MASS SPECTRAL IMAGING OF CLINICAL SAMPLES USING DEEP LEARNING

MARCO BRILLANTI

IMPERIAL COLLEGE LONDON

DEPARTMENT OF METABOLISM, DIGESTION AND REPRODUCTION

PhD Thesis

DECLARATION OF ORIGINALITY

I declare that the work presented in this PhD titled "Mass Spectral Imaging of Clinical Samples Using Deep Learning" has been carried out by me in the Department of Metabolism, Digestion and Reproduction. The information derived from other scientific literature has been duly acknowledge in the text and referenced in the bibliography.

The work has been seen by my supervisor before presentation.

COPYRIGHT DECLARATION

ACKNOWLEDGEMENTS

I want to thank my supervisors, Prof Robert C Glen and Prof Zoltan Takats, for their help, care, and consistent support throughout my PhD.

I thank Merck KGaA for giving me the possibility and the instruments to achieve this important result.

I thank Dr Paolo Inglese for the support about everything related to mass spectrometry imaging, for introducing me to the statistical analysis of mass spectral data and for sharing his code during the early stage of my PhD.

I want to dedicate this thesis to my parents, in particular to my father that was the pillar of my family, and to my friends, especially Andrea, for their support and care throughout these four years.

ABSTRACT

A better interpretation of tumour heterogeneity and variability is vital for the improvement of novel diagnostic techniques and personalized cancer treatments. Tumour tissue heterogeneity is characterized by biochemical heterogeneity, which can be investigated by unsupervised metabolomics.

Mass Spectrometry Imaging (MSI) combined with Machine Learning techniques have generated increasing interest as analytical and diagnostic tools for the analysis of spatial molecular patterns in tissue samples. Considering the high complexity of data produced by the application of MSI, which can consist of many thousands of spectral peaks, statistical analysis and in particular machine learning and deep learning have been investigated as novel approaches to deduce the relationships between the measured molecular patterns and the local structural and biological properties of the tissues.

Machine learning have historically been divided into two main categories: Supervised and Unsupervised learning. In MSI, supervised learning methods may be used to segment tissues into histologically relevant areas e.g. the classification of tissue regions in H&E (Haemotoxylin and Eosin) stained samples. Initial classification by an expert histopathologist, through visual inspection enables the development of univariate or multivariate models, based on tissue regions that have significantly up/down-regulated ions. However, complex data may result in underdetermined models, and alternative methods that can cope with high dimensionality and noisy data are required.

Here, we describe, apply, and test a novel diagnostic procedure built using a combination of MSI and deep learning with the objective of delineating and identifying biochemical differences between cancerous and non-cancerous tissue in metastatic liver cancer and epithelial ovarian

cancer. The workflow investigates the robustness of single (1D) to multidimensional (3D) tumour analyses and also highlights possible biomarkers which are not accessible from classical visual analysis of the H&E images. The identification of key molecular markers may provide a deeper understanding of tumour heterogeneity and potential targets for intervention.

Table of Contents

List of Abbreviations

3-D	3-dimensional
2-D	2-dimensional
AI	Artificial Intelligence
ANN	Artificial Neural Network
CD	Contrastive Divergence
Cer	Ceramide
CNN	Convolutional Neural Network
CSR	Complete Spatial Randomness
DBI	Davies-Bouldin Index
DBN	Deep Belief Network
DESI	Desorption Electro-Spray Ionisation
DL	Deep Learning
DR	Dimensionality Reduction
H&E	Hematoxylin and Eosin
IR	Infrared
k-NN	k-Nearest Neighbour
LDA	Linear Discriminant Analysis
m/z	Mass-To-Charge ratio
MALDI	Matrix-Assisted Laser Desorption Ionisation
MS	Mass Spectrometry
MSI	Mass Spectrometry Imaging
NMI	Normalised Mutual Information
OPTICS	Ordering Points To Identify the Clustering Structure
PCA	Principal Component Analysis
PE	Phosphatidylethanolamine
PG	Phosphatidylglycerol
PLA	Phospholipase
PS	Phosphatidylserine
RBM	Restricted Boltzmann Machine
RGB	Red-Green-Blue colour channels
ROI	Region of Interest
SIMS	Secondary Ion Mass Spectrometry
SPUTNIK	SPatially aUTomatic deNoising for Ims ToolkitPla
SSAE	Stacked Sparse Auto Encoder
SSIM	Structural Similarity Index Measure
SVM	Support Vector Machine
TGF	Transforming growth factor
TIC	Total Ion Count
t-SNE	t-distributed Stochastic Neighbour Embedding

List of Figures

List of Tables

1 BACKGROUND - AI AND DATA ANALYSIS IN MEDICAL DIAGNOSIS

Artificial Intelligence (AI) has made stunning progress over the last years: voice-controlled digital assistants, visual translation services, self-driving cars, the internet of things (IoT) and other advances have made it into homes and headlines. Most of these developments are based on a new family of methodologies typically described as Deep Learning (DL).

Artificial Intelligence (AI) has recently undergone a renaissance as a branch of computer science concerned with building *smart* machines capable of performing tasks that generally require human intelligence.

One could view the recent developments in AI as advances built on earlier work in machine learning (ML), ML is also being adopted in many sectors to understand, predict, and improve tasks that require e.g. analysis of complex high dimensional data. ML studies and offers computer algorithms that may improve 'automatically' through continuous improvement and by the use of ever larger sources of data. In particular, especially in a research context, the family of ML methods which include Deep Learning (DL) are highly promising.

DL is a technique to perform machine learning inspired by our brain's own network of neurons. However, the human brain seems to recognize the world in a fundamentally different way, learning mostly in an unsupervised manner with minimal amounts of labelled (training) data. Therefore, novel DL methodologies blend disciplines like mathematics, computer science, and neuroscience. This interdisciplinary approach aims to rethink the foundations of artificial and biological intelligence.

DL algorithms use different architectures of networks that go under the name of deep Artificial Neural Networks (ANNs) that have revolutionised many modern technological sectors like Computer Vision (CV), Natural Language Processing (NLP), Time Series Analysis, Generative DL, and many others.

The major breakthroughs in the field were made by Hinton et al with several publications. On 2007 Hinton et al published an article entitled *Boltzmann machine* (Hinton , E. et al., 2007), in 2009 introduced *Deep Belief Networks* (Hinton, E. et al., 2009). Hinton et al. in 2014 also published *Dropout*, an efficient way of training neural Networks (Hinton, E et al, 2014). All these researches have opened the current deep learning era.

Like in many other fields, AI in medical diagnosis has contributed significantly to the evolution of medical informatics and biomedicine. New technologies, tools and computational power have increased the complexity for developing medical diagnostic systems to act as decision support.

The application of DL on imaging (in this specific case medical imaging) or also video (archived and Real-time) is used in medical diagnosis to classify, adapt, learn, and interpret big and complex datasets.

In medical diagnosis, recent work (Esteva, A. et al., 2021) (Wang, S., et al., 2021) (Zhou, S.K., et al., 2021) (Islam, M.M., et al., 2021) (Devunooru, S., et al., 2021) applied DL algorithms combined with a variety of ML methods to highlight tumour sub-structures and metabolic pathways. A combination of this technology or even tailored algorithms is applied to diagnose, classify, and predict different medical conditions assisting clinicians and medical staff in providing a more precise diagnosis for their patients' health conditions.

Therefore, it is clear that the medical diagnostic field is showing an increase in the number of intelligent systems being developed that could also learn from new cases while in operation (Davenport, T. et al., 2019). In the future, as researchers continue to develop DL methods and AI techniques, computer applications in medicine appear to be increasing in complexity, and in the performance of more accurate diagnosis (Richens, J.G., et al., 2020).

With the advent of powerful computing, huge storage capacity of modern machines, and also Cloud Computing Systems (e.g., Google Platform Computing - GCP, Amazon Web Service - AWS, and others) health care organisations and many research groups have started to collect large quantities of data about illness (and wellness) in different ways and formats with extensive metadata to provide the best support for their patients while reducing cost and enhancing the quality of health care.

Critical investments in medical AI methods promote the development of advanced core computer science research and engineering to provide enhanced quality and effectiveness of health care. High-tech diagnostic procedures are improving the quality and longevity of life that can be realised by the collection and in-depth analysis of data collected on health and illness (European Union, 2019).

Methods for learning, mining and visualising from data can provide the foundations for new directions in medicine by the development of tools and insights that identify even subtle (but essential) signals in the clinical data (Islam Md S., et al., 2018) (Lysaght, T., et al., 2019) (Sinkala M., et al., 2020).

Just a few examples on how AI systems are routinely assisting clinicians can be represented by mentioning modern tools designed by biotech companies all over the world. An American company based in San Francisco, called Enlitic, develops DL medical tools to streamline radiology

diagnoses. The DL platform analyses unstructured medical data (X-ray images, blood tests, EKGs, genomics, medical history) to give clinicians better insights into a patient's needs (in a real-time manner). Harvard University's teaching hospital uses AI-enhanced microscopes to scan for harmful bacteria in patient's blood samples at a faster rate than is possible using traditional scanning with 95% accuracy. A final; example is the work performed by a company called H2O.ai that utilises AI to predict ICU (Intensive Care Unit) transfers. This improves clinical workflows and even identify a patient's risk of hospital-acquired infections. It can also predict and detect sepsis quickly mining patients' health data (Wu, M., et al., 2021).

2 CANCER BIOLOGY AND MASS SPECTROMETRY IMAGING (MSI)

2.1 CANCER BIOLOGY

Cancer is a disease characterised by a group of mutated cells growing uncontrollably by disregarding the standard rules of cell division. Carcinogenesis, tumorigenesis, or oncogenesis are similar names used to refer to this process (Garcia-Milian R., 2014).

Healthy cells are continuously exposed to signals, internal and external, that dictate whether the cell should differentiate, divide, or die. Cancer cells develop a degree of freedom from these dynamics, resulting in uncontrolled growth and proliferation, which is typically classified as invasion first and a metastatic phase later, if allowed to continue and spread, it can be fatal. Almost 90% of patients diagnosed with cancer die due to the metastatic process.

Over the past fifty years, spectacular advances in cancer research, have given insight into how the disrupted behaviour of these cells is related to genome mutations. Mutated DNA leads to the production of proteins that disrupt the fragile cellular homeostasis. Successive epochs of mutation and selective expansion of these cells results in the formation of a tumour mass. Subsequent rounds of mutation and expansion lead to tumour growth and progression, which eventually breaks through the basal membrane barrier surrounding tissues and spreads to other parts of the body, generating a phenomenon previously introduced as metastasis. For the sake of precision, not all the cancers induce a metastatic process. The causes that trigger this malignant

process are still under extensive study (Micalizzi, D.S., et al., 2021) (American Association for Cancer Research, 2020) (Fares J, et al., 2020) (Zhao ZM, et al., 2016).

Initiation and progression of cancers do not depend only on a hereditary factor, they also depend on external elements of the environment, such as tobacco, chemicals, radiation, and infectious organisms. However, tumours proliferate only under advantageous conditions; therefore, it is through a metastatic mechanism that the tumours survive and grow, colonising new body locations where there are no restrictions on space or nutrients. Like any other tumours, metastases also show disruption to cellular pathways. However, metastatic tumours display additional unique cellular features, which enable them to change and adapt to their new environments.

Most cancer cells usually die or survive for long periods of time as micro-metastases. These small agglomerations of tumour cells are much harder to detect with classical/standard diagnostic procedures compare to more voluminous agglomeration (bigger metastasis) (Selves J., et al., 2018).

We can refer to cancer as a multifaceted disease. It is different in one patient to another and continuously evolves into a progressively complex interplay of different tumour cells with their changing environment (Marte B, et al., 2013).

Tumours present two different types of heterogeneities; the intra-tumour spatial heterogeneity refers to the remarkable differences shown by multiple tumour specimens obtained from the same patient confirming the presence of sub-structures between geographical regions in the same tumour (Campbell P., et al., 2010) (Wu X., et al., 2012) (Tao Y., et al., 2011). Temporal heterogeneity instead refers to the heterogeneity present between the primary tumour and its distal metastasis (LaBonia G.J., et al., 2016).

Tumour heterogeneity poses a challenge to personalised cancer medicine because a single needle biopsy or surgical excision is unlikely to accurately capture the complete landscape of a patient's cancer (Caprioli R.M., et al., 2010) (Bedard P.L., et al., 2013).

Both type of differentiation, geographical and temporal, can be linked with a histological classification of the tumour tissue. Also, it is possible to refer to a different classification of tumour heterogeneity called phenotypic heterogeneity which is biologically relevant because heterogeneous tumours often behave differently from homogenous tumours, and this can be of prognostic significance (Marusyk A., et al., 2010). Phenotypic heterogeneity arises among cancer cells within the same tumour because of genetic change, environmental differences, and reversible changes in cell properties (Misale S. et al., 2012) (Stanta G., et al., 2016). Human cancer intra-tumour phenotypic heterogeneity has been associated with tumour progression, treatment resistance and metastasis development (Martinez-Outschoorn U.E., et al., 2017).

Determining the regions of tumour heterogeneity and similarity is not only very important to investigate the nature of the dissimilarity of tumours and to classify those into sub-groups, but can be, through topological mapping of the heterogeneity, a crucial tool to understand the possible interactions between those different cell clusters (Ness R.O., et al., 2016).

A possible technical solution is provided by 3D Mass Spectrometry Imaging (MSI), which can capture the different molecular patterns present in sub-regions of the tumour tissue, and it represents a highly promising approach for determining tumour microenvironment heterogeneity (Oppenheimer S.R., et al., 2010) (McCombie G., et al., 2005). Later in this work, we provide additional support to the importance of studying the biological interactions in a 3-dimensional environment (Fonville J.M., et al., 2013) (Yue X., et al., 2016).

2.2 MASS SPECTROMETRY IMAGING (MSI)

Pathology has recently entered the era of personalised medicine and treatments, this brings new expectations concerning the accuracy and precision of tissue-based diagnosis, especially, when quantification of histologic features and biomarker expression is required. For many years, traditional pathology diagnosis, based on visual inspection of histological images, has been considered as ground truth, which is no longer a sufficient concept in contemporary biomarker research and clinical use (Vranic S., et al., 2021).

The application of Machine learning brings the possibility of significant changes in pathology. Computer vision with other imaging techniques provide immediate benefits of increased capacity, automation, precision, and analysis reproducibility (Laurinavicius A., et al., 2012). In this setting, anatomic pathology is becoming increasingly quantitative or analytical, and similar to other scientific areas is moving from analogue to digital, which presents new possibilities to process the signals and retrieving new information that may not be discovered due to the complexity or vagaries of human vision.

ML and DL are powerful methodologies capable of mining huge amount of data. This modern approach opens the possibility to enlarge the scale of biological analysis introducing new possibilities. For instance, the combination of ML/DL with Mass Spectrometry (MS), which can create large datasets that comprehensively profile the chemical environment of tissues, may significantly improve the throughput, accuracy and reliability of diagnosis.

Mass spectrometers use the differences in the mass-charge ratio (*m/z*) of ionised fragments of molecules to separate and quantify them. MS allows quantitation of molecules and their fragments and provides structural information by the identification of distinctive fragmentation patterns.

MSI allows the investigation of the spatial distribution of molecules on tissue surfaces or tissue volumes. The combination of local analysis with molecular characterisation leads to the creation of a 'chemical microscope' which can be used for the direct bio-molecular characterisation of histological tissue surfaces (Ashrafian H., et al., 2021).

The use of MSI allows label-free detection and mapping of a wide range of biological compounds. The presence or absence of these molecules can be the direct result of disease pathology (Amstalden van Hove E.R., et al., 2010).

Multi-modal MSI strategies have made this technique a powerful tool for spatial localisation and identification of elements, pharmaceuticals, metabolites, lipids, peptides, and proteins in biological tissues. The application of MSI approaches for a bio-chemical analysis of tissues have two main advantages over conventional radiography approaches. Firstly, MSI does not require a radioactive label, and secondly, it allows simultaneous detection of the drug compound (if the MSI technique is used to trace and assess the drug action during the treatment of a disease) and metabolites in tissue. MSI can be used to differentiate between drugs and metabolites and also provide histological information in the cancer research field, which makes it a promising tool for finding new biomarkers and investigating proteins through the direct and morphology-driven analysis of tissue sections. On the other hand, the application of MSI techniques is typically applied to dead tissue (based on tissue sample analysis), which represents a (possibly severe) disadvantage of this method. However, this is common with many other standard methods in pathology, such as excising and freezing or chemically fixing tissues.

Recently, the integration of 3D-MSI data with H&E, or with IHC (immunohistochemistry), has allowed relationships between histological and molecular information to be deduced leading to a better interpretation of the volumetric tumour heterogeneity (Lotz J.M., et al., 2017) (Nemes

P., et al., 2007) (Kottke P.A., et al., 2010) (Addie R.D., et al., 2015). The ability of MSI to perform in-depth histological classification demonstrates its ability to acquire clinically relevant data and can be considered the necessary first step of any new diagnostic/prognostic tool (Deininger S.O., et al., 2008) (Jones E.A., et al., 2013) (Hanselmann M., et al., 2009).

MSI datasets can be obtained through the adoption of different analytical methods and techniques. A brief description of the most common methods used is provided here.

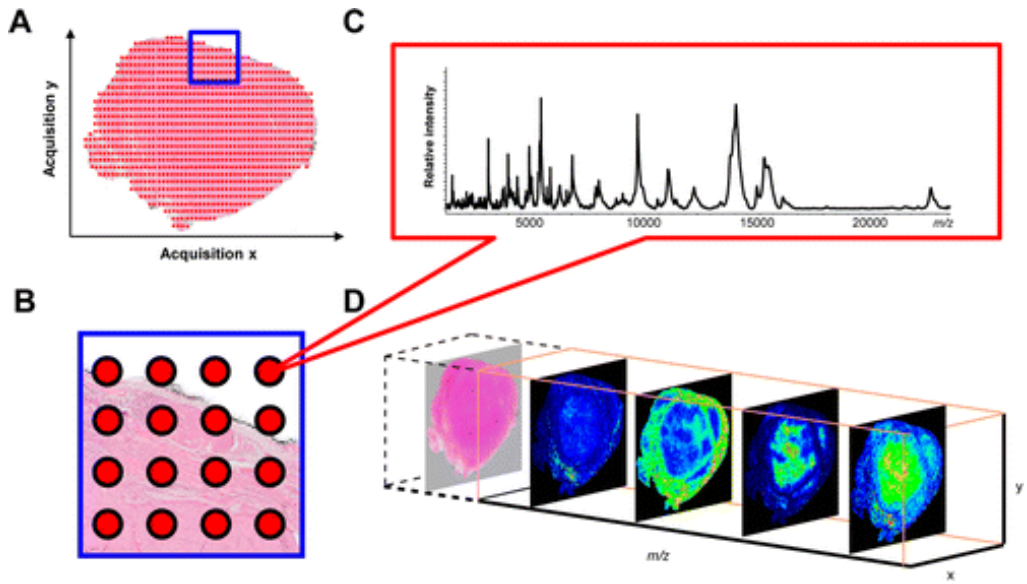


*Figure 2.1 – An illustrative representation of a generic MSI structure (Ruben, D. A., et al., 2015). In section **A** is reported the grid of points where the MS probe will be positioned. The coordinates (x, y) of this position are recorded. Section **B** shows an enlargement of the tissue sample (H&E image) showing the locations where the MS will be acquired. For every of these location with coordinates (x, y) an MS is acquired (regardless the MS technique used). Section **C** shows the MS acquired for a specific location. Finally, section **D** reports the whole MS acquire for the specific location. For every (x, y) the MS acquires information in a specific range of m/z values, which are represented by different images embedded in a single pixel.*

Mainly, the acquisition of an MSI dataset using different methodologies or technologies is based on a shared principle, which is summarised in the following list of technical steps. The mass spectrum is measured over a small area of a sample and represents the molecular composition of that area. Through the iteration of this process, adjacent areas are scanned defining a virtual two-

dimensional grid that covers the entire sample. For every single scan, the coordinates (*x, y*) are registered, hence creating the pixels of a hyperspectral-like image, which is plotted to obtain a raw representation of the MSI.

The technology utilised to acquire an MSI dataset consists, briefly, in the utilisation of a probe that scans the whole sample moving row-by-row or column-by-column, while mass spectra are (continuously or discretely) acquired. Spatial coordinates are recorded using the position of the probe as a reference. When the acquisition of the tissue sample's information is completed, the generated dataset consisting of raw mass spectra and their corresponding spatial coordinates (Figure 2.1). This dataset represents the MS device output and is afterwards analysed using a variety of statistical. ML and AI methods.

As mentioned before, the are several analytical techniques that go under the MSI category. The primary three technologies are DESI-MSI, MALDI-MSI and SIMS-MSI. The significant difference between these techniques is represented by the analytical method used to extract the MSI data and to the technical process used to deliver the samples to the mass spectrometer. These are described below.

2.2.1 MALDI-MSI

One of the most recent and widely adopted MSI techniques is the soft ionisation technique (Hillenkamp F., et al., 1991) matrix-assisted laser desorption ionisation that goes under the name of MALDI (Römpp A., et al., 2015).

The ionisation is based on the interaction of an analyte, which could be, for instance, a tissue sample, embedded in a matrix, with a laser of a defined wavelength focussed on the sample.

MALDI requires a unique sample preparation that consists of the application of a matrix to the sample surface.

The matrix, which is a compound of crystallized molecules, acts like a buffer between the sample and the laser photons. It also helps the ionisation of the sample thanks to a better transfer of the energy carried by the laser (particularly in the UV or IR spectrum). The matrix mixes with the molecular content of the sample and absorbs the ultraviolet light and converts it to heat energy.

This samples preparation procedure is of fundamental importance for the ion extraction outcome, in fact, it highly depends on the homogenous distribution of the matrix on top of the tissue sample (Schwartz S. A., et al., 2003). There are two main techniques used to evenly apply the matrix on top of the sample, the first one consists of spray a small matrix droplet on top of the tissues under analysis, while the other, adopts sublimation (Norris J.L., et al., 2013).

When the sample is properly embedded into the matrix, the application of the laser to the tissue sample surface enables the extraction of matrix clusters combined with the analyte which are then directed toward an analyser.

Nowadays, different extraction techniques are suggested for MALDI application for the pre-treatment of clinical specimens/isolates (Public Health England, 2019) (Clark A.E., et al., 2013). There is not a single best recommended extraction method. Users should ensure that they use an appropriate extraction method to get accurate identification results as well as to demonstrate that molecules profiles remain consistent with database fingerprints. For example, yeasts require a protein extraction procedure to be correctly identified (Croxatto A., et al., 2012). Filamentous fungi still lack standardised extraction protocols (Peng, Y., et al., 2019).

Despite the absence of a standard procedure for the application of MALDI, this technique is widely used in different fields of research. For instance, MALDI-MSI is used to assess drugs

dynamic and metabolism (Lockwood S.Y., et al., 2016) (Rubakhin S.S., et al., 2005) (Castellino S., et al., 2011). It is also applied in different fields like plant biology (Lockwood S.Y., et al., 2016) (Rubakhin S.S., et al., 2005) (Castellino S., et al., 2011). In particular, MALDI was successfully applied to cancer research (Deininger S., et al., 2008) (Gustafsson J.O.R., et al., 2011) (Rauser S., et al., 2010).

The spatial resolution provided by MALDI is directly connected with the diameter of the laser applied to the sample. The resolution achievable is about 10 – 20μm, but recent work has shown that MALDI can generate images even of single cells (Zavalin A., et al., 2012) (Passarelli M.K., et al., 2013) (Boggio K.J., et al., 2011) in the typical mass range (*m/z*) of 2–20 kDa (Singhal N., et al., 2015).

A typical MALDI-MSI processing and data acquisition can take over 30 minutes, limiting its clinical utility for intraoperative diagnostics. However, a recent study shows a *rapid* MALDI-MSI methodology that can be completed in less than 5 minutes (including sample preparation and analysis). This new workflow results compatible with the clinical frozen section procedure (Basu, S.S., et al., 2019).

As mentioned before, MALDI-MSI has demonstrated tremendous potential in different research sectors even if it presents a central technical limit related to the sample preparation and the matrix application. An inhomogeneous matrix layer could produce biased ion spatial patterns that do not represent the exact molecular distributions of the sample and also the matrix chemical composition can react with the molecular composition of the sample analysed creating imprecise quantification of the local ion concentrations. Nevertheless, the lack of standard procedures that guide the use of MALDI for the extraction of certain categories of molecules can lead to problems with reproducibility. These technical limits and their possible solutions are still under development.

2.2.2 DESI-MSI

A second soft ionisation technique adopted for the creation of an MSI dataset is the DESI (Desorption electrospray ionisation), which allows, like MALDI, the detection of mass spectrometry profiles for small mass molecules (e.g., lipids, metabolites) with a spatial resolution of 50-100 μm and an upper mass range detection limit of 2 kDa.

The tissue sample analysed using DESI is sprayed with a high pressure, charged solvent mixture (roughly 5-7 bar). In most of the practical cases encountered in the literature, the solvent consists of a mixture of water and methanol combined in different proportions.

The solvent is sprayed using a probe with a distinct geometrical configuration characterised by the distance and angle from the tissue sample. The probe can scatter the droplets containing both solvent and analyte molecules towards the analyser, where, subsequently, the spectra are acquired at atmospheric pressure. A typical prototype adopted to explain the DESI ionisation mechanism for both small and large molecules has been presented by Takáts et al. (Takáts Z., et al., 2005). The application of DESI for the creation of MSI datasets relies on the interaction of the solvent jet (droplets) with a thin tissue section placed on a flat surface.

One of the main advantages provided by ambient pressure DESI, compared to MALDI, for instance, is that minimum sample preparation is required prior to tissue analysis. This significant advantage makes DESI one of the most flexible techniques for MSI. However, its efficiency strongly depends on the geometrical configuration of the sprayer adopted, and the characteristics of the solvent jet chosen, such as solvent mixture, pressure, flow rate and the electrostatic voltage applied (Takáts Z., et al., 2005) (Tillner J., et al., 2016).

The time required by DESI in order to create an MSI changes in function of the spatial resolution required and the size of the sample under analysis. However, constant study and

improvements are being made to DESI to accelerate the acquisition process (Tillner, J., et al., 2017).

Both DESI and MALDI present the same limit regarding the lack of knowledge in understanding the interactions between the solvent/matrix and the chemical structure of the tissue sample under analysis. For this reason, only the relative abundances of substances within the samples should be considered. The development of techniques for quantitative analysis performed using DESI-MSI is continuing (Swales J.G., et al., 2018) (Taylor A.J., et al., 2018), improving the already high analytical capabilities of this technique.

2.2.3 SIMS-MSI

SIMS represents another conventional technique utilised for the creation of MSI datasets (Honig R.E., 1958) (Liebl H., 1967). This classical method, compared with other MSI methods previously described, provides the highest spatial resolution, but it is usually applied only in small regions of the 2-dimensional sample.

The analysis is based on the bombarding of the sample surface with a high energy ion beam (keV) that produces a secondary ion jet, which is composed of ions from the sample. This second jet is directed into a mass spectrometer to determine the spectral profile associated with the sample (Vickerman J.C., et al., 2011). The energy of the primary ion beam is generally high compared to the covalent bond energies of the analyte molecules, resulting in their fragmentation. This '*aggressive'* method of exaction of ions makes SIMS a '*hard*' ionization technique.

Like MALDI and DESI, SIMS technology can be used to generate images with extremely high resolution (<10μm) (Passarelli K., et al., 2015). However, when performing a comparison between these techniques, SIMS results are more challenging to obtain due to the requirement for

very sophisticated instrumentation and also large tissue samples cannot be scanned. Nevertheless, SIMS also tends to destroy the analytes and so it is more commonly used for elemental analysis.

Also, SIMS requires long acquisition times (roughly between 2 and 4 hours), which is incompatible with the analysis of lipid species on tissue sections (Quentin, P.V., et al., 2015).

Considering the previous stated pro and cons of the SIMS methodology, it does not represent a valid alternative to DESI and MALDI for the extraction of MSI datasets of histological images. However, like DESI and MALDI, SIMS-MSI has undergone constant updates in order to perform quantitative measurements of ion abundances (Wagner M.S., et al., 2008) (Hashimoto H., et al., 2004).

2.3 APPLICATION OF MSI IN CANCER STUDIES

MSI has become a valuable tool in cancer research thanks to the constant endeavour of researchers (in particular physicians, pathologists, and molecular biologists) and engineers in discovering new technologies and tissue analysis methodologies. MSI has the capability to directly connect molecular changes with histology through the generation of molecular maps of tissue sections. MSI aims to transform tissue-based diagnostics underlying biochemistry in the analysed sample. This powerful technique does not substitute classical histology but rather unlock its full capacity, and lead to a superior molecular histology synthesis, when these two diagnostic procedures are combinate. In fact, combining the power of MSI in measuring hundreds of molecules at the cellular level with direct correlation to histological structures can improve pathology field. However, MSI can achieve its goal only when used in combination with state-of-the-art tissue morphology evaluation. In fact, tissue heterogeneity (for instance, between different sample areas) produces differences between the acquired spectra and therefore introduces technical

artifacts. Correct interpretation of molecular maps (or images) obtained with MSI can be only accomplished by extensive understanding of the tissue structure (Li K., et al. , 2020).

Tumours are complex structures that contain numerous different cells (epithelial, endothelial, stromal, vascular, and inflammatory). Thus, a combination of MSI with histology (and histopathologist expertise) is particularly important in data interpretation of samples and results. The detection and characterisation of tumour cells and their environment is possible using MSI, and this can be combined with histological images that can be used to support clinicians in the diagnosis and eventually management of cancer (e.g., classify patients, make therapy decisions, and predict outcomes).

MSI is being widely applied to the *in situ* molecular analysis of cancerous tissues (or cancerous cells) with the objective of identifying tumour margins, classifying primary tumour, assessing the patient response to chemotherapy treatments (and eventually their resistance to the drugs administered), evaluating metastatic status, identifying diagnostic and prognostic biomarkers (Qin, L. & Z., et al., 2018).

In particular, the explanation of the spatial location and abundance of peptides, proteins, lipids, glycans, and drug metabolites in the tumour tissue may improve diagnosis, staging, and personalised treatments. However, the type of molecules (or ions in the context of MS) detected depend on the sample preparation, instrumentation, and acquisition protocols chosen.

As mentioned earlier, the tumour environment depends not only on the tumour cells but also on their interactions with surrounding areas (e.g., stroma, blood vessels, and the immune system) (Kriegsmann J., 2015). Thus, complex signalling is a key aspect of tumour heterogeneity, which can be investigated in detail using MSI techniques e.g., to directly identify and quantify proteins and peptides (Buchberger AR, et al. , 2018).

The metastatic status is usually crucial in determining if e.g. a patient needs to receive chemotherapy treatments or needs surgical removal of local lymph nodes (which often results in noncancer health complications). Therefore, identification of primary tumour biomarkers of metastasis would prevent unnecessary surgery and permit a more personalised therapy (Casadonte, R., et al., 2014). However, pinpointing metastasis through the analysis of the primary cancer is difficult, especially when the location of the primary tumour is unknown (Pavlidis, N., et al., 2016). In this context, a statistical classification model based, for instance, on peptide MSI can be helpful. Casadonte et al. (Casadonte, R., et al., 2014) have developed an MSI statistical approach for determining the metastatic process (breast cancer) using the primary tumour status (pancreatic origin) with an overall accuracy of 83%. Another study, (Balluff, B., et al., 2010) shows how the HER2 (human epidermal growth factor receptor 2) status of gastric tumour can be predicted (with an accuracy of 90%) using peptide MSI expression originating from breast cancers. These studies highlight the potential of MSI in the characterisation of cancers independently to their origin site.

MSI in cancer research is also deployed for the assessment of innate or acquired resistance to chemotherapy cocktails (chemoresistance), which strongly impact the survival rate of cancer patients. Administration of chemotherapy treatments to patients has conventionally been standardised, despite the knowledge that patients' response to these cocktails can be remarkably different due to the patient's genetic diversity and also to tumour heterogeneity. Identification of markers that predict chemoresistance would help save lives and to determine if an alternative treatment course could be more advantageous. Ion a study by Bauer et al. (Bauer, J. A., et al., 2010) the application of MSI is able to discriminate two isoforms (functionally similar proteins that have a similar but not identical amino acid sequence) of the protein defensin, which are of a particular interest because they react differently in breast tumours when exposed to chemotherapy

drugs. Another study (Aichler M., 2013) shows how peptide MSI created a connection between mitochondrial defects and chemotherapy response in patients affected by oesophageal adenocarcinoma. These studies highlight the potential possessed by the MSI techniques in changing classifications of tumours and chemoresponse.

It is now clear that detecting cancer in its early stages (before metastasis) will help increase survival rates and minimise the amount of treatment necessary.

Probably the bigger clinical challenge in the treatment of cancer is the variable response of the disease when patients are under therapy. An important aspect connected to the occurrence of different responses to treatments is the presence of tumour (inter and intra) heterogeneity. Inter-tumour heterogeneity refers to the dissimilarities in cancer characteristics between patients, while intra-tumour heterogeneity refers to the molecular variety within a patient. As mentioned before, in the development of novel cancer treatments intra-tumour heterogeneity is considered a fundamental factor. However, this is a very complex challenge since assessment requires an untargeted molecular analysis technique that takes into consideration both spatial and temporal dynamics of the tumour. In this context MSI may be a powerful tool to perform the required task especially when applied to solid tumours.

Furthermore, the combination of MSI with manual annotations can extrapolate molecular profiles from ROI (regions of interest) and allocate them to tumour cells. This analysis of tissues enables more distinct biomarkers discovery (McDowell, C.T., et al., 2021) (Addie, R.D., et al., 2015) (McDonnell, L. A., et al., 2010) (Schone, C., et al., 2013) for diagnostic purposes (McDowell, C.T., et al., 2021) (Guenther, S., et al., 2015) (Lazova, R., et al., 2012), for prognostic purposes (Heijs, B., et al., 2020) (Elsner, M., et al., 2012), to predict reaction to therapy

(Palubeckaitė, I., et al., 2020) (Aichler M., 2013). These studies provide evidence of the capability of MSI to interpret inter-tumour heterogeneity.

Over the last decade, MSI has already had a significant impact in cancer research, unveiling biomolecular changes linked with oncogenesis, diagnosis, and prognosis.

The combination of DESI-MSI with ML, DL and H&E provides insights into intra-tumour heterogeneity in different cancers (metastatic liver cancer and epithelial ovarian cancer) and is the central area of research of this thesis.

2.3.1 LIPID DESI-MSI IN CANCER RESEARCH

In human cancers studies, DESI-MSI is used for the discrimination of tumour subtypes, assessment of tumour grades, and for the identification of tumour margins (Calligaris, D., et al., 2014) (Eberlin, L.S., et al., 2014). An important application of DESI-MSI shows how this technique is capable of generating lipid profiles for the classification of human brain tumours through lipidomic analysis (Eberlin L.S., 2012). Other lipidomic studies (Abbassi-Ghadi, N., et al., 2014) (Dill, A.L., et al., 2011) (Eberlin L.S., 2012) show how DESI-MSI is able to identify the phospholipids phosphatidylserine (PS), phosphatidylinositol (PI), and phosphatidylethanolamine as biomarkers for distinguishing healthy tissue from tumour (benign and malignant) tissues. This discriminatory potential possessed by DESI-MSI is highlighted in the work of Calligaris et al. (Calligaris, D., et al., 2014), where lipid biomarkers between tumour and healthy tissue in breast (tumour margins) have been identified via including the phosphoinositol PI(18:0/20:4). This was shown to be abundant in the tumour area, while it was almost absent or weak in normal investigation performed by Yang et al. (Yang, L., et al., 2015) corroborates the results above. Also,

using DESI-MSI, Abbassi-Ghadi egt al. (Abbassi-Ghadi, N., et al., 2014) found that PS(20:4/18:0) and PI(18:0/20:4) are significantly lower in concentration in primary tumour tissues.

These examples highlight the use of MSI lipid profiling of tumour tissues in the literature and demonstrate that it can be used for classifying tumour types, with high specificity in differentiating disease grade, and in identifying tumour margins.

2.3.2 MALDI-MSI ON WHOLE-BODY SECTIONS

Generally, MSI techniques are applied to tissue sections (e.g., slices of biopsy), however, there are only a few studies where MSI has been deployed on larger body sections. Here, MALDI-MSI is more advantageous than other techniques. Rohner et al. (Rohner, T.C., et al., 2005) were the first to demonstrate the application of MALDI-MSI for the assessment of drug dynamics in whole-body animal sections. Subsequently, a broad range of drugs have been imaged in whole-body animal sections (Khatib-Shahidi, S., et al., 2006). In the work of Khatib-Shahidi et al. (Khatib-Shahidi, S., et al., 2006) MALDI-MSI has been deployed to concomitantly detect (at various points in time) drugs and their individual metabolite distributions across whole-body tissue sections. Stoeckli et al. (Stoeckli, M., et al., 2007) measured the distribution of a 14C labelled compound in whole rat tissue sections using MALDI-MSI and whole-body autoradiography (WBA) which displayed good quantitative alignment between both techniques. Finally, Trim et al. (Trim, P.J., et al., 2008) compared the distribution of vinblastine (a chemotherapy medication) within whole-body sections using MALDI-MSI and WBA. This research shows how MALDI-MSI is advantageous in separating the drug from an endogenous isobaric lipid. New studies in the application of MALDI-MSI to whole-body section continue to appear (Bingming, C., et al., 2020) (Djambazova, K., et al., 2020).

2.4 UNSUPERVISED LEARNING IN MSI

Previously, we discussed the potential possessed for MSI in histopathology and field of cancer research, and at the same time, highlighted how complex this approach is, in terms of technology and computational procedures required.

Nowadays, modern technologies create MSI datasets of massive size, in fact raw MSI datasets (prior pre-processing) may be of the order of GBs and TBs for a single experiment. These datasets usually contain 10^3-10^6 pixels with 10^4–10^6 *m/z* bins per pixel. Managing these large datasets can be challenging and computationally demanding even for modern computers leading to memory shortages and extremely long calculation times (Verbeeck, N., et al., 2019).

MSI data is commonly explored using unsupervised data analysis approaches, which aim to extract the key trends embedded within the data. Generally, an unsupervised approach is used to reveal structure from features in the data by looking at their relationships (usually in a multidimensional space). Many unsupervised techniques are used in the analysis of MSI datasets (Verbeeck, N., Caprioli, R.M. et al., 2020).

An advantage of an unsupervised approach is that little needs to be known in advance about the data (labelling or prior information on the data). This contrasts with supervised methods, where the data are generally fitted to a pre-classified model.

Typically, unsupervised machine learning methodologies are used for exploratory analysis of MSI datasets and are divided into three main classes: Factorisation, Clustering, and Manifold Learning (Verbeeck, N., Caprioli, R.M. et al., 2020).

The first important class of methods discussed here is Factorisation. These methods take high dimensional MSI datasets and decompose them into a reduced number of trends that represent the analysed data. This *reduced* representation allows the researcher to obtain visual insights into

the substructure of the MSI dataset, and it frequently reveals colocalised and correlated spatial and molecular patterns. On top of these important advantages, factorisation can also provide a low dimensional (less complex) representation of the original MSI dataset, which can be more effectively computed and analysed.

For factorisation methodologies, the most well-known and applied method is PCA (Principal Components Analysis) (Jolliffe, I., et al., 2002). One aim of this technique can be to reduce the dimensionality of a dataset by describing it with by a smaller number of variables, while still maintaining as much as possible the original variance in the data. However, the determination of the *correct* number of components to retain (for a particular application) is a complex task to implement (Peres-Neto, P.R., et al., 2005) especially since there are no analytical solutions.

In addition to PCA other data reduction and feature extraction methods are available. These alternative methods derive efficient (retaining much of the original information) dataset compression while also utilising less memory-expensive computation than PCA. For instance, one of the most used alternatives (with respect to PCA) is the ICA (Independent Component Analysis) technique. ICA is a matrix decomposition that aims to find statistically independent components in the dataset (Jutten, C., et al., 1991) (Comon, P., et al. , 1994). From the operational point of view, PCA and ICA are quite similar, however, ICA requires components to be statistically independent of each other, while PCA derives components to be uncorrelated. The ICA requirement results are more demanding than PCA. In fact, if we consider two variables x and y that are statistically independent, they can also be defined as uncorrelated, however, this property is not mutual, in fact, if x and y are uncorrelated this does not mean that they are also necessarily independent.

Another factorisation technique that can be deployed on MSI dataset (instead PCA) is MAF (Maximum Autocorrelation Factorisation) first proposed by Switzer et al. (Switzer, P. et al., 1984). In numerous studies (Tyler, B.J., et al., 2007) (Henderson, A., et al. , 2009) (Park, J.W., et al., 2009) (Hanrieder, J., et al., 2014) MAF has been deployed as a multivariate analysis methodology on SIMS-MSI.

A weakness of these methods is that they will often introduce (due to mathematical constraints) negative peaks in the spectral domain or negative values in their spatial expression images (pixels). These negative values are difficult to decipher, since there is not a direct physical meaning to them. In order to avoid these technical issues, researchers have embraced *non-negativity constrained matrix factorisation* (NMF) (Lee, D.D., et al., 1999) (Jolliffe, I., et al., 2002) (Van de Plas, R., et al., 2007b). There are several techniques currently available to deploy NMF, however, the predominant one is the MCR (Multivariate Curve Resolution) which is largely applied onto SIMS-MSI datasets.

The second class of unsupervised learning methods adopted for an MSI analysis is Clustering. These algorithms are extensively applied for exploratory purposes. They provide a low dimensional description of the high dimensional molecular content embedded in an MSI dataset by grouping pixels with similar mass spectral profiles (similar chemical content).

Generally, in MSI applications, clustering algorithms are used for spatial segmentations, in fact, a cluster in the spectral space corresponds to a segmentation in the image space. Spatial segmentation provided by clustering methods is of particular interest in pathology and clinical research as it is capable of determining/highlighting subregions in the tissue sample under analysis.

On the other hand, some applications of clustering methods (for MSI datasets analysis) have focused on grouping data along the spatial domain, and segmenting along the spectral domain

with the objective of clustering ion images with a similar spatial expression. However, most MSI clustering applications just focus on spatial segmentation.

Today, many clustering/segmentations methods are available. Theoretically, there are methods that require the number of expected clusters as a parameter (e.g., k-means), while others do not (e.g., DBSCAN). Clustering techniques have not only been applied to MSI data (Bemis, K.D., et al., 2016), but they have also been modified in accordance with the nature of MSI dataset by, for instance, taking account of the spatial neighbourhood of a spectrum (Alexandrov, T., et al., 2010) (Alexandrov, T., & Kobarg, J.H. , 2011). Considering the type of MSI data under analysis and the biological problem targeted, clustering algorithms shows large differences in terms of performance (Jones, E.A., et al., 2011) (Sarkari, S., et al., 2014).

Theoretically, clustering algorithms can be divided in two main groups: Hierarchical Clustering (e.g., DBSCAN, OPTICS, K-means) and High Dimensional Data Clustering (HDDC) (Bouveyron, C., et al., 2007). However, regardless of the method used, much of the work in MSI clustering uses "*hard segmentation*" which means that pixels can only belong to a single cluster.

Verbeeck, et al. (Verbeeck, N., et al., 2019) imply that the application of a dimensionality reduction (DR) technique before deploying a clustering procedure can remove noise variation from the MSI dataset and, therefore, improve the smoothness of the resulting spatial segmentation.

In the study of Alexandrov et al. (Alexandrov, T., & Kobarg, J.H. , 2011) the spectral similarity of pixels is integrated with their spatial proximity in the tissue, which can also be visually assessed from a progression of pixels through the tissue slices. However, while the combination of spatial information can be advantageous for smoothing segmentation results to achieve a better visual interpretation, it must be noted that these techniques should be applied with care (Verbeeck, N., et al., 2019).

In chapter 3 we will describe the two main clustering algorithms deployed during this work, which are DBSCAN and OPTICS.

The final class of unsupervised learning methodology described here is Manifold Learning. Projection (mapping) of a high dimensional MSI dataset onto a low dimensional space generally utilises PCA, ICA, and NMF. However, the linear nature of these techniques makes them less efficient in the interpretation of the nonlinear structure of MSI datasets. In this case, nonlinear manifold learning techniques can be employed (Cayton, L., et al. , 2005) (Tenenbaum, J.B., et al., 2000) (Roweis, S.T., et al., 2000).

The main idea is that the dataset, in reality, lies on a lower-dimensional manifold that is integrated in the high dimensional feature space (Cayton, L., et al. , 2005). In this context, if the high dimensional data can be represented in a linear subspace (low dimensional), linear techniques (e.g., PCA) can reveal that subspace exactly. However, if the data cannot be represented with a linear subspace but with a nonlinear subspace, manifold learning methods need to be deployed to efficiently capture that subspace.

Human visualization is mainly limited to 2D or 3D representations, in fact, higher dimensional visualisation becomes challenging (de Oliveira, M.C.F., et al., 2003). Manifold learning techniques are effective in compressing high dimensional data into a lower dimensional space making them useful in the provision of a visual representation of data without the necessity of overlying complex lower dimensional plots. Therefore, manifold learning techniques provide a more concise and insightful representation of a high dimensional space. Also, linear techniques focus on keeping different points far apart in the low dimensional representation, while manifold learning techniques such as t-distributed stochastic neighbourhood embedding (e.g., t-SNE) focus on projecting similar data points close together in the low dimensional space (this phenomenon is

called the *crowding problem* which is addressed in later). More effective low dimensional representation (Van Der Maaten L, et al., 2008) leads to mode effective clustering of similar points.

In the following chapter (Chapter 3) we will describe the two main manifold learning techniques deployed here: t-SNE and Parametric t-SNE.

2.5 MSI STATISTICAL ANALYSIS

The MSI datasets, regardless of the technique used, consist of a considerable number (typically over 10k, depending on the size of the sample under analysis) of spectral profiles (in pixels), constituting a continuous curve representing the dataset that contains the *m/z* ratio and signal intensity (equivalent to the number of detected ions).

MSI datasets from the instruments are usually classified as *raw*, and they are not suitable for direct statistical analysis for various reasons:

- The continuous curve contains redundant information that inflates the dimensionality of the *raw* MSI dataset. This redundant information is mainly generated by values close to zero between data that represents genuine peaks.
- The observed *m/z* values are shifted around the true values due to the introduction of detector errors (*intra-spectral shift).* Also, the introduction of electronic and thermal noises generates random fluctuations that could mislead the detection of the presence of ions (*noise fluctuations*) for that particular *m/z* value.
- Intra-spectral intensity fluctuations for the same ions could be introduced. This depends on the variations in the electronic response of the detector to the same ion concentrations. Therefore, the observed peak heights could show quantitative differences.

These briefly described technical and analytical problems that are required to be addressed before the MSI dataset analysis begins. The cleansing process is called *pre-processing* and represents a crucial part of any image analysis process.

Currently, there are several automated or semi-automated algorithms available written with different programming languages, to perform this procedure (Gibb S., et al., 2012) (Bemis K.D., et al., 2015) (Race A.M., et al., 2016) (Veselkov K., et al., 2018).

The pre-processing procedure is composed of different steps that mainly aim to assess and improve the quality of the data and cleanse the dataset from noise or data that does not relate to the problem at hand.

Peak-detection and *peak-matching* represent challenging pre-processing steps (even though some techniques, like PCA for instance, may not require them). These are of critical importance for a successful analysis. Wrongly detected or mismatched peaks can compromise the analysis results.

Peak-detection is the process used to identify those peaks that are connected with real ions in the sample to allow measurement of their concentration. On the other hand, peak-matching is the process that allows removal of fluctuations of the peak position around the accurate *m/z* value, in this way processed *m/z* values represent those peaks detected at the same accurate *m/z* value for the same ions in all the spectra (in an ideal scenario). Practically, the same *m/z* values can be linked with different ions (isomers and isobaric groups) making the identification process (assignment of ions labels to *m/z* values) a very challenging task.

The pre-processing procedure generates a *pre-processed* MSI dataset which is typically represented as vectors of common peaks assembled in a matrix where the rows represent the pixel

indices and the columns the variable indices. This refinement of the MSI dataset represents the baseline dataset for the analysis performed in this work.

As described previously, the datasets are analysed using supervised and unsupervised methods. Supervised learning methods applied to MSI datasets often consists of multivariate models with the objective of defining a statistical relationship between the pixel vectors (categorised as observations) and an already defined set of labels that represents the property of interest. Automated tissue classification can be an example of supervised learning where the model is fitted to pre-classified (usually by a histopathologist) regions of the tissues. It is generally based on manual annotation of pixels in a histology image, such as: tumour tissue, background, stroma, mucosa or connective tissue. The automated classification is performed by the projection of the histopathologists annotations, available on the optical H&E image, to the corresponding areas of the MSI dataset. The application of a e.g. a cross-validation strategy can assess the performance of this analysis in re-classification of training and test sets.

