Using Self-Organising Maps to Cluster Complex Biological Data

Daniel West
MA (Hons) MB BChir

Master of Science

University of York
Computer Science
May 2021

Abstract

Cancer is a common disease during the modern age which requires accurate detection and prediction of its development. Prostate cancer is an interesting form as it is rarely fatal, yet requires surgical excision to remove, which itself may have adverse effects. Therefore, it is important to assess correctly each patient to minimise risk from cancer progression and from treatment side effects.

Raman spectroscopy is an analytical technique which has gained interest in the analysis of biological specimens, as it is a robust technique which produces distinct molecular signals which can be used to identify biomolecules. The sheer volume and dimensionality of spectral data necessitates computational analysis: this work covers the use of self-organising maps for investigating such data.

Self-organising maps are a machine learning technique which spot patterns and reduce dimensionality in high dimensional datasets in an unsupervised manner. Their use can help to discern clusters within the dataset which may not be readily apparent.

The use of self-organising maps to analyse Raman spectral data from human cell samples is an underexplored area of research. This work forms a feasibility study for the use of self-organising maps for such an application, and shows that they are able to correctly cluster cancer and non-cancer samples from a blinded dataset with optimum parameters. Moreover, the optimised SOM shows delineation into three clusters, one of normal prostate data and two of prostate cancer data. Analysis of these clusters shows spectral differences related to lipid composition, an observation which has been linked to more aggressive cancer progression.

Contents

List of Tables

List of Figures

List of Algorithms

Acknowledgements

Throughout the realisation of this thesis I have had excellent support from my supervisors. I have to thank Professor Susan Stepney for all of her kind words, logical guidance, interesting discussion, and belief in my ability. I would like to thank Dr Yvette Hancock for being encouraging, engaged, and a source of support, particularly with the more complex and interesting aspects of quantum physics.

Dr Simon O'Keefe has been very helpful throughout my research, and has been able to advise on information to benefit my studies. Dr Angelika Sebald has been a source of great intellectual discussion, productive distraction, and excellent food and company. Thank you both.

I would like to thank Carl Banbury and his research group for kindly sharing their data with me, and for discussing their methods with me when I had questions.

I would like to extend extreme thanks to my family for their unwavering support and keen interest in my work. My Mum, Jackie, my Dad, Steven, and my Stepmum, Sue, are always happy to learn of my progress and to ensure I am enjoying myself, both inside and outside of work. My partner, Mark, has supported me in ways he may not even realise, is always interested to learn more, and has kept my passion for my research alive.

Finally, I'd like to thank my cats, Nala and Rory, for being excellent company and comfort during my experimentation and writing. They've “helped” to test my ability to find bugs in my code every time they walked across the keyboard. Their input in video research meetings was equally welcome.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Chapter 1

Introduction

This work investigates the use of self-organising maps (SOMs) as a diagnostic and prognostic tool in the setting of prostate cancer. Prostate cell samples are analysed using Raman spectroscopy, and the complex signals generated are fed into the SOM algorithm to produce clusters of like samples based on the presence or absence of disease. The use of SOMs to analyse Raman spectral data from human cancer cells is underexplored, and this work forms a feasibility study of this application.

Cancer is a disease in which the body moves from a normal state to one of disorganisation via the acquisition of genetic alterations. Such aberrations take time to become established and result in altering internal behaviours of cells. The earlier these changes are discovered, the more chance there is to prevent further progression of the disease, or to remove the bulk of diseased tissue before it grows.

Clinical methods of analysis currently consist of histology (direct observation of cells under the microscope) and epidemiological studies of disease trends and management outcomes. Imaging techniques are widely used to aid the diagnosis of cancers, and give information at the whole body level—for this the cancer tissue must be large enough to be seen at the resolution of a whole body scan.

Interest has grown in newer technologies to investigate the processes within cells which lead to disease states. Genetic analysis of patients and populations highlights genes which may predispose to malignancy or which are commonly affected during cancer establishment. Analysis of the resultant proteins produced and subsequent impact on cellular behaviour is performed with the hope of understanding the cellular dysfunction and ways to circumvent it. Spectroscopic analyses can give information at the molecular level, and the sheer volume and dimensionality of data gathered from spectroscopy renders computational analysis necessary.

Prostate cancer in humans has been chosen as a disease model due to its societal importance, biological interest, and the unique clinical question it poses: whether or not to treat. The majority of prostate cancers are slow growing and do not infiltrate surrounding tissues, and treatment decisions take into account the benefits on symptoms as well as life expectancy. In the small subset of aggressive cancers, there is a need to intervene at the earliest possible time to sustain life, and so good prognostic as well as diagnostic accuracy is required.

Chapter 2 covers the relevant biological, physical, and computational background of this project. It introduces the complex nature of the human cell and discusses cancer development under a complex systems approach. The necessary aspects of Raman spectroscopy are explained, and the SOM algorithm is discussed. Review of current literature shows the SOM to have potential for the classification of biological tissues based on Raman spectroscopy.

Chapter 3 outlines the research question which forms the basis of the project, and Chapter 4 discusses the materials and methods used.

Chapter 5 discusses the development of the MySOM module, a python program which inherits the SOM training features of MiniSom [101] and incorporates tools for plotting figures and analysing Raman spectral data. It uses examples based on common datasets in machine learning to show that it can produce meaningful results and to reveal how the SOM process works.

Chapter 6 reviews the use of MySOM to analyse Raman spectra generated from pig eyes, following the publication of this data [11]. Although MySOM uses a different SOM architecture and parameters from [11], it is able to cluster the data in a comparable manner to the published results, showing that it can be used to cluster Raman spectral data from a biological source.

Chapter 7 uses Raman data from prostate cell lines to test the SOM's ability to cluster by disease status, which the SOM does as expected. The sensitivity of the SOM method to changes in parameters is also demonstrated.

Chapter 8 covers the analysis of a blinded dataset of samples from normal prostate and prostate cancer. An outlying observation is uncovered and examined, found to have a very different spectral shape to the others, and it is removed from the dataset as it greatly impacts the SOM training process. A parameter sweep is performed to find optimum training parameters, and the resultant SOM displays three discrete clusters: one with data from normal prostate, and two with data from prostate cancer. Subsequent analyses of these clusters and their average spectra yield discernible differences in the lipid composition between normal and cancer, and between the two cancer subclusters.

Chapter 9 summarises the results of this project, and discusses ideas for future works.

Chapter 2

Literature Review

Cancer is a phenomenon with a large body of research, from many different fields of science and the humanities. It is the very complex nature of cancer, from the microscopic level of the cell to the large scale nature of population healthcare, which renders it amenable to investigation by different specialties for particular tasks.

The project undertaken in this work investigates biomolecular aspects of prostate cancer and is a cross-disciplinary venture between Biology, Physics, and Computer Science. Firstly, the nature of biological complex systems and the relevant aspects of cancer biology are discussed. Secondly, the physical nature of electromagnetic radiation and the Raman spectra generated from cell samples is covered. Thirdly, computational methods employed to analyse these high dimensionality spectral data are explored.

2.1 The Biological Problem

2.1.1 Biological Systems

Very few biological phenomena exist in isolation, rather within the framework of ever growing hierarchical systems. As technology and understanding of individual biological components has advanced, it has become more and more evident that study of biological systems is key to understanding more complicated biological observations [52].

A system refers to a set of entities which interact to form a whole phenomenon, dissociated from external entities by a boundary [64]. Entities which exist outside of the boundary form part of the system's environment, and may exchange information with the system components. Entities within a system interact with each other and their environment by a defined set of rules, and display typical attributes and behaviour. The behaviour of the system as a whole can be observed, and is distinct from that of its components. Systems can be interpreted from two complementary viewpoints.

The reductionist approach dissects systems into individual fundamental parts and attempts to explain system-wide behaviour as a direct result of component behaviours; that is to say that the total is equivalent to the sum of its parts [35, p. 16].

The complex systems approach views interactions of system entities with each other and their environments as the key contributors to system-wide behaviour [12]. Although individual attributes and behaviours may be studied and understood at the level of individual entities, the vast range of possible interactions and dependence on different states both within the system and within the environment create a higher order behavioural organisation within the system; that is to say that the total is greater than the sum of its parts.

Complex Systems

Many commonly encountered biological systems are complex systems, which display several key complexity conferring properties, as reviewed well in [12].

Emergence refers to the relationship between the entities within a system and system behaviour as a whole. System-wide behaviour is difficult to predict from entities in isolation, as this behaviour only emerges when component entities are brought together. Such systems may be nonlinear as changes in individual entities may not have a proportional influence on system-wide behaviour.

Interdependence describes the relationships between entities within a complex system. If one entity undergoes change, it may impact other entities, and potentially system behaviour as a whole. Feedback loops may exist, where the behaviour of an entity impacts upon itself, either directly or via another agent. These loops are common in biological systems, whereby system parameters must be kept within narrow margins to optimise system behaviour.

Biological Organisation

Combining reductionist and complexity theory, hierarchical reductionism [22, p. 21] explains biological systems in terms of hierarchies of systems. Each level of the hierarchy results from a reductionist view of the fundamental scales of activity, and within each level a complex system of interacting entities exists. The system within each level comprises entities which are the result of the behaviours of the system in the level below, and behaviours which impact the entities of the system in the level above.

This hierarchy can be extrapolated from subatomic particles to the universe, and so a subset of the hierarchy relevant to the work discussed is given in Table 2.1.

Hierarchical Level	Example
Biomolecule	Peptide
Macromolecule	Protein
Organelle	Nucleus
Cell	Myocyte
Tissue	Muscle
Organ	Heart
Organ system	Cardiovascular System
Organism	Human

Table 2.1: Biological hierarchy from biomolecule to organism.

2.1.2 Cancer as a Complex System

Cancer is an abnormal growth of cells within an organism, caused by a change in the normal function of cells [36]. Cancerous cells sustain their transformation via usurping the supply of resources and maintaining an environment which supports cancer growth [37]. This appropriation develops due to changes at the genetic level, which affect the chemical reactions of a cell and ultimately its behaviour, and changes in the microenvironment (the surrounding collection of cells, proteins, and messenger molecules). These altered processes lead to rapid growth, local tissue invasion, and distant (metastatic) spread.

These changes are caused by alteration to the internal cellular system and its surrounding environment, representing a complex adaptive system [61]. Histologic changes are recognised in many cancer types, reflecting perturbation of normal tissue architecture. However, precisely how genetic, biochemical, and microenvironmental changes at lower hierarchical levels affect a group of cells is difficult to predict, and further studies aimed at investigating the emergent properties of these complex adaptive systems are required.

Review of the different aspects of cancer at each hierarchical level is important to understand the importance of cancer as a clinical problem. The following sections discuss key aspects of cancer biology, as relevant to this study into prostate cancer, from the human scale down to the cellular scale, where the main focus of the thesis resides.

2.1.3 Prostate Cancer

Prostate Cancer Epidemiology

In the UK, prostate cancer is the most commonly diagnosed cancer, with a lifetime diagnosis rate of 1 in 8 men [98]. It has an annual incidence of 48500 cases [97], and an annual death rate of 11500 [98]. Risk factors for prostate cancer include advancing age, ethnicity, obesity, and family history.

As prostate cancer is a relatively common disease amongst men, it poses an

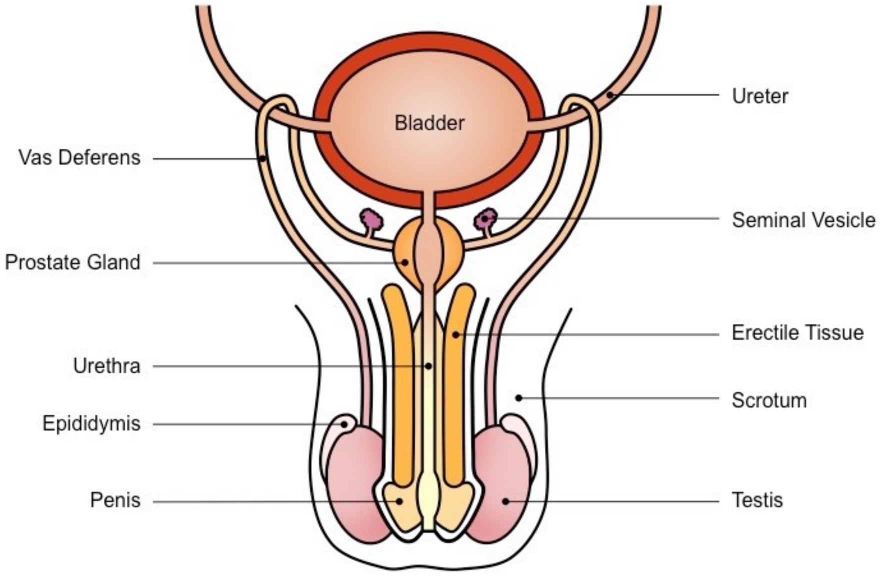


Figure 2.1: Diagrammatic representation of the human male reproductive system. Image reproduced from `ib.bioninja.com.au/standard-level/topic-6-human-physiology/66-hormones-homeostasis-and/male-reproductive-system.html`

interesting clinical dilemma: most men with the disease die with it, whereas only a small proportion die from an aggressive advanced form [84]. Furthermore, the disease is heterogeneous, with the prostate often containing several cancer foci developing at different stages.

The Human Male Reproductive System

The male reproductive system functions to produce and to propagate spermatozoa, short-lived cells which contain half the amount of DNA of somatic cells. They must fuse with similarly haploid ova within females in order to survive and to grow into offspring. Spermatozoa are produced in the testes and propelled along the vasa deferens, through the urethra, and ejected into the vagina of the female during sexual intercourse. Along their route they mix with nutrient-rich fluids from the seminal vesicles and prostate, which are required for survival in the female genital tract. An overview of the human male reproductive system is given in Figure 2.1, and a thorough review of the physiology of the male reproductive system can be found in [69].

2.1.4 Clinical Classification of Prostate Cancer

The Prostate Gland

The prostate is a glandular structure which consists of two main tissue types. The glandular tissue comprises of luminal and basal cells, with occasional interspersed neuroendocrine cells. The structural stromal tissue comprises of fibroblasts, smooth

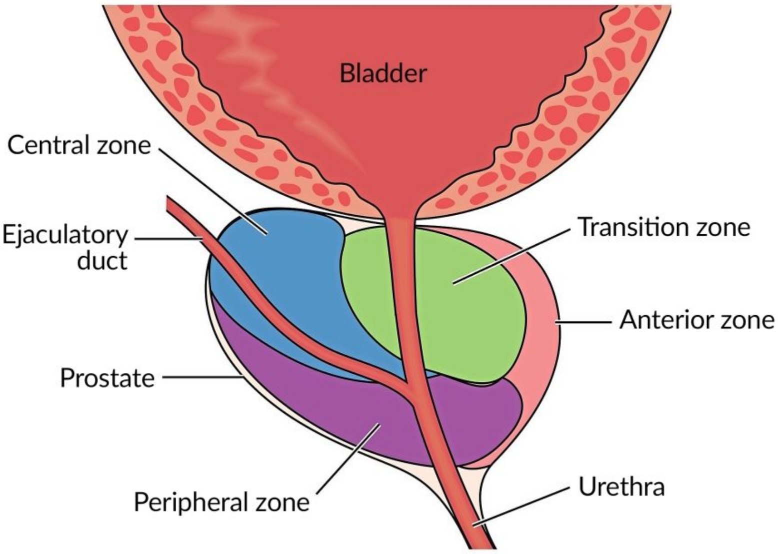


Figure 2.2: Diagrammatic representation of the human prostate showing the four anatomical zones—anterior, central, peripheral, and transition.
Image reproduced from `thestar.com.my/lifestyle/health/2019/07/11/men-prostate-challenges`

muscle cells, and blood vessels.

