P. & Hummon, A. (2015), 'Proteomic challenges: sample preparation techniques for microgram-quantity protein analysis from biological samples', *International journal of molecular sciences* **16**(2), 3537–3563.

Finney, G. L. (2012), Tools and Analyses for Differential Label-Free Proteomics Using Mass Spectrometry, PhD thesis.

Foley, J. P. & Dorsey, J. G. (1983), 'Equations for calculation of chromatographic figures of merit for ideal and skewed peaks', *Analytical Chemistry* **55**(4), 730–737.

Fraley, C. & Raftery, A. E. (2002), 'Model-based clustering, discriminant analysis, and density estimation', *Journal of the American statistical Association* **97**(458), 611–631.

Goebel-Stengel, M., Stengel, A., Taché, Y. & Reeve Jr, J. R. (2011), 'The importance of using the optimal plasticware and glassware in studies involving peptides', *Analytical biochemistry* **414**(1), 38–46.

Goldstein, M. & Uchida, S. (2016), 'A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data', *PloS one* **11**(4), e0152173.

Golubev, A. (2017), 'Exponentially modified peak functions in biomedical sciences and related disciplines', *Computational and mathematical methods in medicine* **2017**.

Grossmann, A., Kronland-Martinet, R. & Morlet, J. (1990), Reading and understanding continuous wavelet transforms, *in* 'Wavelets', Springer, pp. 2–20.

Hartigan, J. A., Hartigan, P. M. et al. (1985), 'The dip test of unimodality', *The annals of Statistics* **13**(1), 70–84.

Hassanat, A., Almohammadi, K., Alkafaween, E., Abunawas, E., Hammouri, A. & Prasath, V. (2019), 'Choosing mutation and crossover ratios for genetic algorithms—a review with a new dynamic approach', *Information* **10**(12), 390.

Hernandez, N. (2001), 'Small nuclear rna genes: a model system to study fundamental mechanisms of transcription', *Journal of Biological Chemistry* **276**(29), 26733–26736.

Herrero, J. & Dopazo, J. (2002), 'Combining hierarchical clustering and self-organizing maps for exploratory analysis of gene expression patterns', *Journal of Proteome Research* **1**(5), 467–470.

Hinneburg, A. & Gabriel, H.-H. (2007), Denclue 2.0: Fast clustering based on kernel density estimation, *in* 'International symposium on intelligent data analysis', Springer, pp. 70–80.

Hood, L. & Rowen, L. (2013), 'The human genome project: big science transforms biology and medicine', *Genome medicine* **5**(9), 79.

Hossin, M. & Sulaiman, M. (2015), 'A review on evaluation metrics for data classification evaluations', *International Journal of Data Mining & Knowledge Management Process* **5**(2), 1.

Huang, A. (2008), Similarity measures for text document clustering, *in* 'Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand', Vol. 4, pp. 9–56.

Ioannidis, J. P. & Bossuyt, P. M. (2017), 'Waste, leaks, and failures in the biomarker pipeline', *Clinical chemistry* **63**(5), 963–972.

Ivanov, A. R., Colangelo, C. M., Dufresne, C. P., Friedman, D. B., Lilley, K. S., Mechtler, K., Phinney, B. S., Rose, K. L., Rudnick, P. A., Searle, B. C. et al. (2013), 'Interlaboratory studies and initiatives developing standards for proteomics', *Proteomics* **13**(6), 904–909.

Jebari, K. & Madiafi, M. (2013), 'Selection methods for genetic algorithms', *International Journal of Emerging Sciences* **3**(4), 333–344.

Jović, A., Brkić, K. & Bogunović, N. (2015), A review of feature selection methods with applications, *in* '2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)', IEEE, pp. 1200–1205.

Kalambet, Y., Kozmin, Y., Mikhailova, K., Nagaev, I. & Tikhonov, P. (2011), 'Reconstruction of chromatographic peaks using the exponentially modified gaussian function', *Journal of Chemometrics* **25**(7), 352–356.