The combination of supervised analysis with a univariate statistical technique can lead to the identification of significantly abundant peak variables in one of the tissue classes which is very useful for the elucidation of possible molecular/biological mechanisms associated with the different tissue types.

On the other hand, unsupervised learning methodologies do not require any additional external information, and they can define the classes of interest from analysis of the MSI dataset itself. The advantage of unsupervised analysis, compared to supervised analysis, is that it can identify new classifications that are free from user-defined labels. For example, tissue sub-regions could be detected automatically by the algorithm, to achieve automatic segmentation and then to allow interpretation and understanding of the tumour heterogeneity.

This powerful method of analysis can improve performance, avoiding time-consuming manual annotations, and may. discover additional patterns (clusters) in the MSI datasets that are new (not defined by the histopathologist for example) with respect to a classical visual investigation. However, the freedom provided by the application of an unsupervised analysis creates significant difficulties (Von Luxburg U., et al., 2012) due to the lack of a reliable ground truth.

Clustering algorithms lead to results that may vary unexpectedly between samples, even sharing the same mathematical (statistical) properties, so it is pivotal that the researcher (user) validates the biological meaning of the clustering (tissue segmentation) obtained. This dependency from the user does not make the diagnostic procedure completely automated.

3 LINEAR SVM, CLUSTERING AND DIMENSIONALITY REDUCTION METHODOLOGIES

The diagnostic workflow described in the next chapter (Chapter 4) deploys and compares several algorithms for different purposes. The first exercise aims to classify pixels (supervised learning methods) through the labelling provided by the histopathologist, a second exercise performs the dimensional reduction (DR) of the high dimensional MSI dataset, and finally, a final set of algorithms are used for clustering. In this chapter we will provide the theoretical background on the methodologies deployed in this work.

3.1 Linear SVM

In machine learning, support vector machines or support vector networks (SVM) are supervised learning models with associated learning algorithms that analyse data used for regression, classification, and also outlier detection. These characteristics make SVM one of the most popular models in Machine Learning field.

Given a set of training data where each is marked as belonging to one or the other categories, the training algorithm of an SVM builds a model that assigns unseen data (new examples) to one of the categories, making it a non-probabilistic binary linear classifier.

An SVM model represents the data points in space, mapped so that the points that belong to separate categories are clearly divided by a gap (margin) as wide as possible. The region bounded by these two hyperplanes is called the "margin", and the maximum-margin hyperplane is

the hyperplane that lies halfway between them. Then, new data points are mapped into that same space and, depending on which side of the gap they fall, predicted to belong to one of the categories.

Classification of data is a regular task in machine learning. As described earlier, the classification task aims to decide which class a new data point will belong. In the case of support vector machines, or in general, in the case of a linear classifier, a data point is viewed as a *p*-dimensional vector, and we want to know whether we can separate such points with a *(p-1)*-dimensional hyperplane.

Many hyperplanes might classify the data. Basically, the algorithm must define the hyperplane such that the distance from it to the nearest data point on each side is maximised. If such a hyperplane exists, it is identified as the maximum margin hyperplane. Also, the linear classifier goes under the name of the maximum margin classifier.

Formally, an SVM builds a hyperplane (or set of them) in a high-dimensional (or infinite-dimensional) space, which can be used for classification and/or regression. Intuitively, a good separation is reached by the hyperplane that has the greatest distance to the nearest training-data point of any of the classes (called a functional margin). In general, the larger the margin the lower is the generalisation error of the classifier.

We are given a training dataset of m points of the form:

$$(\vec{x}_1, y_1), \dots , (\vec{x}_m, y_m)$$

where the y_i are either 1 or −1, each indicating the class to which the point $\vec{x}_I$ belong. Each $\vec{x}_i$ is a p-dimensional real vector. We want to find the "maximum-margin hyperplane" that divides the group of points $\vec{x}_i$ for which $y_i = 1$ from the group of points for which $y_i = -1$, which is defined so that the distance between the hyperplane and the nearest $\vec{x}_i$ point from either group is maximized. Any hyperplane can be written as the set of points $\vec{x}_i$ satisfying:

$$\vec{w} \cdot \vec{x} - b = 0 \tag{3.1}$$

where $\vec{w}$ is the (not necessarily normalized) normal vector to the hyperplane. This is much like the Hesse normal form, except that $\vec{w}$ is not necessarily a unit vector. The parameter $b/_{\| \vec{w} \|}$ determines the offset of the hyperplane from the origin along the normal vector $\vec{w}$.

Computing the (soft margin) SVM classifier amounts to minimizing an expression of the form:

$$\left[\frac{1}{m}\sum_{i=1}^{m} max\big(0,1 - y_i(w \cdot x_i - b)\big)\right] + \lambda\|w\|^2 \tag{3.2}$$

Where λ determines the trade-off between expanding the margin size and certifying that the x_i lie on the correct side of the margin. Therefore, for sufficiently small values of λ, the second term in the loss function will become unimportant, hence, it will behave similarly to the hard-margin SVM.

It is now clear that there are two methods of classification that go under the name of a linear SVM method. The first one is the *Soft Margin Classification*, which provides a more flexible classification method. The objective is to find a good balance between keeping the margin as large as possible and limiting the margin violations. The second method is called *Hard Margin Classification* that strictly imposes that all the instances be off the margin. This method only works if the dataset is linearly separable and has no outliers.

The original algorithm (Vapnik, V., 1963) builds a linear classifier. However, years later, Boser et al. (Boser, B.E., et al., 1992) suggested a method to build a nonlinear classifier using the kernel trick (Aizerman, M.A., et al., 1964) to maximum-margin hyperplanes. This updated algorithm results in quite a similar result to the original one proposed by Vapnik, except that every dot product is then replaced with a non-linear kernel function, which maps a function from its

original space into another space via integration (integral transformation). This allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space. The transformation may be nonlinear.

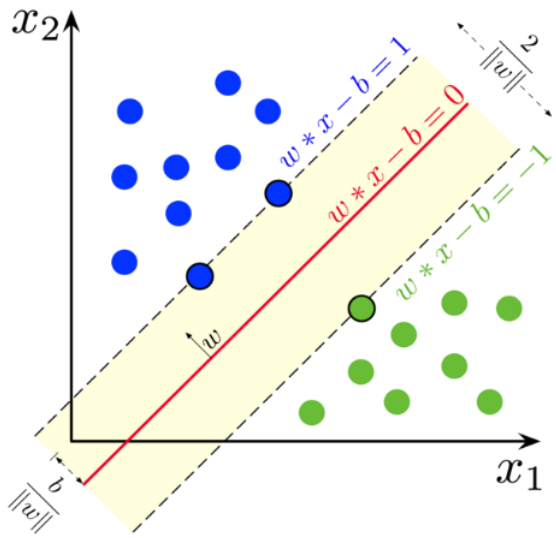


Figure 3.1- Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.

3.2 CLUSTERING ALGORITHMS – DBSCAN and OPTICS

In this section, we will give a brief description of the clustering methods employed: the "Ordering points to identify the clustering structure" (OPTICS) (Ankerst M., et al., 1999) and the "Density-based spatial clustering of applications with noise" (DBSCAN) algorithm (Ester M., et al., 1996).

Density-based clustering algorithms aim to recognise agglomerations of points, which are defined as clusters, that are close and consist of an important number of members (Daszykowski, M. and Walczak, B., 2010). There are two critical parameters consider:

1. ε, which indicates the minimum radius for considering a neighbour

2. *MinPts,* denotes the minimum number of members to define a cluster

Formally speaking, given a dataset $X = \{\boldsymbol{x}_i\}_{i=1,\dots,n}$, a data point $\boldsymbol{x}_p \in X$ is defined to be *directly density-reachable* from another point $\boldsymbol{x}_q \in X$, if and only if the following two conditions are meet:

1. $\boldsymbol{x}_p \in \mathcal{N}_\varepsilon(\boldsymbol{x}_q)$
2. $|\mathcal{N}_\varepsilon(\boldsymbol{x}_q)| > MinPts$

In the first condition, $\mathcal{N}_\varepsilon(\boldsymbol{x}_q)$ represents the neighbourhood of the data point $\boldsymbol{x}_p$ with a radius ε centred in $\boldsymbol{x}_q$, while the second condition implies that the cardinality of the neighbourhood is greater than $MinPts$.

These two conditions can be applied on agglomeration of points and not just to a pair of data points. If we consider $\boldsymbol{x}_p, \boldsymbol{x}_q \in X$ as two data points. They are defined *density-reachable* if there is a sequence of data points starting with the $\boldsymbol{x}_p$ and ending with $\boldsymbol{x}_q$. Furthermore, that are all *directly* density-reachable with regard to the parameters $MinPts$ and ε.

Generally, if there is a third data point $\boldsymbol{x}_o \in X$ that is density-reachable with $\boldsymbol{x}_p$ and $\boldsymbol{x}_q$, while these two points are not density-reachable, then they are called *density-connected*.

Considering a generic set of density-reachable data points, $D \subseteq X$, a *density-based cluster* is defined as a subset $C \subseteq D$ of points that fulfil the two conditions below:

1. $Maximality$: $\forall \boldsymbol{x}_j, \boldsymbol{x}_k \in D$, if $\boldsymbol{x}_j \in C$ and $\boldsymbol{x}_k$ is *density-reachable* from $\boldsymbol{x}_j$ with respect to $MinPts$ and ε, then $\boldsymbol{x}_k \in C$
2. $Connectivity$: $\forall \boldsymbol{x}_j, \boldsymbol{x}_k \in D$: $\boldsymbol{x}_j \in C$ is *density-connected* with $\boldsymbol{x}_k$ with respect to $MinPts$ and ε

In more descriptive language, the *Maximality* condition guarantees that the cluster contains the maximum number of density-reachable data points, while the *Connectivity* condition assures that all the points in the cluster are density-connected.

The data points that do not belong to any clusters, because they do not fit the conditions stated before, are labelled as *noise*.

The OPTICS algorithm is a density-based clustering method, like the DBSCAN, which is another extensively used clustering method. Formally speaking, given a dataset $X = \{\boldsymbol{x}_i\}_{i=1,\dots,n}$ of data points, DBSCAN assigns to each point two quantities:

1. *Core-distance*
2. *Reachability-distance*

Let now take $\boldsymbol{x}_s \in X$ as a generic data point and let assume that the parameters ε and $MinPts$ have been set already. At this point, the *core-distance* is determined as

$$\textit{core-distance}_{\varepsilon,MinPts}(\boldsymbol{x}_s) = \begin{cases} UNDETERMINED & \text{if } |\mathcal{N}_\varepsilon(\boldsymbol{x}_s)| < MinPts \\ \varepsilon & \text{otherwise} \end{cases} \tag{3.3}$$

On the other hand, the *reachability-distance* is outlined between each pair of data points $\boldsymbol{x}_p, \boldsymbol{x}_q \in X$, as below:

$$\begin{aligned} &\textit{reachability-distance}_{\varepsilon,MinPts}(\boldsymbol{x}_p, \boldsymbol{x}_q) \\ &= \begin{cases} UNDETERMINED & if\ |\mathcal{N}_\varepsilon(\boldsymbol{x}_q)| < MinPts \\ \max\left(\textit{core-distance}_{\varepsilon,MinPts}(\boldsymbol{x}_p), \|\boldsymbol{x}_q - \boldsymbol{x}_p\|\right) & \text{otherwise} \end{cases} \end{aligned} \tag{3.4}$$

where $\| \,.\, \|$ indicates the Euclidean distance between points.

As a first step the OPTICS algorithm sorts the data points through a routine process called *ExpandClusterOrder* that, for each data point, determines the neighbourhood $\mathcal{N}_\varepsilon$, while calculates the value of $\textit{core-distance}_{\varepsilon,MinPts}$ and sets $\textit{reachability-distance}_{\varepsilon,MinPts} = UNDETERMINED$.

Subsequently, the data point is introduced into a seed-list called *OrderSeeds*. At this point, if the examinate data point is a *core* object, which means that it satisfies the condition $|\mathcal{N}_{\varepsilon}(\,.\,)| > MinPts$, then the *directly* density-reachable data points are pushed into *OrderSeeds*, sorted by their reachability-distance values to the closest core object.

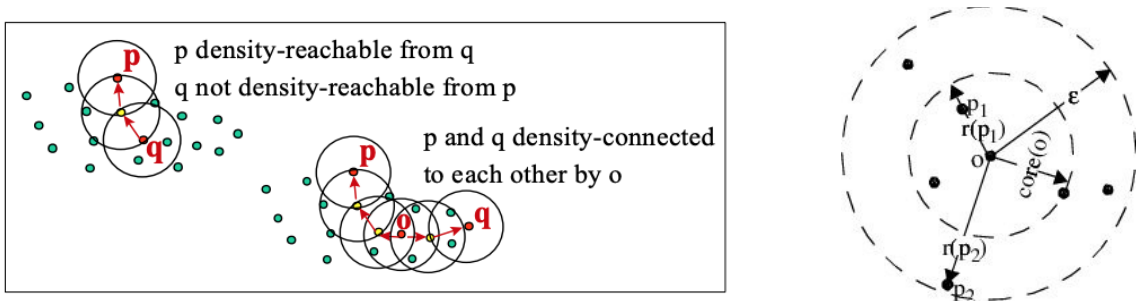


Figure 3.2 – On the left are reported the *Density-reachability and connectivity*, while on the right are showed the *Core-distance(o)* and *reachability-distances r(p1,o), r(p2,o) for MinPts=4 (Ankerst M., et al., 1999)*.

At each iteration, the component of the *OrderSeeds* with the smallest reachability-distance is selected, and its core-distance calculated. Then, all its neighbours are added to *OrderSeeds*, where they are ordered by their reachability-distance.

Finally, the OPTICS algorithm returns the order of the data points showing their reachability properties, which are represented by the reachability-distances to their nearest core object. The sorted list of data points (by their reachability-distances) can be plotted into a *reachability plot* (Figure 3.3). This plot allows visualisation of the density-related structure of the data efficiently. In particular, the reachability plot shows the groups of densely connected data points as valleys.

Extracting clusters from the *reachability plot* can be done manually by selecting a range on the *x*-axis after visual inspection or by setting a threshold on the *y*-axis which produces a similar the result a similar to a DBSCAN clustering with the same ε and $MinPts$ parameters.

There are also different algorithms that try to detect the valleys by steepness, knee detection, or local maxima automatically. However, the clustering obtained these methods are usually hierarchical and cannot be achieved using a single DBSCAN run.

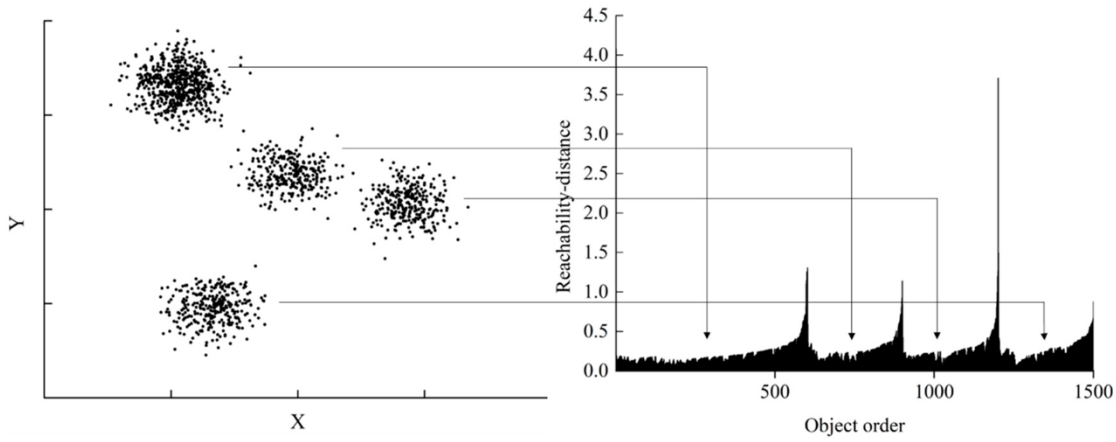


Figure 3.3 - Example of a reachability plot (right) calculated from a set of 2-dimensional data points (left). The plot reveals the four regions with a higher density of data points. They can be clustered by setting a threshold for the reachability distance equal to about 0.9.

Both OPTICS and DBSCAN are density-based clustering methods, however, there are some differences connected to the use of these two algorithms that need to be highlighted:

- Memory and computational costs: OPTICS requires more memory than DBSCAN as it maintains a priority queue to determine the next data point which is closest to the point currently being processed in terms of Reachability Distance. It also requires more computational power because the nearest neighbour queries are more complicated than radius queries in DBSCAN.
- Fewer parameters: OPTICS does not need to maintain the ε parameter, which is only given in the pseudo-code to reduce the time taken. This leads to the reduction of the analytical process of parameter tuning.

- OPTICS does not segregate the given data into clusters like DBSCAN. It produces a Reachability distance plot, and it is upon the interpretation of the researcher to cluster the points accordingly (using ε for instance).
- OPTICS results are relatively insensitive to parameter settings. Good results are obtained if parameters are just "*large enough*" ($MinPts$ for instance).
- OPTICS does not have an as well-defined concept of noise as DBSCAN.

3.3 DIMENSIONALITY REDUCTION METHODOLOGIES

3.3.1 t-SNE

Today, there are several unsupervised dimensionality reduction techniques available. Among them, commonly used methods are t-SNE and PCA (Abdelmoula W.M., et al., 2016) (Fonville J.M., et al., 2013).

In this section, we give an initial description of the mapping techniques and t-SNE, describing its limits in the analysis of complex high dimensional spaces, like MSI.

The DR technique t-SNE aims to determine a map between a high dimensional space (also called the *original space*), and a low dimensional space usually composed of two or three dimensions (also called the *target space*), while simultaneously preserving specific statistical properties between the data points.

Using a lower-dimensional space has several advantages. The most relevant is that it is possible to perform a faster analysis and reduce the computational overhead. Another significant advantage is the possibility to plot the dimensionally reduced space and therefore easily capture (visualise) previously hidden properties of the original dataset.

Let us formalise how a general DR technique works. The scope of these mathematical methods is to determine a map

$$\phi: X \rightarrow \mathbb{R}^{p'}$$

Where $X = \{\boldsymbol{x}_i\}_{i=1,\ldots,n}$ is n data points of a p-dimensional dataset with $p \gg 3$ and $p' \leq p$, such that a set of conditions are satisfied, α on X. The condition α characterises the properties of the map and must be carefully chosen depending on the specific type of analysis performed. For example, PCA determines the projections of the original data over a new set of axes (*loadings*) that are calculated as linear combinations of the original variables. The loadings are defined such that they decreasingly capture the variance of the original data.

The DR techniques can be categorised into two main groups *parametric* and *non-parametric* defined as a function of the ϕ's structure chosen. The DR can be defined as *parametric* when the projection map ϕ is expressed analytically, for instance, a Gaussian with fitted mean and variance. When the map (projection) does not have a global expression, but its value is fitted on each original data point or on their neighbourhoods, the DR is kwon as *non-parametric*. The difference between these two classes of DR techniques is crucial when we want to perform the projection of X into $\mathbb{R}^{p'}$, especially for new X that were not available during the first estimation of the map ϕ , for example unseen data.

Theoretically, a parametric mapping ϕ can project these new data points in a straightforward manner because of its analytical nature. In fact, in this situation, the projection in $\mathbb{R}^{p'}$ of new datapoints ($\boldsymbol{y}$) would be simply calculated as $\phi(\boldsymbol{y})$. This property is not valid for a non-parametric ϕ as an analytical expression is not available, therefore, a different approach and further approximations are required (Bengio Y., et al., 2003).

The t-SNE algorithm first introduced by Van Der Maaten *et al.* (Van Der Maaten L, et al., 2008) considers a set $X = \{\boldsymbol{x}_i\}_{i=1,\dots,N}$ of N data points represented by C-dimensional vectors (with $C \gg 3$) belonging to a metric space T.

Specifically, one of the objectives of the analysis is to find an 'accurate' representation of the pairwise statistical relationships in a lower dimensional space T' with dimension $d \in \{2, 3\}$.

t-SNE provides a method to project data points onto T', preserving the probability of finding their original neighbours in the projected neighbourhoods in T'. For this purpose, given a data point $\boldsymbol{x}_i \in T$, the conditional probability of finding another data point $\boldsymbol{x}_j \in T$ in its neighbourhood, is calculated as normally distributed

$$p_{j|i} = \frac{\exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_k\|^2}{2\sigma_i^2}\right)} \tag{3.5}$$

where $\| . \|$ represents the Euclidean distance. The variance σ_i^2 is set such that the *perplexity* (defined as 2^H, where $H = -\sum_j p_{j|i} \log_2 p_{j|i}$ is Shannon entropy) assumes a certain value (specified by the user). Perplexity can be interpreted as the expected number of neighbours per data point (Van Der Maaten L, et al., 2008). To make the calculation simpler a symmetric join is used:

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N} \tag{3.6}$$

In this way it is possible to consider the conditional probabilities $p_{j|i}$ and $p_{i|j}$ having the same values.

In T' (the lower dimensional space), the pairwise probabilities are not modelled as a Gaussian as in T, but as a Student-t distribution with one degree of freedom,

$$q_{ij} = \frac{\left(1 - \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2\right)^{-1}}{\sum_{k \neq l}(1 - \|\boldsymbol{y}_k - \boldsymbol{y}_l\|^2)^{-1}} \tag{3.7}$$

where $\boldsymbol{y}_i$ is the projection of the generic data point $\boldsymbol{x}_i$ in T'.

t-SNE uses a Student-t distribution to attenuate the *crowding problem,* which is an effect that arises due to the different degrees of freedom between T and T'. To better understand how the crowding problem works we can consider a set of neighbour points in the high dimensional space T. These points are considered to be equidistant from a particular data point and they lie on one of the axes. When we project these points from T onto a lower dimensional space T' they are going to occupy a more extensive region ending up with the space being filled with data points, with the consequent loss of structural relationships between distant neighbours in T. In other words, when the t-SNE maps a high dimensional dataset in two dimensions (3-dimension), the area (volume) occupied by the data points in two-dimensional space, which originally have high distances in the high dimensional space, will not be sufficiently larger. Therefore, the area used by the data points in two dimensions have low distances compared to the high dimensional data. This is expected to the fact that size scales up exponentially as dimensions increase.

Through the application of a Student-t distribution, which is a heavy-tailed distribution, it is possible to attenuate the undesirable crowding problem. The moderately distant neighbours (corresponding to a smaller p_{ij}) would be pushed farther away in the lower dimensional space (corresponding to a smaller q_{ij}). Using this approach, the t-SNE technique maps the T data points onto T' in such a way that the neighbourhood probabilities become similar each other.

Mathematically, we can say that t-SNE aims to minimise the Kullback-Leibler divergence (KL) (3.8) between the two distributions via a gradient descent approach where the gradient of the cost function (the Kullback-Leibler divergence) is (3.9)

$$KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{3.8}$$

$$\frac{\delta C}{\delta \boldsymbol{y}_i} = 4 \sum_j \left\{ (p_{ij} - q_{ij})(\boldsymbol{y}_i - \boldsymbol{y}_j) \left(1 - \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2\right)^{-1} \right\} \tag{3.9}$$

The expression of the gradient suggests that the low-dimensional data points move under the effect of a set of attractive/repulsive elastic forces until a *local* minimum is reached.

t-SNE uses the pairwise distances between data points to fit a Gaussian kernel and a Student-t kernel in the two spaces and tries to minimise their differences. With the term *kernel* we refer to the shape of the function that is used to take the average of the neighbouring points. For instance, a Gaussian kernel is a kernel with the shape of a Gaussian (normal distribution) curve.

The t-SNE model generated can't be used for future unseen observations which therefore represents a limitation, especially when we want to interpret new data or, as in this work, re-use the model for the analysis of a different (or expanded) MSI dataset.

t-SNE is a DR technique that can be categorised as non-parametric (Gisbrecht A., et al., 2015).

The t-SNE technique can be reinterpreted in order to become a parametric DR technique and provide a better interpretation of high dimensional MSI data space.

3.3.2 Restricted Boltzmann machine (RBM)

A RBM is defined as a stochastic generative model based on ANNs (Smolensky P., et al., 1986) (Hinton G. E., et al., 2002) (Munz E.D., et al., 2017) and represents the building block for the construction of the parametric version of t-SNE.

The architecture of an RBM consists of two layers of neurons, the visible layer $\boldsymbol{v}$ and the hidden layer $\boldsymbol{h}$, fully connected each other. A typical RBM's structure can be seen in figure 3.3.

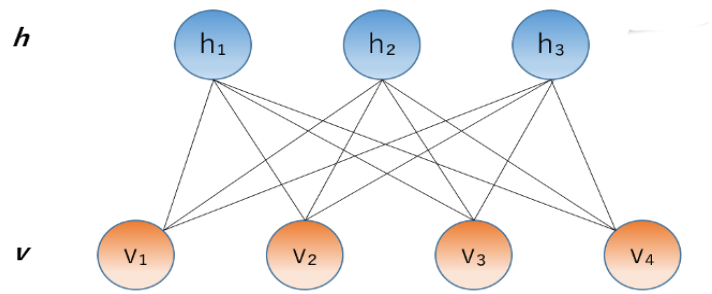


Figure 3.4 – An illustrative representation of a generic RBM structure. In the fully connected bipartite graph the Visible Layer ***v*** *and the Hidden Layer* ***h*** *are also visible.*

For the sake of simplicity and to better utilise the estimation algorithm offered by the RBM, we consider that the input variables are Bernoulli-distributed or binary-valued even though the input variables of an RBM can be Gaussian-distributed (Hinton G.E., et al., 2012).

Formally, considering $X = \{\boldsymbol{x}_i\}_{i=1,\dots,n}$ a set of n observations where each are represented by a p-dimensional feature vector $\boldsymbol{x}$ and generated by a random process that involves an unobserved p'-dimensional (often $p' < p$) variable $\boldsymbol{z}$ (*latent variable*).

Formalising what we have just described is equivalent to saying that each observation $\boldsymbol{x}$ can be modelled by a conditional probability distribution over the latent variables

$$p(\boldsymbol{x} \mid \boldsymbol{z})$$

In this way, two main results can be fulfilled at the same time. The first one is that the approximate knowledge of the latent variable and the conditional probability distribution creates a compression of the observed data. The second result is that the new data points can be generated from $p(\boldsymbol{x} \mid \boldsymbol{z})$ making the stochastic model *generative.*

An RBM aims to estimate the conditional probability $p(\boldsymbol{x} \mid \boldsymbol{h})$. In order to achieve this, we can define $X = \{\boldsymbol{x}_i\}_{i=1,\dots,n}$ with $\boldsymbol{x_i} \in [0,1]^p$ as a set of Bernoulli-distributed p-dimensional observations, which are randomly generated by a process with a set of $p' < p$ hidden variables $\boldsymbol{h}$. Every observation (p-dimensional) is considered as an input; therefore, the visible layer $\boldsymbol{v}$ is represented with p neurons and it consists of exactly $\boldsymbol{x_i}$. The hidden layer $\boldsymbol{h}$ and has p' units (neurons) ($p' < p$) with values assigned and calculated as a non-linear function of the linear combination

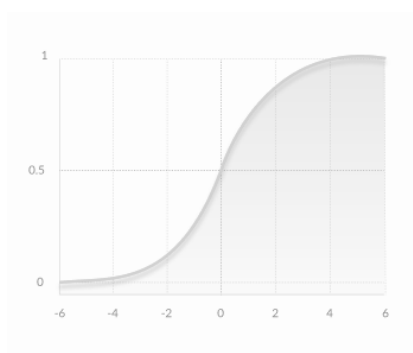
$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{v} + \boldsymbol{a} \tag{3.10}$$

Where $\boldsymbol{W} = [w_{rs}], (r = 1, \dots, p;\ s = 1, \dots, p')$ is a *weight* matrix, and $\boldsymbol{a} = [a_s], (s = 1, \dots, p')$ is a visible-hidden *bias* vector.

In the construction of ANNs (or in deep learning in general) there are several non-linear activation functions that can be used, such as sigmoid, tanh, ReLU, Leaky ReLU, and others. In the construction of an RBM is commonly used a *sigmoid* function which is defined as follows

$$\sigma(\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{x}}} \tag{3.11}$$

And characterised by the following trend



The combination of equation (3.10) with (3.11) produces

$$\boldsymbol{h} = \sigma(\boldsymbol{W}\boldsymbol{v} + \boldsymbol{a}) \tag{3.12}$$

Considering the nature of the RBM, which is defined as a bipartite graph it is possible to repeat the previous procedure starting from the hidden layer. The visible activation function, in this situation, will be

$$\boldsymbol{v}' = \sigma(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{h} + \boldsymbol{b}) \tag{3.13}$$

where $\boldsymbol{W}^{\mathrm{T}}$ is the transpose of the weight matrix and $\boldsymbol{b}$ is a hidden-visible bias vector.

At this point, it is clear that this back-and-forth process between the visible layer and the hidden layer can be iterated multiple times (*epochs*) which is also the primary process utilised during the training of these ANNs.

As mentioned before the purpose of an RBM is to estimate the conditional probability $p(\boldsymbol{x} \mid \boldsymbol{h})$. This can be achieved by taking into consideration the fact that the RBMs are ANNs that belong to a wider class of networks called Hopfield networks (Hopfield J.J., et al., 1982) which are characterised by an internal energy function that gives the joint probability of all the possible pairs of visible layers and hidden layers. For the RBMs the internal energy is represented by the formula (4.10) and the joint probability is the equation (3.15)

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\sum_{i} b_i v_i - \sum_{j} a_j h_j - \sum_{i,j} v_i h_j w_{ij} \tag{3.14}$$

$$p(\boldsymbol{v}, \boldsymbol{h}) = \frac{e^{-E(\boldsymbol{v},\boldsymbol{h})}}{\sum_{\boldsymbol{h},\boldsymbol{v}} e^{-E(\boldsymbol{v},\boldsymbol{h})}} \tag{3.15}$$

and, in particular, the probability associated with the visible units is given by the marginal distribution

$$p(\boldsymbol{v}) = \sum_{\boldsymbol{h}} p(\boldsymbol{v}, \boldsymbol{h}) \tag{3.16}$$

3.3.3 Learning process

When we are talking about the learning procedure of an ANN, we are referring to the process that leads to the determination of the *weights* associated with connections established between the network layers, both visible and hidden, with the objective of closely resembling the observations. It can be mathematically formalised and expressed in terms of maximisation of the probability $p(\boldsymbol{v})$.

$$\frac{\partial \log p(\boldsymbol{v})}{\partial w_{ij}} = \langle v_i h_j \rangle - \langle v'_i h'_j \rangle \tag{3.17}$$

where $\langle . \rangle$ represents the expectation value. The weights are updated after each observation through the use of a critical parameter called the *learning rate*, ϵ.

$$\Delta w_{ij} = \epsilon\left(v_i h_j - v'_i h'_j\right) \tag{3.18}$$

A similar formula is used to update the biases

$$\begin{aligned} \Delta b_i &= \epsilon(v_i - v'_i) \\ \Delta a_i &= \epsilon(h_i - h'_i) \end{aligned} \tag{3.19}$$

The procedure described is called *Contrastive Divergence* (CD). In this particular example, since we use only one generative step the CD procedure is denoted as CD_1.

3.3.4 Combination of RBMs and creation of an DBN

As mentioned previously, the RBM represents the building block for the construction of parametric t-SNE. Unfortunately, the simple architectural structure of this single neural network

limits the complexity of the distributions that it can estimate. In order to overcome this limitation, a combination of RBMs can be deployed and subsequently stacked to generate a more complex and highly non-linear mapping that represents the observations.

The combination of several RBMs generates a deep neural network called a *deep belief network* (DBN) which can manage a more complex dataset as an input, reconstructing and generating the observed data points. A DBN is organised as a set of bottom-up *stacked* RBMs. The DBN hidden layer (of the bottom layer) works as an input for the visible layer (the top layer) (Figure 3.4) and so on.

The DBN training process follows the same rules established by Hinton et al. (Hinton G.E, 2006) (Bengio Y, et al., 2007). First of all, the RBMs are fitted separately through a process called *pre-training* previously proposed by Larochelle (Larochelle H., et al., 2009).

This process aims to find a better starting point to train the DBN in terms of the connection weights. The first RBM uses as an input the observation; therefore, the number of neurons, which compose the visible layer, is equal to the features of the input, and later goes through the training process as described before. The output of the first RBM is then used as an input for the second RBM. Considering this kind of connection between subsequent RBMs, the number of neurons of the hidden layer of the first RBM are equal to the number of neurons of the visible layer of the second RBM. The process carries on like this until the last RBM is pre-trained. The process described goes under the name of the *greedy layer-wise* method.

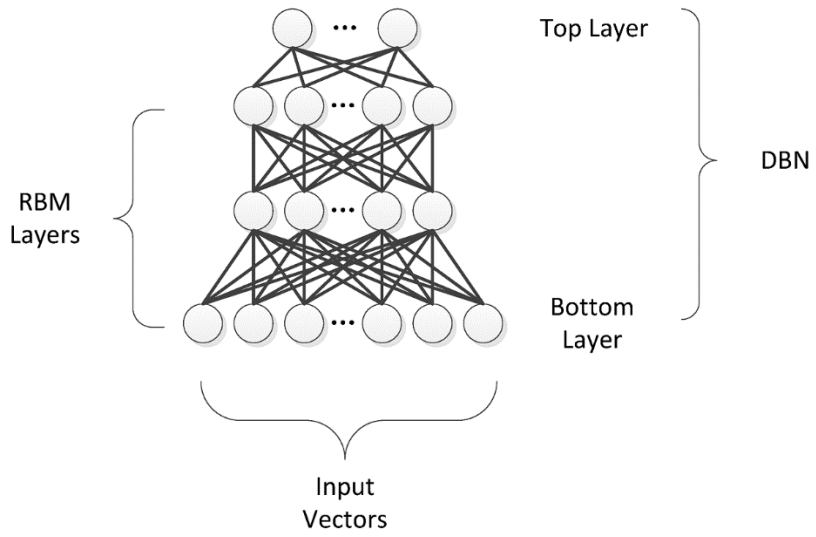


Figure 3.5 - A schematic representation of the structure of a DBN originated by the stacking of two RBMs. The visible layer of the second RBM receives its activations from the hidden layer of the first RBM (bottom layer) as input. Afterwards, they are combined (stacked) in single neural network architecture. The iteration of this process leads to the combination of a few RBMs together generating the final DBN structure.

Training deep ANNs is a challenging task due to the occurrence of a phenomenon called *vanishing gradient.* This means that the weights in layers close to the input (visible) layer are not updated in response to errors calculated from the training dataset. In deep ANNs the number of hidden layers is increased, therefore, the amount of error information propagated back to previous layers is drastically reduced. This means that the weights connected to hidden layers close to the output layer are updated normally, while the weights associated to hidden layers close to the visible layer are not updated (or minimally). In general, this process prevents the training of very deep ANNs. In this context, the technique of greedy layer-wise pretraining provides a way to construct multi-layered (deep) ANNs while only training shallow networks.

After this procedure, the pre-trained RBMs are stacked together to form a single network called a DBN. This network goes through a second training process based on gradient descent, called *fine-tuning,* which has the objective of performing the dimensionality reduction. The DR is

accomplished by decreasing the number of neurons. Practically, starting from the bottom RBM (first layer) to the top one (last or deeper layer) the number of neurons per layer is decreasing. The architecture built using this methodology generates a *parametric* DR technique (Hinton G E., et al., 2006) that maps, using non-linear functions, the input data onto a lower-dimensional space.

3.3.5 Parametric t-SNE

In this section, we will combine the properties possessed by t-SNE and the DBN in order to generate *parametric* t-SNE. This DR technique was first introduced by Van Der Maaten (Van Der Maaten L., et al., 2009) to overcome the limitations of classical t-SNE.

The fitting procedure for the parametric t-SNE is in part similar to the previous section regarding the DBN. Recapping, the RBMs are trained individually; subsequently, all the fitted RBMs are stacked together to create the DBN, and finally, fine-tuned with the aim of determining the optimal weights for a t-SNE mapping. The architecture of the DBN is similar to the architecture of an Autoencoder which is composed of two main parts. The first part is called *Encoder* and it is composed with all the visible layers of the RBMs previously built. The second part is called *Decoder* and it has a mirrored structure of the encoder. The decoder is used to calibrate and assess the performance of the DR process executed by the encoder. It achieves this by the reconstruction of the high dimensional space starting from the latent space (lower dimensional space) and, subsequently, comparing the reconstructed space with the original one. At this point, the mirrored part (decoder) of the DBN architecture is discarded, recovering only the 'half' neural network consisting of the layers from the input to the deepest hidden layer (encoder). Finally, as the last step, a t-SNE layer is stacked on top of the neural network, and the model is fine-tuned; therefore, minimising the t-SNE cost function.

Mathematically, if we consider a set $X = \{\boldsymbol{x}_i\}_{i=1,\dots,n}$ of $p-$ dimensional ($p \gg 3$) observation (input) and the parametric DR mapping provided by the DBN

$$f: T \to T'$$

This means that the low dimensional representation of a generic observation is given by

$$\boldsymbol{y}_i = f(W, \boldsymbol{x}_i) \tag{3.20}$$

Where W are the weights. The final cost function for t-SNE is based on the equation 4.5

$$\frac{\delta C}{\delta \boldsymbol{y}_i} = 4 \sum_j \left\{ (p_{ij} - q_{ij}) \times \left(f(W, \boldsymbol{x}_i) - f(W, \boldsymbol{x}_j) \right) \left(1 - \left\| f(W, \boldsymbol{x}_i) - f(W, \boldsymbol{x}_j) \right\|^2 \right)^{-1} \right\} \tag{3.21}$$

where also q_{ij} (3.7) reflects the DBN formulation

$$q_{ij} = \frac{\left(1 - \left\| f(W, \boldsymbol{x}_i) - f(W, \boldsymbol{x}_j) \right\|^2 \right)^{-1}}{\sum_{k \neq l} (1 - \| f(W, \boldsymbol{x}_k) - f(W, \boldsymbol{x}_l) \|^2)^{-1}} \tag{3.22}$$

The deepest layer of the parametric t-SNE has a linear activation function which can obtain a more stable output from the network. This solution is a common technique adopted in many computer vision problems that require the application of Convolutional Neural Networks (CNN) (Zhou D.X., et al., 2020) (David R., et al., 2020) (Mostavi M., et al., 2020). The training of this deep network can be performed using a subset of the available data and the map obtained can be applied (e.g., for validation) to the hold-out data points (Van Der Maaten L., et al., 2009).

3.4 STATEMENT OF THE PROBLEM

In the previous introductory chapters, we have emphasized the potential possessed by MSI techniques in a clinical research context, and at the same time, we have pinpointed a series of technical difficulties connected to the analysis of the high dimensional complex datasets produced by MSI methodologies. Several studies have shown encouraging results (Fonville JM., et al., 2012)

(Alexandrov T., et al., 2013) (Inglese P., et al., 2019) on how to manage MSI datasets. The work described in this thesis aims to contribute to this field utilizing modern technology, in this case deep learning, for the extraction and identification of possible tumour biomarkers. Furthermore, this work provides tangible evidence of the superiority of a 3-dimensional tissue analysis (3D-MSI) compared to a classical 2-dimensional analysis (single image analysis).

The unsupervised 3D-MSI exploratory analysis (described in the following chapters) utilises parametric t-SNE since it has the properties necessary to provide highly nonlinear dimensionality reduction. A visual representation (a projection) of complex 3D-MSI datasets (using dimensionality reduction and clustering) can help with the identification of patterns and subgroups of observations, however the very large dimensionality that characterises MSI datasets can limit this approach. Therefore, nonlinear DR techniques are applied (Fonville J.M, et al., 2013) (Thomas S.A., et al., 2016) (Inglese P, et al., 2017) in order to compress and project the high-dimensional statistical information into a lower-dimensional space suitable for visualisation (3D or 2D).

Subsequently, in order to identify sub-structure in the dimensionally reduced MSI datasets and to highlight associated regions in the H&E images, clustering algorithms are deployed. In this work, we will show for a series of novel datasets that cluster analysis (after proper validation) can identify subsets of highly co-localised ions that are intimately connected with the molecular mechanisms occurring in the analysed tumour. In particular, a focus on lipidomics reveals underlying patterns of metabolism that can separate regions of cancerous tissue from non-cancerous tissue allowing a more in-depth analysis of the chemical differences between different regions of the samples.

3.4.1 Organisation of this Thesis

- *CHAPTER 4*: We describe the analysis of a 3D-DESI-MSI metastatic liver cancer dataset. The workflow is divided into three main sets of operations: the first includes the pre-processing of the MSI dataset, the co-registration of the images (MSI and H&E), the construction of the 3-dimensional MSI dataset, the manual annotation of pixels (tumour, healthy and background), and finally the classification (supervised learning procedure) of all pixels using the labels (annotated pixels). The second set of operations, which is the most consistent part of the entire workflow, refers to the application of unsupervised learning methods to the 3D-MSI dataset. First of all, the construction of the parametric t-SNE (RBMs, DBN, t-SNE) to perform a highly nonlinear dimensionality reduction using a deep ANNs approach, and, secondly, the application of clustering algorithms (OPTICS, DBSCAN, K-NN) to highlight subregions in the tumour tissue. The last set of operations deployed includes a univariate statistical analysis of the informative content of the clusters and, subsequently, the identification of the ten most correlated ions per cluster. In this last section, a wide and deep description of the function of lipids in the cells' mechanism is also provided, expressing particular emphasis on the clustering analysis results that led to the identification of possible biomarkers that are associated with metastatic liver cancer. The workflow is shown to be more effective than classical methods of analysis (e.g. PCA), and the classical approach of visual inspection of the H&E (2-dimensional analysis). This work is built with the support of a previous study (Inglese P, et al., 2017) and shows that the designed workflow can identify density-based clusters in high complex 3D-MSI dataset unveiling important biological insights.

- *CHAPTER 5*: The work described in this chapter aims to demonstrate and assess the robustness of the workflow introduced in chapter 4 and also provides strong evidence of the superiority of a 3-dimensional analysis (3D-MSI) compared to the classical 2-dimensional image analysis (single tissue slice). In order to achieve the just stated conclusion the analysis of one-hundred and twenty-three (123), 3-dimensional tissue models are performed. Chapter 5 is divided into two main sections: the first section focuses on the creation of the 123 3-dimensional tissue models, composed with a reduced number of slices, using the *original* model (analysed in chapter 4) as a source of data and the subsequent mining of these datasets using the workflow introduced in chapter 4. The second section focuses on the evaluation and comparison of the results (clustering comparison) by similarity indices (Rand Index) calculation. Finally, in this section, the assessment of the robustness of workflow is performed through the analysis of five 3-dimensional tissue models built with randomly shuffled pixels.

- *CHAPTER 6*: In this chapter, additional applications and further assessment of the workflow introduced in chapter 4 are reported. Here the workflow is applied on a 3D-DESI-MSI dataset built using human ovarian cancer tissue slices. The work described in this chapter aims to provide further solid proof of the flexibility and potential of DL applications on MSI datasets. Chapter 6 is dived into three main sections. The first section is focused on pre-processing of the 3D-MSI dataset workflow empathising the peak detection and peak matching processes. We will also evaluate a novel DR technique called UMAP (Smets T., et al., 2019). The second section of this chapter shows the unsupervised analysis of the 3D-DESI-MSI dataset (application of the parametric t-SNE and subsequent

clustering). This section provides also a deep insight into the biological mechanisms that occur in the discovered substructure of ovarian cancer providing also a list of relevant molecules that could represent possible biomarkers. A general description of the role of lipids in oncogenesis provides a link from the detected biomarkers to oncogenesis. Finally, the third section of chapter 6 provides another strong evidence of the superiority of a 3-dimensional analysis compared to the classical 2-dimensional image analysis corroborating also the results obtained in chapter 5.

4 APPLICATIONS OF DEEP LEARNING TO 3D-MSI OF METASTATIC LIVER CANCER

4.1 INTRODUCTION

In the introductory chapters, we saw that Mass Spectrometry Images (MSI) represent the local spatial distributions of ions projected into the tissue sample analysed (regardless of the techniques used to generate them), and we have also highlighted how the datasets generated should be filtered and pre-processed to maximise the amount of useful information. Filtering the MSI dataset can help to reveal regions characterised by common ions (deduced from their mass), which are assumed to be an indication of a specific sub-type of the tissue sample. The identification of tissue subregions, especially the ones related to the presence of tumour, helps us to explore the molecular content of the disease and improve our knowledge of the mechanisms that occur in that specific environment. Indeed, we extend and apply these principles to the investigation of tumour-related ion spatial distributions (lipids).

In this chapter, we will describe an investigational workflow based on deep learning for dimensionality reduction that helps us to partitioning the mass spectrometry data (clustering), revealing the heterogeneity of the tumour tissue and leads us to improved biochemical analysis of the tumour area. The choice of this methodological approach (unsupervised analysis techniques) is a consequence of the lack of prior knowledge of the ions common to the different sub-types of tissue. Also, the complexity and the high dimensionality of the MSI dataset requires a highly non-

linear technique to extract the complex relationships from the data that are an essential feature of the tumour.

Improving our understanding of the mechanisms within the tumour tissues may enable the design of more targeted treatment strategies with the improved clinical outcomes and, of course, patient life expectancy.

The application of deep learning for dimensionality reduction on 3D-MSI and the subsequent clustering of the latent space obtained has already been used previously (Inglese P, et al., 2017) and represents the architecture on which we base this research. However, with this work we want to explore the applications of the DL workflow designed by analysing different tissue samples (metastatic liver cancer in this chapter and epithelial ovarian cancer in chapter 5) and investigate further the biological mechanisms of different tumour types (especially based on the detection of lipids) revealing characteristic *fingerprints* of the tumour tissue. Furthermore, we suggest that a 3-dimensional approach to the analysis of MSI provides a more reliable way to interpret the MSI datasets compared to the *classical* two-dimensional analysis largely used in histopathology.

On the other hand, the lack of a ground truth makes the application of unsupervised analysis methods, for these purposes, extremely difficult and tricky (Von Luxburg U., et al., 2012). For this reason, a careful validation of the clusters generated during the analysis needs to be performed. It is straightforward to extract clustering that do not reflect the real hierarchy in the tissue samples analysed. In the literature, there are available examples that show how to, partially, overcome the possibly misleading interpretation of the clustering results. For example, Balluff et al. (Balluff B., et al., 2015) have used a technique called *consensus*, which is based on the application of different clustering techniques on MALDI-MSI datasets. The authors applied various methods to an MSI

dataset, and the most frequent partition obtained was considered a "*correct"* result. In this work, in order to validate the robustness of the clustering outcomes, we will apply a methodology based on co-expression networks. This approach helps us to identify the number of significant clusters that could be used to segment the tissue sample. In fact, the application of unsupervised learning leads to an additional difficulty represented by the number of *expected* clusters (Handl J., et al., 2005), which is not known *a priori*.

Another important feature of this work is represented by the visualisation of the data structure, utilising the application of a dimensionality reduction technique on the 3D-MSI dataset and the creation of a 2-dimension latent space. In fact, the possibility to plot the latent space and the subsequent visualisation of the clusters found on top of the single tissue slice images represents a valuable way to overcome the challenges generated by clustering algorithms. Visual inspection can help to increase the confidence in the distribution of the detected clusters in the low-dimensional space. The morphological characteristics of the projected clusters on the tissue regions can help to discriminate between real and arbitrary results.

One of the most extensively used DR method in the literature is Principal Components Analysis (PCA), which is an orthogonal linear transformation method that projects the data points onto a vector space whose basis captures most of the data variance. As we will see in this chapter, one of the significant limitations of the PCA, is, in fact, its linearity (Alexandrov T, et al., 2012) (Wijetunge C.D., et al., 2014). We will also investigate how non-linear methods (e.g. deep learning methods) can be more powerful than linear methods. However, they often need approximations or specific assumptions in order to reduce the dimensionality of unseen (out-of-sample) data (Bengio Y., et al., 2003). Artificial Neural Networks have recently undergone a renaissance due to significant algorithmic improvements and the availability of high-performance computing. ANNs

applied to complex high dimensional datasets (e.g 3D-MSI) may result in a more flexible and reliable DR of unseen data. In order to achieve this, we have used a combination of neural networks and a powerful non-linear dimensionality reduction method, called t-distributed Stochastic Neighbour Embedding (t-SNE).

In the next sections, we will describe the MSI dataset pre-processing methodology, the deep learning workflow (Inglese P, et al., 2017) adopted as scaffolding for our analysis, we will also see how to extract clusters from the low-dimensional representation of the MSI dataset and how they lead, with further statistical analysis, to biochemical hypotheses generation. Finally, we will interpret the results obtained.