The prostate is divided anatomically into four zones (Figure 2.2), an anterior fibromuscular zone and three glandular zones (central, transition, and peripheral) [72], each of which contains different combinations and ratios of each prostate cell type. Androgens, the male sex hormones, stimulate the prostate gland to grow and to produce prostatic fluid. This fluid contains specialised enzymes, citrate, and zinc required for spermatozoa to survive [66].

Cancer most commonly arises in the peripheral zone, an observation which likely relates to the specific combination of cells and the microenvironment found in this region [42]. The variation in prostate cancer incidence and distribution, even within the same prostate, confers the need for an accurate diagnostic (is the disease present) and prognostic (how severe is the disease) tool in order to inform management strategies.

Gleason Grading System for Prostate Cancer

The current Gleason grade system [73] stratifies patients based on prostate histology, the cellular architecture visible under an optical microscope. Gleason originally devised a score from 1 to 5, with increasing number correlating to more disturbed prostate architecture and invasion of surrounding tissues [34]. Prostate cancer has a heterogeneous nature as it can develop and grow in different locations within a patient's prostate. For this reason, the two most common cancer cell patterns observed within a patient's prostate tissue are scored based on Gleason's grading, and added to give the patient a final score between 2 and 10. In order to simplify

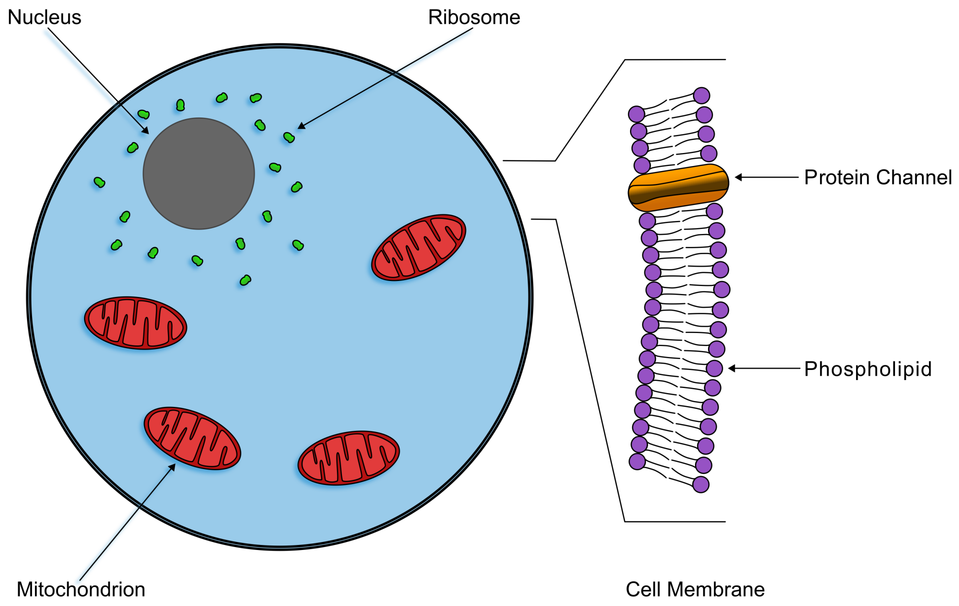


Figure 2.3: Diagrammatic representation of a human cell. The cell is bound by a membrane consisting of a phospholipid bilayer which forms a boundary with the surrounding environment and controls passage of substances into and out of the cell through specialised channels. The nucleus is the location of genetic information in the form of deoxyribonucleic acid (DNA), and the surrounding ribosomes are where this genetic code is read to produce proteins. Mitochondria are the areas of energy production within the cell.

classification, the International Society of Urological Pathology developed the grade group system, which combines Gleason grades with similar clinical outcomes and prognoses into 5 grade groups [27].

The histology-based Gleason grade system relies on how cells look, but gives no information into the internal changes and biochemical reactions which underpin malignant transformation. In order to gain more information of these subcellular processes, it is necessary to use an analytical technique, and work is currently being conducted to investigate the use of Raman spectroscopy to analyse prostate samples [8, 19, 49, 80, 85]. Such research is also being performed at the University of York.

2.1.5 Cell Biology

The Cell

The functional unit of life is the cell, a complex system of biological macromolecules and organelles contained within cytoplasmic fluid and bounded by a membrane. Each organelle is a distinct compartment responsible for a process fundamental to survival, propagation, and normal function of the cell. The relevant organelles found in human cells considered below are the nucleus, cell membrane, cytoplasm, mitochondria, and ribosome—information is adapted from [3]. A diagrammatic representation of a cell containing these organelles is shown in Figure 2.3.

The nucleus houses genetic material in the form of deoxyribonucleic acid (DNA), a large polymer of nucleotides containing the bases adenine, cytosine, guanine and thymine. The order of these bases forms the genetic code of the cell, instructions for it to produce all its necessary constituents and machinery. This DNA must be replicated faithfully when the cell undergoes division, as any changes are passed to the next generation and could alter normal function. In order to ensure DNA is protected, it is wound tightly around chromatin proteins when not being used for replication or synthesis.

The cell membrane consists of a phospholipid bilayer with hydrophobic lipid tails facing its centre and hydrophilic phosphate groups facing its edge and forming an interface with the cell's internal and external media. Within the membrane there are proteinaceous channels which allow transport of polar molecules; hydrophobic molecules can diffuse through the membrane directly. This configuration allows complete segregation of the cell's internal components from its environment, with the ability to exchange small molecules for use as chemical reagents or signals between the cell and its surroundings—acting as the boundary of the cell complex system. Due to differing concentrations of polar molecules on either side of the membrane, a potential difference exists across it which can be used as a source of energy for transporting molecules or configuring change.

The cytoplasm is the main body of the cell, consisting mostly of the liquid cytosol where chemical reactions occur. It also contains the cytoskeleton, a rich network of tubules providing structure to the cell and a scaffold for transport of molecules between organelles and the cell membrane.

Mitochondria are small organelles enveloped in their own phospholipid bilayer, and are responsible for producing large amounts of adenosine triphosphate (ATP). ATP is the universal energy currency of the cell required for many chemical reactions.

Ribosomes are the specialised machinery within cells where protein synthesis occurs. Genetic information in nuclear DNA is read by cellular machinery to produce complementary messenger ribonucleic acid (mRNA). mRNA exits the nucleus and binds to ribosomes, where the sequence of bases is read and proteins are produced based on the genetic code.

Cell Metabolism

Cellular metabolism is a collection of interconnected processes required for survival, in which macromolecules are broken down into their constituent biomolecules and energy (catabolism), and new macromolecules are built from biomolecules with the input of energy (anabolism) [3]. The major catabolic process within animal cells involves the breakdown of the sugar glucose in two stages.

The first stage, glycolysis, occurs within the cytoplasm of the cell and does not require oxygen. A single glucose molecule is broken down to form two pyruvate molecules with a net gain of two ATP.

The second stage, aerobic respiration, occurs within mitochondria and requires oxygen to go to completion. Pyruvate is broken down to acetate, which enters the Krebs cycle by binding to oxaloacetate to form citrate. Citrate undergoes several chemical reactions to produce enzymatic co-factors, reverting to oxaloacetate in the process. Further acetate can then bind and allow the cycle to continue. The enzymatic co-factors enter the electron transport chain, where a series of oxidation and reduction reactions occur to produce ATP. A single glucose molecule is broken down to form six molecules of carbon dioxide and water with a net gain of 36 ATP.

The intracellular changes which occur during malignant transformation comprise genetic changes which alter the quantity and nature of molecules produced, and ultimately the behaviour of the cell. Often the process of cellular metabolism is usurped to allow malignant cells to grow and to divide rapidly. The Warburg Effect [102] refers to the commonly observed phenomenon of metabolic reversal: cancer cells switch from an aerobic metabolism to glycolysis and pyruvate fermentation. Although this process is an inefficient method of energy production, it allows cells to grow rapidly as the molecular skeleton of pyruvate is not broken down and can be incorporated into biomass [37]. Furthermore, rapidly growing tumours may outgrow their blood supply, but can continue to grow if they are not dependent on oxygen.

Prostate Gland Cell Metabolism

The metabolism found in human cells is slightly altered in normal prostate tissue. Citrate is stored and secreted by the prostate, rather than being used in the Krebs cycle for energy production [75]. Prostate cells instead produce energy via the less efficient process of glycolysis [9].

However, in prostate malignancy there is still metabolic reorganisation, as citrate is instead used to fuel energy production via oxidative phosphorylation and for lipid production [9].

The Cell Cycle

Normal cells replicate via the cell cycle, a regulated sequence of processes governing energy production, DNA replication, protein synthesis, and cellular division [86]. Division does not occur indefinitely, and ceases once a cell reaches its maximum number of divisions—the Hayflick limit [40]. These cells enter senescence (non-replicative existence) or undergo apoptosis (programmed cell death).

Some cells escape the Hayflick limit and senescence, becoming immortalised and able to divide for as long as resources allow. This cancerous transformation is accomplished by mutations leading to expression of genes which drive replication, or suppression of genes which prevent replication [67]. Immortalised cells can be produced artificially, by inducing normal cells to express tumourigenic genes and proteins [14].

Cell Lines

Primary cells are those derived directly from the native tissue under investigation, and are thought to reflect the real cellular behaviour of their tissue of origin [33]. Acquisition of such cells is not always practicable, and replication of these native cells *in vitro* can be used to generate cell lines. These cell lines can be cultured to produce subsequent generations until they enter senescence after achieving the Hayflick limit (finite cell lines), or indefinitely provided the appropriate resources (immortalised cell lines) [87].

Immortalised cell lines are useful for research, as they reflect the state of the native cells at the time they are harvested, are cheaper to use than primary cells, can be farmed rapidly, and afford a biological control for cell-based experiments. However, use of cell lines must be carefully monitored, as they do not fully reflect the nature of their primary cells, as they have mutated to become immortal and likely retain other mutations [50]. Furthermore, the more cell lines divide, the more likely mutations are introduced into DNA leading to persistent phenotypic changes in the offspring. Measurements for the cell lines analysed here were acquired sequentially, and so mutations differing between members of the same cell line are assumed to be negligible.

2.2 The Physical Approach

2.2.1 Electromagnetic Radiation

Electromagnetic radiation is classically considered as a perturbation in the electromagnetic field, which can be characterised by either its wavelength or frequency, the product of which is constant [70]. Under quantum mechanics, such radiation is viewed as a collection of photons, discrete packets of energy, which can be similarly described by the frequency with which they rotate or the distance they travel in one rotation (wavelength) [25].

When light interacts with a molecule, it may be absorbed, reflected, or refracted. Absorption of the energy from a photon raises the molecule to a higher energy level. Reflection of a photon returns it back along its path and refraction changes the direction of the photon's path, both without net changes to the molecule's energy level.

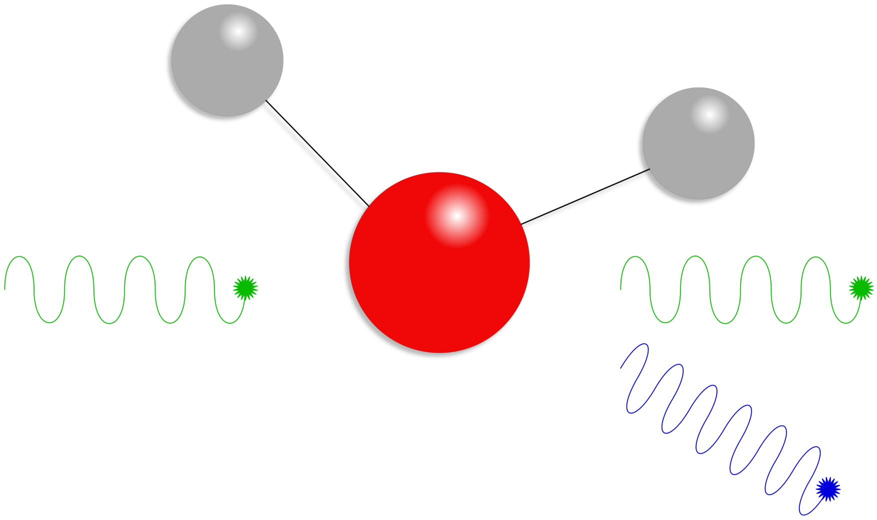


Figure 2.4: Diagrammatic representation of scattering of light. The incident photon (left) interacts with the water molecule, which in turn releases a photon. Elastic scattering occurs where the resultant photon is of the same energy as the incident photon (right)—inelastic scattering occurs when the molecule releases a photon of higher or lower energy than the incident photon (lower right).

Scattering

When an incident photon interacts with a molecule, the energy from the photon may be absorbed, moving the molecule to a higher energy state. Once in this state, the molecule can release a resultant photon travelling in a different direction, and return to a lower energy state. This process of reciprocal absorption and emission of photons via interaction with matter is known as scattering of light (Figure 2.4).

The majority of scattering is elastic (Rayleigh scattering), where the excited molecule returns to its original energy level, and the frequency of the incident and resultant photons is equal. A very small proportion of molecules display inelastic scattering, where the excited molecule returns to a higher (Stokes scattering) or lower (anti-Stokes scattering) energy level and the resultant photon is of a different frequency to the incident photon, as shown in the Jablonski diagram [45] in Figure 2.5. The bold horizontal lines represent the electronic states of the molecule and the pale horizontal lines the vibrational levels within these energy levels.

Elastic scattering produces a resultant photon with equal energy to the incident photon, and hence the same wavelength. Inelastic scattering produces a resultant photon with different energy to the incident photon, and the observed change in energy of the resultant photon is deemed the Raman effect after C. V. Raman who first described the phenomenon [81]. Stokes scattering results in a photon of lower energy and hence a longer wavelength than the incident photon—a red shift. Conversely, anti-Stokes scattering yields a photon of higher energy and thus shorter

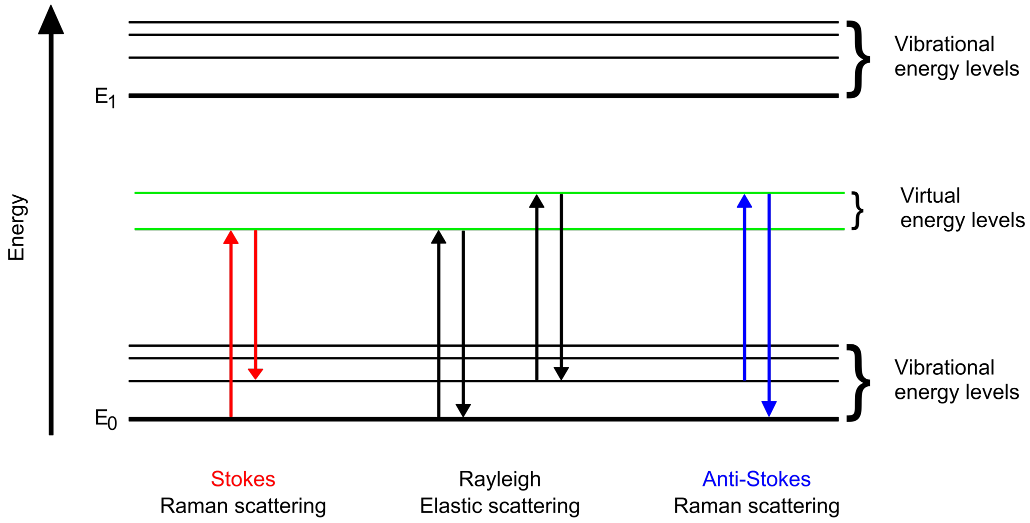


Figure 2.5: Jablonski diagram showing the possible energy transitions of a molecule when light is scattered. Rayleigh scattering (centre) occurs when the excited molecule returns to its previous energy level. Stokes scattering (left) occurs when the excited molecule returns to a higher energy level. Anti-Stokes scattering (right) occurs when an excited molecule returns to a lower energy level.

wavelength than the incident photon—a blue shift. These changes in resultant photon wavelength are shown in Figure 2.6 (adapted from [29, p. 17]).