Karpievitch, Y. V., Polpitiya, A. D., Anderson, G. A., Smith, R. D. & Dabney, A. R. (2010), 'Liquid chromatography mass spectrometry-based proteomics: biological and technological aspects', *The annals of applied statistics* **4**(4), 1797.

Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. (2008), 'Proteowizard: open source software for rapid proteomics tools development', *Bioinformatics* **24**(21), 2534–2536.

Khan, S. S. & Ahmad, A. (2004), 'Cluster center initialization algorithm for k-means clustering', *Pattern recognition letters* **25**(11), 1293–1302.

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S. et al. (2014), 'A draft map of the human proteome', *Nature* **509**(7502), 575.

Kindler, S., Wang, H., Richter, D. & Tiedge, H. (2005), 'Rna transport and local control of translation', *Annu. Rev. Cell Dev. Biol.* **21**, 223–245.

Klir, G. & Yuan, B. (1995), *Fuzzy sets and fuzzy logic*, Vol. 4, Prentice hall New Jersey.

Kriegel, H.-P., Kröger, P., Schubert, E. & Zimek, A. (2009), Loop: local outlier probabilities, *in* 'Proceedings of the 18th ACM conference on Information and knowledge management', ACM, pp. 1649–1652.

Kutbay, U. (2018), 'Partitional clustering', *Recent Applications in Data Clustering* p. 19.

Kuzyk, M. A., Hardie, D. B., Yang, J., Smith, D. S., Jackson, A. M., Parker, C. E. & Borchers, C. H. (2010), 'A quantitative study of the effects of chaotropic agents, surfactants, and solvents on the digestion efficiency of human plasma proteins by trypsin', *Journal of proteome research* **9**(10), 5422–5437.

Langfelder, P., Zhang, B. & Horvath, S. (2008), 'Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r', *Bioinformatics* **24**(5), 719–720.

Latecki, L. J., Lazarevic, A. & Pokrajac, D. (2007), Outlier detection with kernel density functions, *in* 'International Workshop on Machine Learning and Data Mining in Pattern Recognition', Springer, pp. 61–75.

Legrain, P., Aebersold, R., Archakov, A., Bairoch, A., Bala, K., Beretta, L., Bergeron, J., Borchers, C. H., Corthals, G. L., Costello, C. E. et al. (2011), 'The human proteome project: current state and future direction', *Molecular & cellular proteomics* **10**(7), M111–009993.

Lemenkova, P. (2020), 'R libraries {dendextend} and {magrittr} and clustering package scipy. cluster of python for modelling diagrams of dendrogram trees', *Carpathian Journal of Electronic and Computer Engineering* **13**(1), 5–12.

Li, M., Schwartzman, A. et al. (2018), 'Standardization of multivariate gaussian mixture models and background adjustment of pet images in brain oncology', *The Annals of Applied Statistics* **12**(4), 2197–2227.

Lill, J. (2003), 'Proteomic tools for quantitation by mass spectrometry', *Mass Spectrometry Reviews* **22**(3), 182–194.

Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J. & Wu, S. (2013), 'Understanding and enhancement of internal clustering validation measures', *IEEE transactions on cybernetics* **43**(3), 982–994.

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D. & Darnell, J. (2000), The three roles of rna in protein synthesis, *in* 'Molecular Cell Biology. 4th edition', WH Freeman.

Łukasik, S., Kowalski, P. A., Charytanowicz, M. & Kulczycki, P. (2016), Clustering using flower pollination algorithm and calinski-harabasz index, *in* '2016 IEEE Congress on Evolutionary Computation (CEC)', IEEE, pp. 2724–2728.

Ma, Z.-Q., Polzin, K. O., Dasari, S., Chambers, M. C., Schilling, B., Gibson, B. W., Tran, B. Q., Vega-Montoto, L., Liebler, D. C. & Tabb, D. L. (2012), 'Quameter: multivendor performance metrics for lc–ms/ms proteomics instrumentation', *Analytical chemistry* **84**(14), 5845–5850.