4.2 TISSUE SAMPLES AND THE MSI DATASET

The analysis performed in this chapter is based on a 3D-DESI-MSI dataset obtained from human metastatic liver cancer, which is an advanced stage of the disease that has spread into the liver but has originated in other parts of the body. Colorectal cancer previously investigated (Inglese P, et al., 2017) represents a primary cancer. Liver cancer symptoms often do not appear in the early stages. Therefore, adenocarcinoma metastasis tends to be diagnosed at a more advanced stage (Li, X., et al. , 2021).

Human tissue was obtained with informed consent under local ethical approval (14/EE/0024). The biopsy was snap-frozen in liquid nitrogen and stored in a freezer at a temperature of -80°C. Subsequently, the tissue sample was sectioned into parallel slices of 10µm thickness each. One of every tenth section was collected, and their MS acquired. The total amount of slices used in this work was 51.

Mass spectrometry MSI was acquired utilising a Thermo Exactive device, in the negative ion mode over an m/z range of 150 – 1500 *m/z*, with a resolution of 100µm, 1.5 millimetres per minute, and with methanol/water 95/5 as a solvent. Subsequently, the data were converted to imzML (Schramm T., et al., 2012) format using imzML converter v1.1.5.5i.

The tissue samples were organised in 13 layers containing four tissue sections each. Subsequently, the tissue sections were used to perform the H&E (haematoxylin and eosin staining) histological enhancement (Figure 4.2) and examination found the samples to mostly consist of adenocarcinoma metastasis and normal (healthy) liver with areas of connective tissue containing blood vessels and bile ducts. Dr James Kinross provided the specimens, the DESI-MSI was collected by Ms Anna Mroz, and Ms Hiromi Kudo performed the staining at Imperial College London.

4.3 SPUTNIK AND MSI DATASET PRE-PROCESSING

MSI is a powerful analytical technique capable of collecting, in a single run, data on the presence and concentration of thousands of ions which provides a picture of the local chemical properties of the sample tissue under examination. Raw medical images always require a pre-processing procedure to enhance their characteristics and remove the noise added during their acquisition by the e.g. instrument and the solvents. Data on ions that may be of interest is unfortunately acquired in combination with noise or spurious signals (described in the previous chapter) that can reduce the signal-to-noise ratio (SNR).

These unwanted signals can be categorised mainly into two groups, *white noise* and *shot noise*. The white noise is due to the thermal fluctuations of the device, which can create electronic noise that is responsible for variations in the detected signal intensities. The shot noise, instead, is

a random response of the detector, and it can be triggered when the signal intensity is close to the detection limit. Identifying and controlling these sources of noise can be challenging; however, it is essential to find an efficient way to attenuate or even remove these non-informative signals statistically.

The first disruptive effect due to the presence of noisy signals introduces analytical complexity. Spurious signals raise the dimensionality of the data with a consequent increase in the computational complexity of the statistical analysis. On top of this, non-informative signals can affect the quality of machine learning models by reducing their discriminative ability between two or more sub-regions, or even between multiple samples. It is therefore essential to identify (based on the properties of those signals) and subsequently filter the noise since the use of wrong assumptions can lead to the loss of valuable information.

Several filtering tools are now available in the literature (Hall M., 1999) (Guyon I., et al., 2003) (Archibald R., et al., 2007) (Peng H., et al., 2005). Specifically, for unsupervised learning analysis applied to MSI datasets a particular useful technique has been designed by Inglese et al. (Inglese P., et al., 2019) that leads to the attenuation of noisy signals. This technique is called *SPUTNIK (*SPatially aUTomatic deNoising for Ims toolKit). This toolkit is based on the combination of the spatial regularity and the expected localisation of similar signals in the samples, which are two important concepts already in use in similar analyses (Fonville JM., et al., 2012) (Alexandrov T., et al., 2013) (Palmer A., et al., 2016).

The spatial relationship between data in MSI datasets can be extremely useful to validate the distribution and identification of the ion signals. A reasonable property that we should expect is that going from one slice to the next (progression for adjacent tissue slices) where we observe common histology is that we should observe a degree of spatial '*regularity*' of the ion signals.

SPUTNIK (https://github.com/paoloinglese/SPUTNIK) is freely available from the CRAN repository for R, and it consists of three main classes of filters plus a split peak merging tool.

- *Class 1*. MSI datasets contains mixtures of signals coming from both the source of interest and spurious signals. A typical example of this could be the signal generated by the solvent during the acquisition of DESI-MSI data (or a matrix for MALDI-MSI). The class 1 filters act on these uninformative peaks, which can be identified by their geometrical properties. They are also independent of those of the source peak signals.
- *Class 2*. The second class of filters addresses the problem of scattered signals. It is extensively known that fluctuations around the minimum sensitivity level of the detectors can generate noisy signals due to the discretisation of shallow intense continuous signals that randomly activate the detector response. Even if these signals fall into the category of noise, they can still contain and reveal important information about the ions. They should be therefore preserved in the MSI dataset. The application of classical filters could lead to the elimination of this relevant information, but SPUTNIK preserves them offering a more complete and reliable data analysis.
- *Class 3*. Filters based on the statistical analysis of spatial patterns called *complete spatial randomness* (CSR).

The pre-processing pipeline consisted mainly of MS processing (e.g. using SPUTNIK), normalisation, baseline correction, image co-registration, and batch-effect correction. When the pre-processing is complete, the images are ready to be analysed for segmentation and extraction of features of interest (representation and classification). Selection of correlated features leads to dimensionality reduction in the classification task, which improves the computational efficiency and classification performance, as only well-correlated features are used in the classifier.

Considering the complex nature of medical images, learning the latent feature representations by employing deep learning is much more complicated than similar applications in the computer vision and machine learning areas.

In the following paragraphs, we will describe the pre-processing in more detail.

The raw mass spectral profiles were initially centroided and subsequently stored as an imzML file format (Schramm T., et al., 2012). The package MALDIquant available for R (Gibb S, et al., 2012) has been adopted to perform the pre-processing of those imzML files. The centroided peak list corresponding to the image pixels were matched using the command *binPeaks* with the default parameters. At this stage, no intensity normalisation or variance stabilising transformations were applied. After that, using the SPUTNIK command *msiDataset,* a new MSI dataset was created. This dataset contains the matched peak intensities which were also normalised and log-transformed to reduce their heteroscedasticity (the variance of a variable is uneven across the range of values of a second variable that actually predicts it). SPUTNIK was effective in removing the non-informative peak signals reducing the noisy signals (due to the solvent and instrumentation) and improving the baseline.

In more detail, the raw spectra were smoothed using a Savitzky-Golay filter (window width = 7 points, equivalent to about 0.03 m/z, and degree = 3). Peaks were detected using an approach based on the change of sign of the spectra profile first derivative (He P.Q., et al., 2011). Only peaks detected in the range of 500 – 1,300 m/z were analysed. The noise level was estimated using the median absolute deviation of peak intensities (Coombes K.R., et al., 2005). Peaks with intensities smaller than the noise level were discarded. Also, peaks that were found in less than 1% of the entire dataset were considered noise and removed. Peak matching was performed in two steps, within batches and between batches. Firstly, the peaks of the spectra belonging to the four sections

(tissue slices) per in each batch (image) per were matched. Those matched *m/z* value corresponded to the centre of a local histogram fitted on the individual peaks' *m/z* value. Subsequently, the average matched spectra were used as representatives of each batch and matched together using the same procedure. The final set of matched *m/z* values were re-assigned to all the individual spectra. The matched peak intensities were normalised using the median fold-change method (Veselkov K.A., et al., 2011).

Peak matching and normalisation were performed to improve the procedure in terms of computing time and to preserve only the differences due to biological variability. The peak matching results gave 13 spectra (one per batch of four images) (one per slide) where the *m/z* ratio vectors were the result of the peak matching on the individual slide and the intensity of the peaks is the average intensity across the entire slice.

4.4 IMAGE CO-REGISTRATION AND 3D-MSI DATASET BUILDING

The next step in the MSI dataset preparation and 3D tissue model building consists of the co-registration of the MSI images with the H&E images. Image registration is defined as the process of establishing anatomical correspondences between two images, which is a fundamental task in medical image processing (Thompson P.M., et al., 2007) (Lerch J.P., et al., 2008) (Schuff N., et al., 2009) (Leow A.D., et al., 2009) (Frisoni G.B., et al., 2006) (Apostolova L.G., et al., 2006) (Shen D., et al., 2002) (Maintz J.B., et al., 1998).

This procedure, performed used a dedicated MATLAB tool, aims to match very different sources of information, for example, H&E images are high resolution images while MSI are low resolution. The first step consisted in the separation of the four slices incorporated in a single image, the optical images were aligned with the total-ion count (TIC) images from the same sections. The

co-registration of the images, in particular the alignment, was performed by affine transformations such as rotation, translation, and shearing, through gradient descent. Afterwards, the binary version of the optical images, obtained with the application of Otsu thresholding (Otsu N., et al., 1979), were divided into square bounding boxes containing the four largest non-empty regions. The projection of the coordinates of these boxes onto the corresponding TIC images allowed splitting (the MSI were organised in a batch of four per single images) and identify (non-empty regions) of the MS data into the corresponding regions containing the tissue slices. Subsequently, the co-registration of the H&E images was performed through affine transformations. The optical images were co-registered using the previous image as a template. For this reason, the first section remained unchanged. These affine transformed images were used afterwards to register the MS data spatially. The ultimate affine transformation was applied to all the ion images to produce a set of aligned MSI spectra.

Subsequently, all those images were stacked one on top of another to create a 3-dimensional tissue model with voxels of 100 μm^3. The model obtained is intended to reproduce the 3-dimensional topology of the biopsy.

As a final step, in order to remove possible batch effects, the 'removeBatchEffect' command from the 'limma' package for R (Smyth G.K., et al., 2005) was employed with the batches corresponding to the 13 acquisition slides. A visual inspection of the TIC distributions confirmed that the batch effect was present and corrected (Figure 4.4 C). At the end of the pre-processing procedure, the data set obtained consisted of 237150 (62x75x51) spectra with 754 ion features. Prof Robert D Goldin manually annotated three classes of regions in the optical images: tumour, healthy tissue (non-tumour), and background (corresponding to the area in the image

outside the tissue area). These sub-regions were projected on the registered DESI-MSI data to match the corresponding ~ 13k mass spectra.

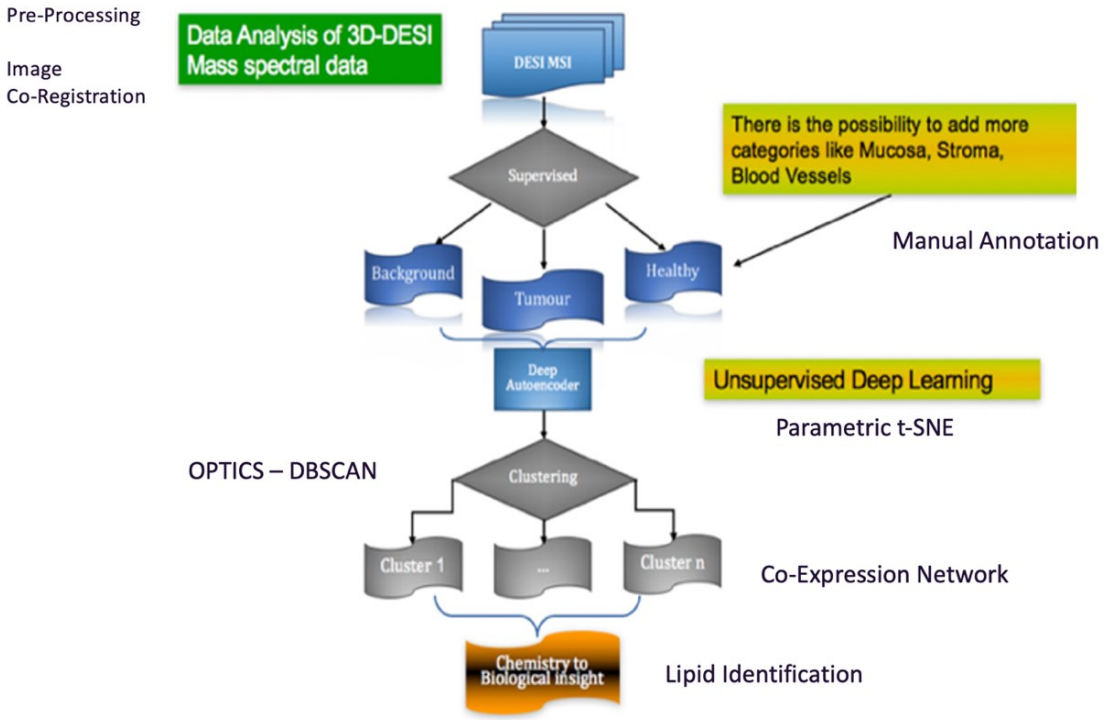


Figure 4.1– Schematic representation of the DL workflow (Inglese P, et al., 2017) . The process starts with the pre-processing of the raw imzML files followed by images co-registration. Finally, the creation of a 3-dimensional tissue model. Consequently, manual annotation of the different tissue area is performed (Background, Tumour, and Healthy). The supervised learning process uses manually labelled pixels to identify the held-out pixels. Pixels belonging to healthy tissue and image background have been discarded.

The tumour pixels are then input to parametric t-SNE for DR. Unsupervised analysis is applied to the 2-dimensional latent space generating clusters. Clusters are then assessed and evaluated by a co-expression network. The list of the highly correlated ions per cluster is finally obtained. The final steps of the workflow consist of ions identification and generation of biochemical insights.

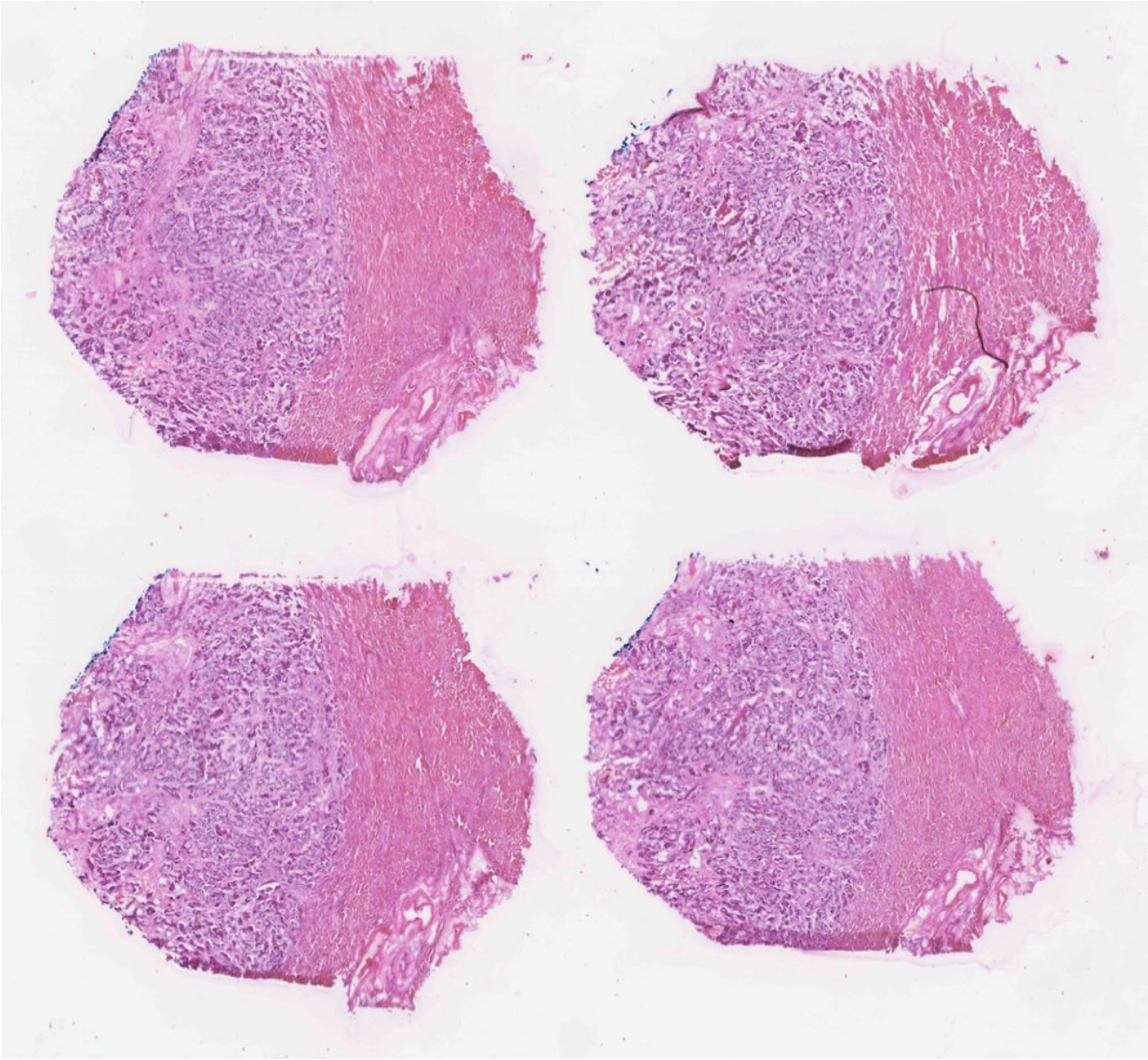


Figure 4.2 - First batch of the four H&E images representing the metastatic liver cancer tissue sample. The samples are organised anticlockwise starting with the first sample in the bottom right corner and the fourth sample located in the bottom left corner in the picture.

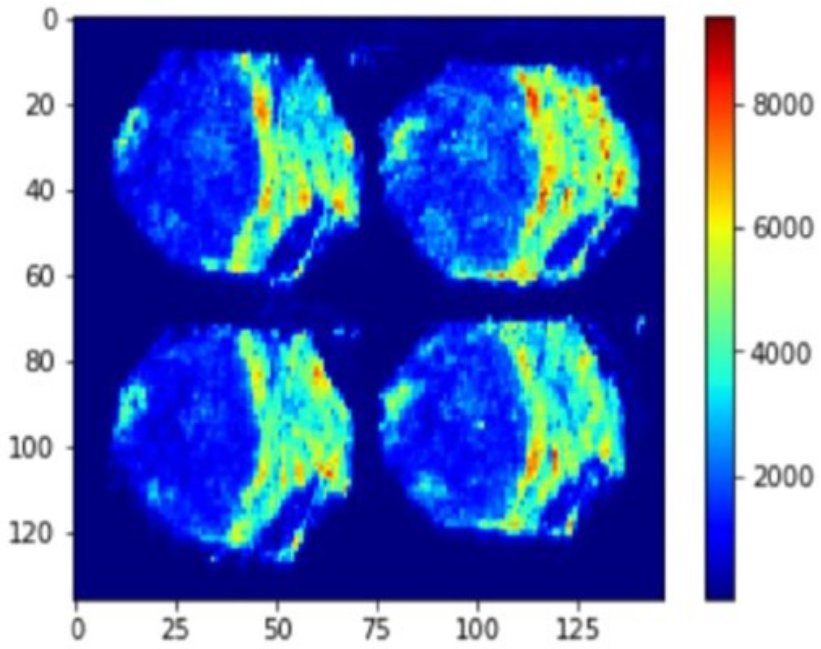
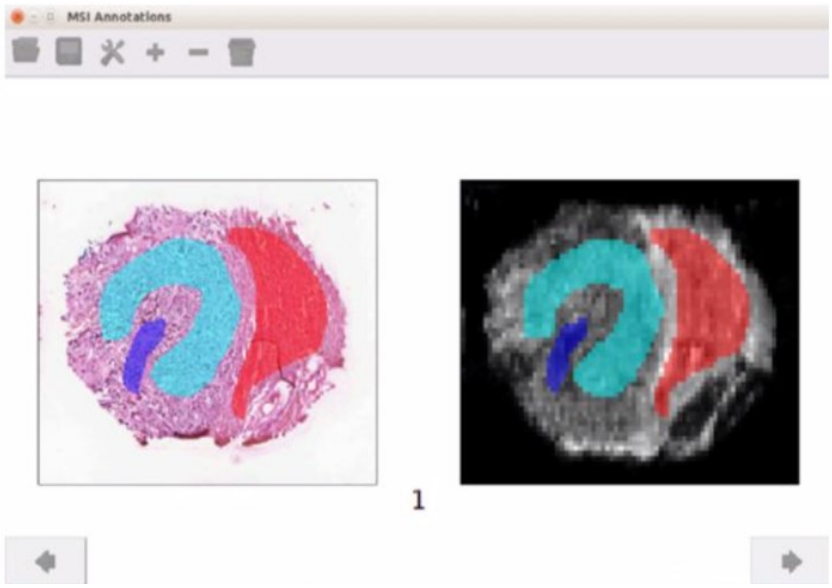
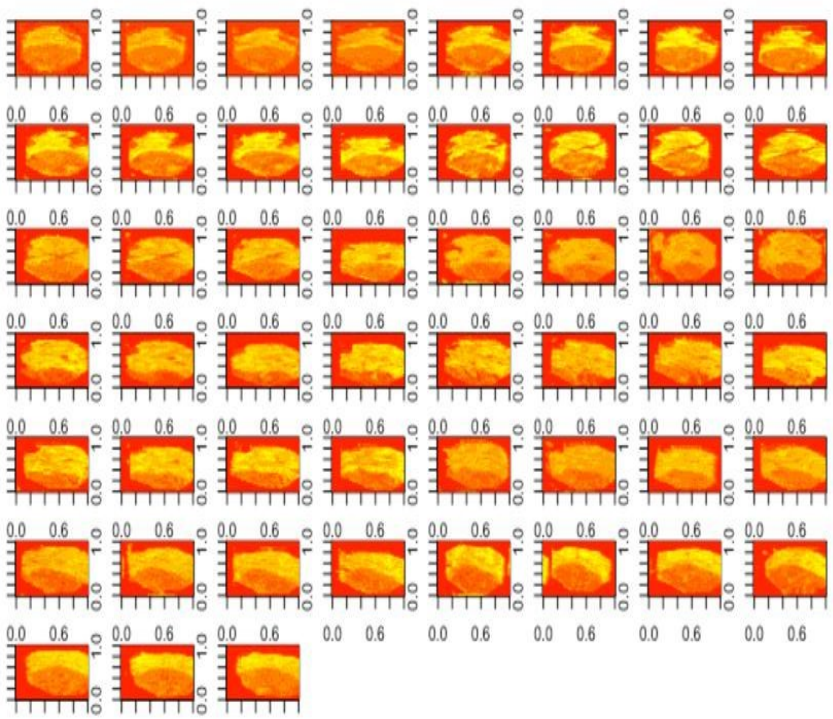


Figure 4.3 - Raw mass spectrometry image obtained from the imzML file format. This picture represents the simple distribution of the spectra embedded into pixels obtained with the application of the DESI-MSI. From a simple visual comparison of Fig 4.2 *and* Fig 4.3 *is visible how different is the resolution of the H&E images from the MSI.*

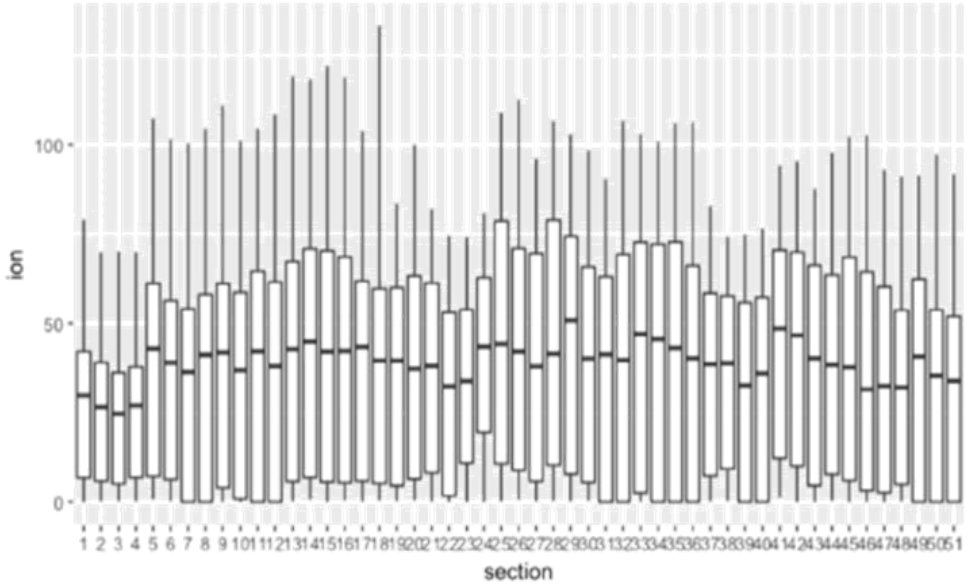
A



B



C



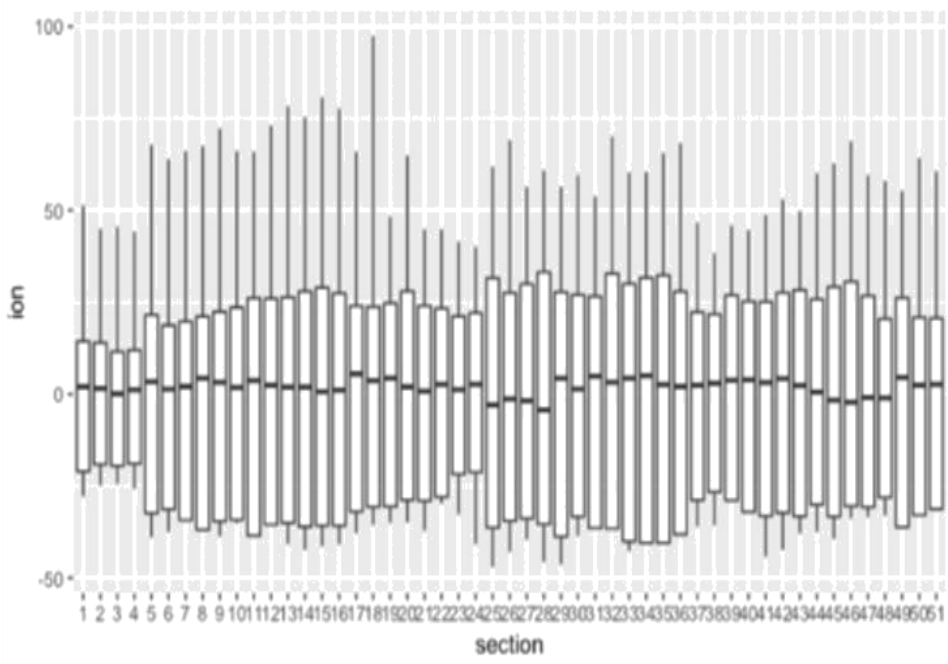


Figure 4.4 – (A) An example of the user interface of the MATLAB script used for the manual annotation. On the screen are reported the two co-registered images, on the left-hand side is reported the H&E image which is used to perform the annotations. On the right-hand side, instead, is displayed the MSI corresponding to the H&E. The coloured areas are an example of how different tissue sub-regions can be highlighted on the H&E images. The same areas are automatically selected on the MSI, creating the connection between the identified tissue sub-structures and their biochemical components (Mass Spectra). (B) and (C) report the results of the batch correction showing the variation of ions before and after the application of the batch correction in all the 51 tissue sections.

4.5 EXTRACTION OF THE TUMOUR MSI DATASET

The next step in the 3D-DESI-MSI dataset analysis is the extraction of the ROI (Region of Interest). To perform the identification of the tissue areas labelled as a tumour, a supervised classifier was trained on the manually annotated spectra (pixels). The model was then used to predict the labels of all the held-out pixels.

Six supervised classifiers were tested, linear SVM (Cortes C., et al., 1995), random forests (Breiman L., et al., 2001), logistic regression, SVC, and two versions of the bagging algorithm, one based on linear SVM and the other based on random forests. All six classifiers were evaluated

through a 30% hold-out cross-validation, repeated five times. All the code was developed in Python using Scikit-learn library.

All the supervised learning techniques tested could predict the held-out pixels with very high accuracy (~0.99), showing a similar performance (Table 4.1). Linear SVM was selected to segment the entire dataset because of its higher accuracy and short computational time.

The spatial map of the three main classes (healthy tissue, tumour tissue, and image background) across the 51 tissue sections was compatible with the morphological properties of the tissue (Figure 4.2, Figure 4.5). As a result of the supervised segmentation, 58,401 spectra were assigned to the tumour class, generating a tumour MSI dataset of [58401x754].

The tumour dataset obtained (after the application of the linear SVM classifier) has very high dimensionality; therefore, the subsequent step in the analysis consists of the application of a DR technique.

Method	***Avg accuracy +/- st. dev.***
Linear SVM	0.99979 (0.00023)
Logistic Regression	0.99948 (0.00026)
Random Forests	0.99960 (0.00032)
SVC	0.99552 (0.00143)
Bagging	0.99975 (0.00031)

Table 4.1 - Performance of the five classifiers on the manually annotated spectra using a 30% hold-out cross-validation, repeated five times.

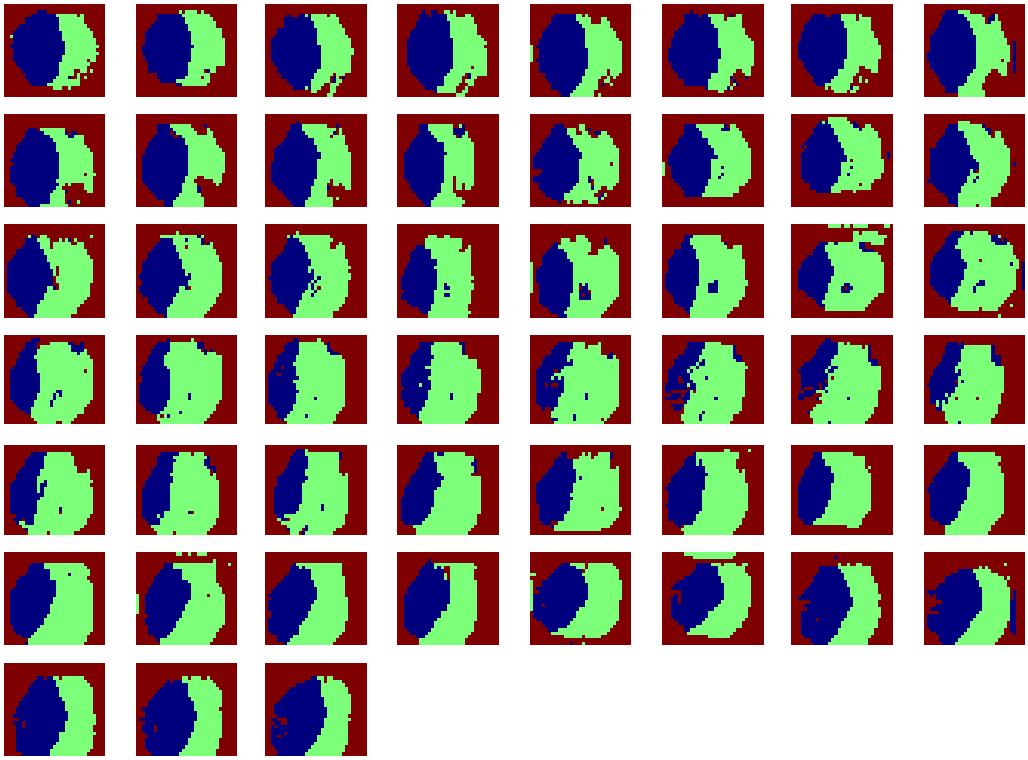


Figure 4.5 - Projection of the three primary tissue labels on the 51 tissue sections (slices). The colours are assigned as follows, background = brown/red, healthy = green, and tumour = blue. The morphology of the three areas is compatible with tissue showing similarities between contiguous sections (i.e. we expect that there is a gradual change across sections in general).

4.6 APPLICATION OF PARAMETRIC t-SNE ON THE TUMOUR DATASET

Different DR algorithms (linear and non-linear) have been compared in Table 4.2 to determine the effectiveness of parametric t-SNE in terms of computational overhead and capacity to highlight complex data sub-structure.

The results indicated that linear techniques were not capable of extracting the high complexity of the features of the dataset. In contrast, non-linear techniques appear to be more useful for DR while conserving the original dataset information. Other techniques, such as Isomap,

MDS, and Kernel PCA, resulted in a RAM (memory) failure due to the high volume of the dataset analysed and to the limited PC hardware available.

Among the non-linear techniques investigated, two provided a model in a reasonable computational time, parametric t-SNE and a variational autoencoder. Both could be used for the unsupervised learning phase (clustering), although the latent space obtained from the autoencoder is more problematic to analyse. It presents data points too densely distributed, making the subsequent clustering procedure difficult to interpret. This result led to the adoption of parametric t-SNE, which appears to be the superior DR technique for this application and was therefore used to determine a low-dimensional mapping of spectra explicitly assigned to the tumour by the linear SVM classifier and to extract a molecular heterogeneity map of the tumour.

Parametric t-SNE was implemented as an 8-layer DBN with a topology of [754 – 500 – 250 – 250 – 1000 – 500 – 250 - 2] neurons and was pre-trained to reconstruct the spectra. Subsequently, a 2-dimensional t-SNE layer was stacked on top of this architecture. The network topology used (not the number of neurons and hidden layers) was inspired by previous work (Hinton GE, et al., 2006) (Van Der Maaten L., et al., 2009) (Inglese P, et al., 2017).

The objective of reducing the dimensionality of the original space to a bidimensional latent space is to generate an easily interpretable map (visually) of the statistical relationships between the spectra (Van Der Maaten L., et al., 2009).

The parametric t-SNE model, trained on 25,000 randomly selected spectra, was subsequently used to map the held-out spectra. Here, exploiting the parametric aspect of the technique, we could generate a DR map utilizing a subset of the available data, with lower computational requirements.

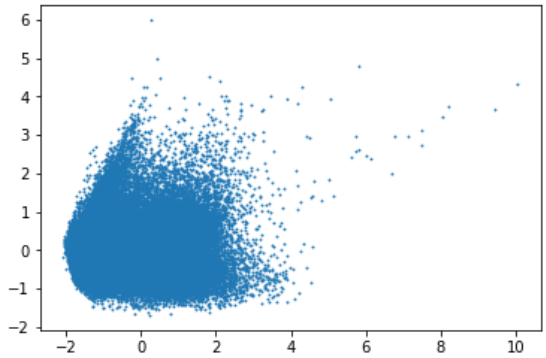
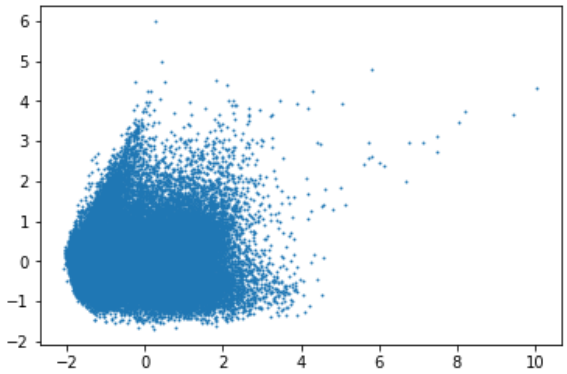
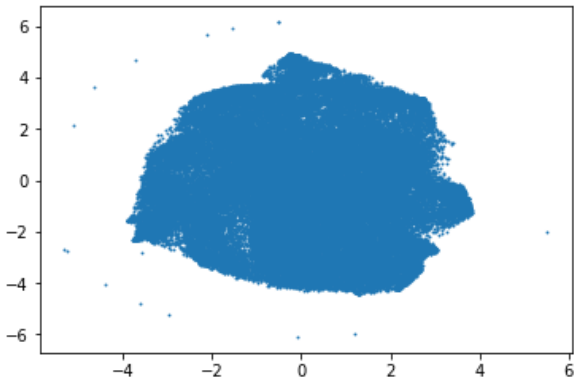
The greedy layer-wise training of DBN was performed using mini-batches of 100 input data, for 35 epochs. Learning rate was set equal to 0.001, and regularisation term was set to 0.0002. All the activation functions were logistic sigmoid (Eq. 3.11), except for the deepest hidden layer, where a linear activation function was used, as described in the previous chapter. Fine-tuning of the t-SNE layer was performed using the Polak-Ribière conjugate gradient with mini-batches of 5,000 input data, for 500 epochs. The t-SNE perplexity parameter was set to 30. The parametric t-SNE model was fitted using a reworked version of the MATLAB code available at https://lvdmaaten.github.io/tsne/code/ptsne.tar.gz.

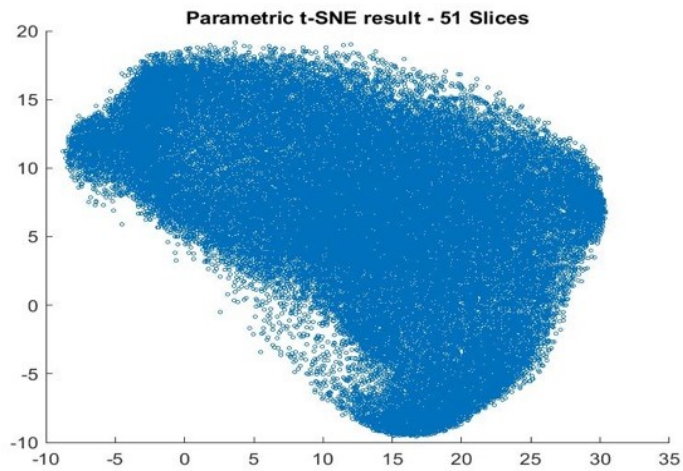
The first qualitative results showed the higher complexity of the parametric t-SNE map, compared to the results obtained by the application of other DR techniques (Table 4.2).

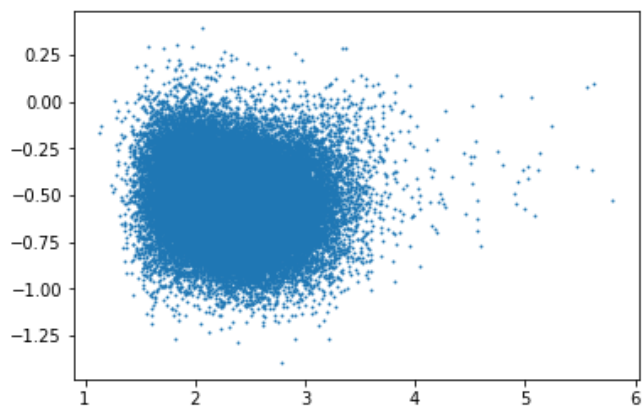
Considering the high complexity of the tumour dataset, the minimum number of dimensions to take into consideration for better compression and interpretation of the data was 411 (with a variance of the 94%) (Figure 4.6). However, we want to perform a visual inspection of the tumour sub-structure; therefore, the aim of the parametric t-SNE was a 2-dimensional latent space.

In the following sections of this study, we will describe how these results can lead to the identification of clusters of spectra with similar profiles.

Table 4.2 – A visual comparison of the 2-dimensional latent spaces generated by nine different dimensionality reduction techniques, linear and non-linear, applied to the tumour dataset.

PCA	
Kernel PCA	Memory fail
Randomize PCA	
t-SNE	
LLE (Local Linear Embedding)	Computational time out

Parametric t-SNE	

MDS (Multidimensional Scaling)	Memory Fail
Isomap	Memory Fail
Variational Autoencoder	

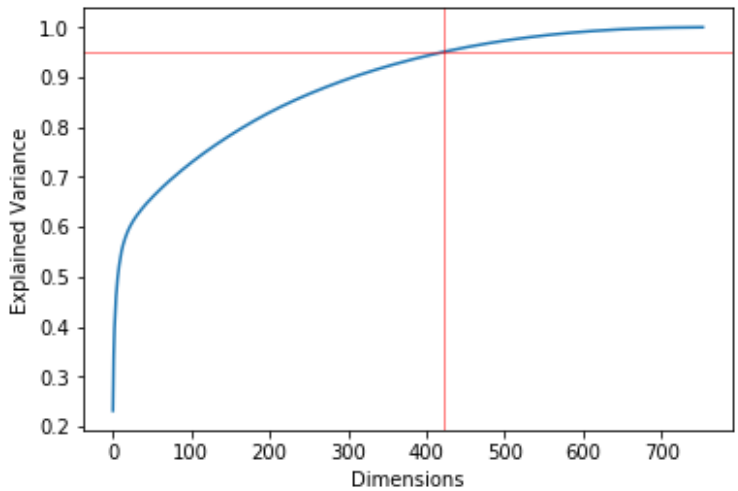


Figure 4.6 – A representative description that is easier to interpret from the high dimensional complexity of the tumour dataset can be represented with a lower-dimensional space of 411 dimensions, which maintains the majority of the information embedded in the high dimensional space (752) accounting for 94% of the variance.

4.7 CLUSTERING OF THE 2-DIMENSIONAL TUMOUR LATENT SPACE

Previously, we have shown how parametric t-SNE is able to produce a detailed low-dimensional representation of the similarity patterns in the spectra. When applied to the tumour spectra, parametric t-SNE places spectra closer with similar profiles generating a map with distinct sub-regions of different density that were absent in the scatter plots of other DR methods (Table 4.2).

Unsupervised clustering has then been applied on the 2-dimensional mapping to highlight and extract groups of highly related data points. Considering the scatter plot of the latent space generated by the parametric t-SNE (Table 4.2), a density-based clustering method is an adequate mathematical approach to accurately detect the clusters.

For the identification of the clusters in the tumour-related spectra, we applied OPTICS and subsequently DBSCAN (described in the introductory chapters) to the 2-dimensional latent space.

For the application of the OPTICS algorithm, we have used 30,000 randomly selected data points with a *MinPts* set to 700, which means that we expected to see groups of spectra larger than this value. As explained previously, the other important parameter to set is ε. Different thresholds, due to different values of ε, determine different partitions; in fact, the lower ε is, the finer are the partitions. For the identification of the adequate ε, a visual analysis of the reachability plot has been adopted. Robust clusters are associated with deep valleys of the reachability plot (Figure 4.7 A). Scanning equally distant thresholds for the reachability-distance, we selected $\varepsilon = 2$.

The same values of *MinPts* and ε have been used for DBSCAN algorithm. The resulting clustering is plotted in Figure 4.7 B.

The final partition was then selected using the minimum value of the Davies-Bouldin index (DBI) (Davies D.L., et al., 1979). Using this method, we determined the presence of three clusters in the tumour dataset (Figure 4.8).

Finally, the cluster's labels identified with the application of the DBSCAN were propagated to the held-out tumour samples, initially identified as noise, as by looking at the closest labelled data point, using the K-nn algorithm. The projection of the cluster labels on the parametric t-SNE latent space is shown in Figure 4.9.

The code used to deploy OPTICS, DBSCAN, and K-nn algorithms is available in the Scikit-learn library for Python.

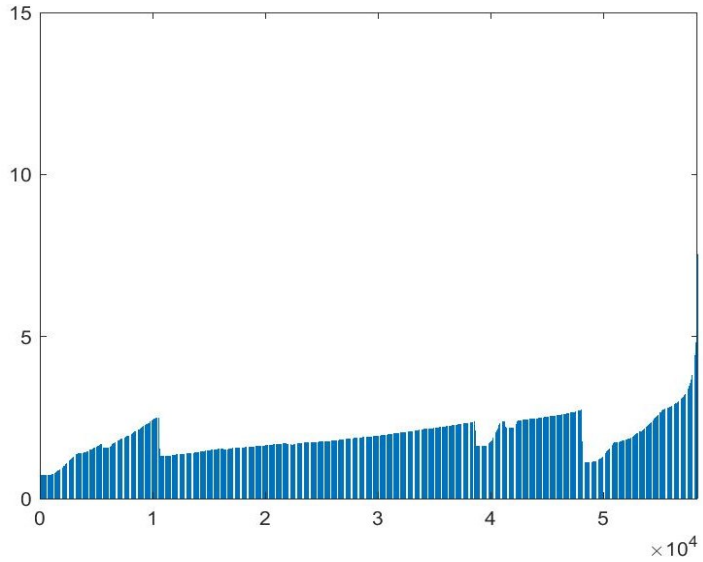
Attention must be paid, again, to the meaning of the data points in Figure 4.9. Each data point corresponds to a spectrum, and it is a pixel in the imaging dataset. We have projected the cluster labels onto the corresponding pixels revealing the spatial distribution of the clusters. These results confirmed that the clusters were compatible with the expected continuity and progression of adjacent tissue sections (Figure 4.10).

However, even if an initial visual inspection of the clustering result points to the conclusion that this partition of the latent space could be correct, we need to realise that the results found are mathematical constructs, therefore further validation is needed. What we have identified as possible partitions of the tissue sample need to effectively represent the tumour microenvironment.

To corroborate the validity of the results obtained, we calculated the ion co-expression network of the tumour-related spectra. The aim was to determine groups of ions that are highly co-localised. Quantitatively, this is equivalent to measuring the correlation between the images associated with each detected peak.

A correlation threshold of 0.87 was applied to a 754 x 754 Pearson's correlation matrix. In this way, only highly correlated pairs of ions were considered. Then, the correlation matrix was used as the adjacency of an undirected force-directed graph (Figure 4.12).

A



B

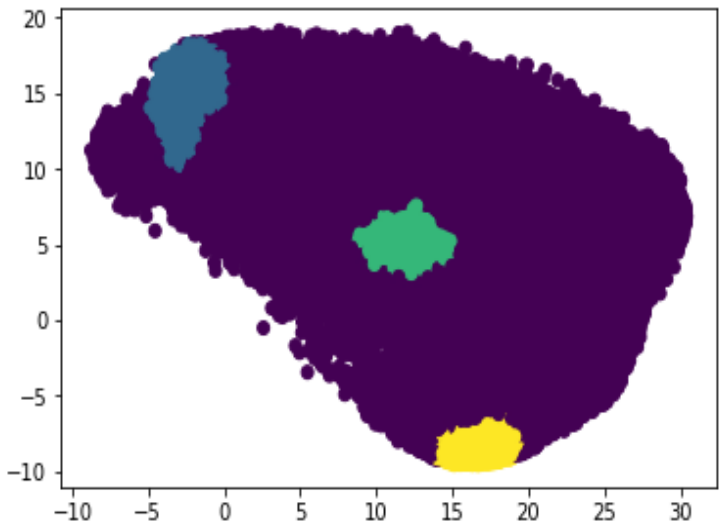


Figure 4.7 - (A) Reachability distance plot of the 2D latent space obtained with the application of the OPTICS (MinPts = 700). (B) Three possible clustering partitions were obtained by the application of the DBSCAN (MinPts = 700, ε = 2). The colours are assigned as follows, noisy points = purple, cluster 1 = blue, cluster 2 = green and cluster 3 = yellow. The other high-density regions did not meet the constraints imposed by MinPts and ε, therefore they were not assigned to additional clusters (noise).

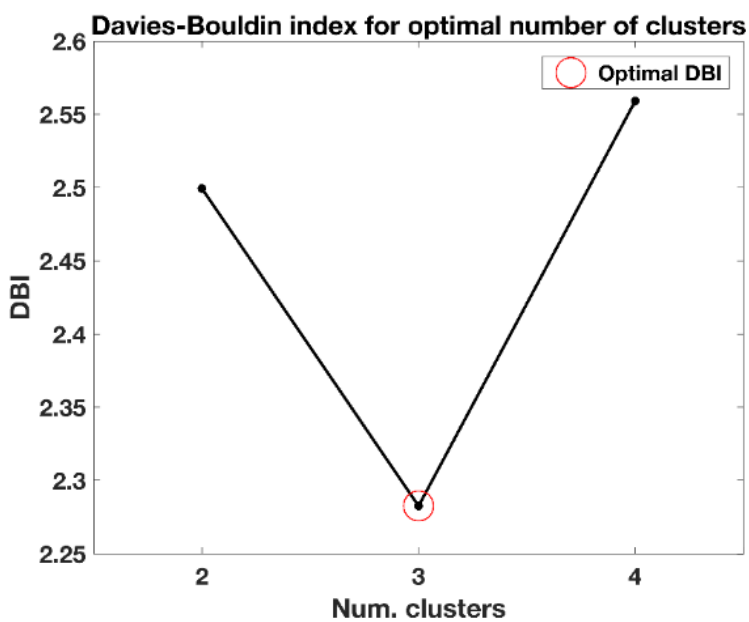


Figure 4.8 - Optimal partition corresponded to 3 clusters (minimum DBI value).