The ability of a molecule to become Raman active depends on its vibration, rotation, and electronic charge, and the energy change which occurs is characteristic for a particular chemical bond [89].

Spectroscopy

Spectroscopy is the study of the interaction of electromagnetic radiation with matter [43]. A spectrum is a generated signal representative of this interaction as a function of the frequency of the radiation. Emission spectra consist of frequencies emitted by matter when it moves to a new energy level, and may consist of lines or bands. Line spectra represent signals from individual atoms emitting photons of specific frequencies, whereas band spectra consist of many lines packed closely together which represent molecules as a nonlinear summation of component atoms, electron spin states, vibration, and rotation.

2.2.2 Raman Spectroscopy

The Raman effect is the uneven scattering of light which occurs when incident light interacts with a molecule and alters its polarisability [26]. The Raman active molecule enters a short-lived virtual higher energy state, and when relaxing moves

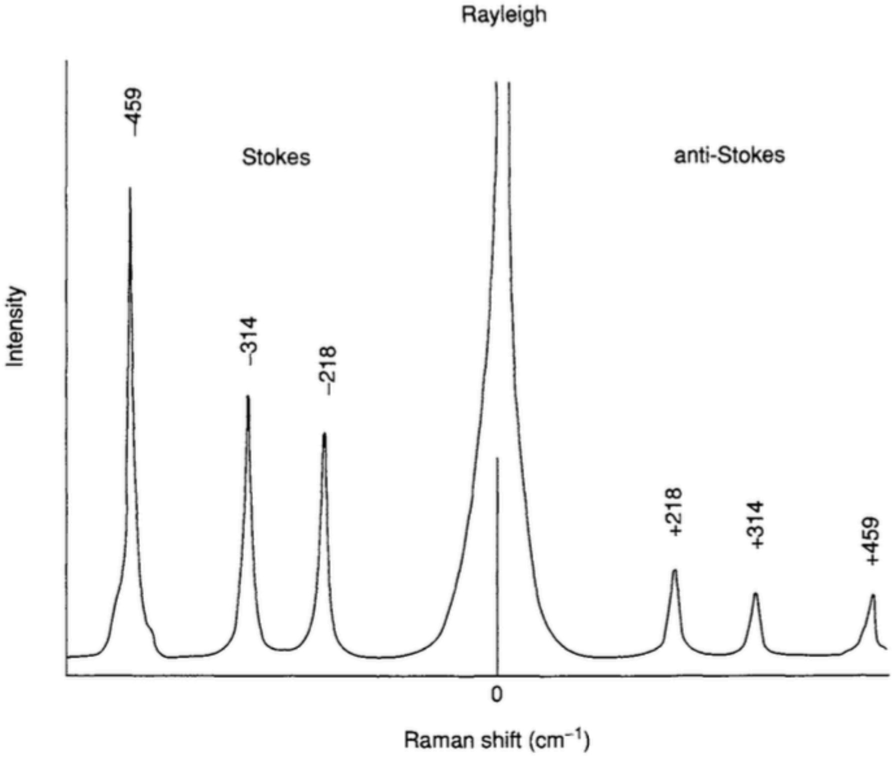


Figure 2.6: Raman spectrum of CCl_4 excited with a 488.0nm laser (adapted from [29, p. 17]). Intensity of the Raman shift observed is proportional to how many photons undergo that shift. Most of the resultant photons undergo no change in energy (Rayleigh scattering, centre). A small proportion of resultant photons lose energy compared to the incident photons (Stokes scattering, left), and a very small proportion of resultant photons gain energy compared to the incident photons (anti-Stokes scattering, right).

to a different energy state than the original, thus releasing a photon of different energy to the incident photon. The change in energy (and hence frequency) of the photons corresponds to the changes in molecular vibration and rotation caused by energy exchange, and is dependent on the vibrational state of the molecule: thus the overall frequency shift and spectra measured are typical of a particular chemical bond.

Stokes scattering is more common as molecules are more likely to be in the electronic ground state compared to an excited state, and most Raman techniques measure this frequency shift. In the field of spectroscopy, as the temporal frequencies of photons encountered are large, it is more common to record a spatial frequency. The wavenumber, ν, describes the number of full cycles of a wave (or full rotations of a photon) which occur in one unit of distance, most commonly cm^{-1}. This is mathematically equivalent to dividing the temporal frequency (f) of the wave by the velocity of the wave (c), which is the reciprocal of wavelength (λ), as shown in Equation 2.1.

$$\nu = \frac{f}{c} = \frac{1}{\lambda} \tag{2.1}$$

The utility of observing the Raman effect as an analytical method results from the fact that the change in resultant frequency corresponds to the molecule itself and is not affected by the wavelength of the incident radiation [81]. By comparing Raman spectra derived from analysing molecules in a test sample against spectra of known chemical signatures, very detailed information can be gathered about the composition of complex structures such as cells, including differentiation between primary cells and cell lines, characterisation of pathological states, and interpretation of cell differentiation [15].

Raman spectra

Raman spectra are generated from measured intensity as a function of the Raman shift observed in units of wavenumber (cm^{-1}). As both the Stokes and anti-Stokes shift give the same information, only one side of the complete spectrum (Figure 2.6) is used for analysis, as shown in Figure 2.7. Spectra are broadly divided into the fingerprint region (800–1800 cm^{-1}), which corresponds to molecular vibrations associated with DNA, proteins, and lipids, and the high wavenumber region (2800–3800 cm^{-1}) which corresponds to vibrations from proteins, lipids, and water [68].

Raman spectral bands are typical for the chemical bonds within an analysed molecule, and standard libraries of spectral shapes generated from known samples have been developed, such as Wiley's KnowItAll [90]. Comparison of experimentally recorded spectral data from an unknown sample with such a standard library has been widely used and validated as a method of analysing the purity of reagents and pharmaceuticals, and more recently to investigate the biochemical composition

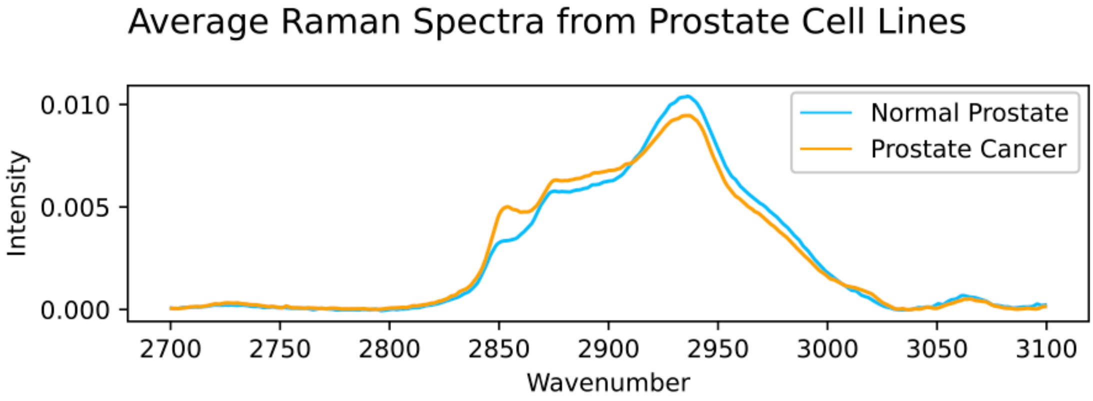


Figure 2.7: Example average Raman spectra generated from analysing normal prostate and prostate cancer tissue samples.

of biological samples [15]. However, the bands obtained when analysing a sample containing many different biomolecules, such as a cell, are not a simple summation of individual bands due to complex, non-linear interactions affected by changes in internal cell parameters. This complexity highlights a key need to analyse Raman spectra from native biomolecules within biological systems, rather than in isolation [32].

Before analysis of spectral data is performed, data are processed to remove noise. Within biological systems, observed variation is often large, and minor differences between samples may represent significant distinguishing features. Therefore, there is a need to ensure that pre-processing does not remove small but significant signals. Common pre-processing techniques include cosmic ray removal, baseline correction, smoothing, Fourier transformations, and data normalisation, reviewed extensively in [32].

Raman Spectra and Cancer Investigation

In recent years, interest in Raman spectroscopy as a diagnostic tool has grown, as it can discern bonds in native molecular structures, therefore it does not require labelling or processing of biological samples. Research is currently investigating the use of Raman spectroscopy to analyse accessible body surfaces, such as the gastro-intestinal tract and skin [20], internal tissues during surgery [20], biopsy samples [62], and blood samples [19]. Great effort has been expended to create a Raman biological standard library, containing details of which molecular structures correspond with Raman spectral bands from both normal and cancer tissues [74, 93], allowing some interpretation of the spectra in terms of constituent biomolecules.

Cui et al. reviewed the developing use of Raman spectroscopy technologies, and how different methods of sampling, both laboratory-based and bedside, could be used for analysing patients [20]. Their conclusions were promising, given the high sensitivity and specificity of Raman spectra to discern normal from cancerous tissue.

Raman spectroscopy has gained interest as a diagnostic and prognostic tool for prostate cancer [49]. Aubertin's group found Raman spectroscopy to have high sensitivity and specificity in being able to distinguish normal from malignant tissue following a supervised machine learning analysis [8].

Circulating hormones stimulate prostate cancer growth, partially by inducing lipid synthesis, and Potcoava's group found that hormone treatment of prostate cells resulted in increased lipid storage within cells, particularly saturated lipids [80]. Several groups have used Raman spectroscopy to investigate the metabolic reprogramming of prostate cancer by analysing lipid levels [1, 85]. Roman's group showed that accumulated intracellular lipid droplets in prostate cancer cells are heterogeneous in both composition and amount, and that X-ray irradiation of the cells leads to depletion of the lipids as part of the cell's damage response [85].

2.3 The Computational Analysis

Once spectra have been gathered from the samples of interest, they must be classified without researchers knowing the ground truth of exactly what to expect. In order to achieve this goal, an unsupervised computational analysis method can be used.

One of the key aims in analysis of spectral data is dimensionality reduction, whereby the many variables in a dataset are reduced to a few composite variables which explain most of the variation. Such feature extraction makes the database easier to interpret as it reduces the redundancy of multiple correlated variables and simplifies onward processing. Principal component analysis (PCA) is a commonly used method of dimensionality reduction and is outlined below. Another example of an appropriate method from machine learning, self-organising maps (SOMs) as used in this thesis, is also discussed.

2.3.1 Principal Component Analysis

Principal component analysis is a commonly used analytical method to investigate datasets with high dimensions, such as the continuous waveforms encountered in spectral data. PCA aims to reduce high dimensionality so that variation is explained by fewer components in order to increase the ease of interpretation and understanding.

PCA finds new variables which are linear combinations of the original variables in the dataset, and which are orthogonal to each other. These principal components should maximise the variance within themselves with the constraint that they must be unrelated to each other, in order that the majority of variation within the dataset is explained by the fewest number of principal components [47]. An image describing this process is given in Figure 2.8.

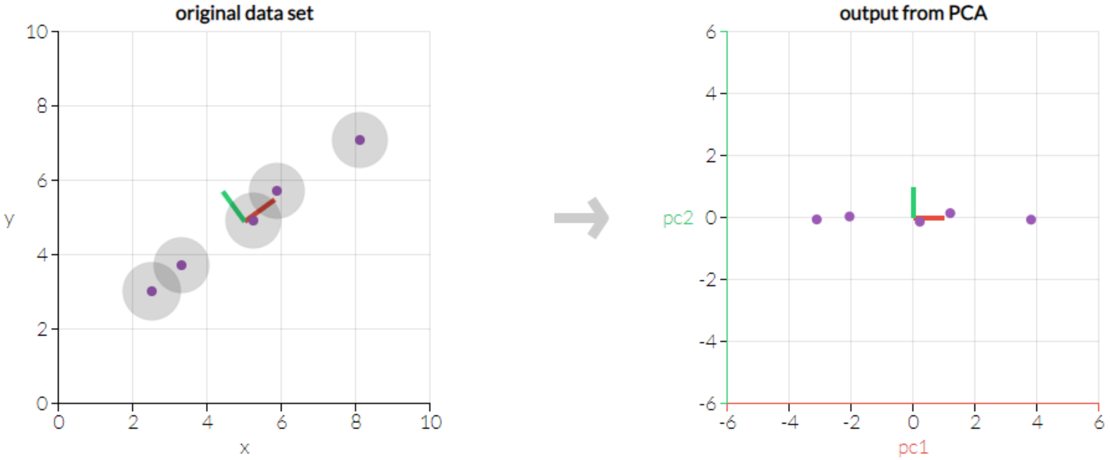


Figure 2.8: Principal component analysis in which principal components pc1 and pc2 have been derived from the variables x and y. Variation within the original dataset is seen within both x and y in the image on the left, whereas the majority of variation is explained by pc1 in the image on the right. In this manner, variation within the dataset has been maintained, but the dimensionality has been reduced. Image adapted from `setosa.io/ev/principal-component-analysis/`.

PCA has several benefits as an analytical method. It is widely understood, used throughout many fields of science and statistics, and is easy and quick to perform with the power of modern computing. It is often used to aid classification problems, as it searches for variables which display high variance and maximise segregation of classes. The resultant principal components can then be analysed with a clustering algorithm such as k-means clustering to reveal the natural groups within the dataset.

PCA has one main caveat when it comes to analysing biological data. Most biological populations are heterogeneous, and so there is large variation both within and between classes. If the within class variance for a population is sufficiently high, PCA aligns its higher order principal components along the axis of within class variation, and the features selected may be completely unrelated to class [18]. Furthermore, to perform k-means clustering, the number of expected clusters must be given before analysis, so there must be some idea of the inherent groups expected to be found within the dataset.

2.3.2 Kohonen Self-Organising Maps

In the 1980s Kohonen introduced a new form of network topological organisation in which input data can be represented by a two-dimensional array [53,54]. These self-organising maps (SOMs) are an elegant way to display high dimensionality data, and allow visualisation of subgroup clusters within the dataset, which may not otherwise be readily apparent. The learning method of SOMs is unsupervised, working to detect inherent patterns within the data and not requiring an expected number of clusters into which to sort data. Furthermore, mathematically SOMs can be considered to be a non-linear version of PCA [41, p. 462], so their analysis of complex biological data may be more appropriate where PCA struggles to cluster data which are not linearly separable.