Martin, H. & Honarvar, F. (1995), 'Application of statistical moments to bearing failure detection', *Applied acoustics* **44**(1), 67–77.

Maugis, C., Celeux, G. & Martin-Magniette, M.-L. (2009), 'Variable selection for clustering with gaussian mixture models', *Biometrics* **65**(3), 701–709.

Maulik, U. & Bandyopadhyay, S. (2000), 'Genetic algorithm-based clustering technique', *Pattern recognition* **33**(9), 1455–1465.

McNicholas, P. D. (2016), 'Model-based clustering', *Journal of Classification* **33**(3), 331–373.

Michalski, A., Cox, J. & Mann, M. (2011), 'More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent lc- ms/ms', *Journal of proteome research* **10**(4), 1785–1793.

Milardi, D., Grande, G., Vincenzoni, F., Pierconti, F. & Pontecorvi, A. (2019), 'Proteomics for the identification of biomarkers in testicular cancer–review', *Frontiers in endocrinology* **10**, 462.

Mirjalili, S. (2019), Genetic algorithm, *in* 'Evolutionary algorithms and neural networks', Springer, pp. 43–55.

Munoz, J. & Heck, A. J. (2014), 'From the human genome to the human proteome', *Angewandte Chemie International Edition* **53**(41), 10864–10866.

Murtagh, F. & Legendre, P. (2011), 'Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm', *arXiv preprint arXiv:1111.6285* .

Nikravesh, M. (2007), Evolution of fuzzy logic: From intelligent systems and computation to human mind, *in* 'Forging new frontiers: Fuzzy pioneers I', Springer, pp. 37–53.

Noga, M., Sucharski, F., Suder, P. & Silberring, J. (2007), 'A practical guide to nano-lc troubleshooting', *Journal of separation science* **30**(14), 2179–2189.

Noller, H. F. (1991), 'Ribosomal rna and translation', *Annual review of biochemistry* **60**(1), 191–227.

Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A. & Mann, M. (2002), 'Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics', *Molecular & cellular proteomics* **1**(5), 376–386.

Parker, C. E. & Borchers, C. H. (2014), 'Mass spectrometry based biomarker discovery, verification, and validation–quality assurance and control of protein biomarker assays', *Molecular oncology* **8**(4), 840–858.

Percy, A. J., Chambers, A. G., Yang, J., Hardie, D. B. & Borchers, C. H. (2014), 'Advances in multiplexed mrm-based protein biomarker quantitation toward clinical utility', *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1844**(5), 917–926.

Petrovic, S. (2006), A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters, *in* 'Proceedings of the 11th Nordic Workshop of Secure IT Systems', sn, pp. 53–64.

Pichler, P., Mazanek, M., Dusberger, F., Weilnböck, L., Huber, C. G., Stingl, C., Luider, T. M., Straube, W. L., Köcher, T. & Mechtler, K. (2012), 'Simpatiqco: a server-based software suite which facilitates monitoring the time course of lc-ms performance metrics on orbitrap instruments', *Journal of proteome research* **11**(11), 5540–5547.

Pino, L. K., Searle, B. C., Bollinger, J. G., Nunn, B., MacLean, B. & MacCoss, M. J. (2017), 'The skyline ecosystem: Informatics for quantitative mass spectrometry proteomics', *Mass spectrometry reviews* .

Poli, R. & Langdon, W. B. (1998), Genetic programming with one-point crossover, *in* 'Soft Computing in Engineering Design and Manufacturing', Springer, pp. 180–189.

Powers, D. M. (2011), 'Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation'.

Premstaller, A., Oberacher, H., Walcher, W., Timperio, A. M., Zolla, L., Chervet, J.-P., Cavusoglu, N., van Dorsselaer, A. & Huber, C. G. (2001), 'High-performance liquid chromatography- electrospray ionization mass spectrometry using monolithic capillary columns for proteomic studies', *Analytical chemistry* **73**(11), 2390–2396.