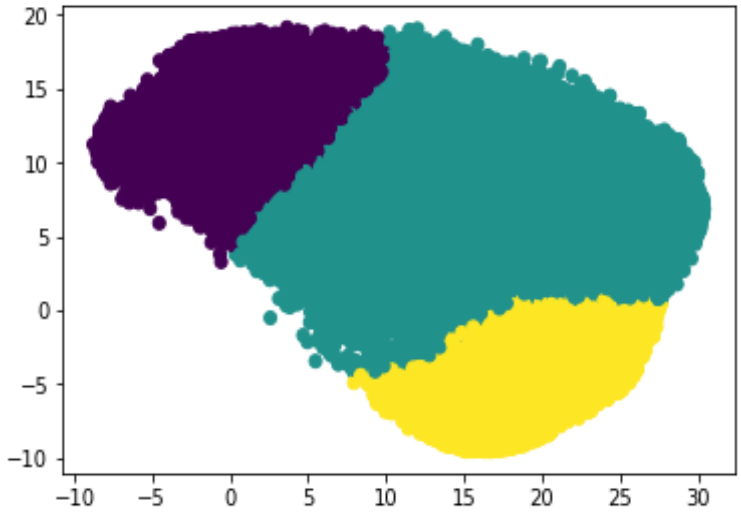


Figure 4.9 - Results of the K-nn algorithm. The data points labelled as noisy points by DBSCAN have now been assigned to the adjacent clusters to do not lose any relevant information connected to the tumour sample. The colours are assigned as follows, ***cluster 1 = purple, cluster 2 = green and cluster 3 = yellow****.*

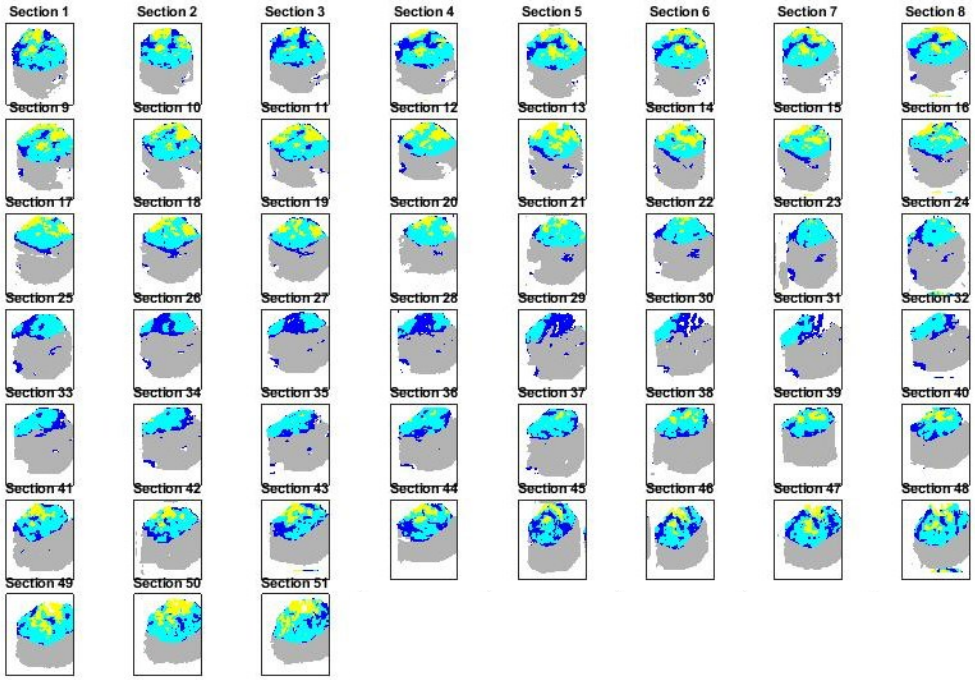


Figure 4.10 - Spatial distribution of the clusters identified by OPTICS. The morphological characteristics of the mapped regions are compatible with the continuity of adjacent tissue sections.

The clusterMaker2 algorithm (Morris J.H., et al., 2011) revealed the presence of eleven disconnected sub-networks, which were spatially distributed in a way that shows three largest sub-networks, the sum-of-intensity images corresponding to their nodes were calculated (Table 4.3). Visual inspection revealed a pairwise correspondence between these three images and the regions associated with the three clusters (Figure 4.12). This correspondence was confirmed by Pearson's correlation between the sum-of-intensity and the cluster images (Table 4.3).

The ten peaks with the largest connectivity degree (sum of their edges) were selected as representative members of the three sub-networks considered. A large connectivity degree is equivalent to saying that the corresponding peak image is more correlated with the other nodes of the sub-network and, also, with the group of peaks sum-of-intensity image.

The relationship between the intensity of the three peaks with the largest connectivity degrees (within each sub-network) and the cluster labels was evaluated using a Kruskal-Wallis test. This test confirmed that the three main peaks per sub-networks were more abundant in the corresponding cluster, previously identified by only visual inspection (Figure 4.10).

4.8 CLUSTERING OF THE HIGH DIMENSIONAL TUMOUR DATASET

Reducing the dimensionality of a dataset leads to an inevitable loss of information (Junfeng An, et al., 2022). In the previous chapters we have studied how parametric t-SNE is a valuable DR methodology. In fact, it effectively reducers the dimensionality of the tumour data set while minimising loss of relevant information. However, here we provide additional proof that supports the importance of reducing the dimensionality of the tumour dataset to cluster and identify the tumour substructure.

In Figure 4.11 below are reported the results of the clustering algorithms OPTICS and DBSCAN applied directly to the whole high dimensional tumour space. Inspection implies that we can see how both the clustering methods struggle to extract a substructure from the dataset. The reachability distance plot (Figure 4.11 – A) calculated with a *MinPts* = 500 does not show any sign of *hills and valleys* which are directly connected with the identification of possible clusters. This result is corroborated by the DBSCAN plots reported in Figure 4.11 – B. Here, we can see that even with two different ε values (0.5 and 2) and *MinPts* = 500 the clustering algorithm is not able to highlight any substructure (clusters) in the high dimensional tumour dataset.

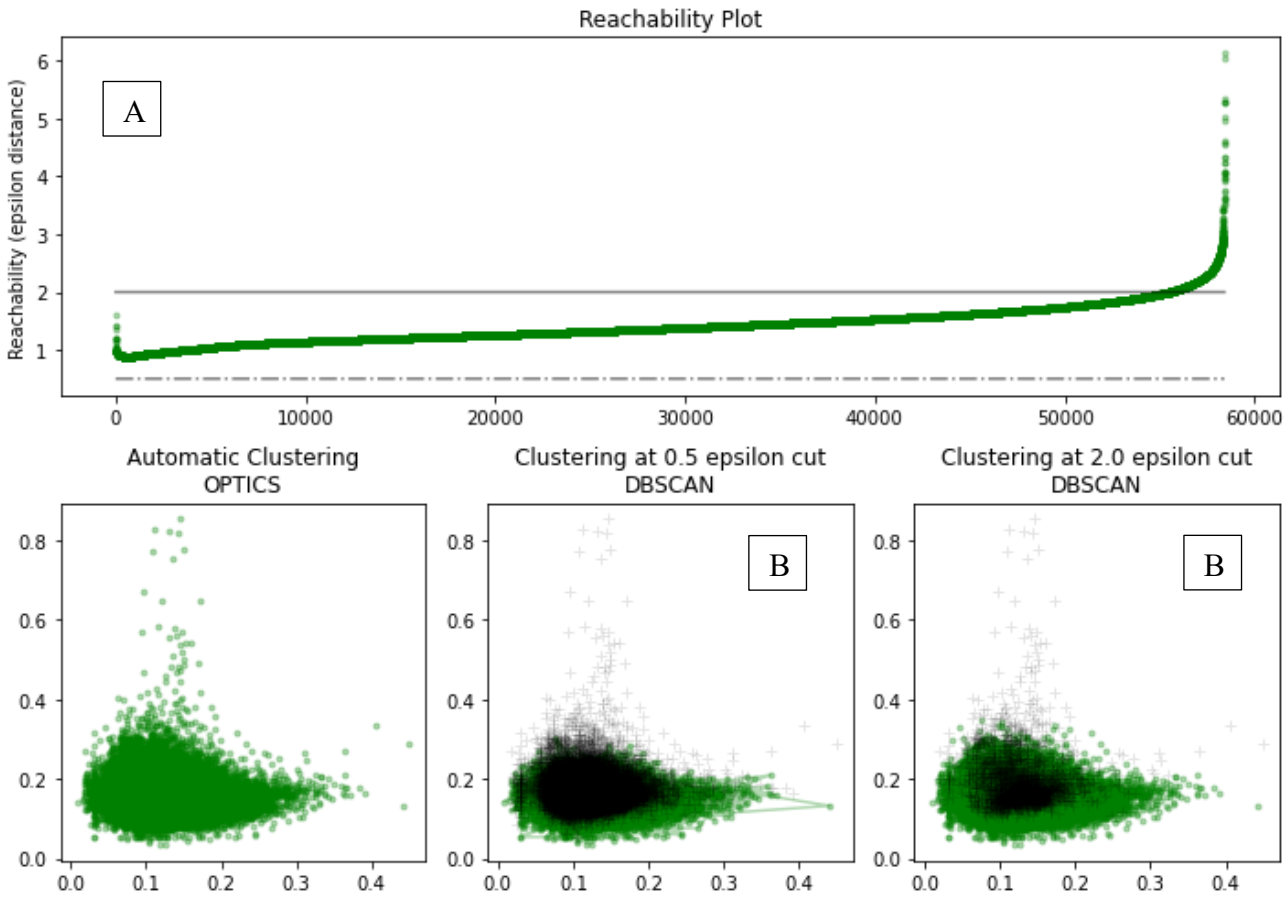


Figure 4.11 - (A) Reachability distance plot of the whole high dimensional tumour dataset. This outcome has been obtained with the application of the OPTICS algorithm (MinPts = 500). (B) Two possible partitions were obtained by the application of the DBSCAN (MinPts = 500, ε = 0.5 and ε = 2).

Therefore, we can conclude that the application of a DR tool, parametric t-SNE in this specific case, is a necessary step during the analysis of a 3D-MSI dataset. The dimensionality reduction enhances the clustering algorithms performance allowing identification of the tumour substructures. Furthermore, the computational overhead needed for the application of OPTICS and DBSCAN on the whole high dimensional dataset is very high, while the application of parametric t-SNE followed by the clustering algorithms makes the process leaner.

4.9 PEAK ANNOTAION AND BIOLOGICAL INSIGHTS

The selected peaks (*m/z* values) were annotated using the METLIN database. The *m/z* values common to all the peaks used in the analysis were searched in the raw data within a window of +/- 5 ppm. Annotations were accepted only if within an error of 5 ppm (Table 4.4).

We observed that PGs and PIs mainly represented cluster one. This result is compatible with previous investigations (Dobrzyńska I., et al., 2005) (Chan E.C.Y., et al., 2009) (Gerbig S., et al., 2012) that showed that this family of phosphatidylglycerol, which belongs to the class of the phospholipids, is associated with rapidly proliferating human cancer. Cluster one also presents an abundancy of the PI-Cer(t18:0/16:0(2OH)), which is categorised as Sphingolipids [SP] and belongs to the main class of Phosphosphingolipids [SP03] and subclass of Ceramide phosphoinositols [SP0303]. The bioactive sphingolipid ceramide molecule has emerged as an antitumorigenic lipid. An abundance of ceramide suppresses cell motility promoted by the epithelial growth factor that is a pro-metastatic factor. In particular, ceramide limits phosphatidylinositol-3-kinase C2β-controlled cell motility in cancer. Therefore, novel anti-tumour treatments use the ceramide molecule as a potential metastasis-suppressor lipid (Kitatani K., et al.,

2016). The annotated ions in the three clusters consisting of ceramides also suggest that necrotic tissue was localised in the corresponding area (Inglese P, et al., 2017).

Table 4.3 - Values of the Pearson's correlation calculated between the sum-of-intensity image of the three main sub-networks and the three OPTICS clusters images.

	Cluster 1	***Cluster 2***	***Cluster 3***
SSI 1	**0.7003**	0.0972	0.1983
SSI 2	0.3566	**0.5223**	0.2842
SSI 3	0.0234	0.0599	**0.4374**

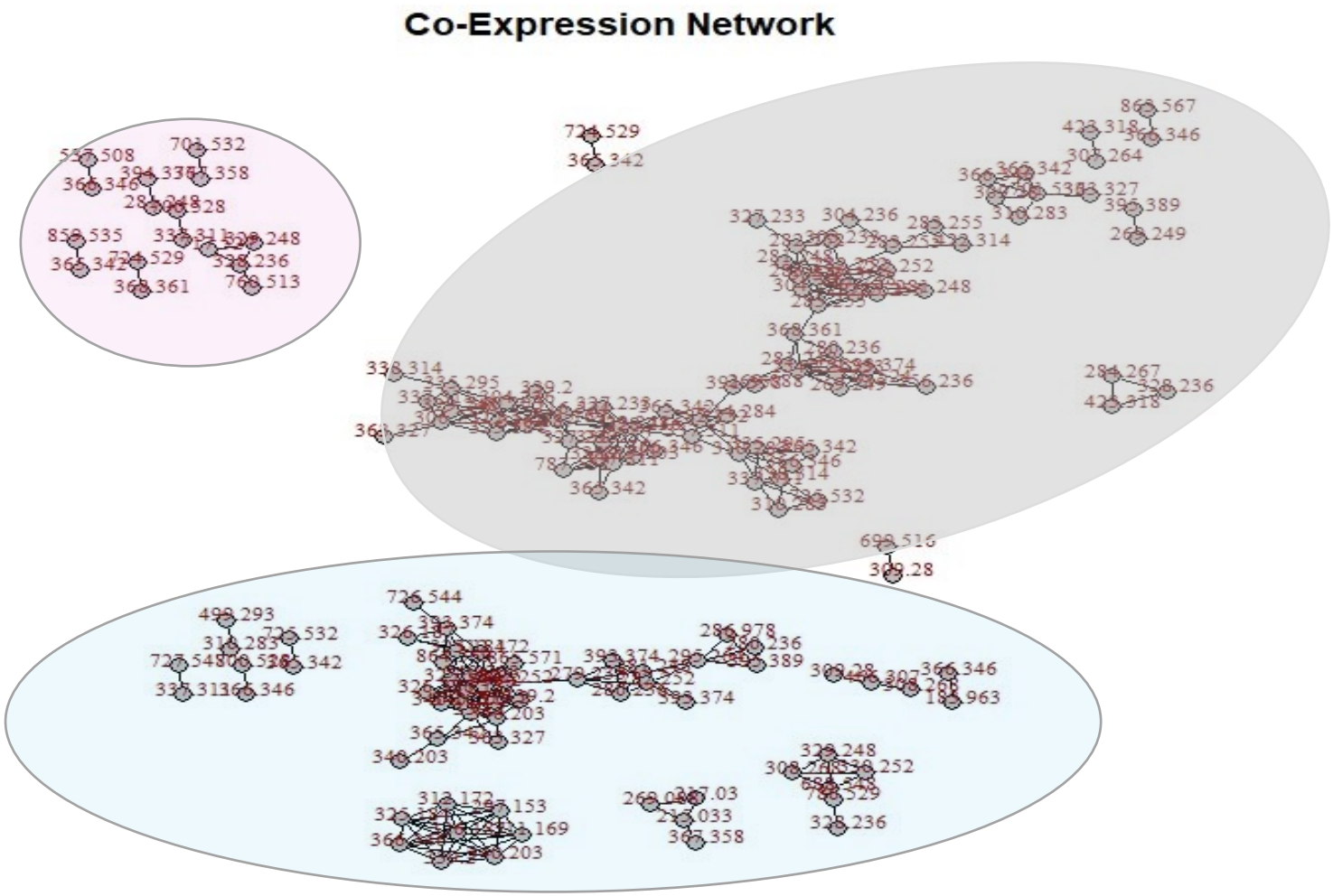
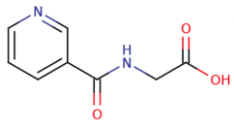
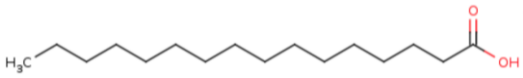
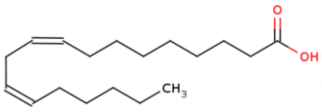
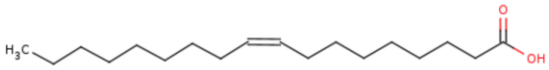
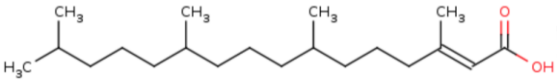
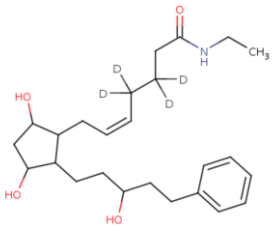
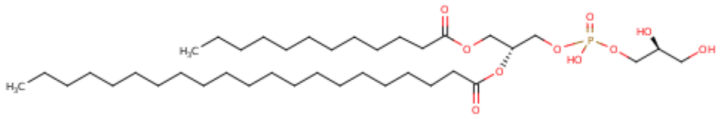
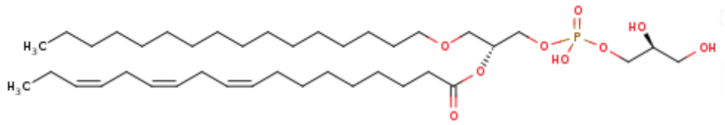
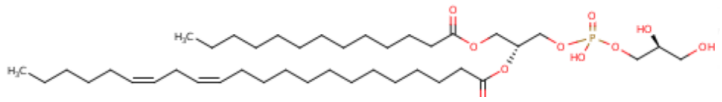
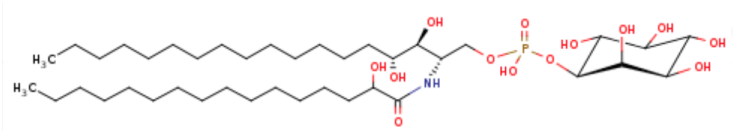
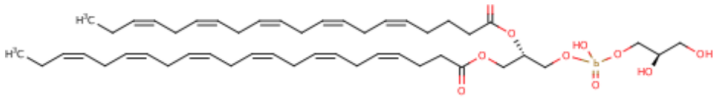
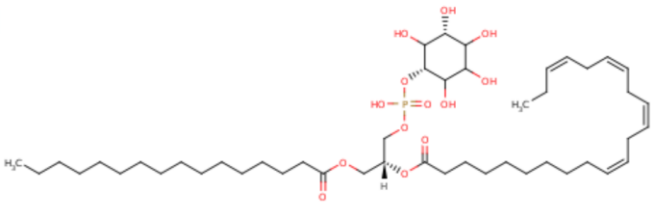


Figure 4.12 - Co-Expression network. The three largest groups of connected ions are coloured in pink, green, and blue, respectively. The sum-of-intensity image of the peaks belonging to these sub-networks is higher in the corresponding regions assigned to the three OPTICS clusters. This correspondence suggests that the clusters are related to similar spatial distributions of the peaks of the three main sub-networks.

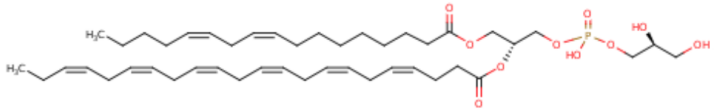
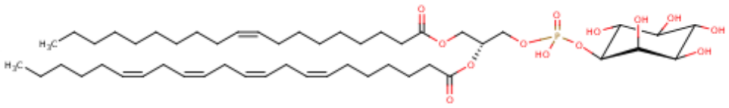
Table 4.4 - The ten m/z values of the clusters with the highest correlation values were annotated using the MATLIN database. Cluster 1 is characterised by a higher abundance of PGs and PIs, whereas cluster 2 shows a higher abundance of PGs and ceramides (Cer), and cluster 3 is mainly made up of PGs.

Cluster 1

m/z	Annotations
217.0296 [M+CL]	Nicotinuric acid 
256.2362	Palmitic acid 
280.2362	Linoleic acid 
282.2518	Oleic Acid 
310.2832	Phytenic acid 
422.3144	Latanoprost ethyl amide-d4 

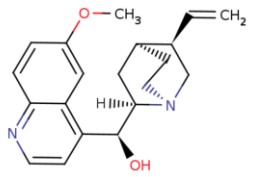
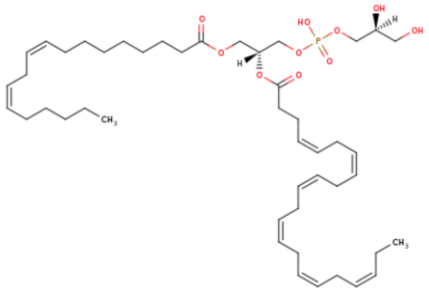
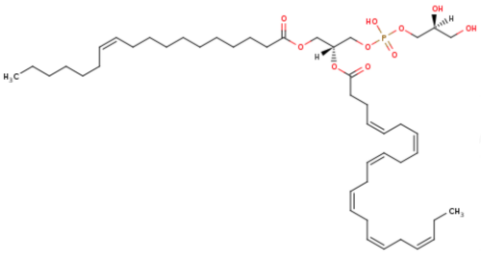
743.5426	PG(12:0/21:0) 
745.5583	PG(O-16:0/18:3(9Z,12Z,15Z)) 
767.5424	PG(13:0/22:2(13Z,16Z)) 
862.5530 [M+CL]	PI-Cer(t18:0/16:0(2OH)) 
864.5693	PG(22:6(4Z,7Z,10Z,13Z,16Z,19Z)/20:5(5Z,8Z,11Z,14Z,17Z)) 
886.5530	PI(16:0/22:4(10Z,13Z,16Z,19Z)) 

Cluster 2

m/z	Annotations
280.2362	Also present in cluster 1
282.2518	Also present in cluster 1
310.2832	Also present in cluster 1
422.3144	Also present in cluster 1
767.5424	Also present in cluster 1
806.4976	PG(17:2(9Z,12Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z)) 
862.5530	Also present in cluster 1
864.5693	Also present in cluster 1
886.5530	Also present in cluster 1
928.5922	PI(19:1(9Z)/22:4(7Z,10Z,13Z,16Z)) 

Cluster 3

m/z	Annotations
256.2362	Also present in cluster 1
280.2362	Also present in cluster 1 and cluster 2
282.2518	Also present in cluster 1 and cluster 2
310.2832	Also present in cluster 1 and cluster 2

325.1841	Quinidine 
422.3144	Also present in cluster 1 and cluster 2
767.5424	Also present in cluster 1 and cluster 2
818.5058	PG(18:2(9Z,12Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z)) 
820.5218	PG(18:1(11Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z)) 
862.5530	Also present in cluster 1 and cluster 2
886.5530	Also present in cluster 1 and cluster 2
928.5922	Also present in cluster 2

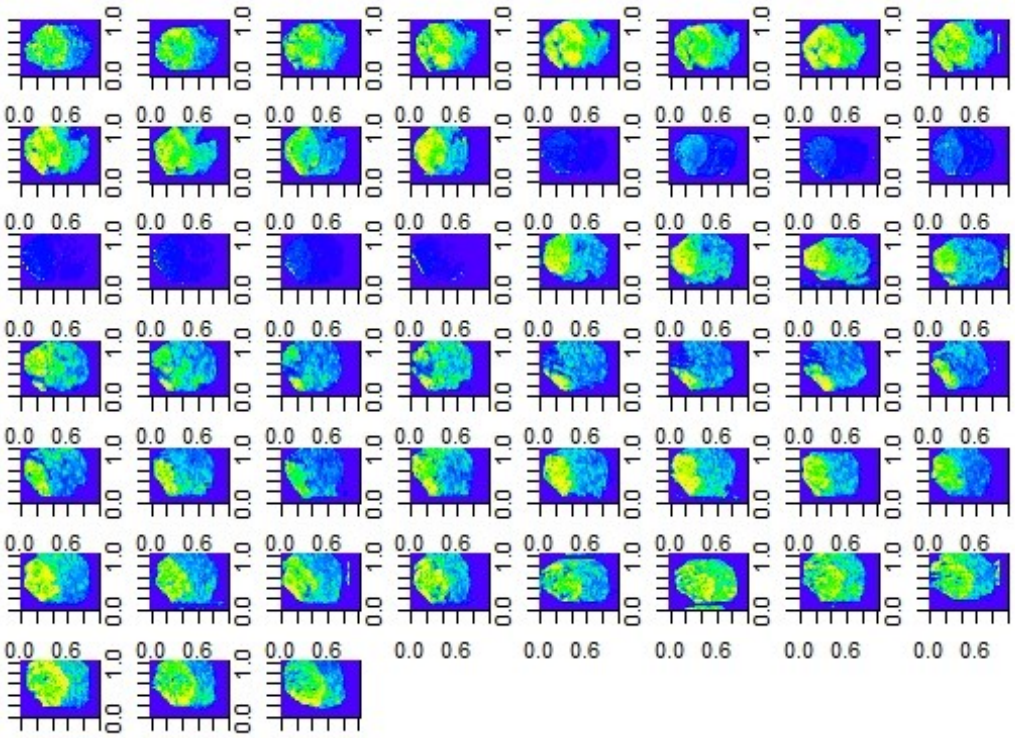
Among the highly correlated molecules detected in cluster one, particular attention has to be paid to the PI(16:0/22:4(10Z,13Z,16Z,19Z)). This lipid is a phosphorylated form of phosphatidylinositol (PI), a phosphoinositide. This class of molecule plays essential roles in membrane trafficking and cell signalling. Numerous studies available in the literature suggest multiple links between the cellular phosphoinositide system and cancer (Bunney T.D., et al., 2010) (Zou Z., et al., 2020). The phosphoinositide signalling system can be seen as a network of enzymes, phospholipid messengers and their binding proteins. The interactions that occur between phosphoinositides and their binding proteins are critical to their regulatory actions (Bunney T.D., et al., 2010). Therefore, the presence of PI(16:0/22:4(10Z,13Z,16Z,19Z)) may be considered as a marker for tumour proliferation. However, this relevant finding needs further investigation that goes beyond the scope of this work.

Cluster two was characterised by a higher abundance of PGs, suggesting the presence of mucus in a mucinous sub-type of cancerous tissue (Gerbig S., et al., 2012) (Mandal M.K., et al., 2013), since PGs have a role as a surfactant. We excluded the bacterial origin of those PGs because of the presence of incompatible large acyl chains ($n > 17$).

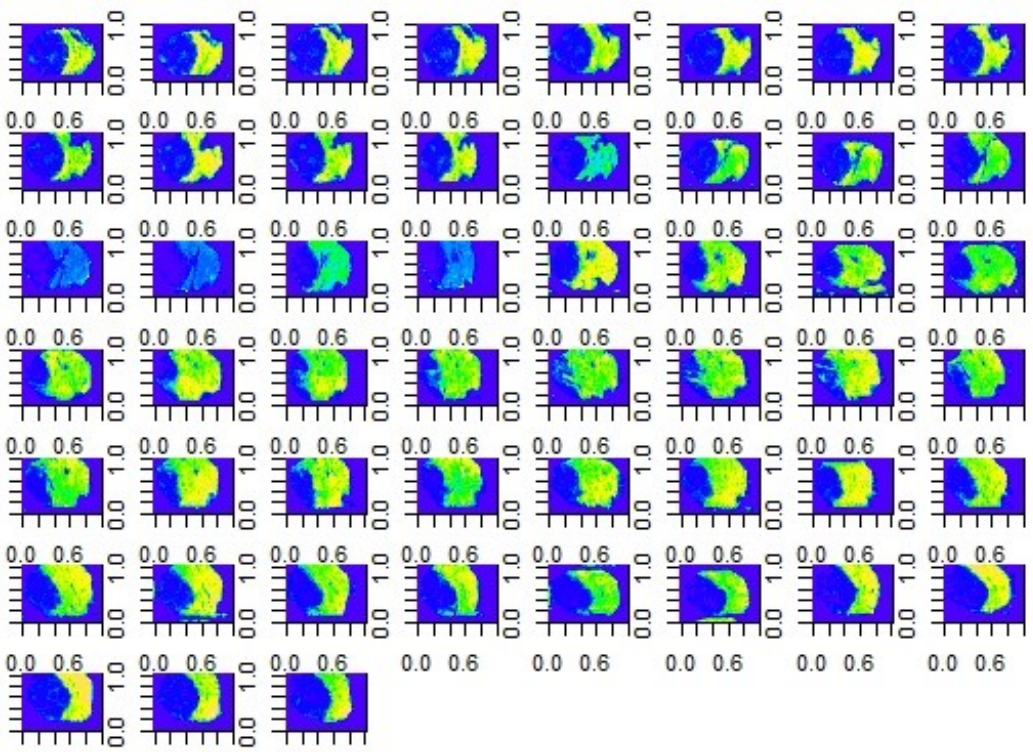
Interestingly, the PG family found in the clusters suggested the hypothesis that a process of apoptosis was occurring in that region of the tissue (Koshiji M., et al., 1998) (Chaurio R.A., et al., 2009) (Birge R.B., et al., 2016). The close localisation of cluster two and three also pointed to a possible signalling mechanism between the two tissue types suggesting that local high activity may be associated with the metastatic-related tissue (Chen L., et al., 2017). It is essential to notice that although PGs are constituents of the membrane of bacterial cells, they have a role as a surfactant in mammalian organisms. For this reason, the observed PGs were probably present in the interstitial space (Inglese P, et al., 2017).

The distributions of the highly correlated ions have been plotted (figure 4.13) to perform a visual comparison (presence and absence) between the healthy tissue and the tumour tissue. The ion distributions allow us to identify those ions that are present just in the tumour tissue and, also, visualise how they grouped in the different tissue sub-regions (Figure 4.13).

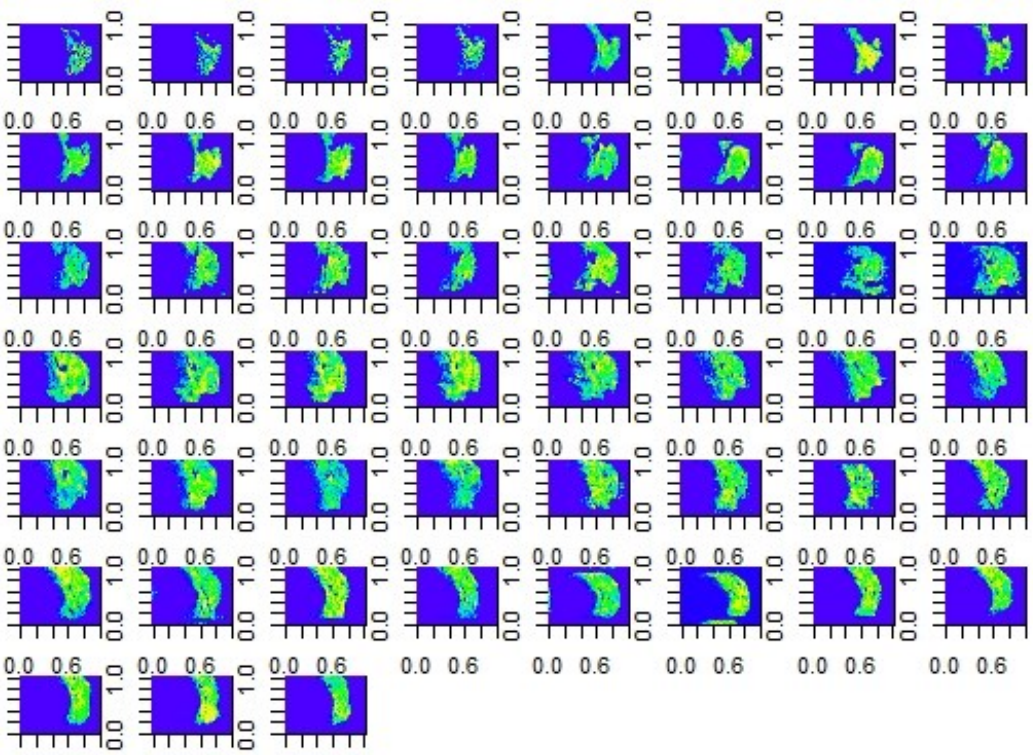
Among the annotated lipids, particular attention needs to be paid to PG*(22:6(4Z,7Z,10Z,13Z,16Z,19Z)/20:5(5Z,8Z,11Z,14Z,17Z))*, which is a phosphatidylglycerol corresponding to 864.5693 *m/z.* This lipid is detectable in the tumour tissue only (Figure 4.13), and it could be classified as a possible metastatic liver cancer biomarker. Although this finding is exciting, further biological study is needed to understand better the role of this lipid plays in oncogenesis, for example, through metabolic and signalling pathway analysis.



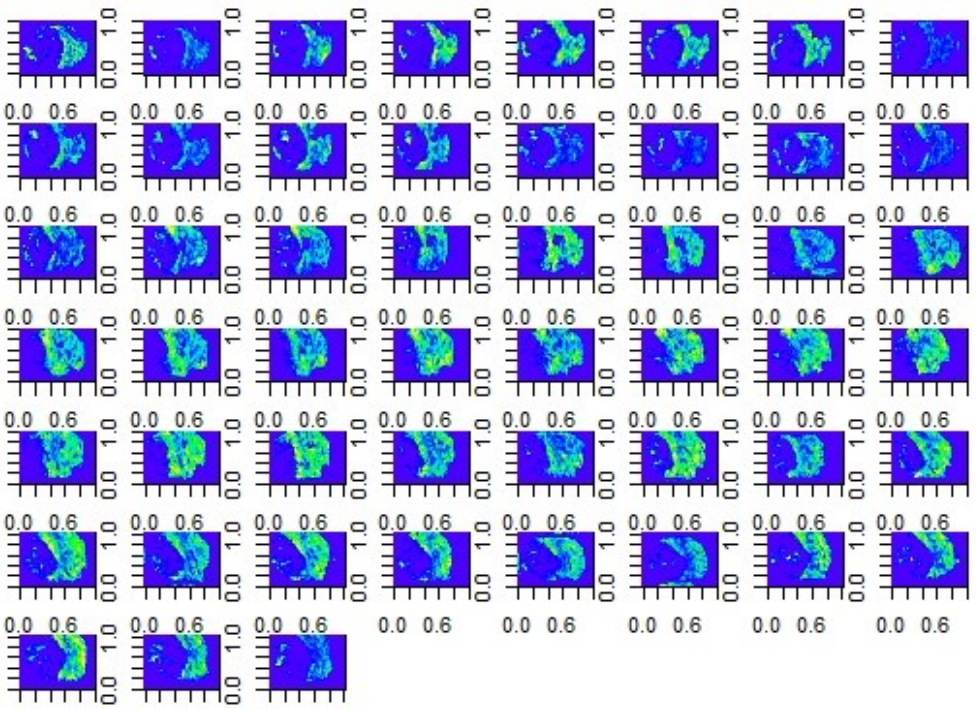
745.5583 m/z [PG(O-16:0/18:3(9Z,12Z,15Z))]



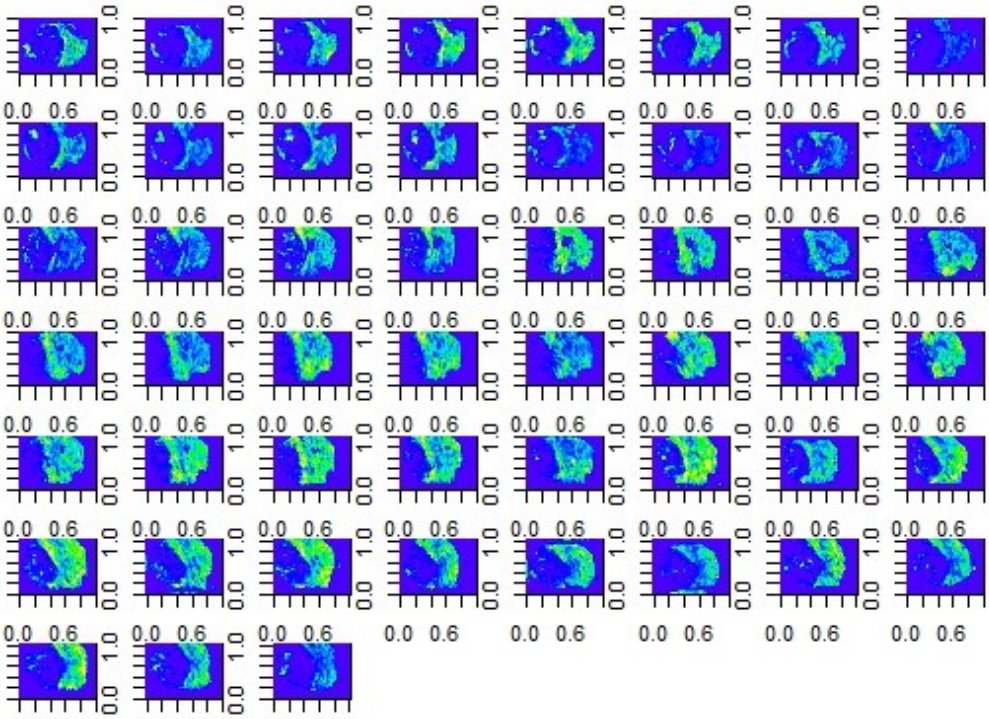
767.5424 m/z [PG(13:0/22:2(13Z,16Z))]



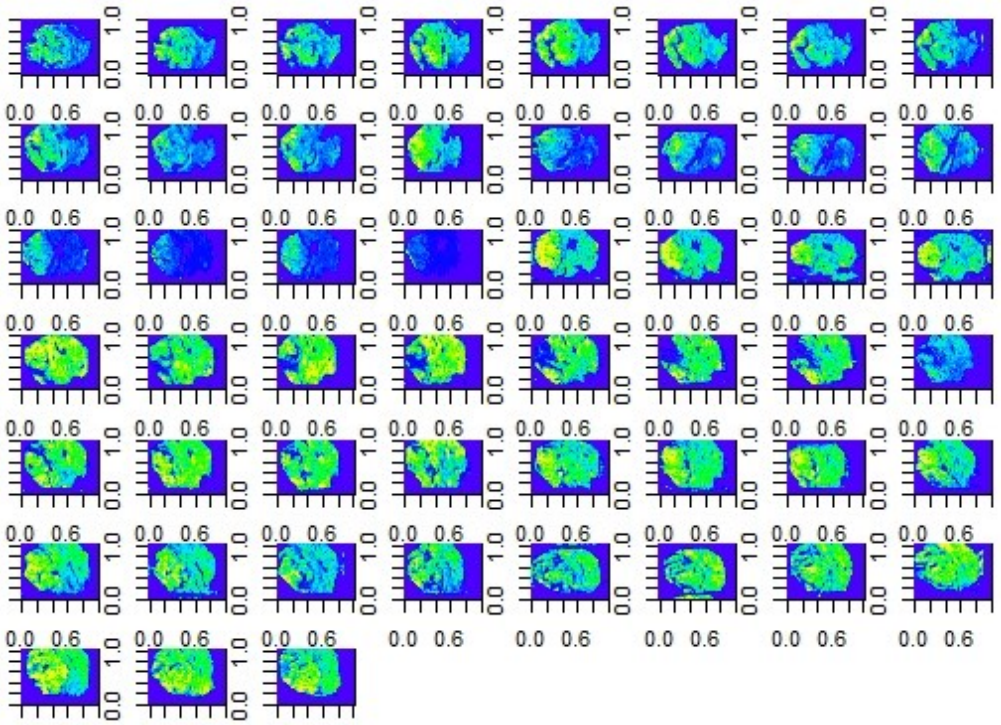
806.4976 m/z [PG(17:2(9Z,12Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z))]



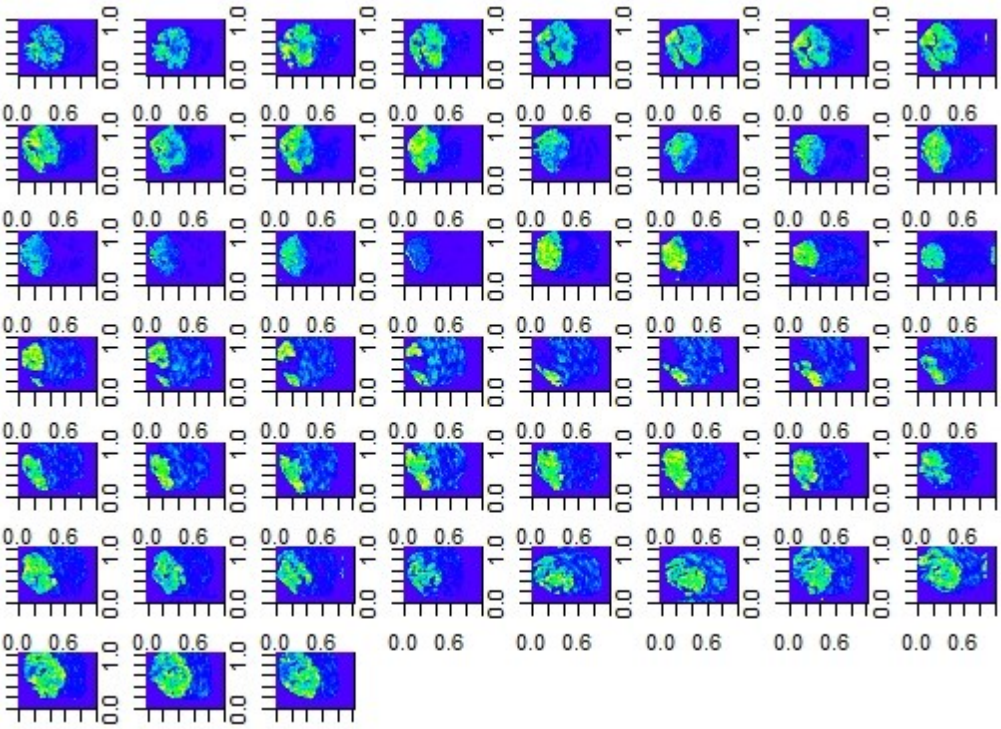
818.5058 m/z [PG(18:2(9Z,12Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z))]



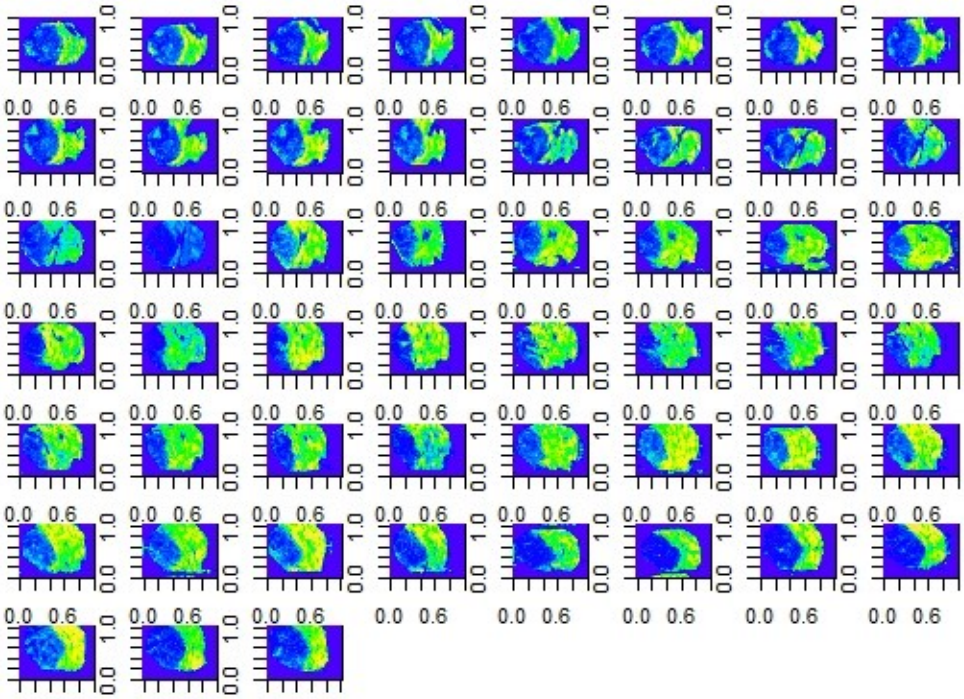
820.5218 m/z [PG(18:1(11Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z))]



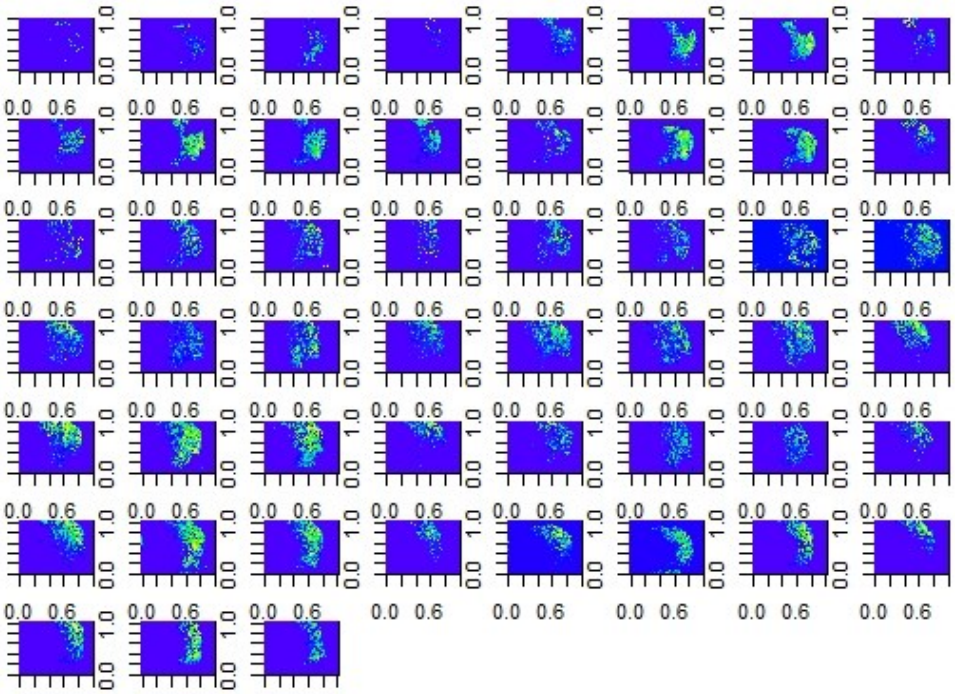
862.5530 m/z [PI-Cer(t18:0/16:0(2OH))]



864.5693 m/z [PG(22:6(4Z,7Z,10Z,13Z,16Z,19Z)/20:5(5Z,8Z,11Z,14Z,17Z))]



886.5530 m/z [PI(16:0/22:4(10Z,13Z,16Z,19Z))]



928.5922 m/z [PI(19:1(9Z)/22:4(7Z,10Z,13Z,16Z))]

Figure 4.13 – Distributions of same of the highly correlated ions in both healthy and tumour tissues.

4.10 LIPIDOMICS IN ONCOLOGY AND CANCER TREATMENTS

At this point, a brief insight into the roles that the lipids play in the oncology field is required. Many lipids are included in the class of phospholipids that define cells and organelles through the construction of membrane structures. The complex metabolism of lipids represents a highly controlled cellular signalling network that is essential for mounting an effective innate immune response.

Phospholipids in innate cells are subject to active regulation by enzymes, whose actions are highly responsive to activation status. Along with their metabolic products, phospholipids regulate a cascade of aspects of innate immune cell biology, including, for instance, shape changes, cell aggregation, blood clotting, and degranulation.

Furthermore, through the hydrolysis process, phospholipids provide substrates for cell-cell communication, enable regulation of haemostasis, immunity response, thrombosis, vascular inflammation (present in cardiovascular disease), and associated comorbidities. Phospholipids themselves are also recognised by innate-like T cells, which are considered essential for recognition of infection or cancer, as well as self-antigens (O'Donnell V.B., et al., 2018).

With few exceptions, cellular membranes, including plasma membranes and internal membranes, are made of glycerophospholipids, which belong to the class of phospholipids. The composition of these molecules is glycerol with a phosphate group and, finally, two fatty acid chains. Glycerol is a 3-carbon molecule that acts as the backbone of membrane lipids. Inside a single glycerophospholipid, fatty acids are bonded to the 1st and 2nd carbons. The phosphate, instead, bonds to the 3rd carbon of the glycerol backbone. Changeable head groups are bonded to the phosphate. 3D models of these molecules reveal their cylindrical shape. This geometry allows glycerophospholipids to line up, side-by-side, to form broadsheets. Glycerophospholipids are the

most copious lipids in the cell membranes. They are hydrophobic and are insoluble in water, but thanks to their unique geometry, they aggregate into bilayers without any energy usage (no ATP required). This process occurs due to their amphoteric nature, with hydrophilic phosphate heads and hydrophobic hydrocarbon tails of fatty acids. In water, these molecules spontaneously align with their heads facing outward and their tails lining up in the bilayer's interior. Consequently, the hydrophilic heads of the glycerophospholipids in a cell's plasma membrane face both the water-based cytoplasm and the exterior of the cell. All the lipids together account for about half the mass of cell membranes.

Cell membranes serve as semi-permeable barriers and guardians allowing some molecules to spread across the lipid bilayer (with or without the investment of energy), but no others. Small hydrophobic molecules, carbon dioxide, and oxygen diffuse across membranes rapidly. Small polar molecules, such as ethanol and water, can also pass membranes, but slowly and with other mechanisms. Other molecules, instead, such as ions and/or large molecules (sugars and amino acids), face limited diffusion across cells membranes. However, those specific molecules can still pass thought cell membranes but only with the use of specific transport proteins (incorporated in the membrane). The cell's membrane has many of these transport proteins, which are highly specific and selective for the molecules they move. When the membrane transport is not immediate (like for water molecules for instance) it often uses energy to enable passage, especially because the transport of their specific molecules usually happens against the concentration gradient, which requires additional ATP usage. The exchange of substances through cells membranes is vital to cell health and maintenance. In fact, it is the membrane barriers and transport proteins that allow the cell to accumulate nutrients in higher concentrations than exist in the environment and, on the other hand, help the cell to dispose waste products (Education, Nature, 2014).