Kohonen's maps are a subtype of artificial neural networks (ANNs), a group of learning algorithms involving a mesh of nodes with edges between them loosely modelling synapses between neurons in the human brain. These neurons carry a weight which alters as learning occurs, and their topology changes to represent the target dataset. ANNs are usually supervised, meaning that during learning there is feedback to say whether they have correctly classified input data [60]. Although Kohonen's maps can be modified for supervised learning, their original realisation was unsupervised, where the network relied solely on segregating data by internal patterns, with no notion of class identity [53].

A note on nomenclature. Within work on ANNs, the terms 'node' and 'neuron' are often used interchangeably to denote the individual units within the network. The term 'node' is used throughout this thesis to describe the units within SOMs, to reduce confusion with traditional ANNs (and because neither artificial network truly reflects the complex nature of biological neuronal function).

Structure and Function of Self-Organising Maps

Kohonen originally introduced the idea of SOMs as a network of connected threshold-logic units which are able to assume the topology of an input dataset [54]. The nodes within Kohonen's maps are arranged in a one- or two-dimensional array, with lateral connections between nodes allowing local feedback loops. This lateral interaction and competition allows the map to become organised, by each node learning to detect a unique pattern from the input dataset [55].

The algorithm attempts to fit an artificial network to a dataset by repeatedly calculating the best matching unit (BMU) of the lattice for each input vector and drawing that node and its neighbours closer to that vector. With each step of the algorithm, the effect a node exerts on its surrounding nodes diminishes, and after many iterations the network assumes the topology of the dataset. The resultant map is much more visually accessible for humans, as demonstrated in Figure 2.9. The steps of the SOM algorithm are outlined in Algorithm 2.1.

Prior to SOM analysis, it is useful to normalise data to improve accuracy of the output, as the normalised input vectors have the same dynamic range [56, p. 115]. A useful method for high dimensionality data is normalisation of the variance of each dimension across the dataset [56, p. 160], and the combination of variance normalisation and a Euclidean measure of distance between vectors is very effective at displaying the relationships between variables in most studies [57].

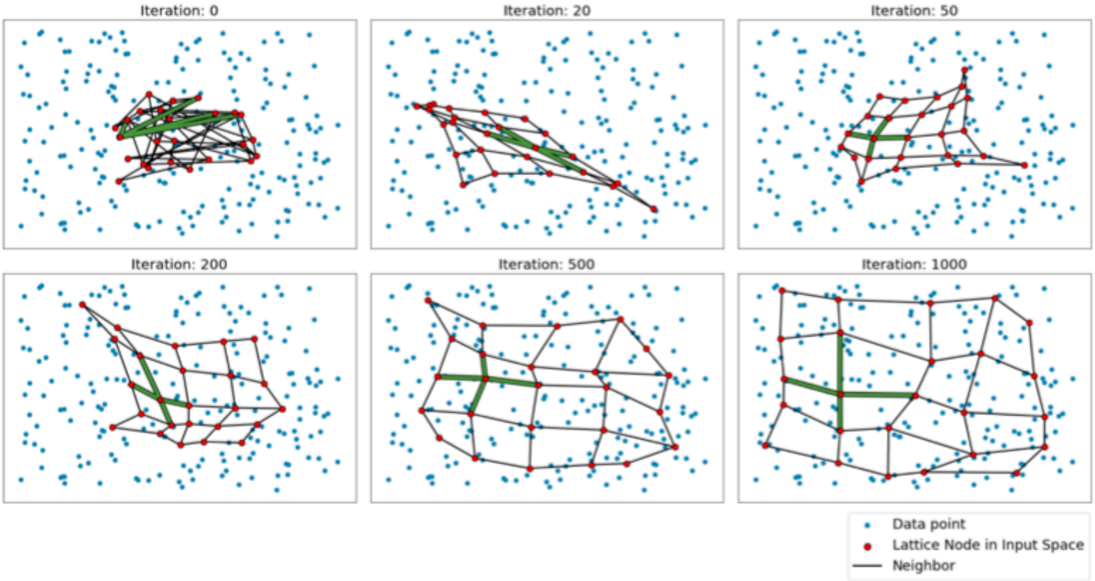


Figure 2.9: Diagram representing the lattice of the SOM assuming the topology of the dataset, adapted from [79]. The red dots are the nodes in the SOM array, the black lines are the connections between neighbouring nodes, and the green lines highlight the neighbours of one specific node. As the number of iterations increases, the shape of the SOM changes from the random starting point to assume that of the dataset.

Algorithm 2.1 Self-Organising Map Process

```
input: input data matrix
initiate weights
for iteration $t$ do
    for row in input do
        calculate euclidean distance to each node
        BMU $\leftarrow$ nearest node
        update BMU weights to better approximate input vector
        update weights of BMU neighbours
    end for
    update neighbourhood function radius, $\sigma(t)$
    update learning rate, $\alpha(t)$
end for
output: SOM
```

2.3.3 Self-Organising Map Parameters

For a SOM to be built, several key parameters must be set: map network topology, configuration, and dimensions; the neighbourhood function; the learning rate; the decay function, and maximum iteration number. Further discussion on optimisation of these parameters can be found in Section 8.4.

Map Network

The array of nodes within the map may be arranged in a one-dimensional line, or higher dimensional lattice [56]. The most common topology is a two-dimensional sheet of nodes arranged regularly, although three-dimensional arrays, arrays with irregularly placed nodes, and dynamic arrays which assume their topology when they receive the training data do exist [58].

The configuration of the lattice nodes may be rectangular or hexagonal, where each node has direct connections with four or six neighbouring nodes, respectively, as shown in Figure 2.10. A hexagonal configuration is often preferred as nodes exert influence over more neighbouring nodes than for a rectangular one [56], although a rectangular configuration may be easier for non-experts to interpret.

The SOM network should be of an appropriate size to display the data well. There is no way to know the most appropriate size before training begins, and the results of training a SOM should be visually inspected to allow trial-and-error derivation of appropriate SOM size [56]. Varying the SOM size can allow finer or coarser resolution of the underlying clusters in the data. Vesanto [100] used the example of $5\sqrt{n}$ nodes (where n is the number of observations in the dataset) when investigating computational complexity of the clustering algorithm—this value is now widely used as a starting SOM array size.

The x and y dimensions of a regular rectangular lattice should be in the ratio of the two highest eigenvalues of the input data autocorrelation matrix, as this configuration makes convergence in learning faster [57].

Neighbourhood Function

The neighbourhood function, $\sigma(t)$, defines a symmetrical region around the BMU, the radius of which decreases monotonically with each iteration step, t [55]. Any surrounding nodes within the neighbourhood at the end of a training step have their weight updated to draw them closer to the BMU and further away from other nodes in the map.

There are several common choices for the form of the neighbourhood function [79]. A Gaussian function may be used where there is a continuous function across the neighbourhood, exerting a larger effect on weight values towards the centre of the neighbourhood and having decreasing impact toward the periphery. A step or

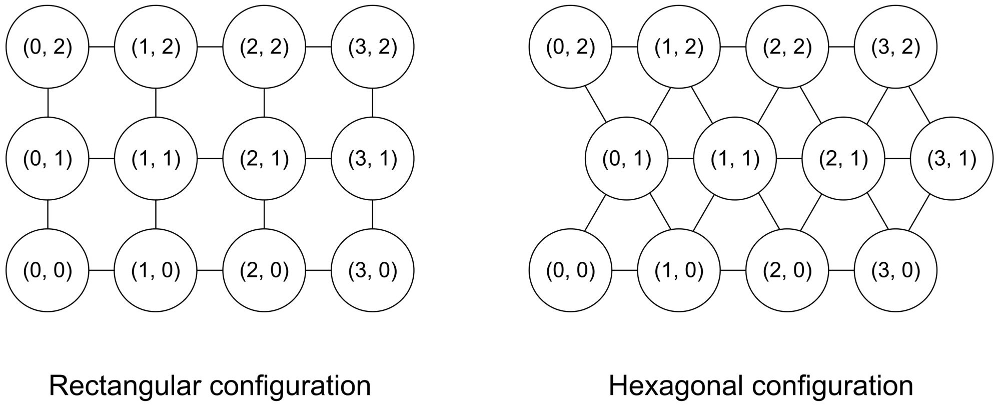


Figure 2.10: SOM lattice configurations may be rectangular (left) in which each node is connected to up to four neighbouring nodes, or hexagonal (right) in which each node is connected to up to six neighbouring nodes.

bubble function may be used where all nodes within the neighbourhood are updated equally. A triangle function uses a combination of bubble and Gaussian form. Early training iterations involve a wide neighbourhood and coarse organisation of data, so the choice of neighbourhood function will impact how data organise themselves, although an optimum neighbourhood function for a dataset cannot be known *a priori*. Due to the iterative decay of the neighbourhood radius, the majority of training iterations involve a narrow neighbourhood which encompasses one or zero surrounding nodes, so the choice of function used is unlikely to impact the later fine organisation phase [79].

The size of $\sigma(t)$ should be appropriate to the size of the map, and may be greater than half the map's diameter [56]. If the starting neighbourhood radius is too small, then the direction of organisation changes across the map and data become clustered in local pockets without global organisation [55].

Learning Rate

The learning rate, $\alpha(t)$, defines how much the weights of nodes in the neighbourhood are affected after each iteration of training [56]. It decreases monotonically with each iteration, t, so its effect on node weights is large at the beginning of training to allow coarse organisation, and small during the last iterations to allow fine grain resolution of clusters [56].

Decay Function

The decay function defines how the neighbourhood function's radius and the learning rate decay with each iteration step. Common functions include linear, exponential, and inversely proportional to t. With maps of up to a few hundred nodes, the actual function used is not crucial, as long as it allows $\sigma(t)$ and $\alpha(t)$ to decrease with increasing t [56]. With larger maps, optimisation of the decay function to minimise learning time and processing power required may be useful. [56].

Iteration Number

SOM learning is stochastic, and so requires many iteration steps to converge [56]. Kohonen suggests using at least 500 times the number of nodes in the network to ensure good statistical accuracy [56, p. 112], although there is no way to ensure that "enough" iteration steps are used.

Computational Complexity

Time studies have shown the complexity of the SOM algorithm to be linear with respect to the number of observations in the dataset, n, linear with respect to the number of dimensions of each observation, k, and quadratic with respect to the number of nodes in the map lattice, l [79]. Therefore, the overall complexity of the algorithm is $O(nkl^2)$.

2.3.4 Investigating the Reliability of Self-Organising Maps

SOM Error Metrics

No single metric can adequately describe the SOM method [28]. Two commonly used metrics, quantisation error and topographic error, assess the SOM's ability to reflect the distribution and topology of the input dataset, respectively.

The quantisation error is the average distance of each data point to its closest node in the lattice. This error expresses how well the SOM is representing the distribution of the input dataset, but has no connection to the topology of the data. Increasing the number of nodes in the lattice decreases quantisation error, but risks overfitting data as the ratio of data points to lattice nodes becomes ≤ 1 [79].

Topographic error is the proportion of input data vectors for which the best and second best matching units are not adjacent in the map network—that is they do not share a direct lateral connection (green lines in Figure 2.9) [79]. This error reflects how well the SOM is representing the topology of the input dataset.

Investigating Reliability

Once a SOM has been generated, it can be interrogated to ensure its results are reliable and valid. De Bodt et al. [23] have developed a toolkit which can be used to assess the reliability of a SOM's representation of the dataset under study. Their first measure is the coefficient of variation in the quantisation error of the map, used to assess if the quantisation error is consistent (and therefore the produced map is a reliable interpretation of the dataset), and to assess if the number of nodes used to make the map is appropriate, as an incorrect choice impacts the metric. Their second novel measure is a test of the reliability of the topographic error. Pairwise comparison of observations over the dataset is performed to investigate whether or not placement of the observations in the same or adjacent nodes of the SOM is significant. This measure tests significance of proximity over a given radius, r, around a node, so for $r = 0$ observations are together if they map to that node, and for $r = 1$ observations are together if they map to that node or one of the eight adjacent nodes in the (rectangular) lattice. If the placement of specific observations together is significant, then it can be inferred that the topology of the dataset has been well preserved, and that the observed clustering is not an artifact of random initialisation of parameters. This measure can be used to support the results of the SOM analysis as being a true reflection of input data topology, and to help to decide whether creating more SOMs with different parameters would be beneficial or necessary.

As SOMs are generated via a machine learning method from a random initial configuration, the same dataset may result in several slightly different SOMs despite using the same parameters. Therefore, an important consideration is the comparison of two or more SOMs. Most comparisons are made by visually inspecting SOMs, although some attempts at statistical analysis have been made. One possibility is use of a dissimilarity index, which is mathematically equivalent to the average difference in representation of the data by two maps [48], useful for analysing two SOMs created from different datasets. Kirt et al. [51] developed a similarity measure by visually inspecting graphs, defining clusters, defining a matrix representation of neighbouring nodes, and finally calculating how much the matrices are identical. This method is particularly useful for analysing two SOMs generated with different datasets and parameters. Mayer's group [71] have developed a method for comparing several SOMs generated from the same dataset based on visual inspection of output mapping. This group use the mean pairwise distance to compare an arbitrarily assigned index SOM against all other SOMs in the set, allowing a measure of how similarly each SOM plots a given data point.

2.3.5 Self-Organising Map Clustering Rationale

The SOM method works by spotting intrinsic patterns within data and clustering like observations together. Kohonen describes this as summarising high dimensional statistical summary data in a low dimensional space [56]. However, once the SOM is trained, it is not readily apparent which features of the input data cause it to cluster how it does. Methods for investigating which components of the input data are important for clustering (specific for that class) include visual inspection of component planes and some novel approaches.

A component plane is the array of weight vectors corresponding to one dimension (component) of the input data vector for each node in the map lattice [56]. These planes can be visually inspected and compared with their SOM, as areas of high signal in the component planes which correspond to clusters in the SOM suggest that this component may be influential in defining cluster membership. Such an approach is commonly used to assess the clusters in SOMs, but may not be feasible for spectral data where there may be thousands of component features and hence component planes.

Rauber and Merkl proposed the LabelSOM method of assigning cluster labels based on the similarity of input data components which map to a given node [82]. Their SOMs were formed using text data mining examples, whereby input datasets contain only the values one and zero to indicate presence or absence of a component within a text document, respectively. Their argument was that the weight of a node is mostly defined by the individual weights from each input observation mapping to that node, and so where components of input data were similar, they would equally contribute to the weight of that node and thus be characteristic of that node across the dataset. However, some features would have weights of zero due to the mass absence of a component mapping to that node, so the method was refined by adding a threshold value for node weight to indicate the minimum importance of a component to be considered.

Tan expanded the LabelSOM method with a hierarchical implementation—HLabelSOM [94]. This method involved labelling the nodes of the map using a threshold weight of 0.5 to define a component feature as descriptive of that node. His program could then create multiple maps at different hierarchical levels, so that if four nodes all map to one component they could be considered as a coalesced node of this feature at a higher level of abstraction.

Although the LabelSOM and HLabelSOM methods are useful in the field of text data mining, their use with Raman spectra is likely limited. The data used to test these two methods were binary, where inclusion or exclusion of a component feature from the input data vector enabled labelling of the nodes. With continuous spectral data, where there are sometimes thousands of feature components with only slightly different values, deciding on an appropriate cutoff value would be difficult.