Razali, N. M., Geraghty, J. et al. (2011), Genetic algorithm performance with different selection strategies in solving tsp, *in* 'Proceedings of the world congress on engineering', Vol. 2, International Association of Engineers Hong Kong, pp. 1–6.

Reddy, C. K. & Vinzamuri, B. (2018), A survey of partitional and hierarchical clustering algorithms, *in* 'Data clustering', Chapman and Hall/CRC, pp. 87–110.

Richardson, J. S. (1981), 'The anatomy and taxonomy of protein structure', *Advances in protein chemistry* **34**, 167–339.

Roointan, A., Mir, T. A., Wani, S. I., Hussain, K. K., Ahmed, B., Abrahim, S., Savardashtaki, A., Gandomani, G., Gandomani, M., Chinnappan, R. et al. (2018), 'Early detection of lung cancer biomarkers through biosensor technology: A review', *Journal of pharmaceutical and biomedical analysis* .

Rudnick, P. A., Clauser, K. R., Kilpatrick, L. E., Tchekhovskoi, D. V., Neta, P., Blonder, N., Billheimer, D. D., Blackman, R. K., Bunk, D. M., Cardasis, H. L. et al. (2010), 'Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses', *Molecular & Cellular Proteomics* **9**(2), 225–241.

Sander, J., Ester, M., Kriegel, H.-P. & Xu, X. (1998), 'Density-based clustering in spatial databases: The algorithm gdbscan and its applications', *Data mining and knowledge discovery* **2**(2), 169–194.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W. & Lin, C.-T. (2017), 'A review of clustering techniques and developments', *Neurocomputing* **267**, 664–681.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. (2017), 'Dbscan revisited, revisited: why and how you should (still) use dbscan', *ACM Transactions on Database Systems (TODS)* **42**(3), 19.

Shewhart, W. A. (1924), 'Some applications of statistical methods to the analysis of physical and engineering data', *Bell System Technical Journal* **3**(1), 43–87.

Shewhart, W. A. (1938), 'Application of statistical methods to manufacturing problems'.

Shi, Y., Xiang, R., Horváth, C. & Wilkins, J. A. (2004), 'The role of liquid chromatography in proteomics', *Journal of Chromatography A* **1053**(1-2), 27–36.

Singh, A., Yadav, A. & Rana, A. (2013), 'K-means with three different distance metrics', *International Journal of Computer Applications* **67**(10).

Spears, W. M. & De Jong, K. A. (1991), An analysis of multi-point crossover, *in* 'Foundations of genetic algorithms', Vol. 1, Elsevier, pp. 301–315.

Srinivas, M. & Patnaik, L. M. (1994), 'Genetic algorithms: A survey', *computer* **27**(6), 17–26.

Tabb, D. L., Vega-Montoto, L., Rudnick, P. A., Variyath, A. M., Ham, A.-J. L., Bunk, D. M., Kilpatrick, L. E., Billheimer, D. D., Blackman, R. K., Cardasis, H. L. et al. (2009), 'Repeatability and reproducibility in proteomic identifications by liquid chromatography- tandem mass spectrometry', *Journal of proteome research* **9**(2), 761–776.

Tautenhahn, R., Boettcher, C. & Neumann, S. (2008), 'Highly sensitive feature detection for high resolution lc/ms', *BMC bioinformatics* **9**(1), 504.

Taylor, R. M., Dance, J., Taylor, R. J. & Prince, J. T. (2013), 'Metriculator: quality assessment for mass spectrometry-based proteomics', *Bioinformatics* **29**(22), 2948–2949.

Trauwaert, E. (1988), 'On the meaning of dunn's partition coefficient for fuzzy clusters', *Fuzzy sets and systems* **25**(2), 217–242.