Glycerophospholipids with sphingolipids have been widely studied due to their pivotal roles not just as central components of cell membrane but also as signalling and regulatory molecules (Chaurio R.A., et al., 2009) (Hannun Y., et al., 2008) (Quehenberger O., et al., 2011) (van Meer G., et al., 2008). On top of this, their structures are incredibly complex, making a complete profile a highly challenging task (Fahy E., et al., 2009).

Many diseases that affect the human body, for instance metabolic, immune, central nervous system disorders, and cancer, relate to an alteration in lipid metabolic enzymes and their pathways. This finding highlights the fundamental role played by lipids in maintaining membrane homeostasis and normal function in healthy cells (Belo-ribi-Djefaflia S., et al., 2016).

Considering the relevance of lipids in the normal and abnormal cell's functionality they are becoming subjects of numerous studies in the oncology field. For example, an intense lipid (and cholesterol) concentration have been shown to be connected to highly proliferative cancer cells, which display either increased the uptake of exogenous (or dietary) lipids and lipoproteins or over-activated endogenous synthesis (lipogenesis and cholesterol synthesis). Excessive quantities of lipids and cholesterol in cancer cells are stored in lipid droplets (LDs). Therefore, high LDs could be treated as a specific sign for the detection of the tumour. Many studies have been conducted in the field showing how lipids are nowadays considered as hallmarks of cancer aggressiveness (Qiu B., et al., 2015) (Accioly M.T., et al., 2008) (Yue S., et al., 2014) (Guillaumond F., et al., 2015) (Bozza P.T., et al., 2010) (de Gonzalo-Calvo D., et al., 2015) (Abramczyk H., et al., 2015).

Other examples of the key role possessed by lipids in oncology are represented here. An increased concentration of PLs (saturated phospholipids) strongly alters signal transduction and protects tumour cells from oxidative damage, like lipid peroxidation, and potentially inhibits the uptake of chemotherapeutic drugs (Rysman E., et al., 2010) (Staubach S., et al., 2011).

In addition to their normal structural roles, lipids manage the cells signal transduction cascades. During this process, lipids can be broken down to generate bioactive lipid mediators that regulate several carcinogenic processes, including cell growth, cell migration and metastasis formation (Kunkel G.T., et al., 2013) (Wang D., et al., 2010) (Nakanishi M., et al., 2013).

The increased PG levels are also found in renal cell and hepatocellular carcinomas (Perry R.H., et al., 2013) (Shroff E.H., et al., 2015). PGs serve as precursors of cardiolipin, which is found almost exclusively in mitochondrial membranes, and it is actively implied in the maintaining of the mitochondrial functionality and membrane integrity. An increase in prostaglandin concentrations in tumour cells not only promotes tumour growth but also coordinates the complex dialogue that occurs between tumour cells and the surrounding stromal area. This connection evades the immune system attack by promoting immunosuppression.

Considering the vital role covered by lipids in the carcinogenesis process, targeting lipids and cholesterol has been the subject of numerous studies. For instance, inhibitor agents that act directly against lipogenic enzymes (FASN, ACLY and ACC) resulted in an effective anti-cancer therapy (Flavin R., et al., 2010) (Hatzivassiliou G., 2005) (Li S., 2013). However, high adverse side effects of FASN-targeting drugs have precluded their clinical development (Bovenga F., et al., 2015). Recently, in a study conducted by Flaveny *et al*. (Flaveny C.A., et al., 2015), avoidance of the toxic side effect of a FASN-targeting drug on colon cancer, if achieved, holds significant promise for cancer therapy.

The development of an anti-tumour therapeutic targeting either the lipid messengers or their carriers between stromal and tumour cells was reported. It consists of the use of the COX-2 enzyme inhibitor (Celecoxib) in order to disrupt PGs synthesis. It revealed a strong antitumoral and antimetastatic effect in various preclinical models (Wang D, 2010) (Xu L., et al., 2014).

Lipidomic studies (preclinical cancer models and clinical trials) provide evidence that supports the crucial role of lipids in tumour growth and metastatic dissemination. Therefore, novel anti-cancer treatments could focus on the disruption of lipid metabolic pathways to achieve an unbalance lipid homeostasis inducing tumour regression and inhibition of metastatic spread, both leading to cancer cell death.

Continued endeavours to identify all the key actors within these different processes may offer novel metabolic targets for cancer treatment. These clinical strategies, based on the tumour dependence towards lipids, may hold promise to treat the most intractable cancers (e.g., pancreatic and lung cancers) which are predicted to become the two deadliest cancers in horizon 2030 (Matrisian L.M., et al., 2016).

As mentioned in the chapter 1, metastasis is the most malignant stage of cancer. Lipid metabolic abnormalities are now increasingly recognised as characteristics of metastatic cells. The accumulation of specific lipid species, such as signalling lipids, may be a causal factor of malignant tumour progression and metastatic behaviour (Luo X., et al., 2018). Accumulation of signalling lipids, including eicosanoids, phosphoinositides, sphingolipids, and fatty acids, alters the cellular biochemical foundation and might be a causal factor of tumour malignant progression and metastasis. Sphingolipid metabolites, such as ceramide and sphingosine, act as important modulators of cell survival, angiogenesis, migration and metastasis (Patmanathan S.N., et al., 2016) (Milstien S., et al., 2006) (Presa N., et al., 2016) (Morad S.A., et al., 2016).

Further elucidation of molecular mechanisms between lipid metabolism and cancer and novel diagnostic procedures like those described in this work are essential in detecting novel biomarkers and therapeutic targets of human cancers. However, a total understanding of the

mechanisms that regulate the dynamic of lipid metabolism in oncogenesis and their role in the conferment the aggressive properties of malignant cancers remain obscure (Luo X, et al., 2017).

4.11 IMPORTANCE OF USING A 3D SPECIMEN FOR THE ANALYSIS

A fundamental feature of this work, building on the work of (Inglese P, et al., 2017), is the construction and analysis of a 3-dimensional block of tissue made with the combination and co-registration of H&E with MSI.

Logically, the third dimension in an additional constraint that allows the creation of a 3-dimensional tissue model that can closely mimic the *original* structure of the tumour (under analysis) and, therefore, generates a more reliable interpretation of mechanisms that occur in the disease. However, these hypotheses need to be corroborated with further analysis, in fact, in the next chapter, we will show more quantitatively the importance of using a 3-dimensional specimen in performing cluster analysis compared to the classical 2-dimensional approach.

This is vitally important as many recent publications are based simply on single surfaces or slices of a sample. The basic question therefore is, are results reproducible and reliable when only one slice is used, and if not, how much additional data (slices) are required?

In the following chapter, we investigate the importance of 3D versus 2D sampling of tumour tissue using the workflows reported earlier to determine if 2D is good enough, or 3D is significantly belter in determining tumour morphology and chemical composition.

4.12 CONCLUSIONS

While gathering information for this chapter, analysing previous work, and implementing the described DL workflow, it became clear that 3D-MSI represents a powerful and innovative

tool to investigate the bio-chemical components and interactions occurring in tumour tissues. The workflow has demonstrated the potential of 3D imaging analysis by deep learning and clustering.

This work, built on previous work (Inglese P, et al., 2017), has shown (through the application of an improved algorithm, better tuning and detailed characterisation of the lipids found) reliable and robust results that led to a detailed picture of the metastatic liver tumour microenvironment (different tumour subgroups of cells characterised by similar mass spectrometry profiles), information which is inaccessible by classical visual inspection of H&E images.

The main component of the analysis workflow is parametric t-SNE. This could efficiently map a high dimensional manifold into a 2-dimensional latent space showing the existence of clusters which were not otherwise visible in high dimensional dataset or through the application of other (linear and non-linear) DR methodologies.

Using the clustering algorithms, OPTICS and DBSCAN, it was possible to identify three clusters. The results were confirmed, at first, visually, and with the application of co-expression network analysis later. The latest revealed the primary ions associated with the sub-regions of the tumour.

For the first time the combination of two sources of data, MSI and H&E, plus the application of this wholly unsupervised and data-driven workflow, allowed us to obtained insights about the biological and chemical interactions (lipidomics) occurring in metastatic liver cancer, and we were also able to distinguish the chemical properties of the tumour sub-regions. Indeed, the study of the most correlated ions in each identified cluster (by co-expression network analysis) helped to obtain a smart biological interpretation of the disease and, eventually, the use of these ions as biomarkers.

The identified biochemical picture of the metastatic liver cancer includes:

- PGs which are associated with rapidly proliferating human cancer.
- Presence of very long acyl chains that excludes bacterial origin and indicates peroxisomal disorders (a group of genetically heterogeneous metabolic diseases that share dysfunction of cellular organelles).
- Presence of ceramides that indicates necrosis or apoptosis activities (degradation of sphingolipids in the necrotic cells).

This innovative workflow could be used to assist histopathologist in the analysis of massive datasets and help them in the identification of specific regions that require human intervention for more detailed characterisation.

investigate if a 3-dimensional tissue model was more robust compared with 2-dimensional MSI tissue model for the identification and characterisation of tumour subgroups. The third dimension introduces topological constraints that, combined with the fact that biochemical interactions are local, can be used to perform a more reliable analysis. This result has crucial importance.

Another conclusion to take in consideration is the high computational power and time that a non-linear technique usually requires to generate a model. Previously, we have stated that this novel DL workflow could be used to help histopathologist during their work, therefore, optimisation and speed up of the entire process is recommended. For example, 4 hours of CPU time (single Intel i7 processor) was necessary to fit the parametric t-SNE model to the training set. Nowadays, the advent of powerful hardware accelerators, like GPUs, could boost the performance of the entire DL workflow.

5 MODEL OPTIMISATION AND ROBUSTENESS ASSESSMENT

5.1 INTRODUCTION

The previous chapter showed how pre-processing, unsupervised dimensionality reduction, and clustering are of critical importance to retrieve the information embedded in complex 3D-MSI datasets. The result is a list of ions that are 'significantly' more concentrated in certain regions than others. These ions are of key importance to interpret the mechanisms that regulate the cancer biology and the interaction of tumours with the surrounding healthy tissues.

However, even though the diagnostic methodology described in chapter 4 is promising, the lack of established knowledge about the precise composition of the analysed samples makes this type of research even more difficult. Cancer is a clear example of an incredibly complex biochemical system, and analysis using an artificial intelligence approach would greatly benefit from an accurate ground truth. Cancerous tissue can be very heterogeneous, showing different phenotypic properties within and between patients. The combination of MSI techniques with unsupervised learning analysis may be an important approach in determining regions of '*metabolic homogeneity*'.

In this chapter, we will show how the creation and selection of a robust tissue model can lead the DL analysis to a better and richer interpretation of the underlying biochemistry in the tissue sample. In particular, we will investigate in more detail whether the adoption of a 3-

dimensional approach can provide a more robust analysis outcome compared to the more common 2-dimensional analysis.

The analysis of the similarity, using a Rand Index, between the pixel spectra (clustering) of one hundred and twenty-three (123) tissue models (3-dimensional and 2-dimensional) is the main contribution of the work presented here.

The underlying assumption is, therefore, that highly co-localised ions are an expression of the local biochemical interactions, which are occurring in a 3-dimensional space. We investigate the advantage of using a 3D tissue model, which allows the study of those interactions within a volumetric distribution, to determine how important the 3rd dimension (and additional data) is to the conclusions.

By providing the overall spatial distribution of the groups of highly co-localised ions by using a third dimension in the construction of the tissue models may lead to improved interpretation of the real dynamics that occur in the tumour.

Another important purpose of the analysis described in this chapter is to evaluate the quantity of data needed in order to properly apply the DL workflow introduced earlier (chapter 4). Specifically, we will evaluate if it is advantageous to use 3-dimensional models constructed using several tissue slices (more detailed tissue sample composed with a high quantity of voxels leading to heavy computation) compared to 'smaller' 3D tissue models (where the number of tissue slices is reduced and therefore the gap between subsequent slices increased). Finally, as a last step in the analysis, we will compare the information content of a full (all the available tissue slices) 3-dimensional approach to the classical 2-dimensional image analysis (single tissue slice) in order to provide robust answers to same crucial questions: is the 2D approach reproducible and robust? How many slices are required to obtain a reliable tissue model? Do the results change qualitatively

or quantitatively? (e.g. clustering, the ranking of the importance of different ions, assignment of tissue sub-types).

5.2 BACKGROUND

As we seen in the previous chapter, the DL workflow has been designed and utilised for the analysis of a 3-D tissue model built with 51 slices (stacked on top of each other) of metastatic liver cancer. Here, to evaluate the robustness of the models, two tests have been performed:

1. Reduction of the dataset by missing out slices to evaluate how much data is needed for reliable results
2. Pixel shuffling

Clearly, as mentioned earlier, a reduced number of slices would be expected to reduce the informational content provided by a more comprehensive 3-dimensional tissue model. Considering, that most literature analyses are of only single tissue slices or even a few pixels per image, we wish to assess how much of a reduction in data content is allowed without compromising the integrity of the results obtained. We will present a systematic analysis of 123 MSI datasets, built with a progressively reduced number of slices, using the DL workflow introduced in chapter 4, and we will compare the results obtained by the calculation of the *similarity (Rand) index*.

5.2.1 SIMILARITY INDEX

Nowadays, there are many mathematical techniques which could be adopted for cluster analysis or for the comparison of results. Examples are the Adjusted Rand Index (ARI), Meila's Variation of Information (VI), Silhouette Index (SI), Dunn Index, Connectivity, Within Cluster

Sum of Squares, Average Distance Within Clusters, Average Distance Between Clusters, Average Cluster Stability, Average Proportion of Non-Overlap (APN), Average Distance (AD), Average Distance Between Means (ADM), Figure of Merit (FOM).

In this work, we have adopted the Rand Index, which is the most straightforward algorithm to implement for the comparison of the different clustering algorithms.

The *Rand index* (Rand W. M., et al., 1971) also called the Rand measure, is an estimate of the similarity between two data clusters.

Given a set of n elements $S = \{o_1, \dots, o_n\}$ and two partitions of S to compare, $X = \{X_1, \dots, X_r\}$, a partition of S into r subsets, and $Y = \{Y_1, \dots, Y_s\}$, }, a partition of S into s subsets, we define the following:

- a, the number of pairs of elements in S that are in the same subset in X and in the same subset in Y
- b, the number of pairs of elements in S that are in the different subsets in X and in different subsets in Y
- c, the number of pairs of elements in S that are in the same subset in X and in different subsets in Y
- d, the number of pairs of elements in S that are in different subsets in X and in the same subset in Y

The Rand index, R, is then defined as:

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}} \tag{5.1}$$

$\binom{n}{2}$ is calculated as n(n-1)/2.

Intuitively, $a + b$ can be considered as the number of agreements between X and Y and $c + d$ as the number of disagreements between X and Y.

Seeing that the denominator represents the total number of pairs, the Rand index consists in the frequency of occurrence of agreements over the total pairs, or it can be considered as the probability that X and Y will agree on a randomly chosen pair.

Similarly, we can also view the Rand index as a measure of the percentage of correct decisions made by the algorithm. It can be calculated using the formula:

$$RI = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.2}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

The value of the Rand index is between 0 and 1, where a value of 0 indicates that the two data clusters do not agree on any pair of points and a value of 1 indicates that the data clusters are perfectly the same.

In mathematical terms, a, b, c, d is defined as follows:

- $a = |S^*|$, where $S^* = \{(o_i, o_j) \mid o_i, o_j \in X_k, o_i, o_j \in Y_l\}$
- $b = |S^*|$, where $S^* = \{(o_i, o_j) \mid o_i \in X_{k1}, o_j \in X_{k2}, o_i \in Y_{l1}, o_j \in Y_{l2}\}$
- $c = |S^*|$, where $S^* = \{(o_i, o_j) \mid o_i, o_j \in X_k, o_i \in Y_{l1}, o_j \in Y_{l2}\}$
- $d = |S^*|$, where $S^* = \{(o_i, o_j) \mid o_i \in X_{k1}, o_j \in X_{k2}, o_i, o_j \in Y_l\}$

for some $1 \leq i,\ j \leq n,\ i \neq j,\ 1 \leq k, k_1, k_2 \leq r,\ k_1 \neq k_2,\ 1 \leq l, l_1, l_2 \leq s,\ l_1 \neq l_2$. (Wikipedia contributors, 2022).

5.3 MODELS' PREPARATION, ANALYSIS, AND SIMILARITY INDEX CALCULATION

In chapter 4, we performed the analysis of a 3D-MSI tissue model built with 51 slices of human metastatic liver cancer. Here, we are creating one hundred and twenty-three (123) new models that we will describe as *reduced models.* The term '*reduced'* has been adopted because we build these 123 tissue models using a subset (reduced number) of slices of the original tissue model (51 slices). Once the reduced models are created, we deploy the DL workflow (Chapter 4) to analyse them. As a final step, the information content of the clustering obtained from the analysis of every single *reduced* model is then compared with the information content of the clustering described in chapter 4.

The 123 reduced models are organised in the following manner:

- 30 models made with 26 slices of the original 51
- 30 models made with 13 slices of the original 51
- 30 models made with 6 slices of the original 51
- 30 models made with 3 slices of the original 51
- 3 models made with just the 1st, 25th, 51st slice of the original 51 slices model.

Every batch of models has been created following the same procedure. For instance, in order to build one of the 30 reduced models with 26 slices, the procedure applied was:

- As a first step, we have built three groups of models. All the models that belong to the first group kept constant the 1st slice of the full model with 51 slices. Instead, all the models that belong to the second group kept constant the 25th slice of the full model with 51 slices. Finally, all the models that belong to the third group kept constant the last slice of the full model with 51 slices, the 51st. In this way, these initial three batches of reduced models are representing, in general, the entire depth of the biopsy.

- As a second step, in the first group of models, for instance, the 1st slice was kept constant, and the following 25 slices, necessary to build 26 slices of the reduced model, were randomly selected among the other fifty (51 slices minus the 1st) in the original 51 slices model.
- In the second group of models, we kept constant the 25th slice, hence the slice in the middle of the original 51 slices tissue sample. To balance the distribution of the remaining 25 slices (necessary to build a reduced model with 26 slices) across all the tissue we have split the whole original dataset (51 slices) into two subsets. Twelve slices were randomly selected from the first subset (from the 1st slice to the 24th), and subsequently, thirteen slices were randomly picked from the second subset of slices (from the 26th to the 51st).
- In the third group of models, the 51st slice of the original dataset was kept constant, and the previous 25 slices (from the 1st to the 50th) were randomly selected among the other previous fifty (51 slices minus the 51st).

This process was then iterated over 10 times for each group of models, generating 30 models made with 26 slices each.

The procedure described was then utilised to build reduced tissue models with a progressively smaller number of slices (13 slices, 6 slices, and 3 slices).

The remaining models are made with the 1st, then 25th, and the 51st slices (single image models or 2-dimensional models). For a total of 123 2D and 3D tissue models.

In order to maintain the third dimension of the *reduced* model, although most of the slices were picked randomly, the positions of those selected slices (compared with the original positions in the space determined by the whole 51 slices dataset), were preserved.

A schematic representation of the methodology adopted to build the 123 reduced models is reported in Figure 5.1, while the workflow presented in this chapter is reported in Figure 5.2.

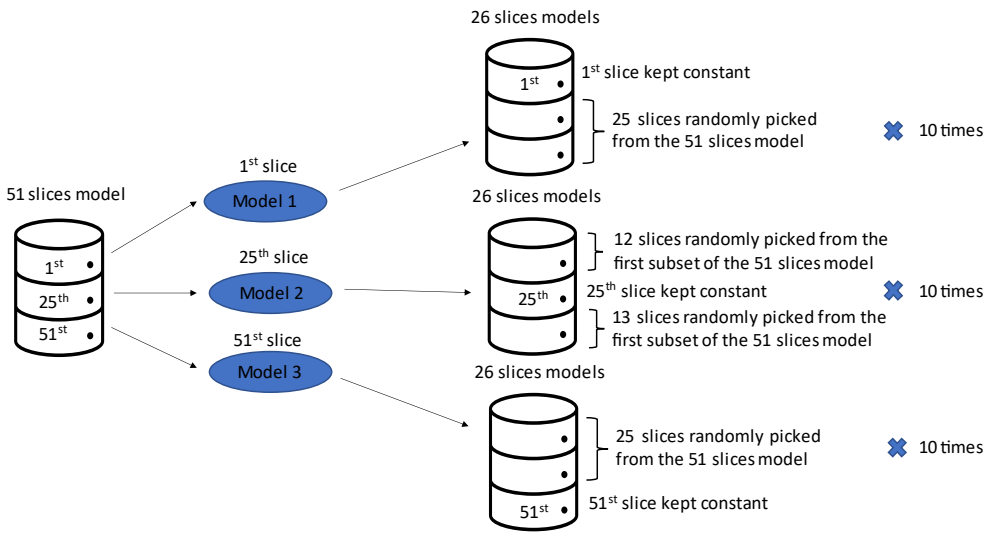


Figure 5.1 - Schematic representation of the methodology adopted to build the 123 reduced 3D tissue models. Here we describe the construction of the reduced 26-slice model.

When all 123 reduced models were finally ready, the DL workflow (chapter 4) was applied to every single model. The application of parametric t-SNE to the reduced models required some adjustment to the architecture. The lower dimensionality of the 123 3D-MSI tumour datasets (due to the smaller amount of tissue slices used to build the model) requires a less deep ANN. Therefore, in order to avoid overfitting, the architecture of the parametric t-SNE method was reviewed and modified accordingly to take into account the volume of the dataset considered.

Finally, the information content of the clustering results in the slices 1st, 25th, and 51st of these 123 reduced models (3 clusters per model), was compared with the information contained in the corresponding three clusters of the same slices (1st, 25th, and 51st) of the original 3D-DESI-MSI model made with 51 slices. For comparison, *similarity indices* (Rand Index) were adopted.

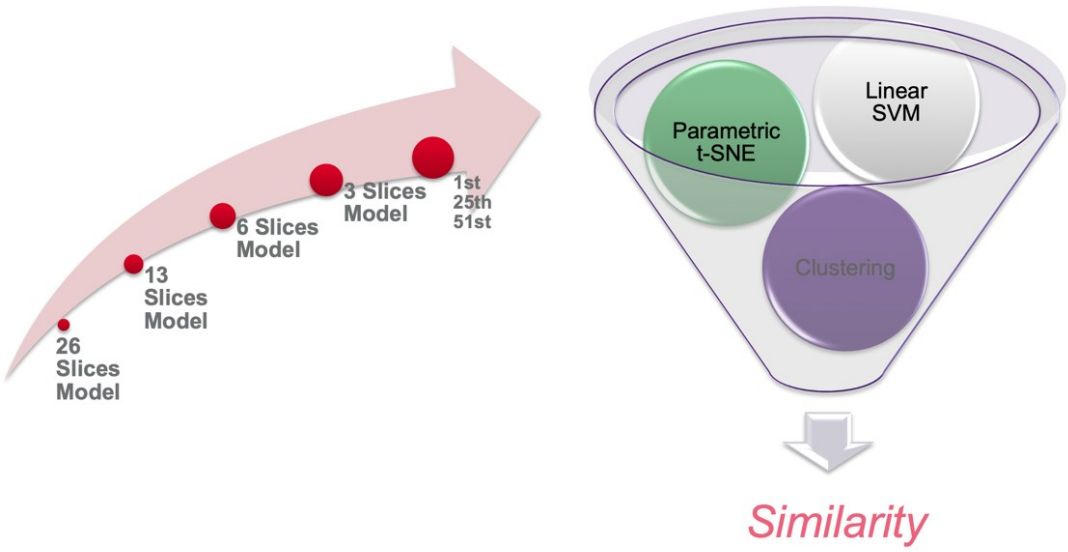


Figure 5.2 - Schematic representation of the operational workflow. The 123 reduced models are input to the DL workflow (chapter 4). The three main steps are highlighted: the supervised learning process performed with a Linear SVM classifier; the DR performed with the parametric t-SNE; the clustering procedure (OPTICS and DBSCAN). Finally, a Rand Index calculation is used to assess the similarity of the results (informative content of the clusters).

5.3.1 SIMILARITY ANALYSIS

As seen before, the output of the DL workflow described in chapter 4 is a segmentation (clustering) of the 2D latent space. Here, to perform a consistent comparison between the output of the analysis performed on a full 51 slices model and the 123 reduced models just created, the DL workflow output has been *'forced'* (setting the OPTICS and DBSCAN parameters accordingly) to generate 3 clusters on any latent space of a reduced models. On top of this, the labelling (number) associated to any clusters has been assigned consistently to the geographical area they are representing. This was performed in addition to the slices that have been kept constant in any single reduced models (the 1st, the 25th, and the 51st slices). A projection of these clusters on those images helped in the identification of a correspondence between clusters.

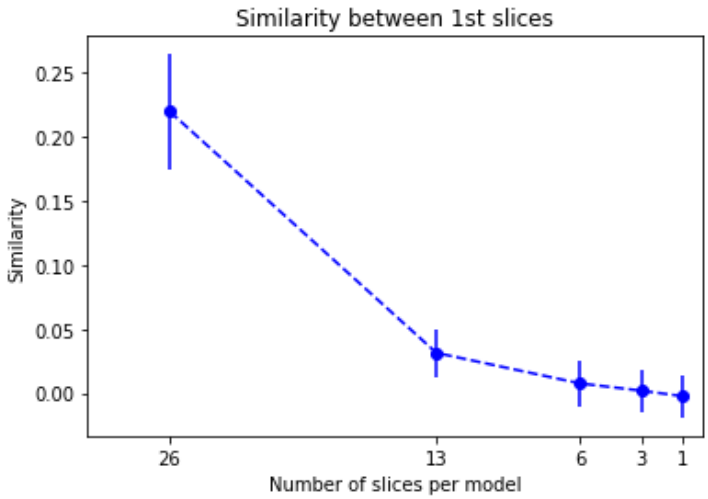
Once a correspondence between clusters that belong to the segmentation of the full 51 slices model and all the clusters obtained from the analysis of the 123 reduced model has been

established (on the same 1st, 25th, and 51st images), we applied the Rand Index to obtain an evaluation of the similarity (of the informative content) between them.

The three scatter plots reported in Figure 5.3 show how the similarity between the three clusters decreases as a function of the number of slices taken into consideration to design the model. From inspection of the graphs, the curves show how the similarity rapidly drops when we compare the informative content of the clustering obtained from the analysis of the full 51 slices model to the clustering of the reduced model built with 26 slices. This first comparison shows average values of 22% similarity when we take into consideration the 1st slice of both models, an average of 33% for the 25th slice, and an average of 28% for the 51st slice.

A subsequent drop can be seen when the model is further reduced from 26 slices to 13 slices. In this case the comparison produces a similarity index (average) of 4% when we take into consideration the 1st slice of both models, an average of 3% for the 25th slice, and an average of 1.5% for the 51st slice.

The similarity index is diminishing toward zero when the number of slices that compose the reduced models reaches a single tissue slice (single image analysis), which represents a 2-dimensional model.



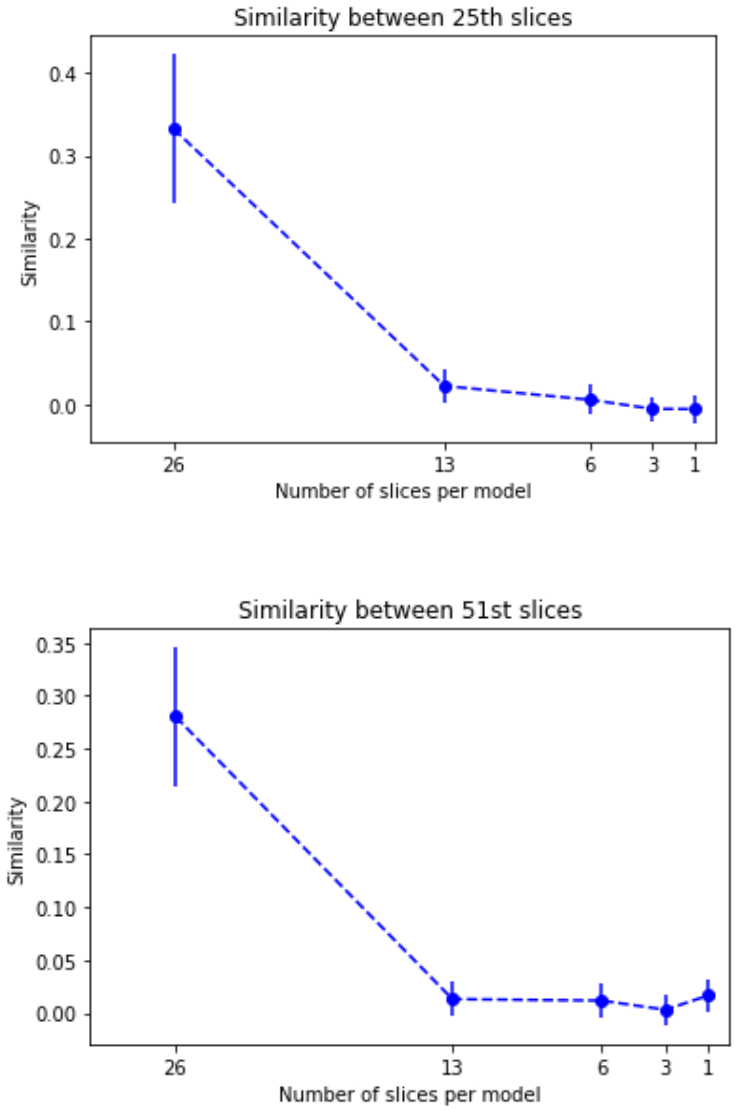


Figure 5.3 - Scatter plots of the Similarity Index (Rand) calculated from the comparison of the informative content of the clustering obtained in the 1st, 25th, and 51st slices with the application of the DL workflow on the 3D-DESI-MSI dataset composed of 51 slices (full model) and the clustering obtained from the analysis of the 123 reduced models in the same slices. The top scatter plot reports the trend of the Rand Index calculated from the comparison of the clustering of the 1st slice. The middle scatter plot reports the trend of the Rand Index calculated from the comparison of the clustering in the 25th (middle) slice. The bottom scatter plot reports the trend of the Rand Index calculated from the comparison of the clustering in the last slice, the 51st.

From the trend of similarity index, it is possible to deduce the importance of having a large amount of data in order to perform a more in-depth tissue analysis and extract more meaningful

information from the MSI dataset. The graphs reported in Figure 5.3, in fact, are representing compelling evidence of the importance of performing a 3-dimensional analysis with a consistent number of tissue slices. Complex tissue models composed of several slices one on another provides a more robust interpretation of the original volumetric distribution of the molecules in the tumours and provides a better picture of its biology. Furthermore, the sequence of slices supports the identification of relevant (highly correlated) molecules, which provides a richer interpretation of the complexity of the analysed sample. This detailed interpretation of the biology can get lost when we simplify the 3-dimensional tissue model by removing slices. This is substantiated by comparison of the similarity index computed between the original model composed with 51-slices and a classical 2-dimensional image (single tissue slice). A more general question may be, when looking at these results, is which slices are truly representative of the biology of the sample? In the absence of a clear decision, it seems that using a combination of slices should better approximate the phenotype of the sample as a whole.

5.4 ROBUSTNESS ASSESSMENT

One of the major concerns that arise from any analysis of a biological system (which is highly complex and dynamic) is that there is the possibility of getting results that look promising but in fact, are not strongly connected with the underlying phenomenon, and may be due e.g. noise or bias in the data. For these reasons extra tests need to be performed in order to validate the link between mathematics (results) and biology.

The 3D tissue model constructed from 51 slices (in chapter 4) due to the correlation between slices and the large amount of data provided represent a stable and robust structure. As was shown in chapter 4, the registration of the images such that the sequence of pixels and their coordinates

are overlapping with similar regions, is of fundamental importance in order to reconstruct the digital picture of those slices (MSI).

The hypothesis here is to assess the stability of the models to the precise matching of the images. One way of investigating the stability of the models is to disrupt the pixel sequence by shuffling them, and subsequently feed the ANN with data that are not in the expected order. With this test, we want to assess the ability of the DL workflow to extract similar information, investigating its capacity and flexibility as a diagnostic procedure (where in the real world, it may often be difficult to precisely align samples). The comparison of the results (clustering) before and after the pixel shuffling would indicate that the results obtained are not influenced by the coordinates and sequence of pixels.

To perform this evaluation, we built five 3D tissue models (51 slices) where the pixels were randomly shuffled by the application of the Numpy shuffling function in Python on the pre-processed MSI dataset. To compare the clustering obtained from the analysis of the shuffled models with the *original clustering* of the 51 slices 3D tissue model, we used the same architecture of the parametric t-SNE and adopted the same parameters for OPTICS and DBSCAN algorithms. As expected, the DL workflow led to an optimal number of three clusters. We maintained the labels provided by the clustering algorithms in order to perform the comparison of the embedded information between the same clusters. Therefore, the comparison between the 15 clusters (three per model) and the 3 clusters of the original dataset have been carried out cluster by cluster with the application of the Similarity Index (RAND). The average inter-cluster similarity is reported in Table 5.1, showing an overall similarity of 76%.

Most of the relevant results have been preserved, demonstrating the robustness of the DL workflow as a diagnostic procedure. This implies that the DL workflow, under these conditions,

can retrieve information that is qualitatively similar to that obtained from the analysis of the structured 3D-DESI-MSI.

Table 5.1 - Average values of the Rand Indices calculated from the comparison of the clustering obtained from the application of the DL workflow (Chapter 4) on five pixels shuffled models

	Cluster 1	**Cluster 2**	**Cluster 3**	**Overall**
Rand Index	78.3 %	78 %	73.3 %	**76.5 %**

5.5 DISCUSSION

In this chapter, we have discussed two sets of tests that have been deployed to evaluate the capacity of the DL workflow to extract relevant information from an MSI dataset when the number of slices used in construction of the model is reduced until there remains a 2-dimensional image, and when the 3D tissue model is based on poorly indexed images with random disruption of the order of the pixels.

The results presented in section 5.3.1 show that the similarity indices obtained from comparisons with the 1st, 25th, and 51st slices are lower, indicating that reducing the number of slices used to build the models removes valuable information and results in the model being unrepresentative of the 3D structure. Classical 2-dimensional image analysis (a single slice model) appears to show instability when compared to models derived from multiple slices and therefore may not always be representative of the overall biology of the sample. This confirmed by the plots of the similarity indices (Figure 5.3).

The second test aims to assess the robustness of the DL workflow when analysing unstructured 3D tissue models by shuffling the pixels in the original dataset. As mentioned earlier in this chapter, the shuffling procedure leads to disruption of the sequence of information and

localisation of ions. The analysis of five 3D tissue models (with the same dimensionality of the original 51-slice 3D-MSI) and the subsequent comparison of the information embedded in the clustering obtained (section 5.4) showed the capacity (76%) of the DL workflow to extract useful information even when the coordinates of the pixels and the order they are actually fed to the neural network is disrupted.

However, pixel shuffling reduces the integrity and robustness of the models obtained from properly indexed 3-D models that have the correct pixel order. It also seems apparent that the DL analysis of the MSI dataset requires several samples (slices) to obtain a robust model. The additional data from many slices is expected to reduce overfitting and aid robust feature extraction and model generation.

5.6 CONCLUSION

In this chapter, we investigated the stability of models that could potentially aid diagnosis. The main conclusions can be summarised as follows:

- Based on the comparison (Rand index) of the clustering results obtained from the analysis of the 3D tissue model built using 51 tissue slices with the clustering obtained in 123 reduced tissue models it appears that the DL workflow (chapter 4) is a flexible and stable procedure for the extraction of diagnostic information embedded in a complex 3D-MSI dataset. The tests highlighted the importance of a 3-dimensional approach. Reducing the number of slices used to build the tissue models impacted negatively on the interpretation of the biology. Figure 5.3 shows how the RAND indices decrease with the reduction of the number of slices in the models. The index tends to zero when we compare the information content of the clustering of the whole MSI dataset (51 slices) with 2-dimensional image

analysis (a single slice model). The result demonstrates the superiority (in terms of the tumour segmentation and biological interpretation) of a 3-dimensional analysis compared to the 2-dimensional analysis.

- Additionally, it is important select sufficient data for the application of DL techniques to imaging. DL methods require a significant amount of data for training purposes, so limiting the amount of available data by reducing the number of slices/the number of pixels (spectra) will reduce the probability of extracting a stable set of important features. This limitation can be improved by the use of parametric t-SNE. Based on the results obtained with the deployment of the second test (pixel shuffling) it appears that the DL workflow can still extract models that have a reasonable similarity index to the full 3D model (76%). This is promising as robust models with limited data may be necessary for use as diagnostics in a real clinical setting.

The hypotheses generated require further set of experiments on larger cohorts of patients. For this reason, the results presented should be considered as preliminary. Nevertheless, the study clearly shows the potential benefits of applying DL on DESI-MSI dataset and the robustness provided by this novel approach.

6 MODEL GENERATION AND ANALYSIS OF EPITHELIAL OVARIAN CANCER DATA IN 3D THROUGH DEEP LEARNING

6.1 INTRODUCTION

In chapters 4 and 5, we described, designed, and tested an unsupervised learning workflow based on the application of DL methodologies for the investigation of the metabolism and biology in different regions of metastatic liver cancer.

In this chapter, we introduce a different MSI pre-processing procedure and refine and reapply the DL workflow to a new 3D-DESI-MSI dataset derived from epithelial ovarian cancer with the objective of extracting relevant information regarding the tumour's microenvironment and, at the same time, show the flexibility of the DL workflow in analysing different sources of MSI data.

At the time of the study, a novel visualisation technique was released, UMAP (Smets T., et al., 2019). Hunce, we wanted to experiment and deploy UMAP on the MSI dataset to assess its potential application as DR technique for clinical research purposes.

Finally, we describe the biochemical mechanisms deduced from the model that characterise ovarian cancer. The longer-term objective is to deduce biomarkers that could indicate targeted treatment strategies to improve clinical outcomes and patient life expectancies.

6.2 UMAP

UMAP is an abbreviation for Uniform Manifold Approximation and Projection, which is a non-linear dimensionality reduction technique. It is visually like t-SNE, but it assumes that the data is uniformly distributed on a locally connected Riemannian manifold (real and differentiable manifold). The metric used is locally constant (or approximately locally constant).

The UMAP method can be adopted for DR and visualisation of MSI data. For this purpose, Smets *et al*. (Smets T., et al., 2019) showed that this DR technique can be beneficial for large (over 100k pixels) MSI datasets. With an almost fourfold decrease in computational overhead, it is more scalable in comparison to the current state-of-the-art methods, such as t-SNE.

A single MSI experiment can lead to gigabytes (Gb) of complex data, in fact, the number of pixels collected in a single experiment, as well as the number of *m/z* bins measured, is ever-increasing with improving technology, which typically can achieve a smaller scale and a greater density of points. In this context, where the complexity and volume of MSI datasets increase, there is a requirement for scalable DR methods to extract the underlying structure of these datasets.

A common issue frequently encountered during the application of DR methods is the selection of the number of features (real or latent) to be retained, each of which contains a part of the total information in the data; therefore, an insufficient number of features results in a loss of information. It is especially relevant where, as in these studies, the objective is to visualise the data in 2 or 3 dimensions. As showed in previous results (chapter 4 and 5) one of the study objectives was the creation of a latent space of two dimensions that could be easily projected on the images in order to visualise the tumour structure. Inevitably, this DR process create a loss of information that can be limited using specific DR methodology.

In chapter 4, we showed the significance of using a non-linear DR technique to extract and preserve the sub-structure of a complex MSI dataset. The non-linearity of the DR method adopted is essential because many biological models are inherently non-linear (Dang T., et al., 2009).

The UMAP algorithm can find an embedding by searching for a low dimensional projection of the data points that has the closest possible equivalent *fuzzy topological structure* (Chang C.L., et al., 1968) (Liu Y.M., et al., 1997). In this work, the UMAP mapping to 2 dimensions was performed using the Python implementation available at https://github.com/lmcinnes/umap.

In order to provide a general description of the UMAP algorithm, we can summarise that this method uses local manifold approximations and assembles their local representations to form a topological representation of the high-dimensional data. The representation of the data in the low-dimensional space is then optimised through the minimisation of the cross-entropy. It is calculated between the two topological representations.

$$C_{UMAP} = \sum_{i \neq j} v_{ij} \log\left(\frac{v_{ij}}{w_{ij}}\right) + \left(1 - v_{ij}\right) \log\left(\frac{1 - v_{ij}}{1 - w_{ij}}\right) \tag{6.1}$$

Here it is useful to recall some resemblance to the Kullback−Leibler divergence (Equation 3.9) used in the classical t-SNE algorithm as a cost function and reported here for convenience:

$$C_{t-SNE} = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) \tag{6.2}$$

Although the similarity in terms of the mathematical expression, the meaning of the two equations is quite different, in fact, it is essential to note that the UMAP does not use the same definitions for v_{ij} and w_{ij}, wherein i and j refer to two objects in the high-dimensional (v_{ij}) and low- dimensional (w_{ij}) space. On the other hand, in the t-SNE cost function, this seeks to minimise

the Kullback−Leibler divergence between the joint probability distribution in the high-dimensional space, p_{ij}, and the joint probability distribution in the low-dimensional space, q_{ij}. The fact that both $p_{ij}$and q_{ij} require calculations over all pairs of points imposes a high computational burden on t-SNE.

Another significant advantage of UMAP is *autocorrelation.* Considering the spatial nature of MSI data, a certain degree of spatial correlation is expected to occur. It means that neighbouring pixels are more likely to be correlated to each other than pixels located at a further distance (Cassese A., et al., 2016). Here, we assume that closely located pixels (to each other) have a more significant spectral similarity or chemical relatedness compared to those pixels that are located further from each other.

6.3 DATA EXTRACTION AND PRE-PROCESSING

The epithelial ovarian cancer tissue samples were obtained with informed consent under local ethical approval (14/EE/0024). The biopsy was snap-frozen in liquid nitrogen and then stored in a freezer at a temperature of -80°C. Subsequently, the specimen was sectioned into parallel slices of 10µm thickness each. One of every tenth section was collected, and their MS was acquired for a total of 10 slices. For this work, we have used nine consecutive slices discarding the first slice acquired due to its low quality (tissue sample folded).

The MSI was acquired utilising a Waters Xevo G2-XS device in the negative ion mode over an *m/z* range of 50 – 1200. The resolution was 100 µm, 10 scan/sec, with a voltage of 4.5 kV. The solvent adopted was a blend of methanol/water 95/5 with a flow rate of 1.5 µL/min. The nitrogen gas pressure was set to 5 bar. Finally, the data were converted to the raw file format.

The tissue samples were organised in 9 layers containing one tissue section each. They were used to perform the H&E (haematoxylin and eosin staining method) histological enhancement (Figure 6.1) and found to mostly consist of epithelial ovarian cancer. Additionally, there is an area of connective tissue. Mrs Kazanc Emine provided the specimen, collected DESI-MSI data and performed staining at Imperial College London.

Data preparation and the subsequent pre-processing workflow applied to this DESI-MSI dataset (that differ to the one applied on the MSI in chapter 4) can be summarised in the following list of computational steps:

- Peak picking and peak detection from the Waters *.raw* files
- Identification of sample-related region-of-interest (ROI) by manual annotation
- Classification of pixels using a supervised learning method
- Mass shift diagnostics and optimal parameters estimation
- Intra-run peak matching
- Testing image quality
- Batch correction (TIC)
- Estimation of blank signal variability
- PCA Images
- Application of the UMAP method

In the following paragraphs a brief description of the steps will be provided.

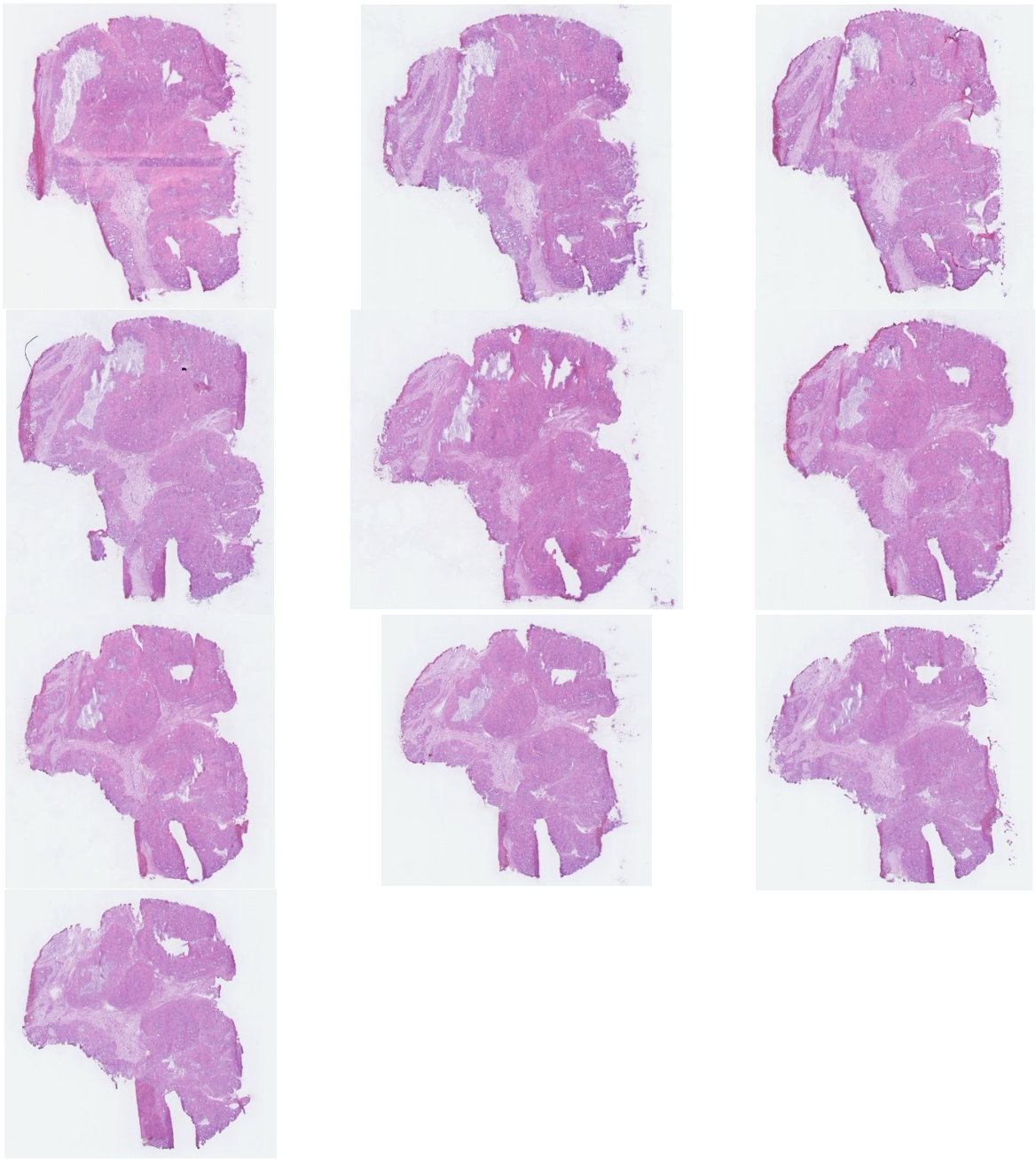


Figure 6.1 - H&E images representing the whole human ovarian cancer tissue samples available for this work.

Peak picking and peak detection phases are based on the same methods that have been used previously in Chapter 4. The raw mass spectral profiles were initially elaborated using the package MALDIquant (Gibb S., et al., 2012). Subsequently, the peak list corresponding to the image pixels

were matched using the command *binPeaks*. No intensity normalisation or variance stabilising transformations were applied. Subsequently, using the computer package SPUTNIK, and the command *msiDataset,* a new MSI dataset was created of the matched peaks. SPUTNIK was also used to remove the non-informative i.e., peak signals with intensities smaller than the noise level were discarded. Also, peaks found in less than 1% of the entire dataset were considered noise and therefore removed.

Peak matching and normalisation were performed to improve the computing time required for the analysis and to preserve only the differences due to biological variability.

The pre-processing workflow begins after the co-registration of the MSI with the H&E images. The 3D tissue model was built by stacking those images one on top of another. The 3D model has voxels of 100 μm^3, and by inspection it reproduces the topology of the biopsy.

Prof Robert D Goldin (Faculty of Medicine, Department of Metabolism, Digestion and Reproduction at Imperial College London) has manually annotated the regions in the optical images associated with tumour, healthy tissue (non-tumour), and background (the area outside the tissue section). Figure 6.2 reports an example of the manual annotation.