2.3.6 Self-Organising Maps and Biological Data

Nuclear Magnetic Resonance Spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy is different to Raman spectroscopy, although there are more examples of its use with SOMs in the literature, where it has been widely used in the pharmaceutical industry for chemical analysis to ensure purity of products. More recently, the ability of SOMs to analyse biological systems has been tested [63]. Human saliva samples were treated with an oral rinse or water, and NMR spectroscopy performed. SOMs were utilised to analyse the dataset and successfully segregated treatment and control groups. SOMs were compared with PCA, and shown to be less affected by variables which were strongly discriminatory for only a small number of samples, which would constitute a poor biomarker to differentiate between the groups. The differences between spectra which the SOMs showed as important for cluster segregation were reviewed, and the biological mechanisms behind the different concentrations of biomolecules seen were postulated.

Raman Spectroscopy

There are few publications on the use of SOMs to analyse live human-derived Raman spectral data. The aim of this work is to test the feasibility of using SOMs to analyse such data, which are inherently complex due to the nature of dynamic cellular systems.

Harris et al. [38] analysed Raman spectra generated from immortalised human-derived normal thyroid cells and a human-derived aggressive thyroid cancer cell line. Using a supervised SOM method they were able to distinguish normal and cancer cells with >90% accuracy. Brazhe et al. [16] analysed primary rat cardiomyocytes by Raman spectroscopy to investigate the different enzyme levels in rod-shaped and round-shaped cells. They used a SOM method to cluster resultant signals which demonstrated separation of the two cell morphologies. Majumdar and Kraft [65] gathered Raman spectra from THP-1 cells, a cell line derived from human acute monocytic leukaemia [96]. They stimulated the cells to differentiate into a different type, and subsequent SOM analysis clustered spectra from cells at different stages during differentiation due to the altered intracellular biomolecules.

There is one example in the literature of SOMs used to analyse Raman spectra recorded from dried sections of pig eye [11]. The SOM successfully clustered spectra based on one of five tissue types of origin (Figure 2.12), although it should be noted that the spectra analysed by this group are themselves easily differentiated by eye (Figure 2.11), which is not the case for fresh prostate tissue samples. This paper demonstrates the ability of SOMs to analyse Raman spectroscopic data gathered from complex biological samples and to uncover more than two natural groups within

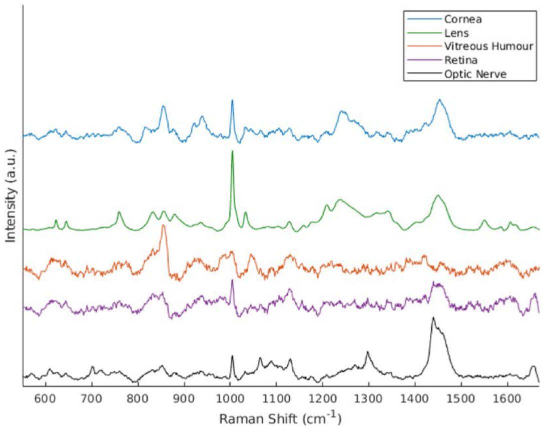


Figure 2.11: Average Raman spectra recorded for each layer of pig eye tissue (adapted from [11]).

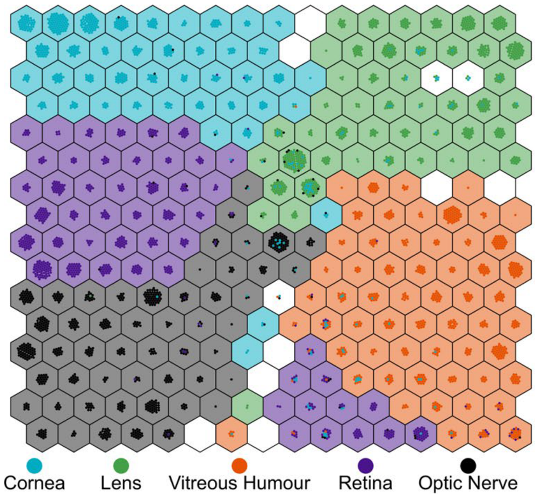


Figure 2.12: Self-organising map built from Raman spectra generated from each layer of pig eye tissue (adapted from [11]).

a sample set. This example is discussed more fully in chapter 6, where it is used to validate the SOM approach used here.

2.3.7 Self-Organising Maps with Supervised Learning

Kohonen mentions that the classification accuracy of SOMs can be increased if a supervised method is used [55]. However, the first principles approach for biology is to investigate blinded data, which helps to reduce bias in the results. Furthermore, if a SOM can cluster complex biological data, such as Raman spectral data from a cell system, in an unsupervised manner successfully, the method may be modifiable to be used for other biological datasets.

Chapter 3

Research Question

As demonstrated above, there is a real need for a reliable test for prostate cancer, one which gives both diagnostic and prognostic information. The current body of evidence supports the idea of using Raman spectroscopy to investigate human cellular samples, and the use of SOMs to analyse the derived spectra. Ultimately, a good test would be able to distinguish between different prostate cell types, such as non-cancer, cancer, another non-cancer disease state, and potentially a transition stage between non-cancer and cancer.

This research project is a feasibility study into the use of self-organising maps in the diagnosis of prostate cancer. The research question is:

"Can self-organising maps distinguish between cancerous and non-cancerous prostate cells?"

Chapter 4

Materials and Methods

The materials and methods used throughout this project are summarised here. The acquisition, storage, and use of data are discussed, and the methods for recording results are outlined. Good software engineering techniques are used throughout this work, and evaluation of their method is covered below.

4.1 Data Management

A data management plan covering details of data storage and use was produced prior to commencing experimental work. Data are stored in hierarchical directories outside of the project code directory, and raw data are read into scripts and written to new files for manipulation if necessary—original data are never edited directly.

4.1.1 Practice Datasets

Practice datasets for testing MySom code include the Iris Flower dataset [31] and a dataset of RGB colour values [17]. The Iris dataset is chosen because it is a commonly used dataset in statistics and machine learning with biologically well defined groups. Furthermore, two of its classes are not linearly separable, enabling assessment of the utility of clustering algorithms. The Colour dataset is chosen as it is visually not difficult to interpret, and class membership is defined by the highest RGB value for each observation, so which colours cluster together may give insight into how the SOM clusters data. Each of these labelled datasets is contained in a single file within the dataset directory.

Raman spectral data on pig eye tissue samples were acquired from the authors of the Banbury paper [11]. These data are organised as a single file per observation, within individual directories for each tissue type, within the dataset directory.

4.1.2 Human Data

The Raman spectral data from prostate cell lines have been gathered in Dr Hancock's laboratory, Department of Physics, University of York. They are provided by Dr Hancock as two datasets, one labelled and one blinded, and each is stored in its own directory. There are no ethical conflicts of interest as data are gathered from commercial cell lines [13, 44].

4.1.3 Data Analysis

The MySom module has been written in Python. It inherits the characteristics of MiniSom [101], and includes code for normalisation of Raman spectral data and production of consistent SOM plots. MiniSom is chosen as the basis for MySom as it is shown to be a popular and versatile package following literature search and review within a recent masters thesis [104].

4.1.4 Appropriateness and Limitations

The proposed analysis methods are reasonable given the body of evidence supporting the use of SOMs to uncover subgroups in large sets of complex biological data. The benefit of the SOM method is that it is unsupervised; there is no assumption of which subgroups are expected, and the algorithm clusters observations without bias.

The main constraints of the method are post-processing analysis: once clusters are defined, analysis of subgroup members is required as it is not readily apparent why the data are clustered as they are. This process is simple if the SOMs separate samples based on expected results, such as cancer and normal for the prostate data, although there is a possibility that other groups may become apparent (perhaps non-cancer disease, transition from normal to cancer, or another unexpected group).

Further to being an appropriate analytical technique, SOMs simplify the visual representation of complex data and are easy to understand, and so lend themselves readily to clinical medicine. A trained SOM can be given new patient data, whose position within the map can be calculated and highlighted. From this image, doctors can understand the significance of the result quickly and use the SOM to form part of the reasoning behind clinical decisions and as a visual aid in patient explanation.

The generated SOMs look for differences in peak intensity bands of the Raman spectra generated from prostate samples. Once these differences are uncovered, potential biological mechanisms underlying the perceived changes in spectral bands can be postulated. However, caution must be utilised, as the spectral fingerprint of each cell reflects both the cellular constituents and environment, and the connections between these complex systems are non-linear—if a given biomolecule is present more in a Raman active state, it does not necessarily mean that its concentration is different.

Once clusters are found, results can be compared with what is known about the prostate samples. This analysis allows a qualitative review of how well the known disease state of the sample matches the output of the SOMs, and any discrepancies can be explored. It is expected that the SOM finds at least as many clusters as clinically classified groups if it is a good method for analysing these data—it may find more groups due to the higher resolution of information gained by Raman spectroscopy (intracellular biomolecular changes) than optical microscopy (tissue architecture). The SOM method can be statistically evaluated by comparison of the clinically classified data with the SOM clusters and subsequent Bayesian statistical and receiver operating characteristic curve analysis.

Once a putative mechanism of action to explain the observed cluster differences is proposed, new biological laboratory experiments can be performed to test these hypotheses, and hopefully to provide new data for further computational analysis.

4.2 Recording Experiments and Results

4.2.1 Logbook

A digital logbook detailing all experiments is kept with the following sections and information:

- What is expected from the code run
- What precisely is done during the code run
    - What code is used
    - What parameters are used
    - What data are used
- Results of the code run
- Discussion of the code run and results
- Addenda
    - Any new information recorded with time and date added

4.2.2 Appropriateness and Limitations

This method of recording experiments is useful, as each stage from conception to execution to discussion is laid out clearly. Recording experiments in this way allows future researchers to follow what has happened, and highlights when information has been added.

4.3 Good Software Engineering Practice

4.3.1 Version Control System

All code is stored in a GitHub repository at `github.com/thenakedcellist/prostate`. This allows the module to be freely available for others to use and quick recovery of older versions if refactoring causes loss of code.

4.3.2 Test Environments

The code for the MySom module is accompanied by a full test suite run using pytest [59]. This practice reduces the chance of a runtime error due to the code itself, and ensures any errors are dealt with using simple known test cases so that code can be optimised during refactoring before use with real data.

Chapter 5

Preliminary Experiments

This chapter discusses the first short experiments performed to check that a module based on MiniSom [101] works as expected. A fully operable tool chain requires a program that can access relevant data files within the data directory, load the correct data in a usable format, normalise the data appropriately, analyse the data using MiniSom, and produce output SOMs and graphs.

The MySom module is written to integrate the methods of MiniSom with these required functionalities. It is tested using the Iris flower dataset [31] and a RGB colour dataset [17], commonly used datasets for classification problems in statistics and machine learning. MySom is able to correctly cluster the three species of iris within the Iris Flower dataset, including being able to mostly segregate data from two species which are not linearly separable [31]. MySom is also able to correctly cluster the colour data into groups based on the highest RGB byte value, and the resultant SOM layout gives some insight into the way in which clusters are formed.

5.1 MySom Module

The MySom module unifies all the tasks necessary for building and interrogating SOMs, and is used for all subsequent analyses. It inherits the features and functions of MiniSom, and extends its capability to include tools for data normalisation, as shown in algorithm 5.1, and plotting of output SOMs. It uses a rectangular lattice configuration to simplify the output for non–computer scientists. The full MySom pseudocode is found in Appendix A. The variables required for the code to run are:

- Map dimensions: x and y
- Number of variables recorded for each observation: k
- Starting neighbourhood radius: $\sigma(0)$
- Starting learning rate: $\alpha(0)$
- Decay function
- Configuration of map array: rectangular or hexagonal
- Random seed: integer or None

Algorithm 5.1 MySom Normalisation and Self-Organising Map Plots

```
input: Raman spectral data
function NORMALISE DATA($A_{m,n}$)
    for $A_{m,n}$ do
        $x_{i,j} \leftarrow x_{i,j}/\|a_{i,*}\|_F$        ▷ divide by Frobenius norm
    end for
    return $A_{m,n}$
end function

procedure MAKE SOM U-MATRIX(input)
    NORMALISE DATA(input)
    SOM TRAIN($input$)
    MAKE SOM U-MATRIX
end procedure
output: SOM u-matrix
```

An example SOM built with synthetic data is shown in Figure 5.1. The output u-matrix is coloured such that darker areas represent nodes which are closer to their neighbours, and lighter colours those which are less densely packed, in order to give a visual representation of data distribution across the map. This plot is useful for discerning the border between clusters, where regions of less densely packed nodes reflect fewer mapped data points than more densely packed regions, such as the low nodal density stripe down the centre of this map. The input data have been overlaid, each data point lying on the node which is closest to it, with a random jitter to ensure that data points do not overlap. Therefore, the position of a point within a node on the map does not reflect actual distance from other points mapped to the same node.

An example density plot derived from the SOM in Figure 5.1 is shown in Figure 5.2. This plot shows the density of the input data across the map space, with dark orange regions corresponding to areas of high density and pale orange areas corresponding to areas of low density. This plot is useful for interpreting where the centre of a cluster lies, as this information is not as readily apparent in the u-matrix image where nodes are shown of uniform size in a regular array and their colour is used to represent distance from one another in the map space. In this case there are two data foci in the lower left and upper right regions of the map, separated by the stripe of low nodal density across the map centre seen in Figure 5.1.

5.2 Iris Flower Dataset

The Iris flower dataset was collated by Edgar Anderson [5] following measurement of 100 iris plants on Quebec's Gaspé peninsula and 50 iris plants native to the Southern United States. His sample consisted of 50 individual plants from each of the three distinct species, *Iris setosa*, *Iris versicolor*, and *Iris virginica* [6]. Anderson worked

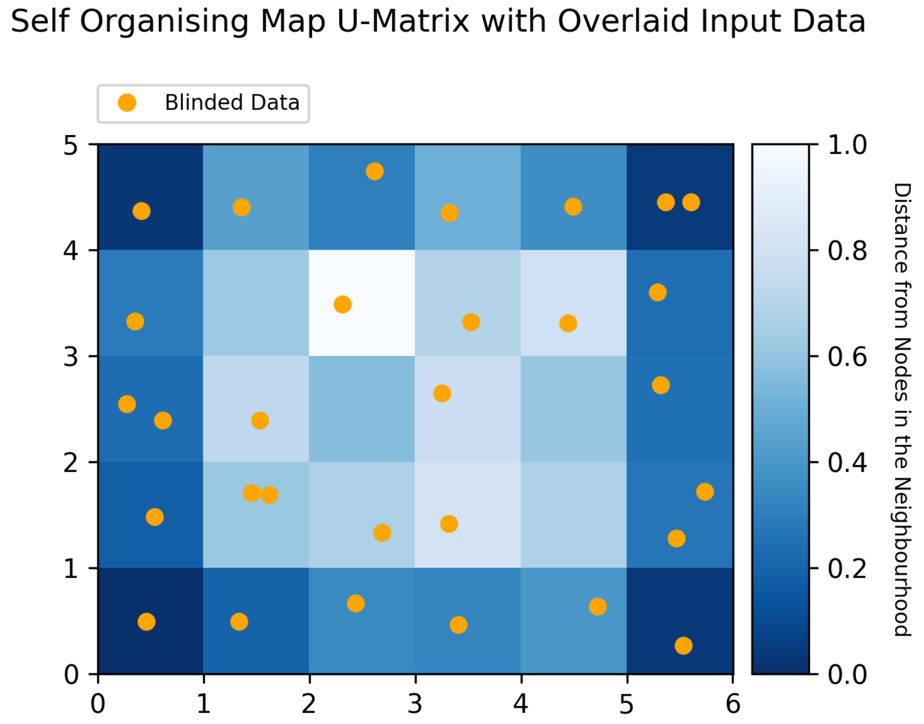


Figure 5.1: A SOM built with synthetic data to highlight key features of the SOM layout. The square nodes of the map array are coloured to reflect their distance from their neighbouring nodes, such that darker regions of the map represent areas of densely packed nodes. The lightly coloured nodes (2, 3), (3, 2) and (3, 1) form a stripe of low nodal density which may represent a border between clusters. Each orange dot is one of the input data observations placed on the array node which maps to it most closely in the map space. A random jitter is added so that multiple observations within one node do not overlap and obscure one another—relative position within a node does not correspond to distance between those observations in the map space.