Tsou, C.-C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A.-C. & Nesvizhskii, A. I. (2015), 'Dia-umpire: comprehensive computational framework for data-independent acquisition proteomics', *Nature methods* **12**(3), 258.

Velmurugan, T. (2012), 'Efficiency of k-means and k-medoids algorithms for clustering arbitrary data points', *Int. J. Computer Technology & Applications* **3**(5), 1758–1764.

Visser, N., Lingeman, H. & Irth, H. (2005), 'Sample preparation for peptides and proteins in biological matrices prior to liquid chromatography and capillary zone electrophoresis', *Analytical and bioanalytical chemistry* **382**(3), 535–558.

Wang, H., Shi, T., Qian, W.-J., Liu, T., Kagan, J., Srivastava, S., Smith, R. D., Rodland, K. D. & Camp, D. G. (2016), 'The clinical impact of recent advances in lc-ms for cancer biomarker discovery and verification', *Expert review of proteomics* **13**(1), 99–114.

Wang, X., Chambers, M. C., Vega-Montoto, L. J., Bunk, D. M., Stein, S. E. & Tabb, D. L. (2014), 'Qc metrics from cptac raw lc-ms/ms data interpreted through multivariate statistics', *Analytical chemistry* **86**(5), 2497–2509.

Whitley, D. (1994), 'A genetic algorithm tutorial', *Statistics and computing* **4**(2), 65–85.

Wildsmith, K. R., Schauer, S. P., Smith, A. M., Arnott, D., Zhu, Y., Haznedar, J., Kaur, S., Mathews, W. R. & Honigberg, L. A. (2014), 'Identification of longitudinally dynamic biomarkers in alzheimer's disease cerebrospinal fluid by targeted proteomics', *Molecular neurodegeneration* **9**(1), 22.

Xu, D. & Tian, Y. (2015), 'A comprehensive survey of clustering algorithms', *Annals of Data Science* **2**(2), 165–193.

Xue, B., Zhang, M., Browne, W. N. & Yao, X. (2015), 'A survey on evolutionary computation approaches to feature selection', *IEEE Transactions on Evolutionary Computation* **20**(4), 606–626.

Yang, Y. & Kamel, M. (2003), Clustering ensemble using swarm intelligence, *in* 'Proceedings of the 2003 IEEE Swarm Intelligence Symposium. SIS'03 (Cat. No. 03EX706)', IEEE, pp. 65–71.

Yoder, J. & Priebe, C. E. (2017), 'Semi-supervised k-means++', *Journal of Statistical Computation and Simulation* **87**(13), 2597–2608.

Zadeh, L. A. (1988), 'Fuzzy logic', *Computer* **21**(4), 83–93.

Zhang, J., Xu, Y., Xia, Y. & Li, J. (2008), 'A new statistical moment-based structural damage detection method', *Structural Engineering and Mechanics* .

Zhang, W., Chang, J., Lei, Z., Huhman, D., Sumner, L. W. & Zhao, P. X. (2014), 'Met-cofea: a liquid chromatography/mass spectrometry data processing platform for metabolite compound feature extraction and annotation', *Analytical chemistry* **86**(13), 6245–6253.

Zhang, W. & Zhao, P. X. (2014), Quality evaluation of extracted ion chromatograms and chromatographic peaks in liquid chromatography/mass spectrometry-based metabolomics data, *in* 'BMC bioinformatics', Vol. 15, BioMed Central, p. S5.

Zhang, Y., Fonslow, B. R., Shan, B., 15 Baek, M.-C. & Yates III, J. R. (2013), 'Protein analysis by shotgun/bottom-up proteomics', *Chemical reviews* **113**(4), 2343–2394.

Zhu, X., Wang, Y., Li, Y., Tan, Y., Wang, G. & Song, Q. (2019), 'A new unsupervised feature selection algorithm using similarity-based feature clustering', *Computational Intelligence* **35**(1), 2–22.