The manually annotated pixels are pivotal for the identification and extraction of the ROI (Region of Interest). In order to perform the identification of the whole tissue areas labelled as a tumour, a supervised classifier was trained on the manually annotated spectra. As before (chapter 4), the adopted classifier was the linear SVM (Cortes C., et al., 1995) which performed with very high accuracy (~0.99). The spatial map obtained by the application of the linear SVM across the nine tissue slices was compatible with the morphological properties of the tissue. An example of the linear SVM output is reported in Figure 6.3.

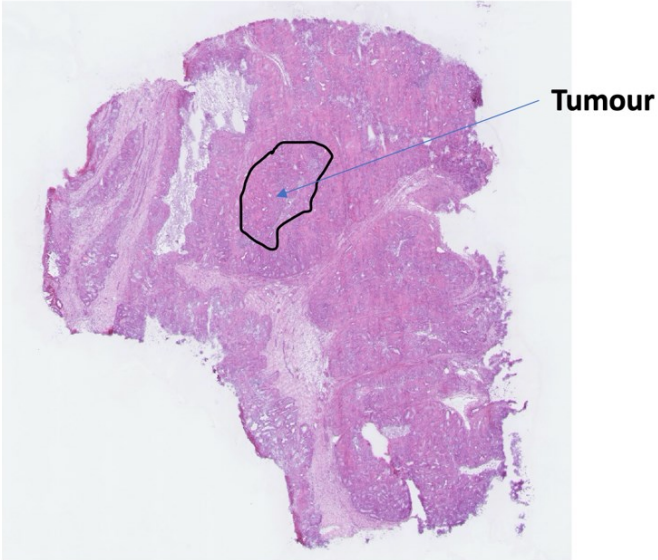


Figure 6.2 - The consultation with the histopathologist led to the identification of the tumour region on the tissue specimen. The tumour spreads evenly in the right-hand side of the sample and it almost entirely composes the sample. The highlighted area represents an example of a manual annotation provided.

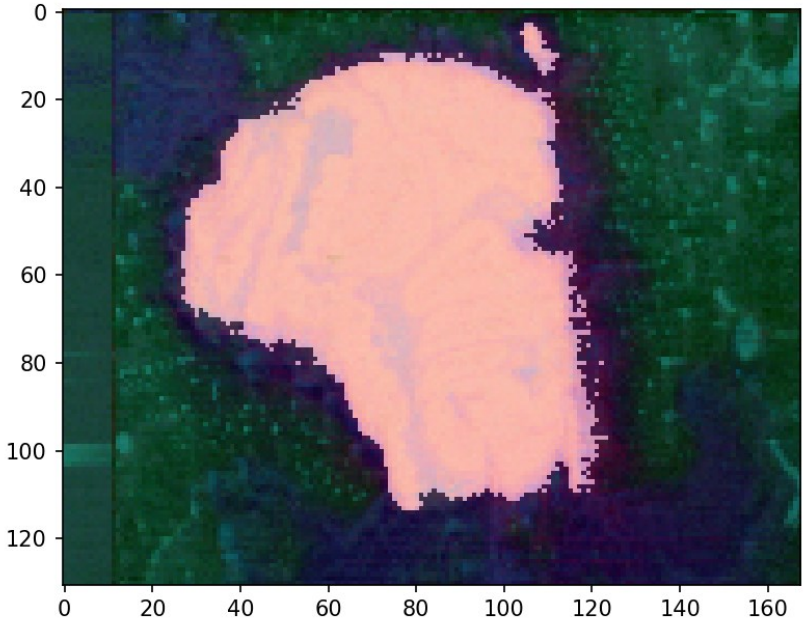


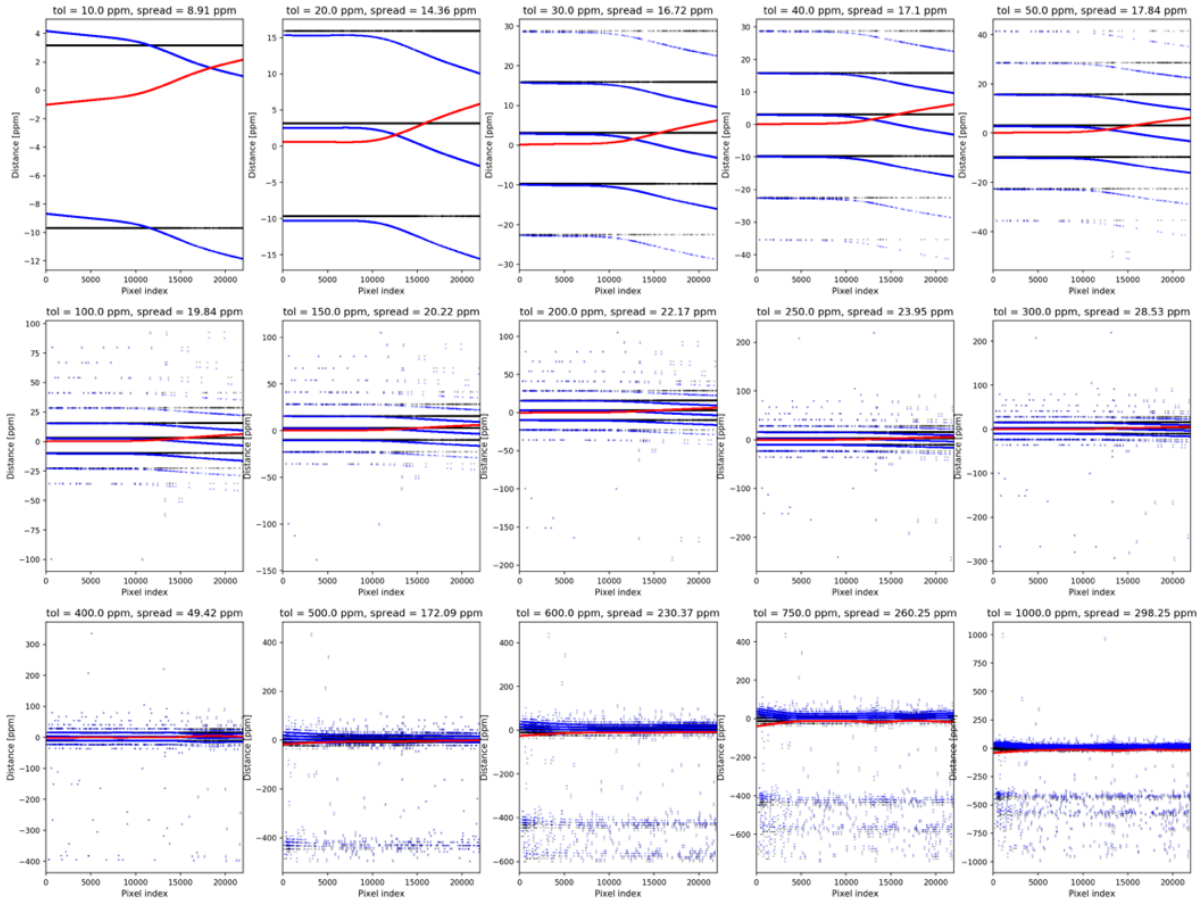
Figure 6.3 - Example of the classification obtained by the application of the linear SVM. The result accurately reflects the histopathology diagnosis. Almost all the pixels have been classified as tumour (light pink).

The mass shift across all pixels is estimated using the theoretical mass of a set of reference ions which are *255.233 m/z* (Palmitic acid), *303.233 m/z* (Eicosatetraenoic acid) and *855.5499 m/z*, which represents an ion (not identified) that is present in the whole tissue sample (Inglese, P. & Strittmatter, 2018).

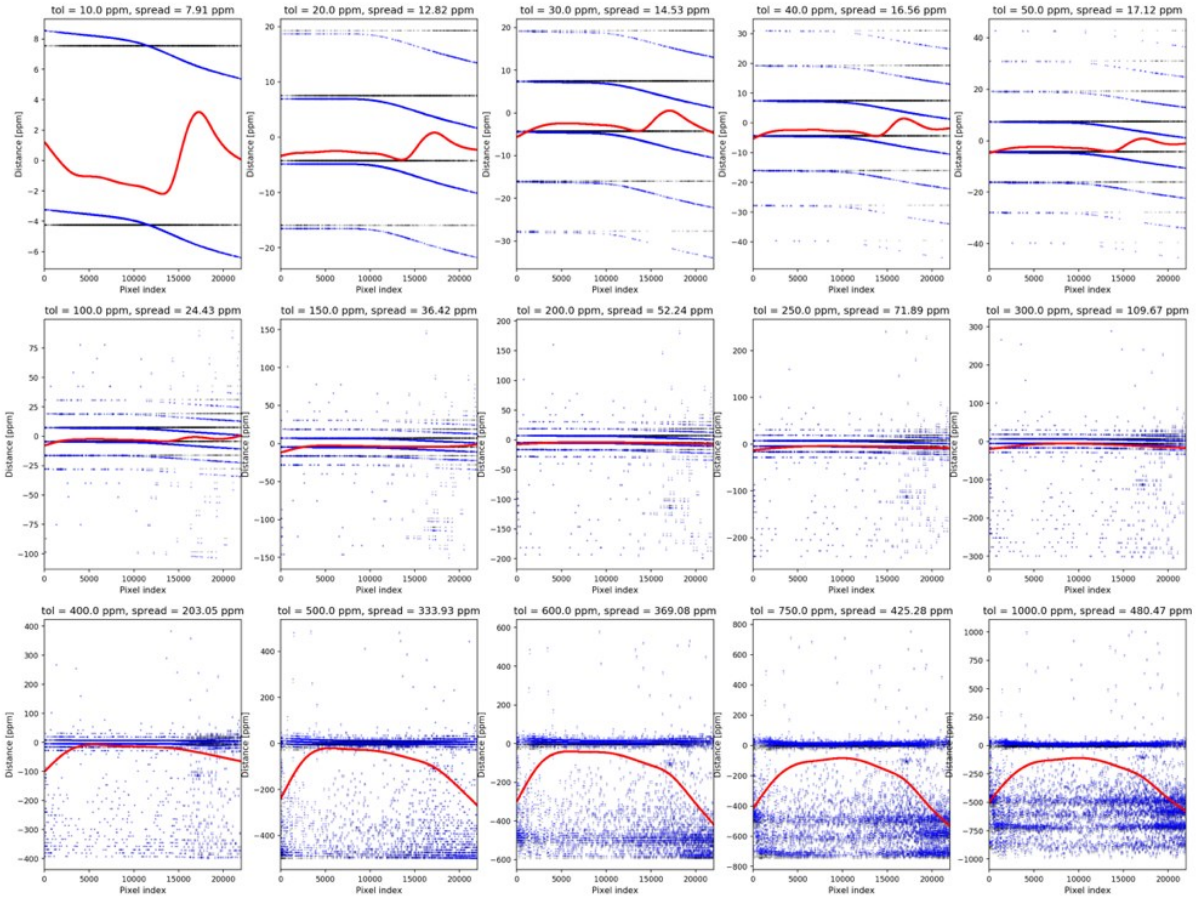
The distances between the matched mass and the theoretical value in ppm have been calculated and reported in figure 6.4. Matched peaks are searched in increasing size windows from 10 ppm until 1000 ppm, around the theoretical mass. Figure 6.4 reports the search window size (*tol*) and the spread of the distances of the matched peaks (*spread*). The latter is estimated from the global trend of the mass shifts and is plotted in red. The values of *tol* and *spread* are essential for the re-calibration and peak matching algorithms. Black dots (figure 6.4) represent raw shifts while red lines represent global fitted trend after a first step re-calibration (which will be used for a second round of re-calibration), finally, blue dots represent corrected shifts after the first step of re-calibration.

The calibration process is a very useful methodology that helps the classification of ions and decreases misleading distributions of these ions across different tissue slices.

255.233 m/z negative



303.233 m/z negative



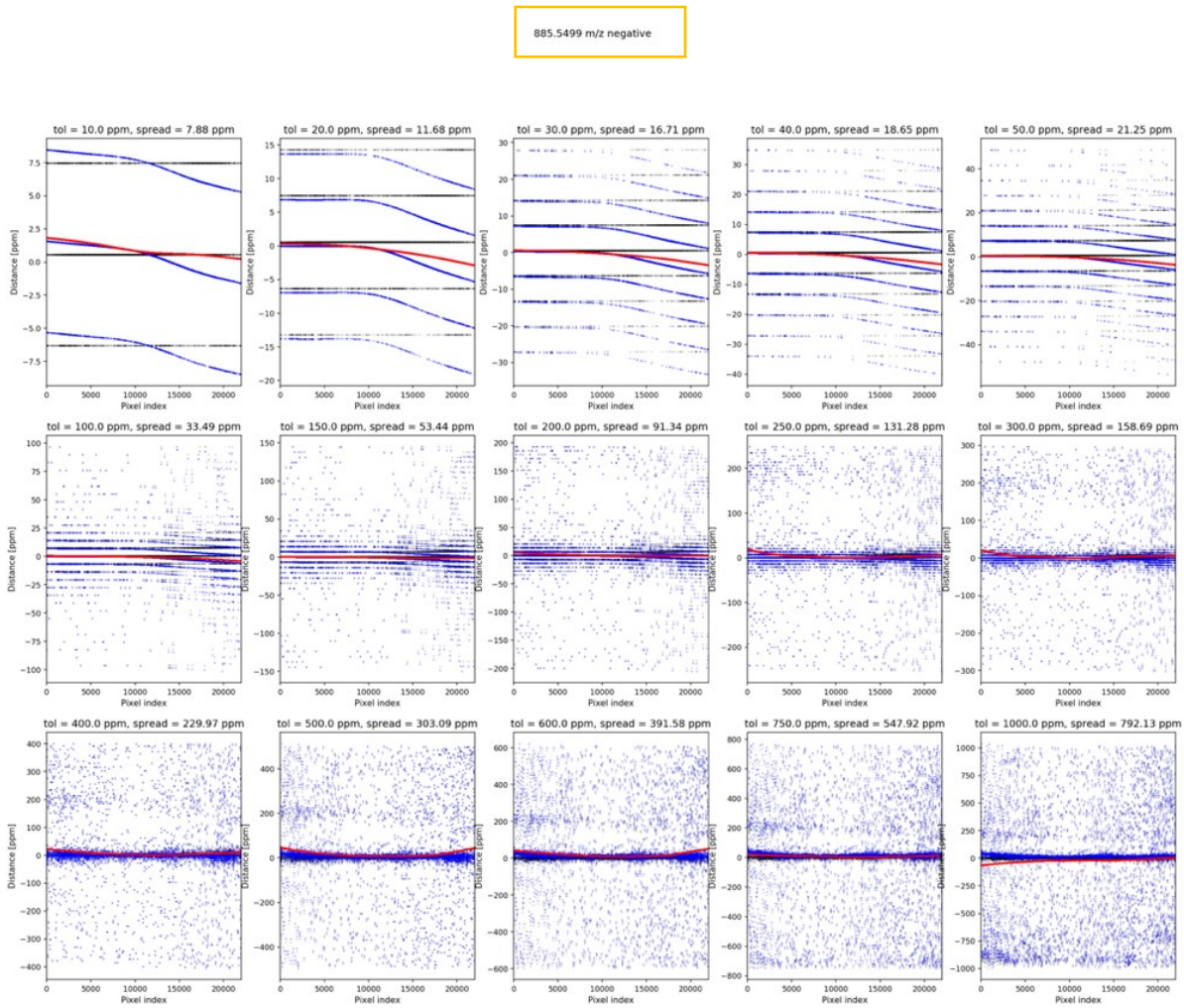


Figure 6.4 - The diagrams above show how the mass shift across all pixels is estimated using the theoretical mass of Palmitic acid (255.233 m/z). Each plot shows the distance between the matched mass and the theoretical value in ppm. Matched peaks are searched in increasing size windows from 10 ppm to 1000 ppm around the theoretical mass. The title of each plot reports the search window size (tol) and the spread of the distances of the matched peaks (spread). The spread is estimated from the global trend of the mass shifts (plotted in red). The middle picture shows the mass shift across all pixels is estimated using the theoretical mass of the Eicosatetraenoic acid (303.233 m/z). The bottom picture shows the mass shift across all pixels is estimated using the theoretical mass of 855.5499 m/z.

Once the pixel classification and a mass shift diagnosis has been deployed, the intensities of the reference ions need to be evaluated bearing in mind that knowing the expected spatial distribution of one ion helps to estimate the correct mass shift. In order to support this analysis in Figure 6.5 are reported a second batch of plots that shows the intensities of the reference ions.

Since the spatial distribution does not change with search windows larger than 100 ppm, we can consider this as a reasonable estimate of the mass shift. In the case of palmitic acid, we expect this percentage to be close to 100%.

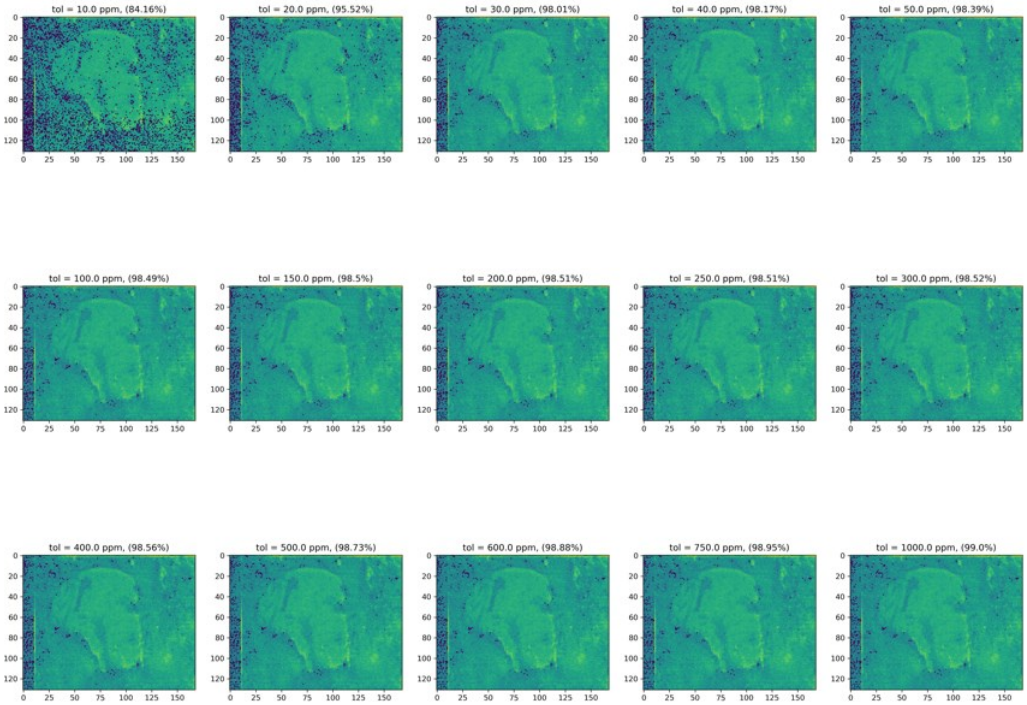
An estimate of the optimal *m/z* shift is presented in Figure 6.6, which shows the percentage of pixels containing the reference peak, for the various tolerance values. The vertical line corresponds to a relative variation of the percentage smaller than 5%. It is assumed that this represents the start of the plateau, which means that with larger tolerances, the corresponding image remains almost identical.

At the end of the pre-processing pipeline, the quality of the matched peaks and peak intensities are tested. A set of references are used to test the mass accuracy. If their mass error is greater than 50 ppm, the peak is considered absent. The formula of the coefficient of variation is:

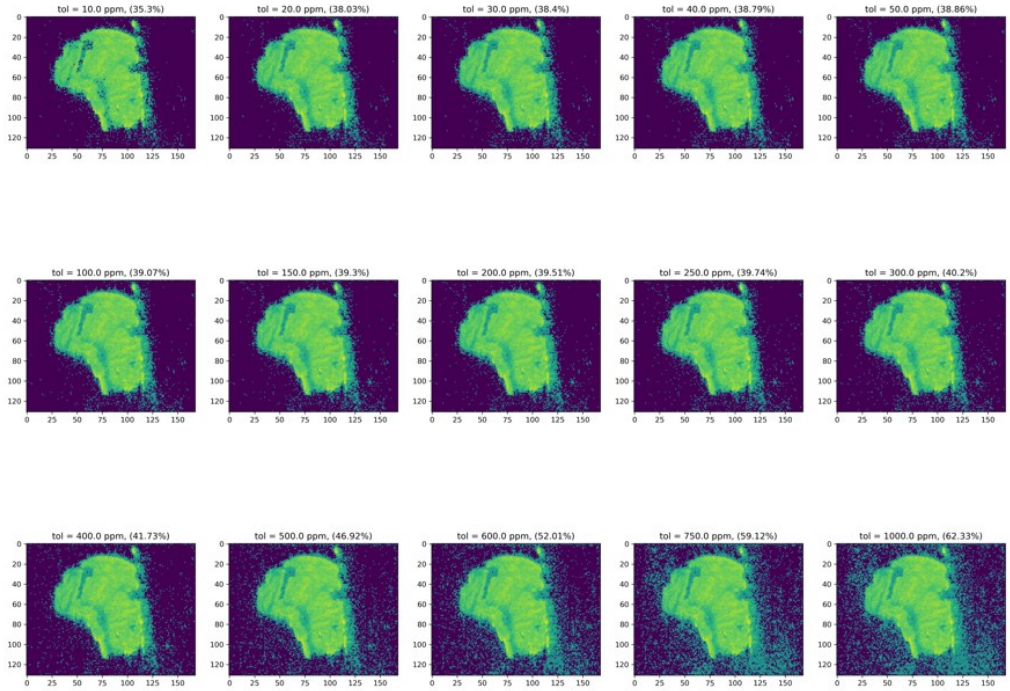
$$CV = \frac{Q_3 - Q_1}{Q_1 + Q_3} \tag{6.1}$$

where Q_1 and Q_3 are the first and third quartiles, respectively.

255.233 m/z negative



303.233 m/z negative



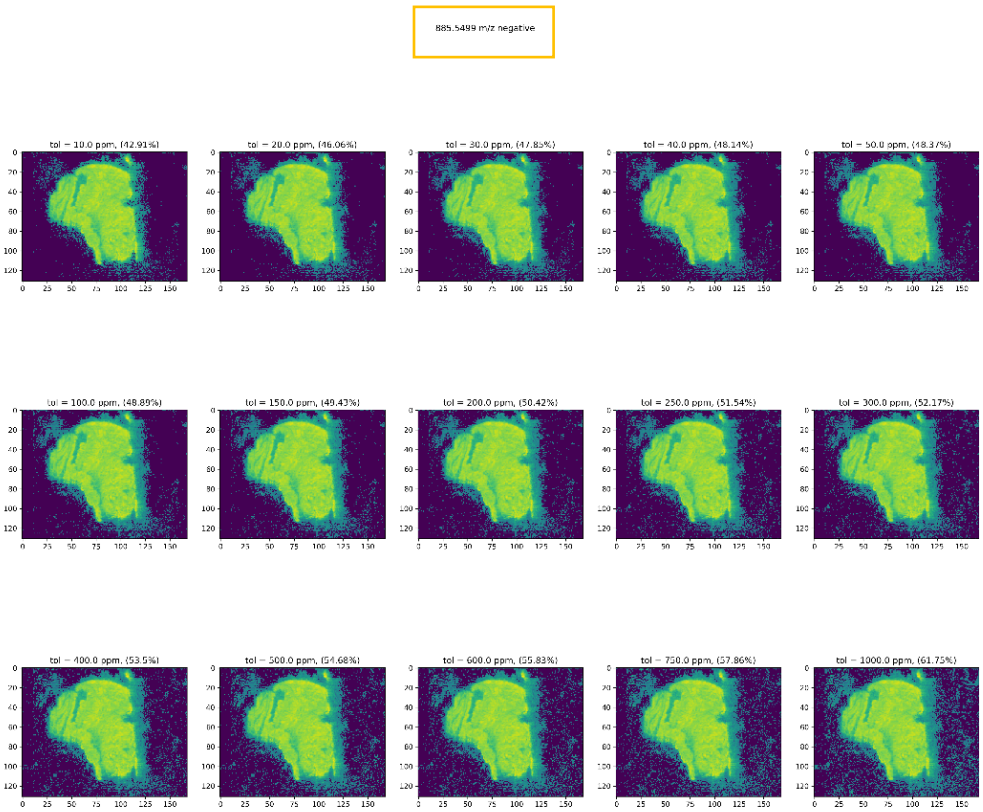


Figure 6.5 - The top image represents the spatial distribution of the matched Palmitic acid. It is expected to be found in the entire image, both in the sample and in the background area. The image in the middle shows the spatial distribution of the matched Eicosatetraenoic acid. The bottom images report the spatial distribution of the matched 855.5499 m/z, which is expected to be found in the entire tissue sample.

The image quality assessed is performed using equation 6.1 and is plotted in figure 6.7. It shows the ion image scatteredness versus the mean peak intensity highlighting the relationship between the quality of the image of each peak (scattered or regular) and the mean value of the peak. The scatter plot reports the relationship between the mean intensity of a peak with the quality of the corresponding image.

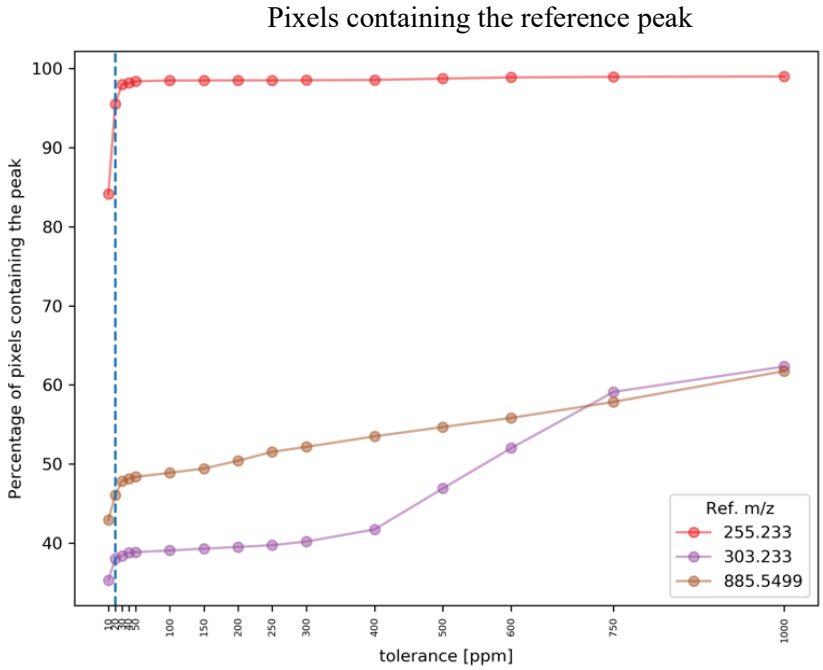


Figure 6.6 - The plot shows the percentage of pixels containing the reference peak, for the various tolerance values used. The vertical line corresponds to the plateau (relative variation of the percentage < 5%).

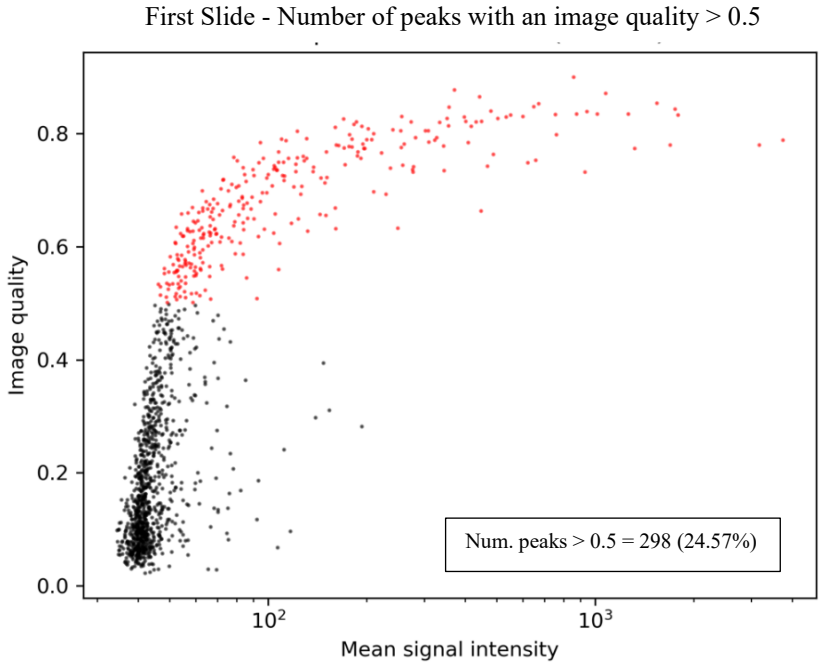


Figure 6.7 - Relationship between the quality of the image of each peak and the mean value of the peak. Each data point represents a peak image. Mean peak intensity is in the range below 1.0e4. The red data points represent high-quality peak images (24,57%).

The idea is that low-intensity signals, which are closer to the sensitivity level of the instrument, should be highly scattered and therefore associated with lower quality. In general, it suggests that zero-intense peaks cannot be considered absent, but they should be considered having a concentration too low to be detected by the instrument. The proportion of peaks with low intensity and high scatteredness may be used to determine the quality of the run. The image quality is defined in a range of [0, 1].

The pre-processing workflow proceeds with further investigation of the distribution of the peak intensities. The resulting plot (regarding the 1st slice of the 3D-DESI-MSI dataset) is reported in Figure 6.8. This plot represents the overall distribution of peak intensities showing the median and interquartile range of each matched peak intensity. Peaks with a coefficient of variation equal to zero are plotted in red. The plot should be used as representative of the intensity of the signal across the sample, and it can help to understand which *m/z* range peaks have been detected and the overall intensity.

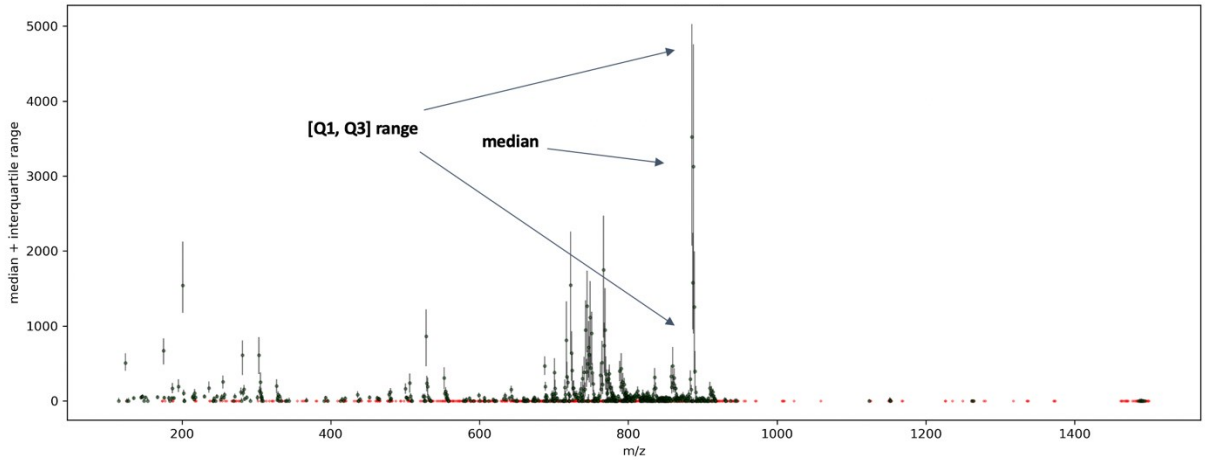


Figure 6.8 - A plot showing the distribution of the peak intensities for the first image of the dataset. The plot can be used to compare different runs to determine optimal parameters. The red points represent the median intensity of constant peaks (interquartile range equal to 0).

The workflow applied to pre-process the dataset and assess the quality of the data also produces the plot reported in Figure 6.9 where the images correspond to the detected reference *m/z* and the error of the detected peak *m/z* in ppm.

Detected reference *m/z*

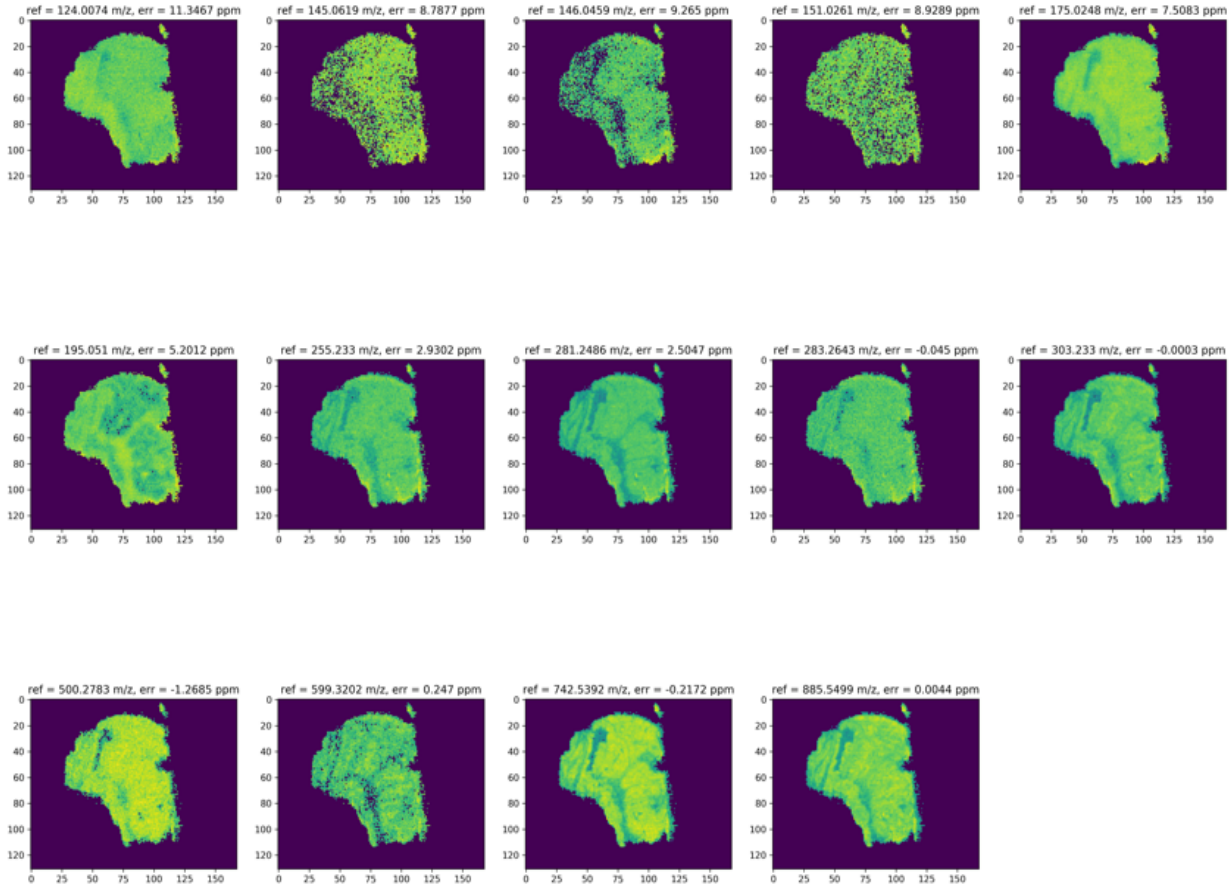


Figure 6.9 - The images correspond to the detected reference m/z. Every picture reports the reference m/z value and the error of the detected peak m/z in ppm

In chapter 3, we discussed the role of the solvent in MS technology and how it affects the information (blank signal variability), the Total Ion Chromatogram (TIC) has been assessed. A TIC is a chromatogram created by summing up intensities of all mass spectral peaks belonging to the same scan. The mean TIC signal is calculated in the first three columns of each row as a

representative measure of the variation along the rows. Particular attention must be paid in the selection of the number of columns to use for the TIC calculation (Figure 6.10). This number must be small; otherwise, there is the risk of getting samples pixels involved. Analogous measures are calculated along the top and bottom columns. In Figure 6.11 are reported the results of the TIC analysis where the comparison between lines represents the difference between slices, whereas variations within the same line represent the intra-run signal instability.

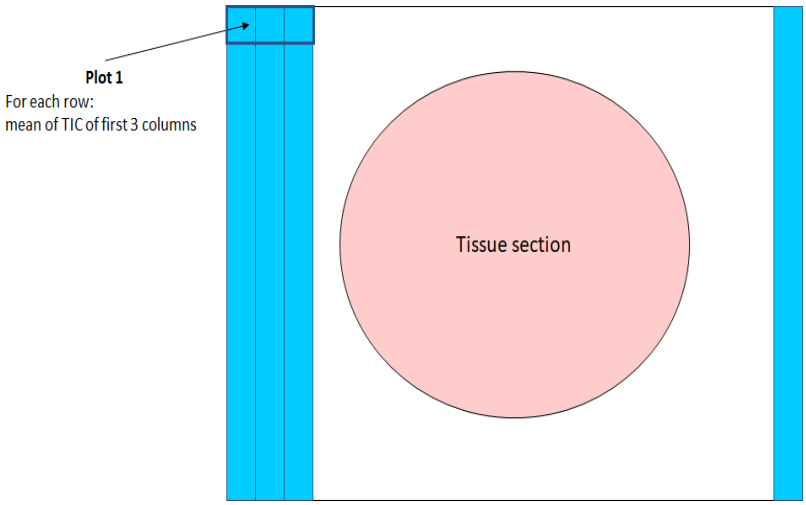


Figure 6.10 - Graphical interpretation of the process adopted for the calculation of the TIC (Blank signal variability the Total Ion Chromatogram). The mean TIC signal is calculated in the first three columns of each row as a representative measure of the variation along the rows. Analogous measures are calculated along the top and bottom columns.

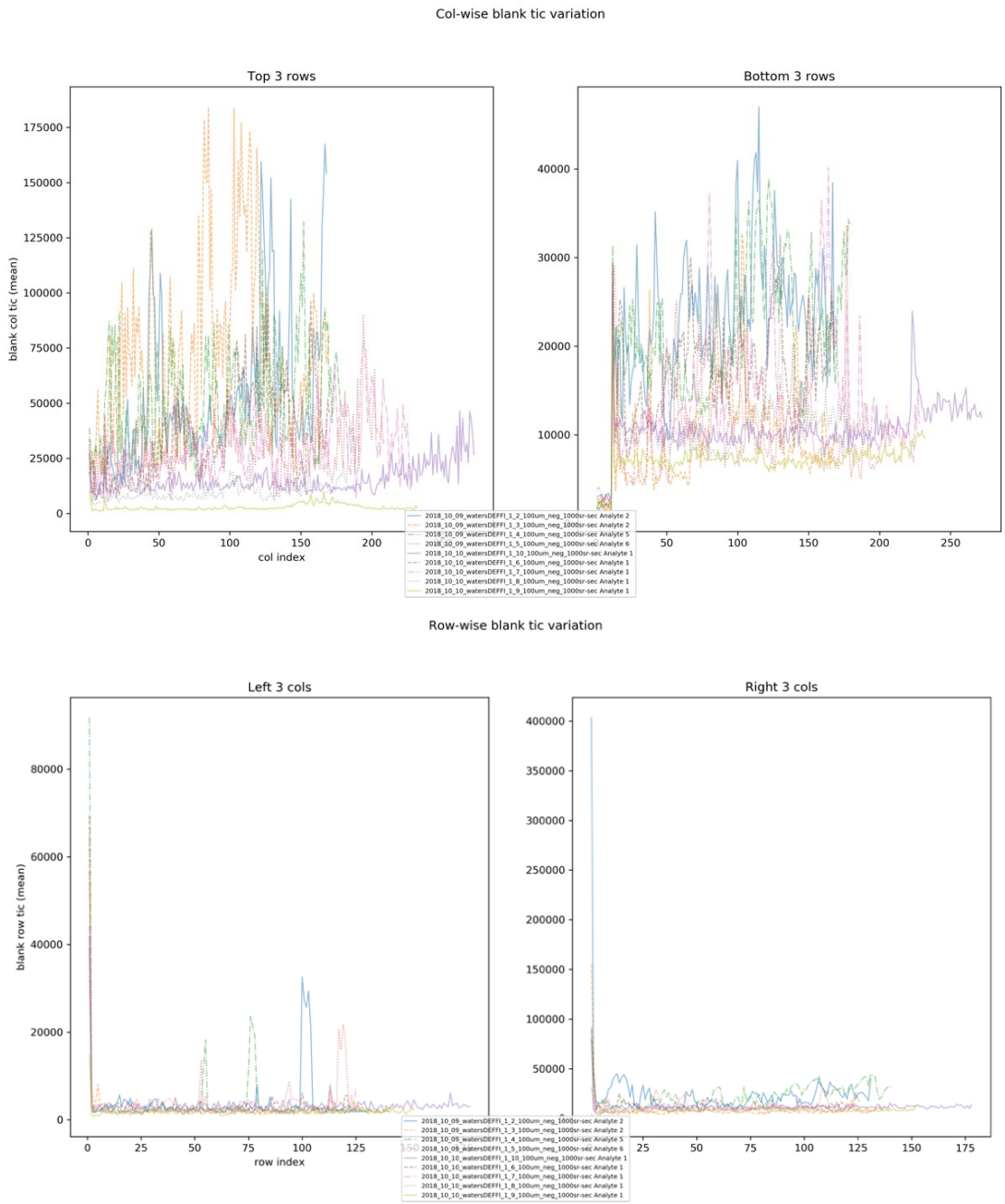


Figure 6.11 - Each line represents the variation of the mean TIC signal (three pixels left and right) along the rows for the different runs (vertical variability). Therefore, each line shows the intra-run blank signal variability

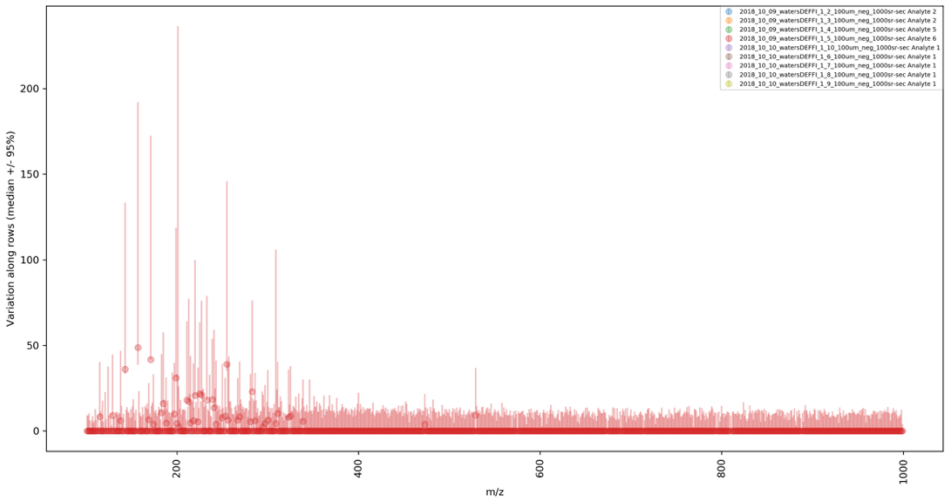


Figure 6.12 - The plot represents the variability along the rows of the individual peaks. The peaks are determined by uniform binning at 1 Da. This plot shows which peaks (or regions) of the blank signal vary significantly along the rows indicating which are less stable.

Previously, in section 6.2, we described the UMAP algorithm recently published (Smets T., et al., 2019). Here an application of UMAP on the 3D-DESI-MSI of the epithelial ovarian cancer is reported. The comparison and replacement of the parametric t-SNE with the novel UMAP methodology as a DR technique goes beyond the scope of this work as the objective is to test and, therefore, reapply the same DL procedure used in previous chapters. However, it was enlightening to study the potential of the UMAP approach and compare the results obtained to a classical methodology such as PCA. An example of the result is reported in Figure 6.13. The results suggest that the application of UMAP provides a more detailed representation of the low dimensional space of the tumour tissue. The colour maps used in the PCA and in the UMAP are automatically generated. Therefore, the same colour of both the images represents different tissue structures. The scales adopted are the same. By a simple visual interpretation of the UMAP image it is possible to identify a high number of sub-structures in the tissue sample suggesting that UMAP may be a powerful tool to reveal additional details in the samples and is a subject for future studies.

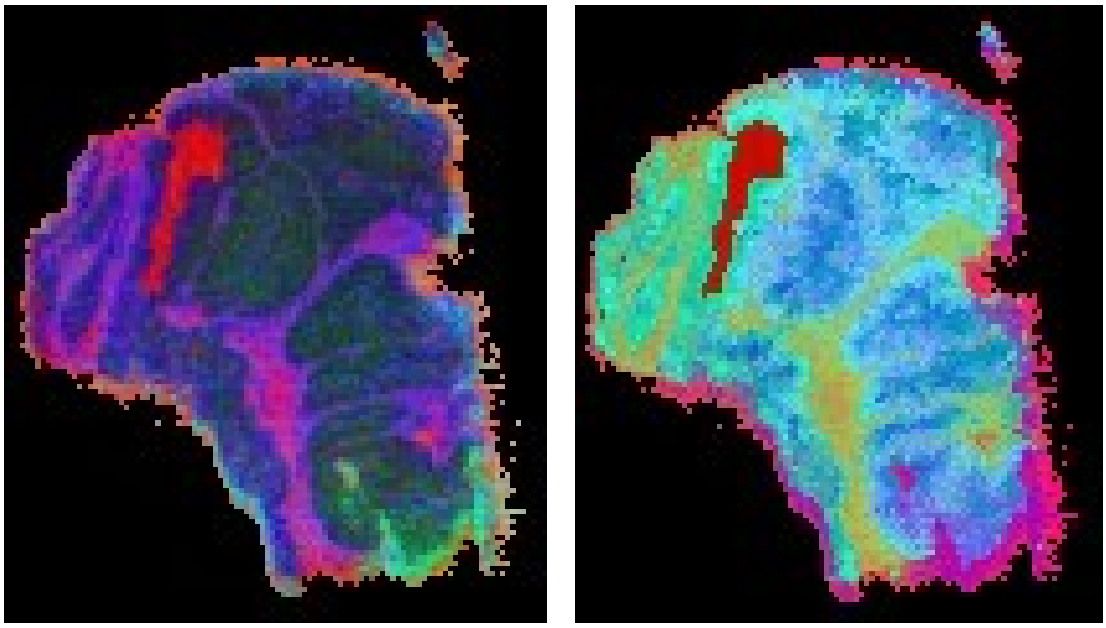


Figure 6.13 - The left picture reports an example of a PCA image (1st slice of the MSI dataset) while the picture on the right shows the same tissue sample analysed with the UMAP algorithm.

6.4 UNSUPERVISED ANALYSIS

In chapter 4, we described the application of a DL workflow for the investigation of a 3D-DESI-MSI dataset of metastatic liver cancer. In this section, we validate the flexibility of the DL workflow (with adequate technical adjustments, which will be discussed later) through the analysis of the pre-processed (section 6.3) 3D-DESI-MSI of epithelial ovarian cancer.

At this point, the analysis of the tumour dataset labelled with the linear SVM consists of the application of parametric t-SNE, clustering methods (OPTICS and DBSCAN), and finally, the calculation (univariate statistical analysis) and identification (with the METLIN database) of the highly correlated ions (per cluster).

The 3D ovarian cancer tissue model complexity appears lower (as expected) than that of the 3D-MSI associated with metastatic liver cancer (Chapter 4) due to the smaller number of

images available to build the tissue model (9 tissue slices instead of 51). Therefore, the architecture of the parametric t-SNE applied is less deep and consists just 4-layers with [400 – 200 – 100 – 2] neurons. The adopted learning rate is Lr = 0.001 with weights W = 0.002. The model fitting was achieved with an iteration of 30 epochs. The parametric t-SNE architecture adopted was chosen through a trial-and-error procedure. The 4-layer network showed a good balance between computational time and projection of the MSI dataset onto the latent space. The output is a 2-dimensional latent space of the tumour dataset (Figure 6.14).