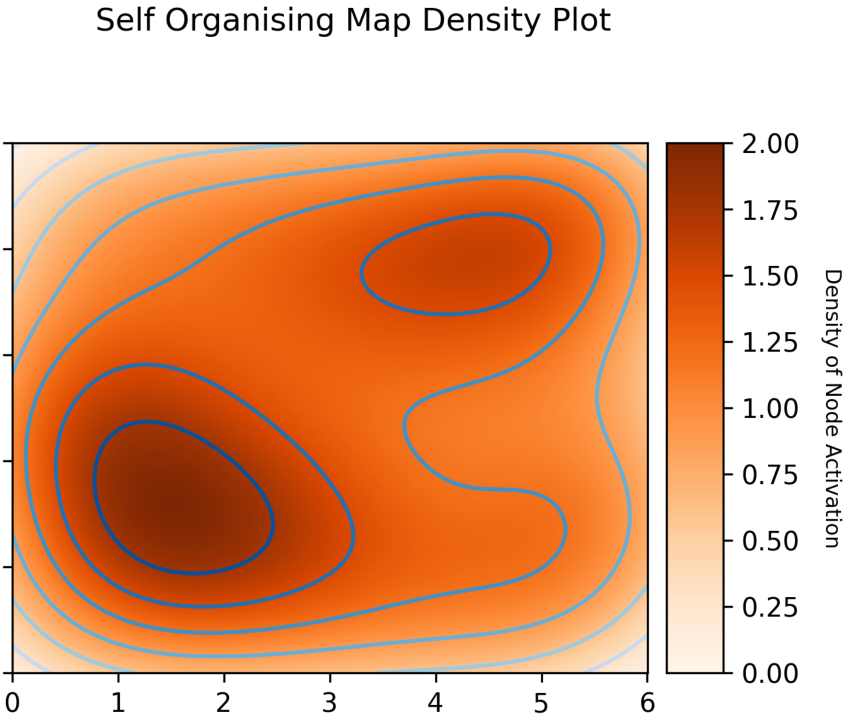


Figure 5.2: The density plot of the SOM built with synthetic data shown in Figure 5.1. The darker regions of the plot represent areas where data points are densely packed in the map space, with the lighter regions representing areas of data sparsity. Two foci of data density are seen centred around (1, 1) and (4, 3), suggesting that there are two clusters within this dataset. The stripe of low nodal density across (2, 3), (3, 2) and (3, 1) in Figure 5.1 corresponds to the paler region between the two data foci shown here, which confirms this is a border between clusters.

Species	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)
I. setosa	5.1	3.5	1.4	0.2
I. versicolor	7.0	3.2	4.7	1.4
I. virginica	6.3	3.3	6.0	2.5

Table 5.1: Subset of the Iris Flower dataset adapted from Fisher's paper [31].

closely with statistician Ronald Fisher, who used the data set to demonstrate his method of linear discriminant analysis (LDA) [31]. LDA is a method used to classify groups based on differences between linear combinations of statistical characteristics, and has been developed for dimensionality reduction of high dimensional data [46].

The Iris Flower dataset has become very popular in the fields of statistics and machine learning for testing the ability of classification algorithms, as it is a simple dataset with low dimensionality [24]. *Iris setosa* is linearly separable from the other two species, whereas the variances in measurements observed for *Iris versicolor* and *Iris virginica* overlap, so the ability of a classification algorithm to correctly group these observations can be used as a measure of its utility.

5.2.1 Materials and Methods

The dataset is a 150×5 array in which each row contains observational data from a single iris flower (Table 5.1). It is a labelled dataset, with the first column containing the species classification and the subsequent columns containing measurements from the flower's petal and sepal.

The entire Iris Flower dataset is analysed with a rectangular array configuration and the following parameters:

- Map dimensions: 8×8
- $\sigma(0)$: 3.0
- $\alpha(0)$: 1.0
- Random seed: 1

5.2.2 Results and Discussion

In the SOM u-matrix (Figure 5.3) there is a stripe of low nodal density spanning from the lower centre to centre right of the map, representing an area of sparsely packed nodes separating the two densely packed blue regions. This stripe most likely represents a clear divide between clusters as organised on the map. The map is trained unsupervised, and the input data are subsequently overlaid to show which nodes match to them most closely. *Iris setosa* forms the lower right cluster separated from the other data points by the stripe of nodal low density. The other two species appear to be somewhat well separated, with *Iris versicolor* spanning the upper right

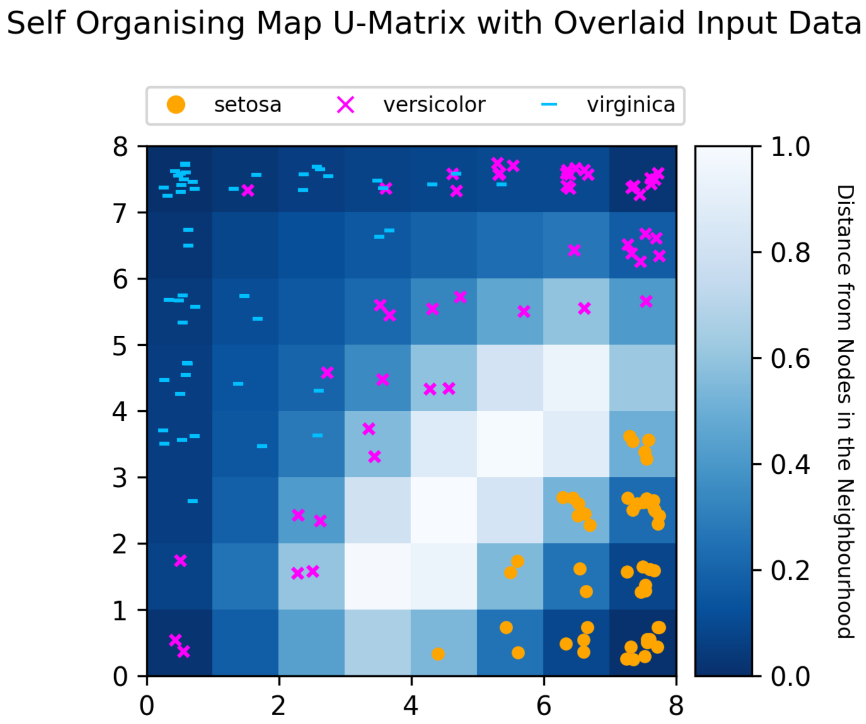


Figure 5.3: SOM u-matrix trained on the entire Iris Flower dataset. There is a stripe of low nodal density extending from (3, 0) to (7, 4) which cuts the map into one small region at the lower right and one large region to the left. This stripe likely represents a border between clusters.

of the map and along the low density stripe, and *Iris virginica* inhabiting the upper left of the map.

Figure 5.4 shows the density plot for the SOM in Figure 5.3. There are three foci of data density, one centred around (7, 1) corresponding to the *I. setosa* cluster, one centred around (1, 7) corresponding to the *I. virginica* cluster, and one centred around (7, 7) corresponding to the *I. versicolor* cluster. The focus of data for *I. setosa* is much more dense than the foci for the other two clusters, and this is because the data are spread across only ten nodes in one corner of the map, whereas data for the other two clusters are spread across a larger number of nodes.

The data are being classified as expected by the SOM. *Iris setosa* is located over ten densely packed nodes and completely separated from the other two species by a large gap of map space (Figure 5.3). The map region containing *Iris versicolor* and *Iris virginica* shows a wider distribution of the input data, possibly signifying that there was difficulty in fully separating these two groups—this is unsurprising as Fisher's original paper states that these two groups are not linearly separable and their variances overlap [31]. There does appear to be a focus for each species in the upper left and upper right corners of the map, with overlap around the border between the two groups, and an observation of *I. versicolor* within the *I. virginica* cluster on node (1, 7).

Ultimately, this example shows that the toolchain works, the MySom module can

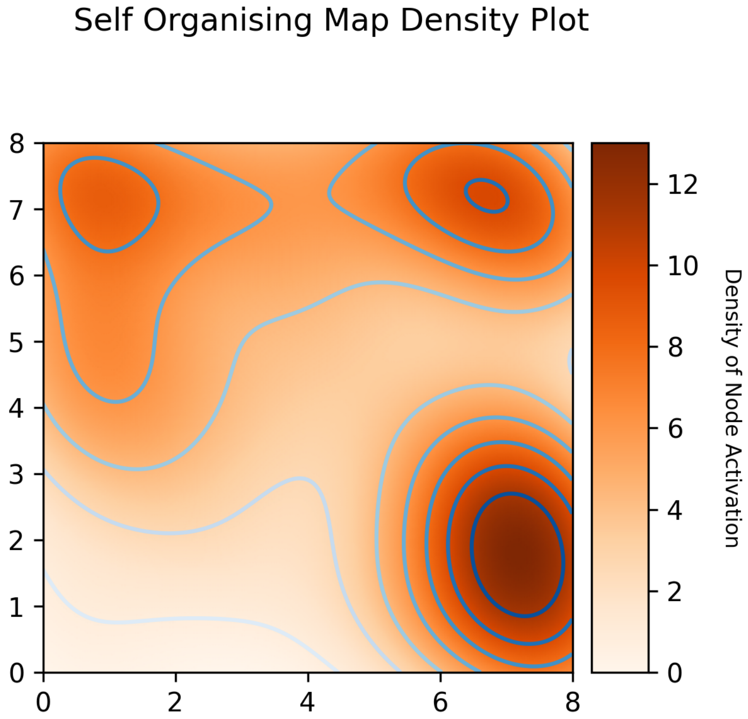


Figure 5.4: Density plot for the SOM shown in Figure 5.3. A focus of high data density is seen centred around (7, 1), with two less dense foci centred around (1, 7), and (7, 7). The high density focus at the lower right corresponds to the cluster for *Iris setosa* in the SOM, and the two less dense foci represent the centre of clusters of the other two iris species. These foci are less dense because the same number of observations are spread over more map nodes than for *I. setosa*.

be used to analyse the data and to produce expected results. Ponmalai and Kamath used SOMs to investigate the Iris Flower dataset (classes were named arbitrarily), and found similarly that one class was easily separated from the others by a stripe of low nodal density on the SOM, and that the other two classes were positioned very closely together without a clear border between them [79]. They concluded that each species class was grouped into a single cluster, and that there was some crossover between the clusters of the two linearly inseparable classes, as found here.

5.3 Colour Classification

The Colour dataset is derived from a GitHub repository containing a database of colours, their names, their hex code, and their RGB byte code [17]. Each three-dimensional vector of RGB values is assigned a categorical colour label, based on which RGB component has the highest byte value. If a single component has the highest value, the colour is labelled red (R), green (G), or blue (B); if two components have an equally high value, the colour is labelled the additive secondary colour cyan (C), magenta (M), or yellow (Y); and if all three components are of equal value, the colour is labelled black (K).

5.3.1 Materials and Methods

The Colour dataset is a 865 × 4 array in which each row contains data for one colour. The first column contains the categorical colour identifier, and the second, third, and fourth columns contain the byte value of the red, blue, and green colour components, respectively (Table 5.2). The number of observations within the dataset belonging to each class and the class derivation are:

- 481 Red $(R) : R > G \wedge R > B$
- 130 Green $(G) : G > R \wedge G > B$
- 193 Blue $(B) : B > R \wedge B > G$
- 14 Cyan $(C) : G = B \wedge G > R$
- 18 Magenta $(M) : R = B \wedge R > G$
- 15 Yellow $(Y) : R = G \wedge R > B$
- 14 Black $(K) : R = G = B$

The entire Colour dataset is analysed with a rectangular array configuration and the following parameters:

- Map dimensions: 12 × 12
- $\sigma(0)$: 3.0
- $\alpha(0)$: 1.0
- Random seed: 1

Categorical Colour Identifier	Red Byte Value	Green Byte Value	Blue Byte Value
R	163	38	56
G	164	198	57
B	93	138	168
C	0	255	255
M	139	0	139
Y	132	132	130
K	0	0	0

Table 5.2: Subset of the Colour dataset adapted from a GitHub repository [17]. The categorical colour identifiers refer to red (R), green (G), blue (B), cyan (C), magenta (M), yellow (Y), and black (K).

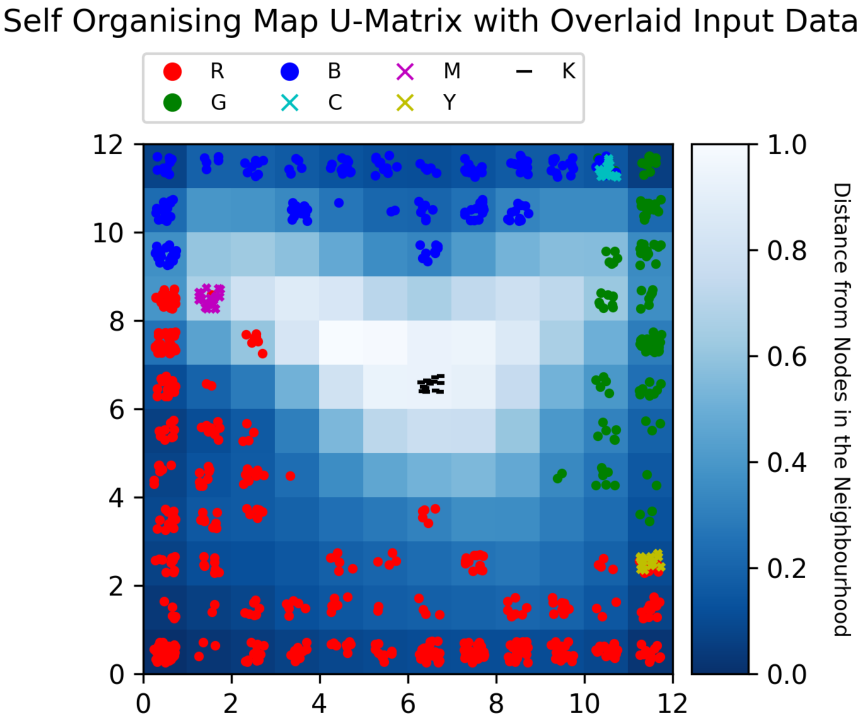


Figure 5.5: SOM u-matrix trained on the entire Colour dataset—overlaid input data are coloured according to categorical colour label. The primary colours are spread across densely packed regions of the map: red (R) over a large region at the lower left, green (G) at the centre right, and blue (B) at the upper centre. The secondary colours each map to one single node between the clusters for their respective primary colour constituents: cyan (C) at (10, 11) between green and blue; magenta (M) at (1, 8) between red and blue; and yellow (Y) at (11, 2) between red and green. The tertiary colours, labelled black (K), are located at (6, 6), a central node in a region of very low nodal density, signifying that these data points are very different from the others.