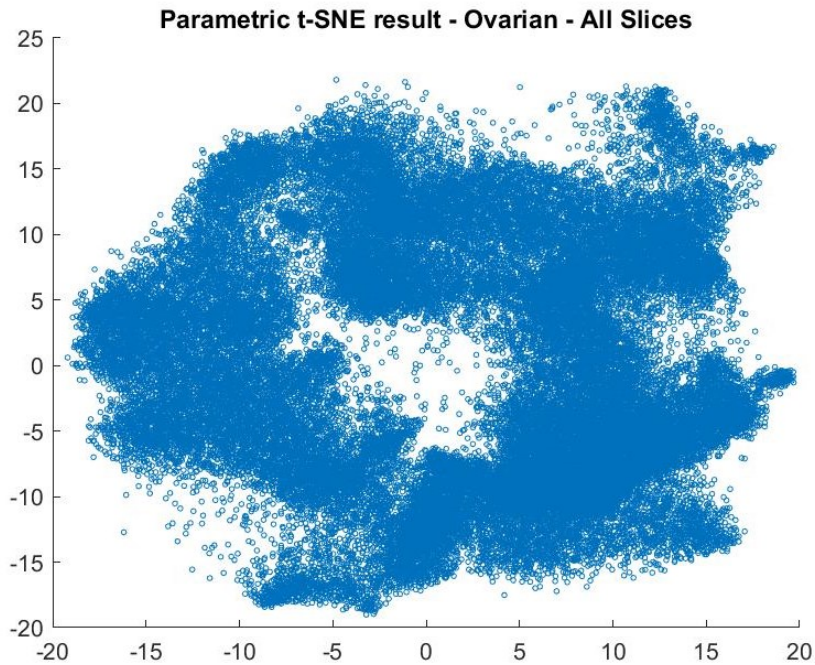


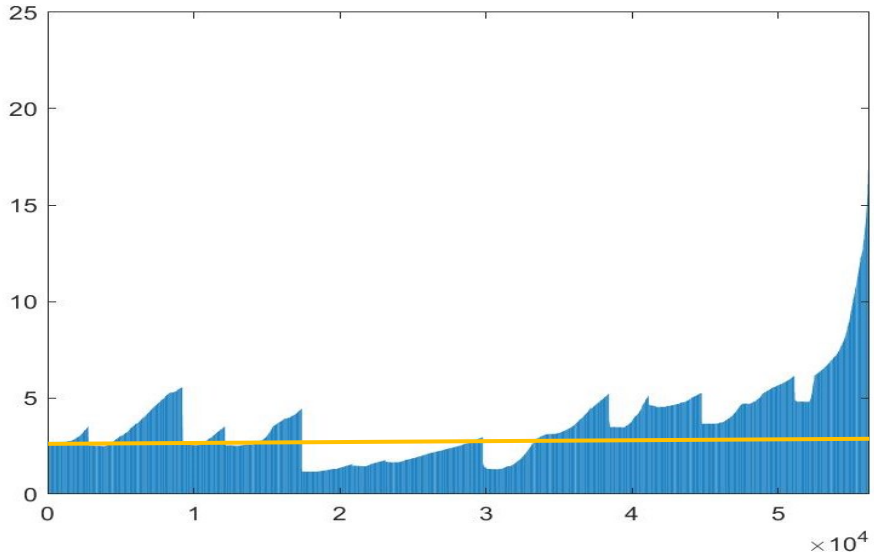
Figure 6.14 - Scatter plot of the 2D space obtained with the application of parametric t-SNE on the 3D-DESI-MSI of the ovarian cancer dataset.

As stated in the results shown in Chapter 4, the application of parametric t-SNE as a highly non-linear DR technique provides a detailed interpretation of the sub-structure of the MSI datasets. Also, in this analysis, the application of parametric t-SNE on the epithelial ovarian cancer dataset leads to a clear identification and characterisation of the sub-structure of the tumour tissue. In fact,

even with an initial simple visual analysis of the latent space (Figure 6.14), we can capture dense areas of data points (clusters).

The clustering of the tumour 2D latent space was performed using OPTICS and DBSCAN. The parameters chosen for clustering are *MinPts* = 1000 and ε= 2.3. Subsequently, the K-nn algorithm allowed the extension of the labelling of the clusters to the noisy data points, which were not dense enough to be considered a cluster. The reachability plot obtained with OPTICS and the final clustering result (after K-nn) are reported in Figure 6.15.

Finally, the application of univariate statistical analysis for the identification of the highly correlated ions per cluster is performed. The top correlated ions, with a correlation greater than 95%, are reported in Table 6.1.



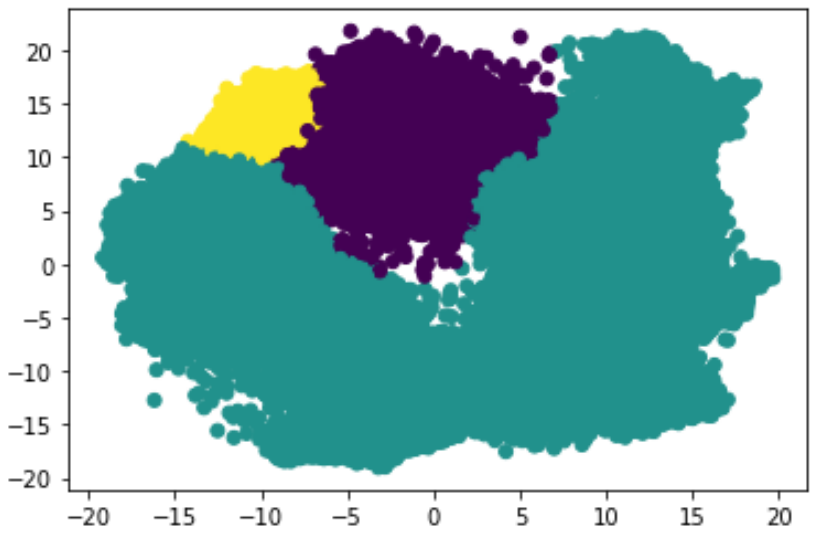


Figure 6.15 - The top image reports the reachability plot obtained with the application of the OPTICS algorithm on the 2D tumour latent space. The bottom picture shows the result of the K-nn algorithm applied on the DBSCAN clustering result.

Table 6.1 - List of the m/z values associated with the highly correlated ions detected in the 3 clusters identified in the tumour 2D latent space.

Cluster 1	**Cluster 2**	**Cluster 3**
678.3993	826.5693,799.6671,790.5417	812.5443
819.4556	802.4823	970.717
1151.694 726.5846	717.5816	796.584
605.3504	792.5531	972.7332
906.5465	768.553	859.5333
535.3017 1264.813	556.3005	688.5472
863.5255	888.5679	788.5383
736.494 701.5084	743.5425	660.514
771.6353	885.5499	785.6528, 836.53789
801.483	833.4474	872.5678

6.5 LIPID IDENTIFICATION

The importance of lipids in cellular signalling and structure is evident (Paragraph 4.10). Lipids are grouped in different families and classes depending on their molecular structure and structural organisation.

The discussion of how these important molecules are classified taking into account their biochemistry and biological role goes beyond the scope of this work; however, lipid classes can be summarised as follow:

- DG – Diacyglycerol. Second messenger signalling lipids.
- PA - Phosphatidic Acids. Lipids that act as biosynthetic precursor of all acylglycerol lipids in the cell.
- PG – Phosphatidylglycerol. Cell membrane lipid.
- PE – Phosphatidylethanolamine. Structural lipid.
- PI – Phosphatidylisositol. Signalling lipid, cell signalling and membrane trafficking.
- PC – Phosphatidylcholine. Mayor component of biological membranes.
- PS – Phosphatidylserine. Cell membrane component and cover a key role in the apoptosis process.
- GlcCer - Glucosylceramide.
- Ganglioside - Molecule composed of a glycosphingolipid which has important roles in the cells' membranes and for cell adhesion and cell-cell interactions.

The identification of the lipids was carried out using a 3-ppm research window on the METLIN database. The identifications of the highly correlated ions associated with the *m/z* values listed in Table 6.1 are reported here in Table 6.2.

Table 6.2 - Identification of the lipids associated with the highly correlated ions per cluster.

	CLUSTER 1		
Rank	*m/z*	error [ppm]	name
1	678.39927 [M+Cl]-	*0*	PA(12:0/20:3(8Z,11Z,14Z))
2	819.45563 [M+Cl]-	*1*	PE(18:4(6Z,9Z,12Z,15Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z))
3	1151.69392 [M+Cl]-	*1*	Ganglioside GM3 (d18:0/12:0)
4	726.58459 [M-H]-	*1*	GlcCer(d14:2(4E,6E)/20:0(2OH))
5	605.35041 [M+Cl]-	*0*	PG(22:0/0:0)
6	906.54651 [M+Cl]-	*0*	PS(20:2(11Z,14Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z))
7	535.30167 [M-H]-	*2*	PS(18:4(6Z,9Z,12Z,15Z)/0:0)
8	1264.81276 [M-H]-	*1*	Ganglioside GA1
9	863.52551 [M-H]-	*2*	PS(22:6(4Z,7Z,10Z,13Z,16Z,19Z)/20:4(5Z,8Z,11Z,14Z))
10	736.49395 [M-H]-	*2*	PS(12:0/20:1(11Z))
11	701.50839 [M-H]-	*0*	PS(O-16:0/14:1(9Z))
12	771.63526 [M-H]-	*2*	GlcCer(d18:2/21:0)
13	801.48296 [M+Cl]-	*3*	PS(12:0/22:4(7Z,10Z,13Z,16Z))

	CLUSTER 2		
Rank	*m/z*	error [ppm]	name
1	826.56934 [M-H]-	*3*	PI-Cer(t18:0/18:0)
2	799.66707 [M-H]-	*1*	GlcCer(d18:2/23:0)
3	790.54171 [M-H]-	*0*	PS(18:1(9Z)/18:1(9Z))[U]
4	802.48234 [M+Cl]-	*1*	PS(12:0/22:4(7Z,10Z,13Z,16Z))
5	717.58155 [M-H]-	*1*	PE-Cer(d14:1(4E)/24:0)
6	792.55311 [M-H]-	*0*	PS(15:0/20:5(5Z,8Z,11Z,14Z,17Z))
7	768.55301 [M-H]-	*0*	PS(13:0/20:3(8Z,11Z,14Z))
8	556.30052 [M-H]-	*2*	PS(20:5(5Z,8Z,11Z,14Z,17Z)/0:0)
9	888.56791 [M-H]-	*0*	PS(22:2(13Z,16Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z))
10	743.54254 [M-H]-	*1*	PS(13:0/18:2(9Z,12Z))
11	885.54991 [M+Cl]-	*3*	PS(18:1(9Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z))
12	883.53574 [M-H]-	*3*	PS(22:4(7Z,10Z,13Z,16Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z))

	CLUSTER 3		
Rank	*m/z*	error [ppm]	name
1	812.54426 [M+Cl]-	*0*	PS(16:0/18:1(11Z))
2	970.71703 [M+Cl]-	*2*	PI-Cer(d18:0/26:0)
3	796.58398 [M-H]-	*1*	PS(15:0/20:3(8Z,11Z,14Z))
4	972.73321 [M-H]-	*0*	PI-Cer(t18:0/22:0(2OH))
5	859.53329 [M-H]-	*1*	PS(20:2(11Z,14Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z))
6	688.54722 [M-H]-	*1*	GlcCer(d15:2(4E,6E)/18:0)
7	788.53829 [M-H]-	*3*	PS(18:1(9Z)/18:1(9Z))[U]
8	660.51401 [M-H]-	*4*	PE-Cer(d14:1(4E)/20:0)
9	785.65278 [M-H]-	*0*	CerP(d18:1/26:1(17Z))
10	836.53789 [M-H]-	*2*	PS(18:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z))
11	872.56784 [M-H]-	*0*	PS(20:3(8Z,11Z,14Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z))

The main classes of lipids found are:

1. A high presence of PSs. Phosphatidylserine is a cell membrane component and has a vital role in the apoptosis process. The role of PSs in oncogenesis and immunology has been extensively studied. They are associated with immunosuppressive factors within tumour microenvironments and the ability to block them to enhance patients' antitumor immune responses (Raymond J., et al., 2015). The ability to detect ovarian cancer at early stages is low, and most cases are diagnosed as advanced disease. Since tumour cells express PSs on their plasma membrane, it is possible to predict that tumours might secrete PS-positive exosomes into the bloodstream that could be a representative biomarker for ovarian cancer (Lea J., et al., 2017). There are several articles in the literature that link the presence of PSs to the development of ovarian cancer and how they could be treated as possible biomarkers (Belzile O., et al., 2018) (Manjarika D., et al., 2018) (Birge R., et al., 2016).
2. Ceramides which are bioactive lipids that mediate cell death in cancer cells, and ceramide-based therapy are now being tested in phase I clinical trials as a cancer treatment (Zhang

X., et al., 2018). Usually, ceramides act as a potent tumour suppressor that causes apoptosis and cell cycle arrest. Glucosylceramide (GlcCer) is of particular interest, especially because it interacts with other ceramides downregulating their 'positive' behaviour (Ishibashi Y., et al., 2013).

3. An abundance of PEs. The role of Phosphatylethanolamine in ovarian cancer is not clear yet. However, in the publication of Zhou *et al.* (Zhou X., et al., 2012), this class of lipids could be considered as biomarkers in the diagnosis of prostate cancer.
4. Gangliosides are molecules composed of a glycosphingolipid. They have an essential role in cell membranes, cell adhesion and cell-cell interactions. Gangliosides are also crucial in the oncogenesis process (August L.T., et al., 2018), in fact, there is a keen interest in the literature in regards of the adoption of Gangliosides as a novel cancer drug treatment (Jinyi L., et al., 2018).

6.6 SIMILARITY INDEX

In chapter 5, we performed the similarity analysis of clusters through the calculation of the Rand Index. In this chapter, we perform a similar analysis on the 3D-DESI-MSI of ovarian cancer. The objective is to further corroborate the utility of a 3-dimensional approach compared to the classical 2-dimensional approach.

For this analysis, twenty-three (23) 3D-MSI tissue models were built using a reduced number of slices:

- Ten models built using four tissue slices.
- Ten models built with two slices.
- Three models consisting in a single 2-dimensional image.

The 23 models have been investigated with the DL workflow described earlier in chapter 4. Finally, the information content of the clustering found in each model was compared with that obtained from the whole model composed of 9 slices (described in the previous sections of this chapter).

Every batch of models has been created following the same procedure. For instance, in order to build the model with four slices, slices were randomly chosen among the available nine (full dataset), and then, the process was iterated ten times.

In order to preserve the biopsy structure, the original progression of the slices was maintained (their structural order). The application of the parametric t-SNE to the new models has required some adjustment to the architecture. The lower dimensionality of the 3D-MSI tumour datasets, due to the smaller amount of tissue slices used, requires a reduction of the number of hidden layers that compose the parametric t-SNE or even a reduction of the number of neurons per layers.

All the latent spaces obtained were segmented by OPTICS and DBSCAN algorithms and *forced* to produce three clusters.

Finally, the similarity index calculation was performed to compare the same clusters of the 23 reduced models with the corresponding clusters obtained from the investigation of the whole 3D-DESI-MSI composed of 9 slices. The scatter plot reported in Figure 6.16 shows how the *similarity index* between clusters decreases as a function of the number of slices taken into consideration to construct the reduced models.

The curve in the graph shows how the similarity drops from the model composed of 9 slices to the model built with 4 slices producing a similarity value of 17.8%. Almost the same value of the Rand index can be obtained from the comparison of the clustering of the whole model with the

clustering provided by the analysis of reduced model built with 2 slices (similarity of the 18.5%). Finally, we see a steep drop of the similarity value when the model is composed of just one slice. In this occasion, the similarity is 9.9%.

This result reinforces the importance of having a large amount of data in order to perform a more in-depth tissue analysis (more representative) and how critical is the introduction of a third dimension in this kind of analysis. Adding slices gives more weight to this hypothesis as the model 'converges' with more slices. Otherwise, it varies a lot between different slices.

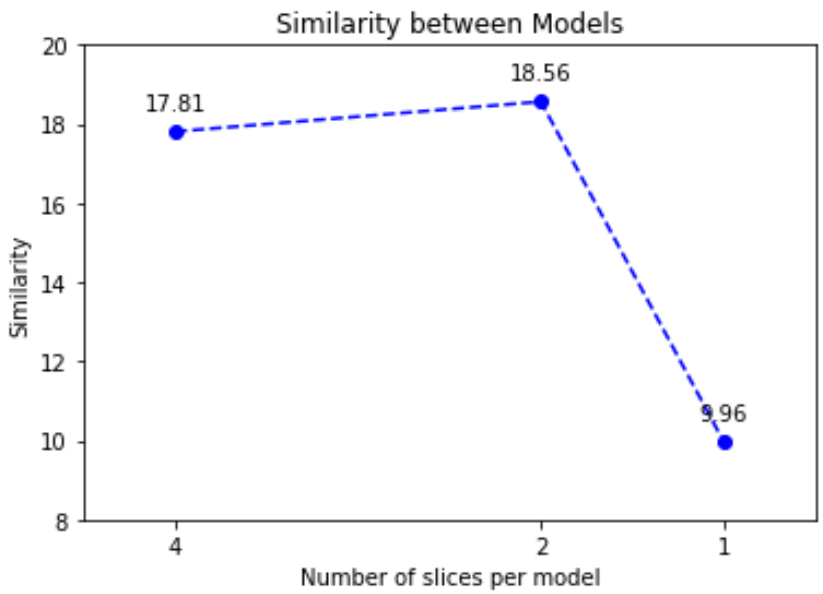


Figure 6.16 - Scatter plot of the Rand Index. The x-axis reports the number of slices taken into consideration for the construction of the reduced 3D tissue models. The y-axis reports the % of similarity indices.

In this study, when comparing the results shown in chapter 5, the similarity index values are higher. This may be a function of the large difference in the dimensionality between the 3D-DESI-MSI of ovarian cancer and the 3D-DESI-MSI of metastatic liver cancer. In fact, in a smaller dataset (like the ovarian cancer dataset investigated in this chapter) the probability is high of finding similarity between the full-size model (9-slices) and the reduced models. This is less

probable in bigger MSI datasets, like the metastatic liver cancer investigated in chapter 4, which was composed of a higher number of tissue slices (51).

6.7 CONCLUSION

The analysis described in this chapter, together with the work presented in the previous sections of this thesis (chapters 4 and 5), investigates the application of DL to 3D-DESI-MSI and shows how segmentation, classification, and investigation of the major metabolic components of tumour sections can indicate underlying mechanisms of disease. This approach represents a powerful and innovative tool to investigate the bio-chemical components and interactions occurring in tumour tissue sub-structures.

The application of the DL workflow on two different 3D-DESI-MSI datasets, ovarian cancer, and metastatic liver cancer, highlights the flexibility and the potential of this investigational procedure. The underlying metabolic processes can be deduced based on the presence of metabolites that correlate with cancerous regions of the tumour. Traditional investigation methods which rely on visual inspection may not show subtle differences between tumour metabolism and morphology.

Most recent published work in the area uses only single images (slices) to deduce the nature of the tumour. When working in three dimensions, it is possible to compare the stability of models as the number of slices is increased, going from a single slice to several slices that span all the tumour morphology. The tests performed, using the DL workflow, demonstrated the increased stability of models when using multiple slices across the full three dimensions of the tumour when compared the classical 2-dimensional image analysis. We demonstrated that 3D-DESI-MSI analysis is a more robust approach to identify and analyse the presence of tumour subgroups of

cells characterised by similar mass spectrometry profiles. Furthermore, the third dimension introduces topological constraints that, combined with the fact that biochemical interactions are typically localised (and naturally occur in 3-dimensions), leads to robust segmentation of the tumour and also provides a better interpretation and understanding of tumour mechanisms and behaviour.

Finally, the clustering analysis with the identification of the most correlated ions per cluster are an expression of the local biochemical components of the tumour, and their interpretation can provide a robust hypothesis describing the mechanisms involved in the analysed samples. This result has crucial importance to achieve a more in-depth insight into the biochemical mechanisms occurring in ovarian cancer. The application of unsupervised learning techniques to 2D-MSI data may not be reliable, as shown by the similarity results, which indicated significant variability in models produced from only single slices compared to multiple slices across the entire tumour.

During the analysis described in this chapter, we also have the possibility to test a more recent DR technique, UMAP. It is through comparison with the classical PCA that we can assert that the application of UMAP provides a more detailed representation of the low dimensional space of the tumour tissue. Indeed, even by a simple visual interpretation of the UMAP image it is possible to identify a high number of sub-structures in the ovarian cancer sample.

The work described in this chapter concludes with insights into the biochemical interactions occurring in ovarian cancer and identifies the chemical/metabolic properties of the tumour sub-regions. The study of the most correlated ions in each cluster provides a detailed picture of the presence and distribution of key metabolites. The use of these ions as biomarkers in ovarian tumour tissue may lead to diagnostic opportunities (Koundouros N., et al., 2020). Important ions for classification include Phosphatidylserine (PS) which may provide a biomarker

for the detection of ovarian cancer at early stages. The presence of Glucosylceramide (GlcCer) was detected. This acts as a potent tumour suppressor that causes apoptosis and cell cycle arrest. Finally, the presence of Gangliosides which play an essential role in cell membranes, was identified. These are essential for cell adhesion and cell/cell interactions. These compounds cover relevant roles in oncogenesis.

This diagnostic approach can give insights that are not captured by classical visual inspection of H&E images. Indeed, this workflow could be used to assist histopathologists in the analysis of H&E and to identify and characterise specific tumour regions.

Considering the nature of the technology applied and the complexity of the dataset under examination, it becomes clear that having a quite significant amount of data is crucial for the success of the analysis. However, one of the significant technical difficulties encountered during the development and testing of the DL workflow described in this work was the very long training time necessary for the ANNs to create a dimensionally reduced model. The computational time required from the parametric t-SNE to create a latent space can represent a limit to the application of this technique for a real-time 3D tissue investigation. Fortunately, nowadays, an essential aspect of the research deploying DL has given to the possibility of using pre-trained ANNs. This approach can drastically diminish the amount of data needed for training and, subsequently, the computational time (Kamnitsas K. et al. , 2018), making the application of this workflow faster.

7 CONCLUSIONS

In this study, we have presented, investigated, tested, and validated a novel computational workflow designed for unsupervised analysis of 3D-MSI datasets. The sequence of methods described, and the algorithms applied have demonstrated a robust process to extract and interpret the complex tumour sub-structure.

The combination of multiple unsupervised learning algorithms, such as DR processes based on DL methodologies (parametric t-SNE), density clustering algorithms (OPTICS and DBSCAN) and the subsequent calculation of similarity indices (Rand Index) have provided deeper insights into the information content of 3D-DESI-MSI datasets. The workflow also creates a concrete basis for future work and suggests the superiority of a 3-dimensional approach compared to the classical 2-dimensional image analysis.

7.1 NON-LINEAR DR AND CLUSTERING OF A 3D METASTATIC LIVER CANCER MODEL

Visual analysis of H&E histology images is considered as the gold standard for the clinical diagnosis of cancers. However, the technological advances in the computer vision sector that we have witnessed in recent years suggest that a clinical application of these technologies can provide us with a better understanding of cancer's biology and lead to personalised medicine in much shorter time. It is in this context that this work makes a contribution to clinical research.

In fact, we have demonstrated how the application and combination of computer vision algorithms with 3D-MSI datasets can be used to build diagnostic procedures which can augment the pathology field providing insights into the metabolic sub-structure of the tumour.

The main component of the exploratory potential of this unsupervised modelling is the application of a DR technique, which generates an approximate representation of the statistical properties of the MSI dataset allowing the identification of patterns in the data that can classify tumour regions and give insights into their metabolic composition.

This methodology combined with the application of density-based clustering algorithms gives an easy-to-interpret and robust insight into the complex structure of the dataset.

In chapter 4, we compared different algorithms that could be deployed for classification and DR purposes. We demonstrated that linear DR methods, such as PCA for example, which is a standard method used for this kind of exploratory analysis, is not capable of identifying the similarity patterns among the spectra of a large 3D-DESI-MSI dataset. On the other hand, a non-linear DR approach, such as the DL-based parametric t-SNE, can highlight the presence of clusters of spectra (pixels) with similar peak intensity patterns. Parametric t-SNE is a flexible technique that can be trained on a smaller number of pixels compared to other non-parametric methods. It also has the benefit that it can be used to project the DR mapping onto the held-out spectra (unseen data).

The clusters identified by DBSCAN, and OPTICS were found to be significantly associated with biologically related tissue sub-types. Also, a co-expression network analysis showed that the spatial distribution of the three main groups of co-localised ions was consistent with the OPTICS results.

The application of this DL workflow on both 3D metastatic liver cancer and 3D epithelial ovarian cancer models (chapter 6) provided results that revealed the correlations between key metabolites and regions of the tumour through the identification of relevant ions (especially lipids) that could be used as possible biomarkers.

7.2 SIMILARITY INDEX AND 3D ANALYSIS

The study and results described in previous chapters (5 and 6) strongly imply that there are advantages to using the additional spatial constraint of the third dimension, which highlights the importance of using data from a 3-dimensional specimen instead of the classical 2-dimensional image investigation and demonstrate the robustness of the diagnostic procedure adopted (based on DL). Additionally, the use of less data (from few tissues slices to a single 2D image) eventually results in simplistic representations of the tumour metabolism that are often very dissimilar from that provided by more complex 3D models (created with a more significant number of tissue slices).

In order to reach this conclusion, we analysed a total of one-hundred and forty-six (146) 3D-DESI-MSI datasets:

- 123 models built using a reduced number of slices from the original 51-slices model of metastatic liver cancer.
- 23 models built using a reduced number of slices from the original 9-slices model of epithelial ovarian cancer.

The analysis performed is based on the application of several variants of the DL workflow introduced in chapter 4 and the subsequent application of clustering algorithms. The architectures and parameters of the technical methodologies deployed were evaluated and readjusted case by case according to the dimensionality of the 3D-MSI under analysis. The identified clusters (three

clusters per model) were labelled accordingly with the clusters originally obtained from the corresponding full-size dataset and their information content were also compared. This comparison was assessed by the calculation of the Rand Index (similarity index).

The significance of the results obtained indicates that a single 2D tissue section is not capable of providing robust results. Different sections can give substantially different models. A 3-dimensional analysis provides, instead, a more stable interpretation of the biochemistry that occurs in the tumour, and the continuity of the adjacent tissue sections allowed us to discard unrealistic tissue partitions.

In chapter 5, we ran a further test to evaluate the robustness of the DL workflow as an investigation tool. The test was performed by running the analysis on five 3D tissue models composed of 51 slices each, where the pixels were randomly shuffled. The DL workflow appeared to be a robust procedure capable of extracting relevant information from a complex 3D-MSI dataset even when the original structure (progression of pixels and tissue slices) is disrupted.

7.3 ANALYSIS OF A 3D-MSI OVARIAN CANCER DATASET

In chapter 6, we validated the efficiency of the DL workflow running the analysis on a second 3D-DESI-MSI dataset built with 9 slices of epithelial ovarian cancer. We have shown that the DL workflow (introduced in chapter 4) is a flexible and reliable diagnostic procedure that can be easily applied to different 3D-DESI-MSI datasets.

In tandem, we also deployed a different pre-processing workflow (compared to the one adopted in chapter 4) which is able to assess the quality of the images, evaluates the distances between the matched mass and the theoretical masses, and provides the percentage of pixels

containing the reference peak (plateau). We also experimented with UMAP to investigate possible advantages over t-SNE.

The biochemical mechanisms associated with ovarian cancer were investigated in detail, leading to a detailed interpretation of the metabolic sub-structure of this particular disease which highlighted important lipids that can be used in further investigations.

7.4 FUTURE WORK

Future work could be focused on the introduction and design of novel diagnostic procedures derived from analysis of the principal metabolites separating cancerous and healthy tissue. The combination of 3D-MSI with other data sources, such as genomics and transcriptomics, could provide a more comprehensive interpretation and understanding of oncogenesis and disease progression. DL diagnostic procedures could help to stratify patients to enable more detailed clinical decisions to be made and to personalised treatments.

Recently, a cutting-edge technology, iKnife, was invented and developed by Prof. Zoltan Takats. The iKnife is based on electrosurgery, which is a popular technology that uses an electrical current to rapidly heat tissue, cutting through it, and minimising blood loss. During the surgical procedure, the electrosurgery vaporises the tissue, creating smoke that is usually sucked away by extraction systems. Prof Takats recognised that this smoke would be a rich source of biological information; therefore, he thought to connect the electrosurgical knife to an MS. It can reveal information about the state of that tissue instantly providing information that generally takes much longer time to reveal through the application of classical laboratory tests. New studies have shown the potential of the iKnife in recording the characteristics of cancerous and non-cancerous tissues (lung, colon, breast, brain, liver, and stomach) in order to create a reference library. During surgery,

iKnife matches its readings to the ones stored in the reference database determining what type of tissue is being cut, giving a result in less than 3 seconds. This characteristic allows surgeons to carry out procedures with a level of accuracy that was not possible before.

Although the current study focussed on cancer detection, a possible combination of this technology with the one presented in this work (DL-based) could lead to a more profound interpretation and understanding of the cancer biology and to enable diagnosis. The MS database created during the application of the iKnife in several surgical procedures could be used to train an AI system. The combination of machine learning and iKnife could lead to a powerful, robust, and fast surgical tool.

AI will enable an era of quicker, cheaper, and more productive drug discovery. Experts do expect these tools to become increasingly important. The application of AI to drug discovery presents both challenges and opportunities for scientists, especially when the techniques are combined with robotic/automation.

Modern drug discovery has entered the big data era due to the massive datasets available for drug candidates. Pivotal to this shift is the development of AI approaches to implementing innovative modelling based on the dynamic, heterogeneous, and broad nature of drug datasets. As a result, recently developed methods such as DL approaches to provide new solutions to efficacy and safety evaluations of drug candidates have been developed (Hao Z., et al., 2020). The resulting models provide insight into the spectrum of chemical structure, in vitro and in vivo studies, and clinical outcomes. The relevant novel data mining and analysis provide critical support to recent modelling studies. The work described here could be used as an additional data source to evaluate the efficacy of drug therapy at the tissue level, for example, by looking at changes in metabolite composition between treated and no treated tissues.

In summary, the advancement of AI in the big data era has paved the way for future rational drug development and optimisation, which will have a significant impact on drug discovery procedures and, eventually, public health (Hao Z., et al., 2020).

8 BIBLIOGRAPHY

Böhning D., et al. (1992). Multinomial Logistic Regression. *Ann. Inst. Statist. Math. Vol. 44, No. 1.*, 197-200.

Abbassi-Ghadi, N., et al. (2014). Discrimination of lymph node metastases using desorption electrospray ionisation-mass spectrometry imaging. *Chemical Communications*, Vol. 50(28), pp. 3661–3664.

Abdelmoula W.M., et al. (2016). Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data. *Proceedings of the National Academy of Sciences, 113(43)*(doi: 10.1073/pnas.15), 12244–12249. https://doi.org/10.1073/pnas.1510227113.

Abramczyk H., et al. (2015). The role of lipid droplets and adipocytes in cancer. Raman imaging of cell cultures: MCF10A, MCF7, and MDA-MB-231 compared to adipocytes in cancerous human breast tissue. *Analyst 2015, 140.*, 2224–2235. doi: 10.1039/c4an01875c.

Accioly M.T., et al. (2008). Lipid bodies are reservoirs of cyclooxy-genase-2 and sites of prostaglandin-E2 synthesis in colon cancer cells. *Cancer Res 2008, 68.*, 1732–1740. doi: 10.1158/0008-5472.CAN-07-1999.

Addie R.D., et al. (2015). Current State and Future Challenges of Mass Spectrometry Imaging for Clinical Research. *Analytical Chemistry 87 (13).*, 6426-6433. doi: 10.1021/acs.analchem.5b00416.

Addie, R.D., et al. (2015). Current state and future challenges of mass spectrometry imaging for clinical research. *Analytical Chemistry*, Vol. 87(13), pp. 6426–6433. http://dx.doi.org/10.1021/acs.analchem.5b0041.

Aichler M., e. a. (2013). Clinical response to chemotherapy in oesophageal adenocarcinoma patients is linked to defects in mitochondria. *The Journal of Pathology*, Vol. 230, Issue 4. https://doi.org/10.1002/path.4199.

Aichler, M., et al. (2013). Clinical response to chemotherapy in oesophageal adenocarcinoma patients is linked to defects in mitochondria. *The Journal of Pathology*, Vol. 230(4), pp. 410–419. http://dx.doi.org/ 10.1002/path.4199.

Aizerman, M.A., et al. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, Vol. 25, pp. 821–837.

Alexandrov T, et al. (2012). MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinformatics. BioMed Central 13, 13(Suppl 16)*(doi: 10.1186/1471-2105-13-S16-S11), S11. https://doi.org/10.1186/1471-2105-13-S16-S11.

Alexandrov T., et al. (2013). Testing for presence of known and unknown molecules in imaging mass spectrometry. *Bioinformatics 29(18), 29(18)*(doi: 10.1093/bioinformatics/btt388), 2335-42. doi: 10.1093/bioinformatics/btt388.

Alexandrov, T., & Kobarg, J.H. . (2011). Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. . *Bioinformatics (Oxford, England).*, Vol. 27(13), pp. i230–i238. http://dx.doi.org/10.1093/bioinformatics/btr246.

Alexandrov, T., et al. (2010). Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering. *Journal of Proteome Research*, Vol. 9(12), pp. 6535–6546. http:// dx.doi.org/10.1021/pr100734z.

American Association for Cancer Research. (2020). Metastasis Has Multiple Origins and Occurs Early in Tumorigenesis. *Cancer Discov 10(7)*, 903. doi: 10.1158/2159-8290.CD-RW2020-082.

Amstalden van Hove E.R., et al. (2010). A concise review of mass spectrometry imaging. *Journal of Chromatography A Volume 1217, Issue 25*, 3946-3954. https://doi.org/10.1016/j.chroma.2010.01.033.

Ankerst M., et al. (1999). Optics: Ordering points to identify the clustering structure. *in ACM Sigmod Record*, 49–60. doi: 10.1145/304181.304187.

Apostolova L.G., et al. (2006). Conversion of mild cognitive impairment to Alzheimer disease predicted by hippocampal atrophy maps. *Arch Neurol. 63(5)*, 693-699. doi:10.1001/archneur.63.5.693.

Archibald R., et al. (2007). Feature selection and classification of hyperspectral images with support vector machines. *IEEE Geoscience and Remote Sensing Letters, vol. 4, no. 4., 4(4)*(doi: 10.1109/LGRS.2007.905116), 674-677. doi: 10.1109/LGRS.2007.905116.

Asa B.H., et al. (2001). Support Vector Clustering. *Journal of Machine Learning Research 2*, 125-137.

Ashrafian H., et al. (2021). Metabolomics: The Stethoscope for the Twenty-First Century. *Medical Principles and Practice.*, 30.4: 301.

Aslam J.A., et al. (2007). On Estimating the Size and Confidence of a Statistical Audit. *Electronic Voting Technology 7*, 8.

August L.T., et al. (2018). Gangliosides in Health and Disease Vol. 156. *Progress in Molecular Biology and Translational Science*, Chapter 2. Congenital Disorders of Ganglioside Biosynthesis.

Balluff B., et al. (2015). De novo discovery of phenotypic intratumour heterogeneity using imaging mass spectrometry. *The Journal of Pathology Vol. 235, Issue 1.*(doi:10.1002/path.4436), 3-13. https://doi.org/10.1002/path.4436.

Balluff, B., et al. (2010). Classification of HER2/neu Status in Gastric Cancer Using a Breast-Cancer Derived Proteome Classifier. *Journal of Proteome Research* , Vol. 9 (12), pp. 6317-6322, doi: 10.1021/pr100573s.

Basu, S.S., et al. (2019). Rapid MALDI mass spectrometry imaging for surgical pathology. *npj Precis. Onc. 3, 17*, https://doi.org/10.1038/s41698-019-0089-y.

Bauer, J. A., et al. (2010). Identification of markers of taxane sensitivity using proteomic and genomic analyses of breast tumors from patients receiving neoadjuvant paclitaxel and radiation. . *Clinical Cancer Research*, Vol. 16(2), pp. 681–690.

Bedard P.L., et al. (2013). Tumour heterogeneity in the clinic. *Nature 501.*, 355–364. https://doi.org/10.1038/nature12627.

Belo-ribi-Djefaflia S., et al. (2016). Lipid metabolic reprogramming in cancer cells. *Oncogenesis volume 5*, e189. https://doi.org/10.1038/oncsis.2015.49.

Belzile O., et al. (2018). Antibody targeting of phosphatidylserine for the detection and immunotherapy of cancer. *Immunotargets Ther 7.*, 1-14. https://doi.org/10.2147/ITT.S134834.

Bemis K.D., et al. (2015). Cardinal: An R package for statistical analysis of mass spectrometry-based imaging experiments. *Bioinformatics 31(14)., 31(14)*, 2418–2420. doi: 10.1093/bioinformatics/btv146.

Bemis, K.D., et al. (2016). Probabilistic segmentation of mass spectrometry (MS) images helps select important ions and characterize confidence in the resulting segments. *Molecular & Cellular Proteomics.*, Vol. 15(5), pp. 1761–1772. http://dx.doi.org/10.1074/mcp.O115.053918.

Bengio Y, et al. (2007). Greedy layer-wise training of deep networks. *Advances in Neural Information Processing Systems 19*, 153–160. Available at: http://papers.nips.cc/paper/3048-greedy.

Bengio Y., et al. (2003). Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. *Advances in Neural Information Processing Systems (NIPS)*, 177–184. Available at: http://papers.nips.cc/paper/2461-out-of-sample-extensions.

Bingming, C., et al. (2020). Combining MALDI mass spectrometry imaging and droplet-base surface sampling analysis for tissue distribution, metabolite profiling, and relative quantification of cyclic peptide melanotan II. *Analytica Chimica Acta*, Vol. 1125, pp. 279-287, ISSN 0003-2670. https://doi.org/10.1016/j.aca.2020.05.050.

Birge R., et al. (2016). Phosphatidylserine is a global immunosuppressive signal in efferocytosis, infectious disease, and cancer. *Cell Death Differ 23*, 962–978. https://doi.org/10.1038/cdd.2016.11.

Birge R.B., et al. (2016). Phosphatidylserine is a global immunosuppressive signal in efferocytosis, infectious disease, and cancer. *Cell Death and Differentiation, 23(6)*, 962–978. doi: 10.1038/cdd.2016.11.

Boggio K.J., et al. (2011). Recent advances in single-cell MALDI mass spectrometry imaging and potential clinical impact. *Expert Review of Proteomics, 8:5., 8(5)*, 591-604. doi: 10.1586/epr.11.53.

Boser, B.E., et al. (1992). A Training Algorithm for Optimal Margin Classifiers. *Association for Computing Machinery*, doi = 10.1145/130385.130401.

Bouveyron, C., et al. (2007). High-dimensional data clustering. *Comput Statistics Data Anal*, Vol. 52(1), pp. 502–519.

Bovenga F., et al. (2015). Uncoupling nuclear receptor LXR and cholesterol metabolism in cancer. *Cell Metab 2015; 21.*, 517–526. https://doi.org/10.1016/j.cmet.2015.03.002.

Bozza P.T., et al. (2010). Lipid droplets in inflammation and cancer. *Prostaglandins Leukot Essent Fatty Acids 2010; 82.*, 243–250. https://doi.org/10.1016/j.plefa.2010.02.005.

Breiman L., et al. (2001). Random forests. *Machine Learning, 45(1)*, 5–32. doi: 10.1023/A:1010933404324.

Buchberger AR, et al. . (2018). Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. *Anal Chem.*, Vol. 90(1), pp. 240-265. doi:10.1021/acs.analchem.7b04733.

Bunney T.D., et al. (2010). Phosphoinositide signalling in cancer: beyond PI3K and PTEN. *Nat Rev Cancer. 10(5).*, 342-52. doi: 10.1038/nrc2842.

Calligaris, D., et al. (2014). MS imaging for breast cancer margin. *Proceedings of the National Academy of Sciences*, Vol. 111 (42) 15184-15189; DOI: 10.1073/pnas.1408129111.

Campbell P., et al. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature 467*, 1109–1113. https://doi.org/10.1038/nature09460.

Caprioli R.M., et al. (2010). MALDI Imaging Mass Spectrometry – Painting Molecular Pictures. *Malocular Oncology 4*, 529 - 538. https://doi.org/10.1016/j.molonc.2010.09.002.

Casadonte, R., et al. (2014). Imaging mass spectrometry to discriminate breast from pancreatic cancer metastasis in formalin-fixed paraffin-embedded tissues. *Tissue Proteomics and Imaging Mass Spectrometry*, Vol. 14, Issue 7-8, pp. 956 - 964. https://doi.org/10.1002/pmic.201300430.

Cassese A., et al. (2016). Spatial Autocorrelation in Mass Spectrometry Imaging. *Analytical Chemistry 2016 88 (11)*, 5871-5878. doi: 10.1021/acs.analchem.6b00672.

Castellino S., et al. (2011). MALDI imaging mass spectrometry: Bridging biology and chemistry in drug development. *Bioanalysis. Vol. 3, No. 21, 3(21)* , 2427–2441. https://doi.org/10.4155/bio.11.232.

Cayton, L., et al. . (2005). Algorithms for manifold learning. *Univ of California at San Diego Tech Rep*, Vol. 44, pp. 973–980.

Chan E.C.Y., et al. (2009). Metabolic profiling of human colorectal cancer using high-resolution magic angle spinning nuclear magnetic resonance (HR-MAS NMR) spectroscopy and gas chromatography mass spectrometry (GC/MS). *Journal of Proteome Research, 8(1)*, 352–361. doi: 10.1021/pr8006232.

Chang C.L., et al. (1968). Fuzzy topological spaces. *Journal of mathematical Analysis and Applications 24.1*, 182-190.

Chaurio R.A., et al. (2009). Phospholipids: Key players in apoptosis and immune regulation. *Molecules, 14(12)*, 4892–4914. doi: 10.3390/molecules14124892.

Chen L., et al. (2017). CPLA2α mediates TGF-β-induced epithelial-mesenchymal transition in breast cancer through PI3k/Akt signaling. *Cell Death and Disease, 8(4)*, e2728. doi: 10.1038/cddis.2017.152.

Clark A.E., et al. (2013). Matrix-assisted laser desorption ionization-time of flight mass spectrometry: a fundamental shift in the routine practice of clinical microbiology. *Clin Microbiol Rev 26(3)*, 547-603. doi: 10.1128/CMR.00072-12.

Comon, P., et al. . (1994). Independent component analysis: A new concept? *Signal Process*, Vol. 36(3), pp. 287–314.

Coombes K.R., et al. (2005). Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics, 5(16).*, 4107–4117. doi: 10.1002/pmic.200401261.

Cortes C., et al. (1995). Support-Vector Networks. *Machine Learning, 20(3).*, 273–297. doi: 10.1023/A:1022627411411.

Croxatto A., et al. (2012). Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. *FEMS Microbiology Reviews 36, Issue 2.*, 380-407. https://doi.org/10.1111/j.1574-6976.2011.00298.x.

Dang T., et al. (2009). Computing Reachable States for Nonlinear Biological Models. *Computational Methods in Systems Biology. Lecture Notes in Computer Science, vol 5688. Springer, Berlin, Heidel*, 126 - 141. https://doi.org/10.1007/978-3-642-03845-7_9.

Daszykowski, M. and Walczak, B. (2010). Density-Based Clustering Methods. *in Comprehensive Chemometrics*, pp. 635–654. doi: 10.1016/B978-044452701-1.00067-3.

Davenport, T. et al. (2019). The potential for artificial intelligence in healthcare. *Future Healthc J.*, Vol. 6(2), pp. 94-98. doi:10.7861/futurehosp.6-2-94.

David R., et al. (2020). A Deep Learning Convolutional Neural Network Can Recognize Common Patterns of Injury in Gastric Pathology. *Archives of Pathology & Laboratory Medicine Vol. 144, Issue 3.*, 370-378. doi = 10.5858/arpa.2019-0004-OA.

Davies D.L., et al. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2)*, 224–227. doi: 10.1109/TPAMI.1979.4766909.

de Gonzalo-Calvo D., et al. (2015). Intratumor cholesteryl ester accumulation is associated with human breast cancer proliferation and aggressive potential: a molecular and clinicopathological study. *BMC Cancer; 15*, 460. doi: 10.1186/s12885-015-1469-5.

de Oliveira, M.C.F., et al. (2003). From visual data exploration to visual data mining: A survey. *IEEE Trans Visual Comput Graphics*, Vol. 9(3), pp. 378–394.

Deininger S., et al. (2008). MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *J Proteome Res. 7(12). , 7(12)*, 5230–5236. doi:10.1021/pr8005777.

Deininger S.O., et al. (2008). MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *J Proteome Res. 7(12).*, 5230-5236. doi:10.1021/pr8005777.

Devunooru, S., et al. (2021). Deep learning neural networks for medical image segmentation of brain tumours for diagnosis: a recent review and taxonomy. . *J Ambient Intell Human Comput* , Vol. 12, pp. 455–483, https://doi.org/10.1007/s12652-020-01998-w.

Dill, A.L., et al. (2011). Multivariate statistical identification of human bladder carcinomas using ambient ionization imaging mass spectrometry. . *Chemistry. A European Journal.*, Vol. 17(10), pp. 2897 - 2902.

Ding C., et al. (2004). K-means clustering via principal component analysis. *ICML '04: Proceedings of the twenty-first international conference on Machine learning.*, 29-36. https://doi.org/10.1145/1015330.1015408.

Djambazova, K., et al. (2020). Resolving the complexity of spatial lipidomics with MALDI trapped ion mobility spectrometry. *ChemRxiV*, https://doi.org/10.26434/chemrxiv.12331652.v1.

Dobrzyńska I., et al. (2005). Changes in electric charge and phospholipids composition in human colorectal cancer cells. *Molecular and Cellular Biochemistry, 276(1–2)*, 113–119. doi: 10.1007/s11010-005-3557-3.

Eberlin L.S., e. a. (2012). Classifying human brain tumors by lipid imaging with mass spectrometry. *Cancer Research*, Vol. 72(3), pp. 645–654.

Eberlin, L.S., et al. (2014). Molecular assessment of surgical-resection margins of gastric cancer by mass-spectrometric imaging. . *Proceedings of the National Academy of Sciences*, Vol. 111(7), pp. 2436–2441. https://doi.org/10.1073/pnas.1400274111.

Education, Nature. (2014). Retrieved from https://www.nature.com/scitable/topicpage/cell-membranes-14052567/

Elsner, M., et al. (2012). MALDI imaging mass spectrometry reveals COX7A2, TAGLN2 and S100-A10 as novel prognostic markers in Barrett's adenocarcinoma. *Journal of Proteomics*, Vol. 75(15), pp. 4693–4704. http://dx.doi.org/10.1016/j.jprot.2012.02.012.

Ester M., et al. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press*, 226–231. doi: 10.1.1.71.1980.

Esteva, A. et al. (2021). Deep learning-enabled medical computer vision. *Digit. Med. 4, 5*, https://doi.org/10.1038/s41746-020-00376-2.

European Union. (2019). Assessing the impact of digital transformation of heatlh services. *Report of the Expert Panel on effective ways of investing in Health (EXPH). Luxembourg: Publications Office of the European Union.*, doi:10.2875/644722.

Fahy E., et al. (2009). Update of the LIPID MAPS comprehensive classification system for lipids. *The Journal of Lipid Research, 50.*, S9-S14. doi: 10.1194/jlr.R800095-JLR200.

Fares J, et al. (2020). Molecular principles of metastasis: a hallmark of cancer revisited. *Signal Transduct and Targeted Therapy*, 5(1): 28. doi: 10.1038/s41392-020-0134-x.

Flaveny C.A., et al. (2015). Broad anti-tumor ac-tivity of a small molecule that selectively targets the Warburg effect and lipogenesis. *Cancer Cell 2015; 28.*, 42–56. doi: 10.1016/j.ccell.2015.05.007.

Flavin R., et al. (2010). Fatty acid synthase as a potential therapeutic target in cancer. *Future Oncol 2010; 6.*, 551–562. https://doi.org/10.2217/fon.10.11.

Fonville J.M, et al. (2013). Hyperspectral visualization of mass spectrometry imaging data. *Analytical Chemistry 85 (3)., 85(3)*(doi: 10.1021/ac302330a), 1415-1423. doi: 10.1021/ac302330a.

Fonville J.M., et al. (2013). Hyperspectral Visualization of Mass Spectrometry Imaging Data. *Analytical Chemistry 85 (3).*, 1415-1423. doi: 10.1021/ac302330a.

Fonville JM., et al. (2012). Robust data processing and normalization strategy for MALDI mass spectrometric imaging. *Analytical Chemistry 84 (3)., 84(3),*(doi: 10.1021/ac201767g), 1310-1319. doi: 10.1021/ac201767g.

Frisoni G.B., et al. (2006). In vivo neuropathology of the hippocampal formation in AD: a radial mapping MR-based study. *NeuroImage Vol. 32, Issue 1.*, 104-110. https://doi.org/10.1016/j.neuroimage.2006.03.015.

Fullerton A.S., et al. (2009). A Conceptual Framework for Ordered Logistic Regression Models. *Sociological Methods & Research, 38(2).*, 306–347. https://doi.org/10.1177/0049124109346162.

Garcia-Milian R. (2014). *What is difference between tumorigenesis and carcinogenesis?* Retrieved from https://www.researchgate.net/post/What_is_difference_between_tumorigenesis_and_carcinogenesis/533e87b3d039b1941d8b4636/citation/download

Gerbig S., et al. (2012). Analysis of colorectal adenocarcinoma tissue by desorption electrospray ionization mass spectrometric imaging. *Analytical and Bioanalytical Chemistry, 403(8)*, 2315–2325. doi: 10.1007/s00216-012-5841-x.