5.3.2 Results and Discussion

The SOM in Figure 5.5 shows good segregation of the colour data across the map space. Several SOMs were produced with different parameters, which all showed similar segregation of the colour clusters.

The primary colour groups categorised as red, green, and blue are discretely separated to the lower left, centre right and upper centre, respectively, with only minor overlap of green and blue on node (10, 11) at the border between these two groups. It is likely that this overlap exists because the RGB values of the mapped input data are similar enough for this node to be closest to some green and some blue observations. The secondary colours each map to a single node at the border between their constituent primary colours: (10, 11) for cyan, (1, 8) for magenta, and (11, 2) for yellow. The tertiary colours, labelled black, all map to node (6, 6), in an area of low nodal density with surrounding nodes far from their neighbours.

The way in which the colours have been so clearly delineated likely reflects the SOM's method of clustering data. A SOM can be considered a graph of the similarity of statistical relationships of high dimensional data plotted in a low dimensional form [56]. By definition of the labelling process used, each primary colour is very likely to have three different byte values for red, green, and blue (although the two lower values could be the same), each secondary colour must have two values the same, and each tertiary colour must have all three values the same. Therefore, the variance within the RGB values for the primary colours is higher than for the secondary, and the variance for the tertiary colours is zero.

This difference in summary statistics for the classes likely explains why they have clustered as they have, with least variance in the small central region of the map and most variance toward its larger periphery. One caveat of this interpretation is the difference in proportion of the three colour types within the dataset: there are many more primary colours than secondary or tertiary, which will affect the map's ability to form clusters. However, there are similar numbers of secondary and tertiary colours, so it is likely that the difference in variance between these groups has caused their separation on the map.

5.4 Summary

This chapter covers the introduction of the MySom module designed to analyse high dimensional input data and to build SOMs. Two datasets are used to perform baseline tests on MySom showing it to produce expected results and to be suitable for application to real data.

Use of the Iris Flower dataset shows that the SOM can easily distinguish the one species of iris which is linearly separable from the other, and further that it can correctly cluster the two species which are known to be linearly inseparable [31].

This ability to discern between the two similar species of iris shows the SOM to be a useful tool for analysing data in complex non-linear systems.

Use of the Colour dataset shows that the SOM is able to correctly cluster RGB colour data as expected, and also gives insight into how the SOM method works by comparing the statistical summaries of input data vectors. In this manner it is able to retain as much information as possible from the three-dimensional dataset while displaying it in two-dimensional space.

The next chapter evaluates the use of MySom to analyse high dimensionality Raman spectral data from biological tissues, and compares the output to results published from the same dataset.

Chapter 6

Pig Eye Experiment

This chapter covers the transition from using small low dimensionality datasets to a large high dimensionality spectral dataset to test MySom. The Banbury group conducted experiments on eye tissue from pigs, in order to assess the feasibility of using Raman spectroscopy and SOMs as a classification tool [11].

The Pig Eye dataset is restructured for use with MySom and used to test the module's ability to cluster Raman spectral data. A different subset of the dataset to that published is used to test MySom, and the reasons for this difference are discussed. Output SOM results are comparable between the MySom and published methods.

6.1 The Pig Eye Dataset

6.1.1 Original Dataset

The original Pig Eye dataset contains Raman spectral data generated from tissue sections of pig eyes. These samples were collected from eleven pigs, from each of five sites within the eye (cornea, lens, optic nerve, retina, vitreous) [11]. 88 Raman spectra were generated per site from each animal, giving $11 \times 5 \times 88 = 4840$ observations.

The data are organised as a single file per observation within five directories, one for each tissue type. Data in the files are arranged in two columns, the first for wavenumber and the second for measured intensity. This format is different from the Iris Flower and Colour datasets used previously, and the data to be used from the prostate cell lines, which are arranged with each row representing a single observation.

In order to convert the raw data into the appropriate format, a script is written which parses the titles of the individual files within the directory and subdirectories, and creates a data array within a .csv file with the correct format for the MySOM module. An assert statement is used in order to stop the code run should the values

Identifier	1669.551758 (cm^{-1})	1668.565430 (cm^{-1})	1667.580078 (cm^{-1})	1666.593750 (cm^{-1})	⋯
animal_1_cornea_7	5.921497	186.249908	57.344173	8.223427	⋯
animal_2_lens_5	14.322484	51.075981	30.978479	26.05842	⋯
animal_3_optic_nerve_53	36.640327	5.393722	40.284195	11.233159	⋯
animal_4_retina_0	95.318336	0.791617	30.455341	1.325872	⋯
animal_5_vitreous_2	68.758102	65.646217	47.07341	38.555355	⋯

Table 6.1: Subset of Pig Eye dataset for Animals 1 to 6

for wavenumber not match between samples.

At runtime this checkpoint is triggered, as there are two sets of wavenumbers used: one for animals 1 to 6, and a different set for animals 7 to 11. It may be the case that the recording machinery was calibrated between experimental runs, and the author has been contacted and agrees with this suspicion.

6.1.2 Restructure of the Dataset for MySom

After processing, the data are split into two directories based on which value of wavenumber is recorded: data for animals 1 to 6 are in the first directory, and data for animals 7 to 11 are in the second.

Within each directory, data are stored as two files. The first file contains a one-dimensional array, the first column of which contains the identification string 'wavenumber', and subsequent columns which contain values for wavenumber. The second file contains a two-dimensional array, the first column of which contains an identification string containing animal number, tissue type and sample number for that observation, and subsequent columns which contain recorded intensity for the corresponding wavenumber.

The recorded data from animals 1–6 is a 2640 × 1015 array (Table 6.1), and the recorded data from animals 7–11 is a 2200 × 1015 array. The Banbury paper [11] does not explicitly state the instrumental accuracy of their experiment, and data are presented exactly as in the raw files obtained from the group.

6.2 Published Pig Eye Analysis

The Banbury group created a SOM using a subset of the Pig Eye dataset. They randomly removed 25% of observations for testing and built the SOM with the other 75% [11]. These data were removed at the tissue level, and not at the level of individual animals, so it is unknown exactly which observations were used to train the map and whether each animal was equally represented.

According to their published source code [10], the Banbury group built their SOM with a hexagonal array configuration and the following parameters:

- Map dimensions: 16 × 16
- σ: 1.0–0.3

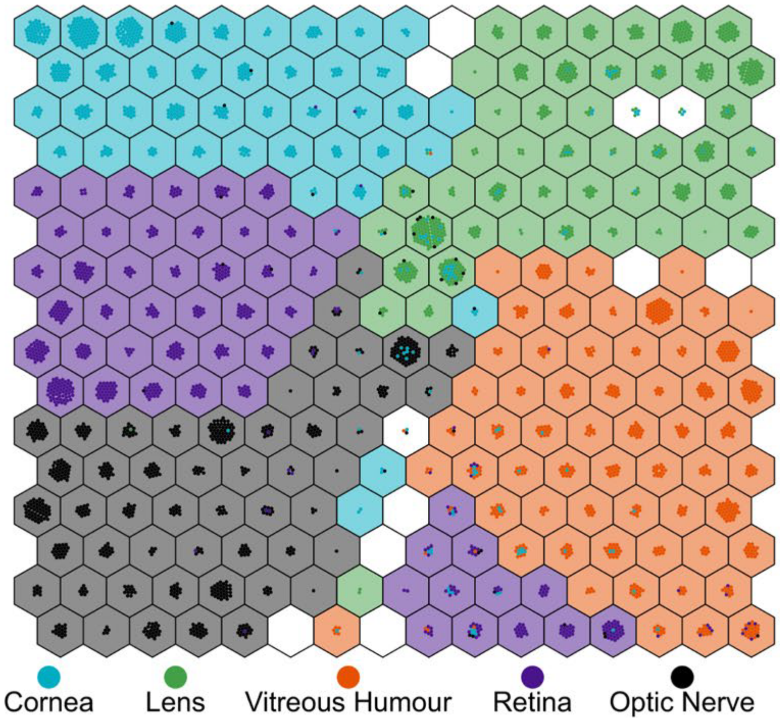


Figure 6.1: The original SOM published by the Banbury group [11]. Each node is coloured to match its modal activating tissue type, and each small dot within a node represents a single observation mapping to that node. There is good delineation of tissue types and an even distribution of modal tissue types.

- α: 0.4–0.1

Ranges are given for neighbourhood radius (σ) and learning rate (α) as the group use a linear decay function for these parameters, tending from maximum to minimum over 10000 iterations.

The map shown in Figure 6.1 shows good delineation of tissue types, although retina appears to segregate into two clusters. In their paper, the group argue that the retina cluster toward the mid bottom of the plot reflects noise in the signal, as several of its nodes and the surrounding nodes are not activated by a clear modal tissue type [11]. The authors do not speculate on this, but there is the possibility that tissue types were not completely separated during sampling, as the retinal layer of the eye is only five cells thick, is bathed on one side by vitreous humour, and surrounds the central optic nerve. A diagram of a human eye cross-section is given for reference in Figure 6.2.

6.3 Re-Analysis of the Data with MySom

Data for the first six animals of the Pig Eye dataset were re-analysed using MySom, using a rectangular array configuration and the following parameters:

- Map dimensions: 16 × 16
- $\sigma(0)$: 1.0

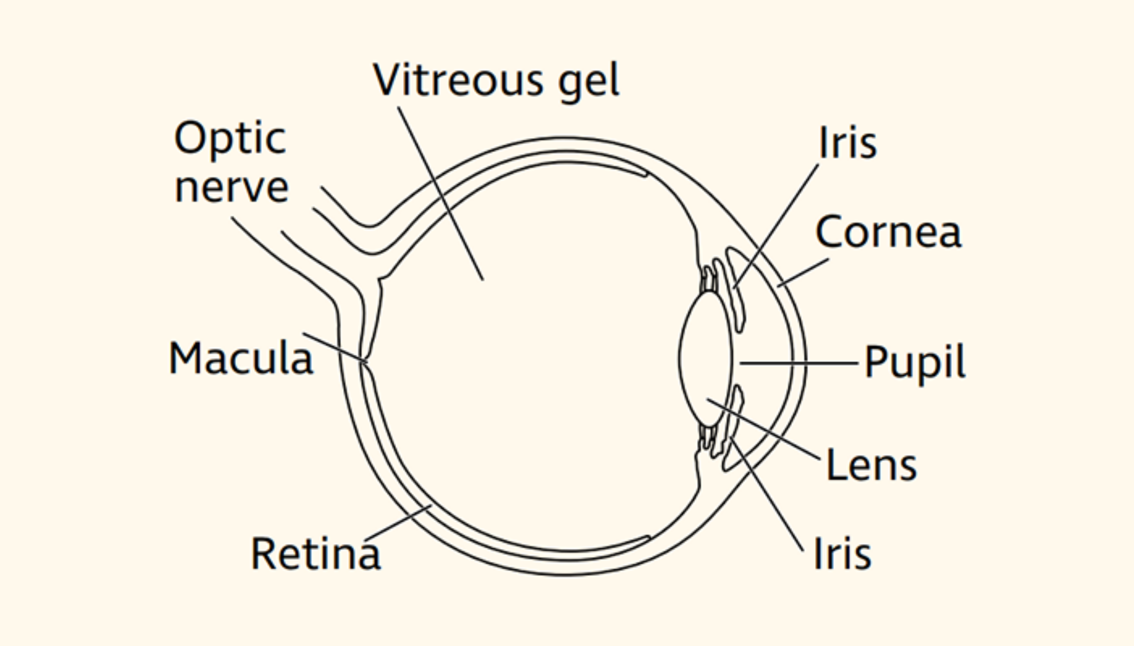


Figure 6.2: Diagram of a human eye cross-section showing the locations of the cornea, lens, optic nerve, retina, and vitreous.
Reproduced from https://www.rnib.org.uk/eye-health-eye-conditions/how-eye-works

- $\alpha(0)$: 0.1
- Random seed: 1

A rectangular lattice is used to test MySOM's ability to produce comparable results with the published data, rather than to replicate the data exactly. This experiment is a test of MySOM's ability to analyse biologically-derived high dimensionality data and to produce meaningful results, as this is ultimately the aim of the main experiments using prostate cell data. The resultant SOMs will ultimately need to be analysed by non-computer scientists, by doctors and patients who may be lay people, and a rectangular lattice is used to aid their ease of interpretation.

6.3.1 Results and Discussion

The SOM u-matrix in Figure 6.3 shows a stripe of low nodal density separating the densely packed nodes of the lower right corner and the rest of the map, likely signifying a cluster in the lower right corner. The rest of the map shows densely packed nodes at the edges and corners, with another stripe of low nodal density traversing from the upper centre to the centre left. These less densely packed areas likely represent edges between clusters. This SOM shows gross separation of the tissue type classes, with discrete clusters defined for the cornea and vitreous. Data for the optic nerve appear to be segregated into a large group at the upper right of the map and a smaller group at the lower centre, with a few single points scattered across the map. Lens has split into two distinct groups, a large one at the lower

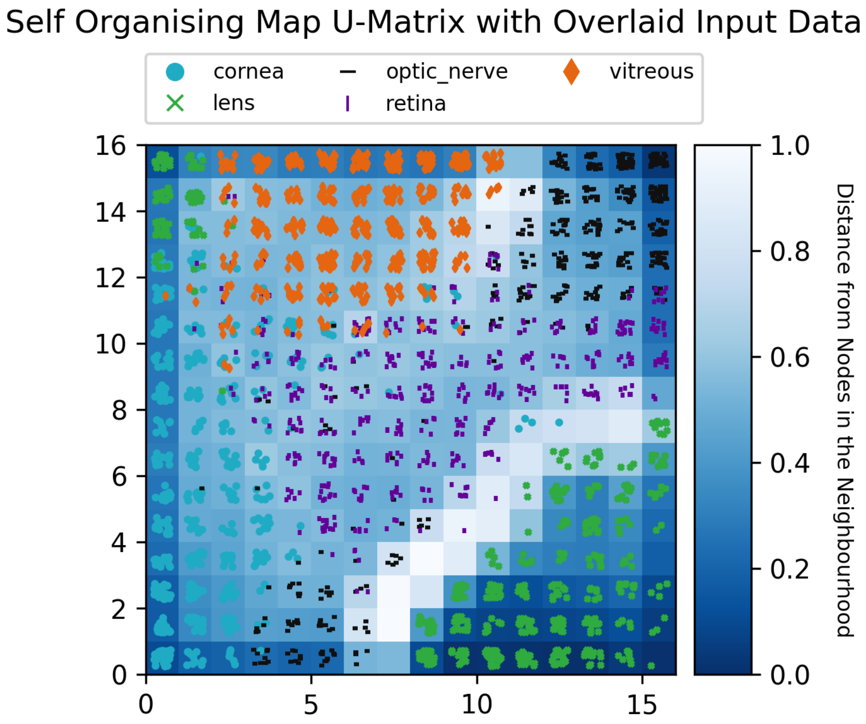


Figure 6.3: SOM u-matrix trained on the first six animals of the Pig Eye dataset with overlaid input data. There is a white stripe across the lower right corner, indicating a large distance between nodes at the lower right and other nodes in the map: this represents a border between the cluster containing lens data in the lower right corner and the rest of the map space. There is another stripe of low nodal density extending down and left from (10, 14), which corresponds to the border between optic nerve, retina, and vitreous.
The bulk of the map is occupied by delineated clusters of each class, with cornea at the left, vitreous at the upper centre, optic nerve at the upper right, retina at the centre, and lens at the lower right. Optic nerve has a second smaller cluster at the lower centre, and there is a small cluster of lens at the upper left. Data for retina are spread over a large number of map nodes of medium density, whereas all other clusters map to an area of more densely packed nodes.