Gibb S, et al. (2012). MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics, Vol. 28, Issue 17.*, 2270–2271. https://doi.org/10.1093/bioinformatics/bts447.

Gibb S., et al. (2012). Maldiquant: A versatile R package for the analysis of mass spectrometry data. *Bioinformatics Vol.28, Issue 17., 28(17)*, 2270–2271. https://doi.org/10.1093/bioinformatics/bts447.

Gisbrecht A., et al. (2015). Parametric nonlinear dimensionality reduction using kernel t-SNE. *Neurocomputing 147*, 71–82. doi:10.1016/j.neucom.2013.11.045.

Guenther, S., et al. (2015). Spatially resolved metabolic phenotyping of breast cancer by desorption electrospray ionization mass spectrometry. . *Cancer Research*, Vol. 75(9), pp. 1828–183. http:// dx.doi.org/10.1158/0008-5472.can-14-2258.

Guillaumond F., et al. (2015). Cholesterol uptake disruption, in association with chemotherapy, is a promising combined metabolic therapy for pancreatic adenocarcinoma. *Proc Natl Acad Sci USA 2015; 112*, 2473–2478. doi: 10.1073/pnas.1421601112.

Gustafsson J.O.R., et al. (2011). MALDI imaging mass spectrometry (MALDI-IMS)-application of spatial proteomics for ovarian cancer classification and diagnosis. (M. D. International, Ed.) *Int J Mol Sci. 12(1)*, 773-794. doi:10.3390/ijms12010773.

Guyon I., et al. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research 3*, 1157-1182. https://www.jmlr.org/papers/v3/guyon03a.html. Retrieved from The Journal of Machine Learning Research, 3(3), pp. 1157–1182.: http://www.jmlr.org/papers/v3/guyon03a.html

Hall M. (1999). Correlation-based Feature Selection for Machine Learning, Methodology. *PhD Thesis - University of Waikato*. Retrieved from The University of Waikato: https://www.lri.fr/~pierres/donn%E9es/save/these/articles/lpr-queue/hall99correlationbased.pdf

Handl J., et al. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics, Vol. 21, Issue 15.*(doi: 10.1093/bioinformatics/bti517), 3201–3212. https://doi.org/10.1093/bioinformatics/bti517.

Hannun Y., et al. (2008). Principles of bioactive lipid signalling: lessons from sphingolipids. *Nat Rev Mol Cell Biol 9*, 139–150. https://doi.org/10.1038/nrm2329.

Hanrieder, J., et al. (2014). Probing the lipid chemistry of neurotoxininduced hippocampal lesions using multimodal imaging mass spectrometry. *Surf Interface Anal*, Vol. 46(S1), pp. 375–378.

Hanselmann M., et al. (2009). Toward digital staining using imaging mass spectrometry and random forests. *J Proteome Res. 8(7).*, 3558–3567. doi: 10.1021/pr900253y.

Hao Z., et al. (2020). Big Data and Artificial Intelligence Modeling for Drug Discovery. *Annual Review of Pharmacology and Toxicology, Vol. 60:1*, 573-589. https://doi.org/10.1146/annurev-pharmtox-010919-023324.

Hashimoto H., et al. (2004). Quantitative TOF-SIMS imaging of DNA microarrays produced by bubble jet printing technique and the role of TOF-SIMS in life science industry. *Applied Surface Science Volumes 231–232.*, 385–391. https://doi.org/10.1016/j.apsusc.2004.03.106.

Hatzivassiliou G., e. a. (2005). ATP citrate lyase inhibition can suppress tumor cell growth. *Cancer Cell 2005; 8.*, 311–321. https://doi.org/10.1016/j.ccr.2005.09.008.

He P.Q., et al. (2011). Self-calibrated warping for mass spectra alignment. *Cancer Informatics, 10,*, pp. 65–82. doi: 10.4137/CIN.S6358.

Heijs, B., et al. (2020). Molecular signatures of tumor progression in myxoid liposarcoma identified by N-glycan mass spectrometry imaging. . *Lab Invest*, Vol. 100, pp. 1252–1261. https://doi.org/10.1038/s41374-020-0435-2.

Henderson, A., et al. . (2009). A comparison of PCA and MAF for ToF-SIMS image interpretation. *Surf Interface Anal.*, VOl. 41(8), pp. 666–674.

Hensher D.A., et al. (2003). The Mixed Logit model: The state of practice. *Transportation 30*, 133–176. https://doi.org/10.1023/A:1022558715350.

Hillenkamp F., et al. (1991). Matrix-Assisted Laser Desorption/lonization Mass Spectrometry of Biopolymers. *Analytical Chemistry 63 (24)., doi: 10.1021/ac00024a002*, 1193A-1203A. doi: 10.1021/ac00024a002.

Hinton , E. et al. (2007). Boltzmann machine. *Scholarpedia*, 2(5):1668.

Hinton G E., et al. (2006). Reducing the dimensionality of data with neural networks. *Science 313(5786)*, 504–507. doi: 10.1126/science.1127647.

Hinton G. E., et al. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation. MIT Press, 14(8)*, 1771–1800. doi: 10.1162/089976602760128018.

Hinton G.E, e. a. (2006). A fast learning algorithm for deep belief nets. *Neural Computation. MIT Press 238 18(7)*, 1527–1554. doi: 10.1162/neco.2006.18.

Hinton G.E., et al. (2012). A practical guide to training restricted Boltzmann machines. *Neural Networks: Tricks of the Trade*, 599–619. doi: 10.1007/978-3-642-35289-8-32.

Hinton GE, et al. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science Vol. 313, Issue 5786*, 504-507. doi: 10.1126/science.1127647.

Hinton, E et al. (2014). Dropout a simple way to prevent neural networks from overfitting. *The journal of machine learning research,*, 15(1), 1929-1958.

Hinton, E. et al. (2009). Deep belief networks. . *Scholarpedia,* , 4(5):5947.

Ho T.K., et al. (1995). Random Decision Forests. *IEEE Computer Society. Proceedings of the Third International Conference on Document Analysis and Recognition Vol. 1*, 278. doi: 10.1109/ICDAR.1995.598994.

Honig R.E. (1958). Sputtering of surfaces by positive ion beams of low energy. *Journal of Applied Physics 29., 29(3)*, 549–555. https://doi.org/10.1063/1.1723219.

Hopfield J.J., et al. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences. National Acad Sciences. 79(8)*, 2554–2558. doi: 10.1073/pnas.79.8.2554.

Inglese P, et al. (2017). Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer. *Chem. Sci., 8*, 3500-3511. doi: 10.1039/C6SC03738K.

Inglese P., et al. (2019). SPUTNIK: an R package for filtering of spatially related peaks in mass spectrometry imaging data. *Bioinformatics 35(1).*(doi: 10.1093/bioinformatics/bty622), 178-180. doi: 10.1093/bioinformatics/bty622.

Inglese, P. & Strittmatter. (2018). Network analysis of mass spectrometry imaging data from colorectal cancer identifies key metabolites common to metastatic development. *bioRxiv 230052*, doi: 10.1101/230052.

Ishibashi Y., et al. (2013). New insights on glucosylated lipids: Metabolism and functions. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids, Vol. 1831, Issue 9*, 1475-1485. https://doi.org/10.1016/j.bbalip.2013.06.001.

Islam Md S., et al. (2018). A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining. *Healthcare 6(2).*, 54. https://doi.org/10.3390/healthcare6020054.

Islam, M.M., et al. (2021). A Review on Deep Learning Techniques for the Diagnosis of Novel Coronavirus (COVID-19). *IEEE*, Vol. 9, pp. 30551-30572, doi: 10.1109/ACCESS.2021.3058537.

Javed A., et al. (2007). On Estimating the Size and Confidence of a Statistical Audit. *Accurate Electronic Voting Technology 7*, 8.

Jinyi L., et al. (2018). Ganglioside GD3 synthase (GD3S), a novel cancer drug target. *Acta Pharm Sin B; 8(5).*, 713–720. doi: 10.1016/j.apsb.2018.07.009.

Jolliffe, I., et al. (2002). Principal component analysis. 2nd ed. *New York: Springer-Verlag New York, Inc.*

Jones E.A., et al. (2013). Imaging mass spectrometry-based molecular histology differentiates microscopically identical and heterogeneous tumors. *J Proteome Res. 12(4).*, 1847-1855. doi:10.1021/pr301190g.

Jones, E.A., et al. (2011). Multiple statistical analysis techniques corroborate intratumor heterogeneity in imaging mass spectrometry datasets of myxofibrosarcoma. *PLoS One.*, Vol. 6(9). e24913. http://dx.doi.org/10.1371/journal.pone.0024913.

Junfeng An, et al. (2022). Cognitive multi-modal consistent hashing with flexible semantic transformation. *Information Processing & Management*, Vol. 59, Issue 1. https://doi.org/10.1016/j.ipm.2021.102743.

Jutten, C., et al. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, Vol. 24(1), pp. 1–10.

Kamnitsas K. et al. . (2018). DeepMedic for Brain Tumor Segmentation. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2016.*, Lecture Notes in Computer Science, vol 10154. Springer, Cham.

Khatib-Shahidi, S., et al. (2006). Direct molecular analysis of whole-body animal tissue sections by imaging MALDI mass spectrometry. *Analytical Chemistry.*, Vol. 78(18), pp. 6448–6456.

Kitatani K., et al. (2016). Ceramide limits phosphatidylinositol-3-kinase C2β-controlled cell motility in ovarian cancer: potential of ceramide as a metastasis-suppressor lipid. *Oncogene*, 2801-12. doi: 10.1038/onc.2015.330.

Kleinberg E M., et al. (2000). On the algorithmic implementation of stochastic discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 5.*, 473-490. doi: 10.1109/34.857004.

Kleinberg E.M., et al. (1996). An overtraining-resistant stochastic modeling method for pattern recognition. *Ann. Statist. 24, no. 6*, 2319-2349. doi:10.1214/aos/1032181157.

Koshiji M., et al. (1998). Apoptosis of colorectal adenocarcinoma (COLO 201) by tumour necrosis factor-alpha (TNF-α) and/or interferon-gamma (IFN-γ), resulting from down- modulation of Bcl-2 expression. *Clinical and Experimental Immunology, 111(1).*, 211–218. doi: 10.1046/j.1365-2249.1998.00460.x.

Kottke P.A., et al. (2010). The Scanning Mass Spectrometry Probe: A Scanning Probe Electrospray Ion Source for Imaging Mass Spectrometry of Submerged Interfaces and Transient Events in Solution. *Analytical Chemistry 82 (1).*, 19-22. doi: 10.1021/ac902102x.

Koundouros N., et al. (2020). Metabolic Fingerprinting Links Oncogenic PIK3CA with Enhanced Arachidonic Acid-Derived Eicosanoids. *Cell. 181(7)*, 1596-1611. doi:10.1016/j.cell.2020.05.053.

Kriegsmann J., e. a. (2015). MALDI TOF imaging mass spectrometry in clinical pathology: a valuable tool for cancer diagnostics (review). *Int J Oncol.*, Vol. 46(3), pp. 893-906. doi: 10.3892/ijo.2014.2788. PMID: 25482502.

Kunkel G.T., et al. (2013). Targeting the sphingosine-1-phosphate axis in cancer, inflammation and beyond. *Nat Rev Drug Discov 2013; 12*, 688–702. https://doi.org/10.1038/nrd4099.

LaBonia G.J., et al. (2016). Drug penetration and metabolism in 3D cell cultures treated in a 3D printed fluidic device: assessment of irinotecan via MALDI imaging mass spectrometry. *Proteomics 16(11-12).*, 1814–1821. doi: 10.1002/pmic.201500524.

Larochelle H., et al. (2009). Exploring Strategies for Training Deep Neural Networks. *Journal of Machine Learning Research 10(1)*, 1–40. doi:10.1109/Tsmcc.2012.2220963.

Laurinavicius A., et al. (2012). Digital Image Analysis in Pathology: Benefits and Obligation. Analytical Cellular Pathology.

Lazova, R., et al. (2012). Imaging mass spectrometry - A new and promising method to differentiate Spitz nevi from Spitzoid malignant melanomas. *The American Journal of Dermatopathology*, Vol. 34(1), pp. 82–90. http://dx.doi.org/10.1097/DAD.0b013e31823df1e2.

Lea J., et al. (2017). Detection of phosphatidylserine-positive exosomes as a diagnostic marker for ovarian malignancies: a proof of concept study. *Oncotarget 8(9)*, 14395-14407. doi: 10.18632/oncotarget.14795.

Lee, D.D., et al. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, Vol. 401(6755), pp. 788–791.

Leow A.D., et al. (2009). Alzheimer's disease neuroimaging initiative: a one-year follow up study using tensor-based morphometry correlating degenerative rates, biomarkers and cognition. *Neuroimage 45(3).*, 645-55. doi: 10.1016/j.neuroimage.2009.01.004.

Lerch J.P., et al. (2008). Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls. *Neurobiology of Aging, Vol. 29, Issue 1.*, 23-30. https://doi.org/10.1016/j.neurobiolaging.2006.09.013.

Li K., et al. . (2020). Microsatellite instability: a review of what the oncologist should know. *Cancer Cell Int.*, 20, 16. https://doi.org/10.1186/s12935-019-1091-8.

Li S., e. a. (2013). TOFA suppresses ovarian cancer cell growth in vitro and in vivo. *Mol Med Rep 2013; 8.*, 373–378. https://doi.org/10.3892/mmr.2013.1505.

Li, X., et al. . (2021). The immunological and metabolic landscape in primary and metastatic liver cancer. *Nat Rev Cancer 21*, PP. 541–557. https://doi.org/10.1038/s41568-021-00383-9.

Liebl H. (1967). Ion microprobe mass analyzer. *Journal of Applied Physics 38., 38(13)*, 5277–5283. https://doi.org/10.1063/1.1709314.

Lin Y., et al. (2006). Random Forests and Adaptive Nearest Neighbors. *Journal of the American Statistical Association, 101:474.*, 578-590. doi: 10.1198/016214505000001230.

Liu Y.M., et al. (1997). *Fuzzy topology. Vol. 9.* World Scientific.

Lockwood S.Y., et al. (2016). Drug penetration and metabolism in 3D cell cultures treated in a 3D printed fluidic device: assessment of irinotecan via MALDI imaging mass spectrometry. *Proteomics 16 (11-12), 16(11–12)*, 1814–1821. doi:10.1002/pmic.201500524.

Lotz J.M., et al. (2017). Integration of 3D multimodal imaging data of a head and neck cancer and advanced feature recognition. *Biochim Biophys Acta Proteins Proteom. 1865(7)*, 946-956. doi:10.1016/j.bbapap.2016.08.018.

Luo X, et al. (2017). Emerging roles of lipid metabolism in cancer metastasis. *Mol Cancer 16.*, 76. https://doi.org/10.1186/s12943-017-0646-3.

Luo X., et al. (2018). The implications of signaling lipids in cancer metastasis. *Exp Mol Med 50.*, 127. https://doi.org/10.1038/s12276-018-0150-x.

Lysaght, T., et al. (2019). AI-Assisted Decision-making in Healthcare. . *ABR*, Vol. 11, pp. 299–314. https://doi.org/10.1007/s41649-019-00096-0.

Maintz J.B., et al. (1998). A survey of medical image registration. *Medical Image Analysis Vol. 2, Issue 1.*, 1-36. https://doi.org/10.1016/S1361-8415(01)80026-8.

Mandal M.K., et al. (2013). Biomolecular analysis and cancer diagnostics by negative mode probe electrospray ionization. *Analyst, 138(6).*, 1682–1688. doi: 10.1039/c3an36554a.

Manjarika D., et al. (2018). A Novel Therapeutic Strategy for Cancer Using Phosphatidylserine Targeting Stearylamine-Bearing Cationic Liposomes. *Molecular Therapy - Nucleic Acids Vol. 10*, 9-27. doi:https://doi.org/10.1016/j.omtn.2017.10.019.

Marte B, et al. (2013). Tumour Heterogeneity. *Nature 501*, 327. https://doi.org/10.1038/501327a.

Martinez-Outschoorn U.E., et al. (2017). Cancer metabolism: a therapeutic perspective. *Nat Rev Clin Oncol 14.*, 11–31 (2017). https://doi.org/10.1038/nrclinonc.2016.60.

Marusyk A., et al. (2010). Tumor heterogeneity: causes and consequences. *Biochim Biophys Acta. 1805(1).*, 105. doi: 10.1016/j.bbcan.2009.11.002.

Matrisian L.M., et al. (2016). The Past, Present, and Future of Pancreatic Cancer Clinical Trials. *Am Soc Clin Oncol Educ Book 35*, e205-15. doi: 10.14694/EDBK_159117.

McCombie G., et al. (2005). Spatial and Spectral Correlations in MALDI Mass Spectrometry Images by Clustering and Multivariate Analysis. *Analytical Chemistry 77 (19).*, 6118-6124. doi: 10.1021/ac051081q.

McDonnell, L. A., et al. (2010). Peptide and protein imaging mass spectrometry in cancer research. *Journal of Proteomics*, Vol. 73(10), pp. 1921–1944. http://dx.doi.org/10.1016/j.jprot.2010.05.007.

McDowell, C.T., et al. (2021). Imaging Mass Spectrometry and Lectin Analysis of N-Linked Glycans in Carbohydrate Antigen–Defined Pancreatic Cancer Tissues. *Molecular & Cellular Proteomics.*, Vol. 20, 100012, ISSN 1535-9476, https://doi.org/10.1074/mcp.RA120.002256.

Micalizzi, D.S., et al. (2021). Translational Regulation of Cancer Metastasis. *Cancer Res.*, 81(3), pp. 517-524. doi: 10.1158/0008-5472. CAN-20-2720. PMID: 33479028; PMCID: PMC7854484.

Milstien S., et al. (2006). Targeting sphingo-sine-1-phosphate: a novel avenue for cancer therapeutics. *Cancer Cell 9.*, 148-150. doi:10.1016/j.ccr.2006.02.025.

Misale S. et al. (2012). Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature 486*, (7404):532-6. doi: 10.1038/nature11156.

Morad S.A., et al. (2016). Short-chain ceramides depress integrin cell surface expression and function in colorectal cancer cells. *Cancer Letters 376.*, 199–204. https://doi.org/10.1016/j.canlet.2016.03.049.

Morris J.H., et al. (2011). ClusterMaker: A multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics, 12(1)*, 436. doi: 10.1186/1471-2105-12-436.

Mostavi M., et al. (2020). Convolutional neural network models for cancer type prediction based on gene expression. *BMC Medical Genomics, 13(Suppl 5).*, 44. https://doi.org/10.1186/s12920-020-0677-2.

Munz E.D., et al. (2017). Psychotherapie in der Psychiatrie. *Nervenheilkunde*, 800–805. doi: 10.1007/s13398-014-0173-7.2.

Nakanishi M., et al. (2013). Multifaceted roles of PGE2 in inflammation and cancer. *Semin Immunopathol 2013; 35.*, 123–137. https://doi.org/10.1007/s00281-012-0342-8.

Nemes P., et al. (2007). Laser Ablation Electrospray Ionization for Atmospheric Pressure, in Vivo, and Imaging Mass Spectrometry. *Analytical Chemistry 79 (21).*, 8098-8106. doi: 10.1021/ac071181r.

Ness R.O., et al. (2016). From Correlation to Causality: Statistical Approaches to Learning Regulatory Relationships in Large-Scale Biomolecular Investigations. *Journal of Proteome Research 2016 15 (3).*, 683-690. doi: 10.1021/acs.jproteome.5b00911.

Norris J.L., et al. (2013). Analysis of tissue specimens by matrix-assisted laser desorption/ionization imaging mass spectrometry in biological and clinical research. *Chemical Reviews 113 (4)., 113(4)*, 2309–2342. doi:10.1021/cr3004295.

O'Donnell V.B., et al. (2018). Phospholipid signaling in innate immune cells. *The Journal of Clinical Investigation 128(7).*, 2670–2679. https://doi.org/10.1172/JCI97944.

Oppenheimer S.R., et al. (2010). A Molecular Analysis of Tumor Margins by MALDI Mass Spectrometry in Renal Carcinoma. *J Proteome Res. 9(5).* , 2182–2190. doi: 10.1021/pr900936z.

Otsu N., et al. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics, 9(1).*, 62–66. doi: 10.1109/TSMC.1979.4310076.

Palmer A., et al. (2016). FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nat Methods 14., 14(1)*(doi: 10.1038/nmeth.4072), 57–60. https://doi.org/10.1038/nmeth.4072.

Palubeckaitė, I., et al. (2020). Mass spectrometry imaging of endogenous metabolites in response to doxorubicin in a novel 3D osteosarcoma cell culture model. *Journal of Mass Spectrometry*, Vol. 55, Issue 4, Sn. 1076-5174. https://doi.org/10.1002/jms.4461.

Park, J.W., et al. (2009). Multivariate analysis of ToF-SIMS data for biological applications. *Surf Interface Anal*, Vol. 41(8), pp. 694–703.

Passarelli K., et al. (2015). Single-Cell Analysis: Visualizing Pharmaceutical and Metabolite Uptake in Cells with Label-Free 3D Mass Spectrometry Imaging. *Analytical Chemistry 87 (13)., 87(13)*, 6696-6702. doi: 10.1021/acs.analchem.5b00842.

Passarelli M.K., et al. (2013). Single-cell imaging mass spectrometry. *Current Opinion in Chemical Biology. Vol. 17, Issue 5., 17(5)*, 854–859. https://doi.org/10.1016/j.cbpa.2013.07.017.

Patmanathan S.N., et al. (2016). Aberrant expression of the S1P regulating enzymes, SPHK1 and SGPL1, contributes to a migratory phenotype in OSCC mediated through S1PR2. *Sci. Rep. 6.*, 25650. doi: 10.1038/srep25650.

Pavlidis, N., et al. (2016). Cancer of unknown primary site. *Proteomics 14(7-8)*, doi: 10.1093/med/9780199656103.003.0059.

Peng H., et al. (2005). Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8)*(doi: 10.1109/TPAMI), 1226–1238. doi: 10.1109/TPAMI.2005.159.

Peng, Y., et al. (2019). MALDI-TOF MS for the rapid identification and drug susceptibility testing of filamentous fungi. *Exp Ther Med*, vol. 18, pp. 4865-4873. https://doi.org/10.3892/etm.2019.8118.

Peres-Neto, P.R., et al. (2005). How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Comput Statistics Data Anal.*, Vol. 49(4), pp. 974–997.

Perry R.H., et al. (2013). Characterization of MYC-induced tumor-igenesis by in situ lipid profiling. *Anal Chem 2013; 85.*, 4259–4262. doi: 10.1021/ac400479j.

Presa N., et al. (2016). Regulation of cell migration and inflammation by ceramide 1- phosphate. *Biochim. Bio-phys. Acta 1861.*, 402–409. https://doi.org/10.1016/j.bbalip.2016.02.007.

Public Health England. (2019). Matrix-assisted laser desorption/ionisation - time of flight mass spectrometry (MALDI-TOF MS) test procedure. *Standards Unit, Microbiology Services, PHE. Issue no: 1.1.*, 1 - 22.

Qin, L. & Z., et al. (2018). Recent advances in matrix-assisted laser desorption/ionisation mass spectrometry imaging (MALDI-MSI) for in situ analysis of endogenous molecules in plants. . *Phytochemical Analysis.*, Vol. 29. doi:10.1002/pca.2759.

Qiu B., et al. (2015). HIF2alpha-dependent lipid storage promotes endoplasmic reticulum homeostasis in clear-cell renal cell carcinoma. *Cancer Discov 2015, 5.*, 652–667. doi: 10.1158/2159-8290.CD-14-1507.

Quehenberger O., et al. (2011). The human plasma lipidome. *The new England Journal of Medicine 365*, 1812-1823. doi: 10.1056/NEJMra1104901.

Quentin, P.V., et al. (2015). Time-of-flight secondary ion mass spectrometry imaging of biological samples with delayed extraction for high mass and high spatial resolutions. *Rapid Communications in Mass Spectrometry*, Vol. 29, Issue 13, pp. 1187-1195. https://doi.org/10.1002/rcm.7210.

Race A.M., et al. (2016). SpectralAnalysis: Software for the Masses. *Analytical Chemistry 88(19)., 88(19)*(doi: 10.1021/acs.analchem.6b01643), 9451–9458. doi: 10.1021/acs.analchem.6b01643.

Ràfols P., et al. (2017). rMSI: An R package for MS imaging data handling and visualization. *Bioinformatics, Vol. 33, Issue 15., 33(15)*(doi: 10.1093/bioinformatics/btx182), 2427–2428. https://doi.org/10.1093/bioinformatics/btx182.

Rand W. M., et al. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association, 66:336.*, 846-850, doi: 10.1080/01621459.1971.10482356.

Rauser S., et al. (2010). Classification of HER2 receptor status in breast cancer tissues by MALDI imaging mass spectrometry. *Journal of Proteome Research 9(4)., 9(4)*, 1854–1863. doi: 10.1021/pr901008d.

Raymond J., et al. (2015). Extracellular Vesicles Present in Human Ovarian Tumor Microenvironments Induce a Phosphatidylserine-Dependent Arrest in the T-cell Signaling Cascade. *Cancer Immunol Res November 1 (3) (11)*, 1269-1278. doi: 10.1158/2326-6066.CIR-15-0086.

Richens, J.G., et al. (2020). Improving the accuracy of medical diagnosis with causal machine learning. *Nat Commun*, Vol. 11, 3923. https://doi.org/10.1038/s41467-020-17419-7.

Rohner, T.C., et al. (2005). MALDI mass spectrometric imaging of biological tissue sections. *Mechanisms of Ageing and Development.*, Vol. 126(1), pp. 177–185.

Römpp A., et al. (2015). Current trends in mass spectrometry imaging. (S. B. Heidelberg, Ed.) *Anal Bioanal Chem 407., doi: 10.1007/s00216-015-8479-7.*, 2023–2025. https://doi.org/10.1007/s00216-015-8479-7.

Roweis, S.T., et al. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* , Vol. 290(5500), pp. 2323–2326.

Rubakhin S.S., et al. (2005). Imaging mass spectrometry: Fundamentals and applications to drug discovery. *Drug Discovery Today Volume 10, Issue 12., 10(12)*, 823–837. https://doi.org/10.1016/S1359-6446(05)03458-6.

Ruben, D. A., et al. (2015). Current State and Future Challenges of Mass Spectrometry Imaging for Clinical Research. *Anal Chem.*, 87(13), pp. 6426-33.

Rysman E., et al. (2010). De novo lipogenesis protects cancer cells from free radicals and chemo-therapeutics by promoting membrane lipid saturation. *Cancer Res 2010; 70*, 8117–8126. doi: 10.1158/0008-5472.CAN-09-3871.

Sarkari, S., et al. (2014). Comparison of clustering pipelines for the analysis of mass spectrometry imaging data. *Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference.*, pp. 4771–4774. http://dx.doi.org/10.1109/embc.2014.6944691.

Schone, C., et al. (2013). MALDI imaging mass spectrometry in cancer research: Combining proteomic profiling and histological evaluation. *Clinical Biochemistry*, Vol. 46(6), pp. 539–545. http://dx.doi.org/10.1016/j.clinbiochem.2013.01.018.

Schramm T., et al. (2012). ImzML - A common data format for the flexible exchange and processing of mass spectrometry imaging data. *Journal of Proteomics, 75(16).*, 5106–5110. doi: 10.1016/j.jprot.2012.07.026.

Schuff N., et al. (2009). MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. *Brain, Vol. 132, Issue 4.*, 1067–1077. https://doi.org/10.1093/brain/awp007.

Schwartz S. A., et al. (2003). Direct tissue analysis using matrix-assisted laser desorption/ionization mass spectrometry: Practical aspects of sample preparation. *J Mass Spectrom. 38(7), 38(7)*, 699–708. doi:10.1002/jms.505.

Selves J., et al. (2018). Immunohistochemistry for Diagnosis of Metastatic Carcinomas of Unknown Primary Site. *Cancers (Basel).*, Vol. 10(4):108. doi:10.3390/cancers10040108.

Shen D., et al. (2002). HAMMER: hierarchical attribute matching mechanism for elastic registration. *IEEE Transactions on Medical Imaging, Vol. 21, no. 11.*, 1421-1439. doi: 10.1109/TMI.2002.803111.

Shroff E.H., et al. (2015). MYC oncogene overexpression drives renal cell carci-noma in a mouse model through glutamine metabolism. *Proc Natl Acad Sci USA 2015; 112*, 6539–6544. https://doi.org/10.1073/pnas.1507228112.

Singhal N., et al. (2015). MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis . *Frontiers in Microbiology*, Vol. 6, pp. 791, doi=10.3389/fmicb.2015.00791 .

Sinkala M., et al. (2020). Machine Learning and Network Analyses Reveal Disease Subtypes of Pancreatic Cancer and their Molecular Characteristics. *Sci Rep 10*, 1212. https://doi.org/10.1038/s41598-020-58290-2.

Smets T., et al. (2019). Evaluation of Distance Metrics and Spatial Autocorrelation in Uniform Manifold Approximation and Projection Applied to Mass Spectrometry Imaging Data. *Analytical Chemistry 91 (9),, 91(9)*(DOI: 10.1021/acs.analchem.8b05827), 5706-5714. doi: 10.1021/acs.analchem.8b05827.

Smolensky P., et al. (1986). Information processing in dynamical systems: Foundations of harmony theory. *in Parallel Distributed Processing Explorations in the Microstructure of Cognition. COLORADO UNIV AT BOULDER DEPT OF COMPUTER SCIENCE*(doi: 10), 194–281.

Smyth G.K., et al. (2005). limma: Linear Models for Microarray Data. *in Bioinformatics and Computational Biology Solutions Using R and Bioconductor.*, 397–420. doi: 10.1007/0-387-29362-0_23.

ssdaa. (n.d.). *dad.* fqqf.

Stanta G., et al. (2016). Tumour heterogeneity: principles and practical consequences. *Virchows Arch. 469(4).*, 371-384. doi:10.1007/s00428-016-1987-9.

Staubach S., et al. (2011). Lipid rafts: signaling and sorting platforms of cells and their roles in cancer. *Expert Rev Proteomics; 8.*, 263–277. doi: 10.1586/epr.11.2.

Stoeckli, M., et al. (2007). Compound and metabolite distribution measured by MALDI mass spectrometric imaging in whole-body tissue sections. *International Journal of Mass Spectrometry.*, Vol. 260(2–3), pp. 195–202.

Swales J.G., et al. (2018). Quantitation of Endogenous Metabolites in Mouse Tumors Using Mass-Spectrometry Imaging. *Analytical Chemistry 90 (10)., 90(10)*, 6051–6058. doi: 10.1021/acs.analchem.7b05239.

Switzer, P. et al. (1984). Min/Max Autocorrelation Factors for Multivariate Spatial Imagery. Ref. SWINSF6.

Takáts Z., et al. (2005). Ambient mass spectrometry using desorption electrospray ionization (DESI): Instrumentation, mechanisms and applications in forensics, chemistry, and biology. *Journal of Mass Spectrometry. Vol. 40, Issue10. , 40(10)*, 1261-1275. https://doi.org/10.1002/jms.922.

Tao Y., et al. (2011). Rapid growth of a hepatocellular carcinoma and the driving mutations revealed by cell-population genetic analysis of whole-genome data. *PNAS, 108 (29).*, 12042-12047. https://doi.org/10.1073/pnas.1108715108.

Taylor A.J., et al. (2018). Exploring Ion Suppression in Mass Spectrometry Imaging of a Heterogeneous Tissue. *Analytical Chemistry 90 (9)., 90(9)*, 5637-5645. doi: 10.1021/acs.analchem.7b05005.

Tenenbaum, J.B., et al. (2000). A global geometric framework for nonlinear dimensionality reduction. Science. *Science*, Vol. 290(5500), pp. 2319–2323.

Thomas S.A., et al. (2016). Dimensionality reduction of mass spectrometry imaging data using autoencoders. *IEEE Symposium Series on Computational Intelligence (SSCI).*(doi: 10.1109/SSCI.2016.7849863), 1-7. doi: 10.1109/SSCI.2016.7849863.

Thompson P.M., et al. (2007). Tracking Alzheimer's disease. *Ann N Y Acad Sci 1097*, 183–214. doi: 10.1196/annals.1379.017.

Tillner J., et al. (2016). Investigation of the Impact of Desorption Electrospray Ionization Sprayer Geometry on Its Performance in Imaging of Biological Tissue. *Analytical Chemistry 88 (9)., 88(9), .* , 4808–4816. doi:10.1021/acs.analchem.6b00345.

Tillner, J., et al. (2017). Faster, More Reproducible DESI-MS for Biological Tissue Imaging. *J Am Soc Mass Spectrom*, Vol. 28(10), pp. 2090-2098, doi:10.1007/s13361-017-1714-z.

Tom D. Bunney and Katan M. (2010). Phosphoinositide signalling in cancer: beyond PI3K and PTEN. *Nature Reviews Cancer volume 10*, 342–352.

Trim, P.J., et al. (2008). Matrix-assisted laser desorption/ionization-ion mobility separation-mass spectrometry imagingofvinblastine inwhole body tissue sections. *Analytical Chemistry*, Vol. 80(22), pp. 8628 - 8634.

Tyler, B.J., et al. (2007). Multivariate analysis strategies for processing ToF-SIMS images of biomaterials. . *Biomaterials*, Vol. 28(15), pp. 2412–2423.

Van de Plas, R., et al. (2007b). Prospective Exploration of Biochemical Tissue Composition via Imaging Mass Spectrometry Guided by Principal Component Analysis. Paper presented at: Pacific Symposium on Biocom. *Paper presented at: Pacific Symposium on Biocomputing 2007*.

Van Der Maaten L, et al. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research 9.*, 2579–2605. doi: 10.1007/s10479-011-0841-3.

Van Der Maaten L., et al. (2009). Learning a Parametric Embedding by Preserving Local Structure. *Artificial Intelligence and Statistics, PMLR 5*, 384–391. .

van Meer G., et al. (2008). Membrane lipids: where they are and how they behave. *Nat Rev Mol Cell Biol 9*, 112–124. https://doi.org/10.1038/nrm2330.

Vapnik, V. (1963). Pattern Recognition Using Generalized Portrait Method. *Automation and Remote Control*, pp. 774-780.

Verbeeck, N., Caprioli, R.M. et al. (2020). Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. *Mass Spec Rev.*, Vol. 39, pp. 245-291. https://doi.org/10.1002/mas.21602.

Verbeeck, N., et al. (2019). Unsupervised Machine Learning for Exploratory Data Analysis in Imaging Mass Spectrometry. *Wiley Online Library (wileyonlinelibrary.com).*, DOI 10.1002/mas.21602.

Veselkov K., et al. (2018). BASIS: High-performance bioinformatics platform for processing of large-scale mass spectrometry imaging data in chemically augmented histology. *Sci Rep 8, 8(1)*(doi: 10.1038/s41598-018-22499-z), 4053. https://doi.org/10.1038/s41598-018-22499-z.

Veselkov K.A., et al. (2011). Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Analytical Chemistry, 83(15).*, 5864–5872. doi: 10.1021/ac201065j.

Vickerman J.C., et al. (2011). *Surface Analysis - The Principal Techniques: Second Edition.* Wiley.

Von Luxburg U., et al. (2012). Clustering: Science or Art? *JMLR: Workshop and Conference Proceedings 27*, 65-79. Available at: http://proceedings.mlr.press/v27/luxburg12a.html.

Vranic S., et al. (2021). The Role of Pathology in the Era of Personalized (Precision) Medicine: A Brief Review. *Acta Med Acad.*, Vol. 50(1), pp. 47-57. doi: 10.5644/ama2006-124.325. PMID: 34075763.

Wagner M.S., et al. (2008). Towards quantitative chemical imaging with ToF-SIMS. *Applied Surface Science Vol. 255, Issue 4., 255(4)*, 992–996. https://doi.org/10.1016/j.apsusc.2008.05.037.

Wang D, e. a. (2010). The role of COX-2 in intestinal inflammation and colorectal cancer. *Oncogene 29.*, 781–788. https://doi.org/10.1038/onc.2009.421.

Wang D., et al. (2010). Eicosanoids and cancer. *Nat Rev Cancer 2010; 10.*, 181-193. https://doi.org/10.1038/nrc2809.

Wang, S., et al. (2021). A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19). *Eur Radiol*, https://doi.org/10.1007/s00330-021-07715-1.

Wijetunge C.D., et al. (2014). Unsupervised learning for exploring MALDI imaging mass spectrometry "omics" data. *Colombo*, 1-6. doi: 10.1109/ICIAFS.2014.7069634.

Wikipedia contributors. (2022, March 13). *Rand index.* Retrieved from Wikipedia, The Free Encyclopedia: https://en.wikipedia.org/w/index.php?title=Rand_index&oldid=1076943999

Wu X., et al. (2012). Clonal Selection Drives Genetic Divergence of Metastatic Medulloblastoma. *Nature 482*, 529–533. https://doi.org/10.1038/nature10825.

Wu, M., et al. (2021). Artificial Intelligence for Clinical Decision Support in Sepsis. *Front Med (Lausanne).*, 8:665464. doi:10.3389/fmed.2021.665464.

Xu L., et al. (2014). COX-2 inhibition potentiates antiangiogenic cancer therapy and prevents metastasis in preclinical models. *Science Translational Medicine Vol. 6, Issue 242*, 242ra84. doi: 10.1126/scitranslmed.3008455.

Yang, L., et al. (2015). Comprehensive lipid profiling of plasma in patients with benign breast tumor and breast cancer reveals novel biomarkers. *Analytical and Bioanalytical Chemistry.*, Vol. 407(17), pp. 5065–5077.

Yue S., et al. (2014). Cholesteryl ester accumulation induced by PTEN loss and PI3K/AKT activation underlies human prostate cancer aggressiveness. *Cell Metab 2014; 19*, 393–406. doi: 10.1016/j.cmet.2014.01.019.

Yue X., et al. (2016). Quantitative Proteomic and Phosphoproteomic Comparison of 2D and 3D Colon Cancer Cell Culture Models. *J Proteome Res. 15(12).*, 4265-4276. doi:10.1021/acs.jproteome.6b00342.

Zavalin A., et al. (2012). Direct imaging of single cells and tissue at sub-cellular spatial resolution using transmission geometry MALDI MS. *Mass Spectrometry Journal Vol.47, Issue 11., 47(11)*, 1473–1481. https://doi.org/10.1002/jms.3108.

Zhang X., et al. (2018). Ceramide Nanoliposomes as a MLKL-Dependent, Necroptosis-Inducing, Chemotherapeutic Reagent in Ovarian Cancer. *Molecular Cancer Therapeutics. Vol. 17, Issue 1.*, 50-59. doi: 10.1158/1535-7163.MCT-17-0173.

Zhao ZM, et al. (2016). Early and multiple origins of metastatic lineages within primary tumors. *Proc Natl Acad Sci USA. 113(8).*, 2140-2145. doi:10.1073/pnas.1525677113.

Zhou D.X., et al. (2020). Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis ISSN 1063-5203.*, 787-794. https://doi.org/10.1016/j.acha.2019.06.004.

Zhou X., et al. (2012). Identification of Plasma Lipid Biomarkers for Prostate Cancer by Lipidomics and Bioinformatics. *PLOS ONE 7(11).*, e48889. https://doi.org/10.1371/journal.pone.0048889.

Zhou, S.K., et al. (2021). A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises. *IEEE*, 1 - 19. doi: 10.1109/JPROC.2021.3054390.

Zou Z., et al. (2020). mTOR signaling pathway and mTOR inhibitors in cancer: progress and challenges. *Cell Biosci 10*, 31. https://doi.org/10.1186/s13578-020-00396-1.

9 APPENDIX – ADDITIONAL CLASSIFIERS AND DR TECHNIQUES

9.1 RANDOM FOREST

The Random Forests algorithm, also called random decision forests, is a learning method primarily used for classification and regression. It operates by building a set of decision trees during training and produces, as an outcome, the class that consists of the mode of the classes (for classification purposes) or mean prediction (for regression) of the individual trees (Ho T.K., et al., 1995) (Breiman L., et al., 2001).

The algorithm for random decision forests was first created by Tin Kam Ho (Ho T.K., et al., 1995) using the random subspace method, which, in Ho et al. formulations, is a way to implement "*stochastic discrimination*" (Kleinberg E.M., et al., 1996) (Kleinberg E M., et al., 2000), which is a method frequently used for pattern recognition purposes. It works by producing weak classifiers and subsequently combining them using the central limit theorem to form a strong classifier.

Random forests can be used to rank the importance of variables in a regression or classification problem in a natural way. The first step in measuring the variable importance in a data set $D_n = \{(X_i, Y_i)\}_{i=1}^{n}$ is to fit a random forest to the data. Subsequently, during the fitting process, the out-of-bag error (OOB is a method of measuring the prediction error of e.g. random forests), for each data point is recorded and averaged over the forest. To estimate the importance of the j-th feature after training, the values of the j-th feature are permuted among the training data,

and the out-of-bag error is calculated again on this just perturbed dataset. The importance score for the feature j-th is calculated by averaging the difference in OOB error, before and after, the permutation over all the trees. The score is then normalized by the standard deviation of these differences. Features which provide large values for this score are ranked as more important than features which provide small values.

For classification purposes, random forests can be viewed as so-called weighted neighbourhoods' schemes which is similar to a k-nearest neighbour algorithm (k-NN) (Lin Y., et al., 2006). The model is built from a training set $\{(x_i, y_i)\}_{i=1}^{n}$ that makes predictions $\hat{y}$ for new points x' by looking at the "neighbourhood" of the point, formalized by a weight function W:

$$\hat{y} = \sum_{i=1}^{n} W\,(x_i, x')y_i \tag{9.1}$$

Since a forest averages the predictions of a set of m trees with individual weight functions W_j, its predictions are:

$$\hat{y} = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} W_j\,(x_i, x')y_i \;=\; \sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m} W_j\,(x_i, x')\right)y_i \tag{9.2}$$

This shows that the whole forest is again a weighted neighbourhood scheme, with weights that average those of the individual trees. The neighbours of x' in this interpretation are the points x_i sharing the same leaf in any tree j. In this way, the neighbourhood of x' depends, in a complex way, on the structure of the trees, and therefore on the training set's structure. Lin et al. (Lin Y., et al., 2006) show that the shape of the neighbourhood used by a random forest adapts to the local importance of each feature.

9.2 LOGISTIC REGRESSION

The logistic model is an extensively used statistical construct which, in its basic form, utilises a logistic function to model a binary dependent variable. Many more complex extensions have been developed (Böhning D., et al., 1992) (Fullerton A.S., et al., 2009) (Hensher D.A., et al., 2003). In regression analysis, the logistic model assumes the name of logistic regression, also known as logit regression. It is a form of binomial regression.

Mathematically, a binary logistic model has a dependent variable with two possible outcomes, such as true or false, or 0 and 1, which can be translated as, e.g. healthy or sick for clinical purposes; represented by an indicator variable. In the logistic model, the log-odds, which is the logarithm of the odds, is modelled as linear combination of one or more independent variables (also called predictors); the independent variables can each be a binary variable or a continuous variable (a real value). The corresponding probability of the value labelled 1 can vary between 0 (certainly the value 0) and 1 (certainly the value 1), hence the labelling. The logistic function converts log-odds to probability (hence the name).

Statistician David Cox first developed logistic regression in 1958. The model itself models the probability of output in terms of input and does not perform statistical classification. However, it can be used to make a classifier, for instance by choosing a cut-off value and classifying inputs with a probability higher than the cut-off as one class, below the cut-off as the other. This is a common way to model a binary classifier. There are several logistic regression models, for example binomial, ordinal or multinomial. The simpler version, the binomial, deals with situations in which the possible outcome for a dependent variable can have only two possible outcomes (0 and 1). Multinomial logistic regression, instead, deals with situations where the outcome can have

three or more possible types (e.g., Healthy tissue vs Tumour tissue vs Background). Finally, ordinal logistic regression deals with dependent variables that are ordered.

Logistic regression aims to measure, by estimating probabilities using a logistic function, the connection between the dependent variable and one independent variable or more. This method can also be seen as a particular case of the generalised linear model. On the other hand, the model of logistic regression is based on entirely different assumptions compared to those of linear regression. First, the conditional distribution $y|x$ is a Bernoulli distribution rather than a Gaussian distribution due to the dependent variable that is binary. Second, the predicted values are probabilities (restricted to (0,1)) therefore the logistic regression predicts the probability of outcomes rather than the outcomes themselves.

A description of logistic regression can start with an explanation of the standard logistic function. This function is a sigmoid function, which takes any real input n, and generates outputs with values between 0 and 1, for the logit, this is interpreted as taking input log-odds and having output probability. The logistic function $\sigma(n)$ is defined as follows:

$$\sigma(n) = \frac{e^n}{e^n + 1} = \frac{1}{1 + e^{-n}} \tag{9.3}$$

Let us assume that n is a linear function of a single explanatory variable x (the case where n is a linear combination of multiple explanatory variables is treated similarly). We can then express n as follows:

$$n = \beta_0 + \beta_1 x \tag{9.4}$$

And the logistic function can now be written as:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{9.5}$$

Note that *p(x)* is interpreted as the probability of the dependent variable equalling a "success" or "case" rather than a failure or non-case.

The model is then fit using the gradient descent method that aims to minimise the Kullback–Leibler divergence (see the DR paragraph for more details). However, other methodologies can be used, such as *One in the Ten Rule*, *Maximum likelihood estimation* (MLE), *Cross-entropy Loss function*, and *Iteratively reweighted least squares* (IRLS).

Goodness of fit in linear regression models is generally measured using R^2, which is called the *coefficient of determination*. It represents the proportion of the variance in the dependent variable that is predictable from the independent variable (or variables).

9.3 BAGGING

Bootstrap aggregating (commonly called bagging) is a machine learning meta-algorithm that aims to improve the stability and accuracy of machine learning methodology used for classification and regression. It also aims to avoid overfitting. Although the bagging is usually applied to decision tree methods, it can be built with any classifier.

Given a standard training set *D* of size *n*, bagging generates *m* new training sets D_i, each of size *n'*, by sampling from *D* uniformly and with replacement. By sampling with replacement, some observations may be repeated in each D_i. If *n' = n*, then for large n the set D_i is expected to have the fraction (1 - 1/e) (≈63.2%) of the unique examples of *D*, the rest being duplicates (Aslam J.A., et al., 2007). This kind of sample is known as a bootstrap sample. Then, m models are fitted using the above m bootstrap samples and combined by averaging the output (for regression) or voting (for classification).

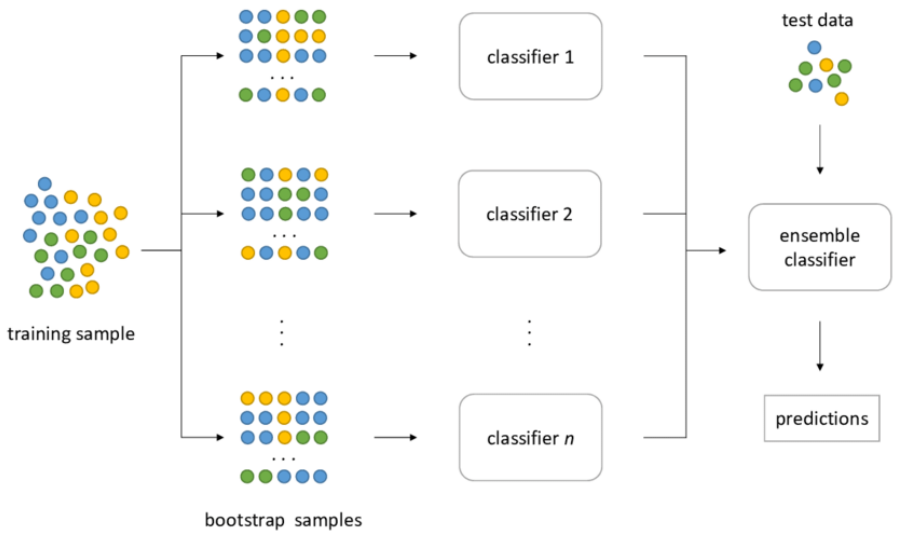


Figure 9.1 – The general structure of the Bagging algorithm from (Javed A., et al., 2007). Before feeding the training sample to the classifier, it is divided into batches, called bootstrap samples, which are subsequently fed to separate classifiers. Then, n models are fitted using the m bootstrap samples and combined by averaging the output (for regression) or voting (for classification).

Bagging leads to improvements for unstable procedures, which include, for instance, ANNs, classification and regression trees, and subset selection in linear regression. On the other hand, it can moderately degrade the performance of stable methods such as K-nearest neighbours.