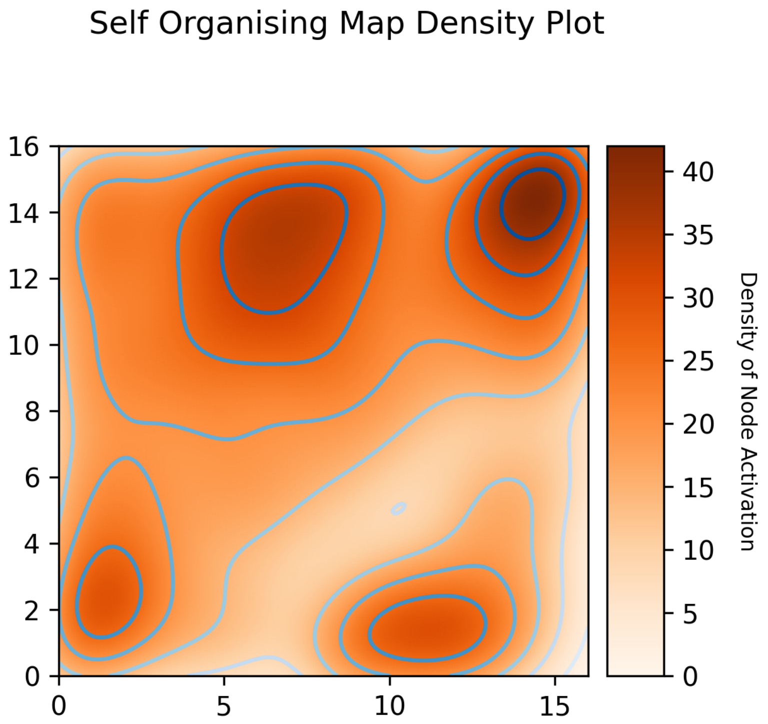


Figure 6.4: Density plot of the data distribution in Figure 6.3 showing four dense groups centred around (7, 13), (14, 14), (1, 2), and (11, 1) corresponding to major cluster centres for vitreous, optic nerve, cornea, and lens, respectively. There is no clear central focus corresponding to retina, which reflects the distribution of these data across many not densely packed nodes in the centre of the map in Figure 6.3.

right and a small one at the upper left. Data for retina appear to be spread across many nodes of medium density within the centre of the map in a single large cluster.

The density plot in Figure 6.4 shows four dense data foci which correspond to cornea, lens, optic nerve, and vitreous, the four clearly segregated tissue types in the published study [11] and in the MySom re-analysis in Figure 6.3. There is no central dense focus corresponding to retina, which is unsurprising given that data for this tissue type were spread widely over medium density areas of the map in Figure 6.3. Furthermore, data for this tissue type are divided into two clusters in the Banbury group's published map (Figure 6.1), so it is likely that these data are more heterogeneous than those of the other tissue types.

6.3.2 Comparison with Published Data

The plots produced by MySom are comparable with the published images from the Banbury group [11] and show that the module works as expected. There are notable methodological differences between the MySOM method and the published data, such as use of a rectangular lattice with an asymptotic decay function here, and use of a hexagonal lattice with a linear decay function in the published method [10]. These difference have been introduced for two reasons: firstly, to test MySOM's ability to analyse complex, biologically-derived spectral data and to produce results comparable with published data without copying them verbatim; and secondly, to show that the methodology that is to be used in the main experiments with Raman spectra gathered from prostate cells will produce meaningful, interpretable results.

MySom can produce SOMs trained on the high dimensional Raman spectroscopic data to successfully cluster observations by tissue type, the low nodal density regions of the map can be used to interpret where borders lie between clusters, and density plots reflecting data distribution across the maps can show central dense foci of data within clusters.

In the Banbury group SOM in Figure 6.1, the data segregate into six clusters, one for each tissue type and a second small cluster for retina, with a few nodes at cluster borders mapping to more than one tissue type. In the MySom output SOM in Figure 6.3, the data are segregated into clusters, one for each tissue type and a second smaller cluster for optic nerve and lens. Similarly to the Banbury group map (Figure 6.1), retina does not seem to cluster into a single region in the MySom map (Figure 6.3) but is rather spread across a large central region. The MySom density plot (Figure 6.4) does not show a central dense focus corresponding to retina, as these data are spread over a large number of nodes such that each node is activated by fewer observations. This result highlights the importance of interpreting SOM u-matrices with overlaid input data (Figure 6.3) and density distribution (Figure 6.4), as this SOM u-matrix alone does not show clear borders between all five tissue groups and the density plot shows only four data foci.

The MySom map shown in Figure 6.3 is not exactly the same as the Banbury group map in Figure 6.1, and there are several reasons as to why this may be. The published map was built using a randomly selected 75% of the original dataset, whereas the MySom map was built using all data from only six animals, and so it is feasible that data segregate differently as they are different subsets. Secondly, the published map uses a hexagonal array configuration, such that node weight changes affect six neighbouring nodes, as opposed to the rectangular configuration employed by MySom in which nodes influence four neighbours. Thirdly, the learning rate and neighbourhood function radius employed by the Banbury group decayed linearly, whereas MySOM uses an asymptotic decay function. The effect of using an asymptotic decay function is that the spread of the neighbourhood function decreases quickly during the early stages of training and more slowly as iteration number increases. This method of training may partially account for the two clusters seen for optic nerve and lens in Figure 6.3, as by random chance these subclusters may have been separated early in the training process and each refined separately with subsequent iterations as the neighbourhood function was too small for partner subclusters to interact.

6.4 Summary

This chapter shows that the MySom module can analyse high dimensionality Raman spectral data gained from biological sources, as it is able to cluster tissue sample data by tissue type. The aim of this chapter was not to faithfully replicate the SOM produced by the Banbury group, but to produce comparable results with MySom to confirm that it can be used with real biological data and that output results are meaningful. The next experiments focus on translating the MySom method to human prostate cell line data.

Chapter 7

Prostate Cell Line Experiment Feasibility Study

The work by the Banbury group concerning tissue samples from pig eyes [11] is a demonstration that SOMs can be used to cluster Raman spectral data. The spectra gathered from their tissue samples are readily separable by eye (Figure 2.11), and it is reasonable to expect that spectral signals would differ between different tissue types as the physical and biochemical differences between them is large.

A more difficult distinction is to be made between samples of the same tissue in a healthy and a disease state. In this instance, the tissue types analysed are the same, and so variation must come from disease status, or inherent heterogeneity within the tissue itself, which could be quantitively assessed on the biomolecular level.

This chapter covers the first experiments performed with MySom and Raman spectral data gathered from prostate cell lines (derived from normal prostate and prostate cancer) at the University of York. The thirty spectra generated from these samples are not easily separable by eye (Figure 7.1). The primary point of this chapter is to show that the SOM method works and clusters Raman data even with a small dataset, and as the method works it is appropriate to continue with a larger unseen dataset. The second point is to show that the output of the SOM is very sensitive to its training parameters and how they change. A parameter sweep is needed to optimise the SOM, but choice of parameters relies indirectly on dataset size, and so optimisation is performed with the full blinded dataset in Chapter 8.

7.1 Prostate Cell Lines

The analysed data are derived from two prostate cell lines, PNT2-C2 and LNCaP. PNT2 is a cell line originating from the healthy prostate epithelium of a thirty-three-year-old man [13], which has been immortalised by viral transfection [21]. During development, the PNT2 cell line developed several subclone lineages; PNT2-C2 is the particular strain used for these experiments [75].

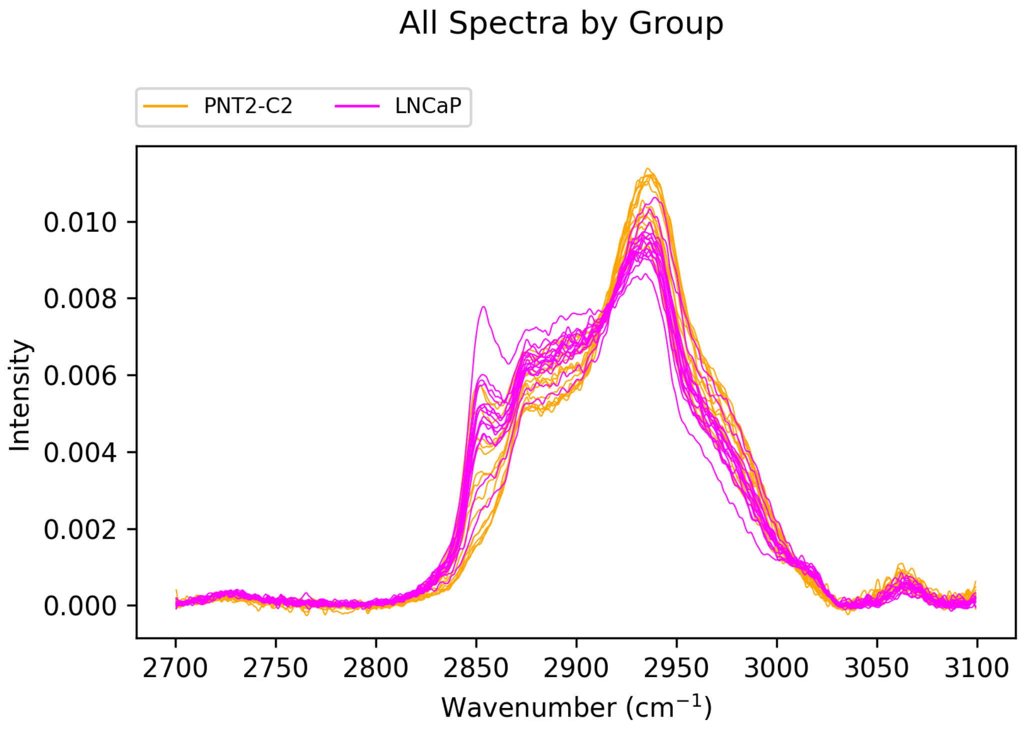


Figure 7.1: Raman spectra generated from thirty cells from two prostate cell lines. Each observation is the Raman spectrum for a single cell nucleus, fifteen from the normal PNT2-C2 prostate cell line and fifteen from the cancer LNCaP prostate cell line. Spectra have been normalised and interpolated, and intensity of observed Raman shift is in arbitrary units. It is difficult to discern unique spectral features between the two groups.

LNCaP is a cell line originating from the left supraclavicular lymph node of a fifty-year-old man, which was the site of a metastatic deposit from prostate carcinoma [44]. It has been used extensively to study prostate cancer *in vitro* [76, 77, 83, 88, 91, 99, 105].

7.2 Raman Spectra from Prostate Cell Lines

The data analysed in this chapter are gathered from Raman spectral analysis of the two prostate cell lines PNT2-C2 and LNCaP. Spectroscopy was performed on the nuclei of fifteen cells from each of these cell lines in Dr Hancock's lab, Department of Physics, University of York. Thirty spectra were generated by recording the intensity of Raman shift observed, and were subsequently normalised and interpolated. These spectra are shown in Figure 7.1, with orange spectra representing PNT2-C2 data and magenta spectra representing LNCaP data.

Within any given system, unique spectral features may be apparent in relative spectral differences, such as the ratio of peak intensities, which can then be compared between systems. It is difficult to discern spectral features which differentiate the PNT2-C2 and LNCaP groups in Figure 7.1 by eye.

	2760.53 (cm^{-1})	2760.91 (cm^{-1})	2761.29 (cm^{-1})	2761.66 (cm^{-1})	$\cdots$
PNT2-C2 high (3)	3.73E-05	4.35E-05	3.76E-05	2.12E-05	$\cdots$
PNT2-C2 high (10)	5.38E-05	5.25E-05	5.07E-05	4.44E-05	$\cdots$
LNCaP_livehigh_6	3.13E-05	3.38E-05	3.65E-05	3.86E-05	$\cdots$
LNCaP_livehigh_14	6.4E-05	5E-05	4.17E-05	4.73E-05	$\cdots$

Table 7.1: Subset of the Known Prostate Cell Line dataset with wavenumber for reference.

7.3 Structure of the Dataset

For SOM analysis, spectral data from the fifteen PNT2-C2 and fifteen LNCaP cell line measurements have been amalgamated into the Known Prostate Cell Line datset of thirty observations.

This labelled dataset is stored as two files, the first holding a one-dimensional array of length 1056, each column containing a value for wavenumber (cm^{-1}). The second file contains a two-dimensional array, the first column of which contains an identification string containing cell line and sample number for that observation, and subsequent columns which contain recorded intensity for the corresponding change in wavenumber in arbitrary units.

The data for the samples are a 30×1057 array, a subset of which is shown in Table 7.1, along with wavenumber for reference. These data are gathered from Dr Hancock's group, Department of Physics, University of York, and they have an instrumental accuracy of ± 3 cm^{-1}.

7.4 SOM Parameter Selection

7.4.1 Changing SOM Parameters

The SOM is built using four parameters: the x and y dimensions of the map, the starting neighbourhood radius, $\sigma(0)$, the starting learning rate, $\alpha(0)$, and the number of iteration steps in the learning process. Changing each value affects how the map develops during training, and optimum values can be tested by varying each parameter in isolation (with all others equal), or for each possible combination of parameter values.

The behaviours of the cell systems captured by the spectral data are inherently complex and display non-linearity, meaning that their response to an input depends on their current state. Optimising each parameter in isolation and grouping the best value for each would likely be ineffective, as each parameter value would be placed in a system with different values for the other parameters than the ones it trained with, and hence a different system state. Therefore, each combination of parameters should be tested to find the optimum parameter set. There are non-exhaustive ways of achieving this, although such methods are not required here as only a small number of combinations are used.

The dataset used here (30 observations) is much smaller than the blinded dataset (285 observations) that is used in Chapter 8, and part of the parameter optimisation relies on size of the dataset (see Section 8.4). Therefore, a complete optimisation analysis is not performed here, but the effects of altering SOM training parameters are highlighted.

Kohonen states that map size cannot be guessed beforehand, that a starting point should be chosen and the dimensions varied by trial-and-error after seeing the results of this plot [57]. Vesanto suggests starting with an initial map size of $5\sqrt{n}$ where n is the number of observations in the dataset [100]. There are 30 observations in this dataset, and $5\sqrt{30} = 27.39$, so map dimensions of $7 \times 4 = 28$ and $6 \times 5 = 30$ are used.

σ should be high enough so that clustering is fast in the first few iterations as the BMU affects at least one node in each direction—if σ is too small, nodes cannot impact surrounding nodes and the SOM takes on a mosaic pattern [56, p. 112]. $\sigma(0) = 1$ is chosen as a value of sigma so that at least one node in each direction surrounding the BMU is affected by the neighbourhood region in the first instance, and $\sigma(0) = 0.5$ is tested to see the effect of a small σ on the output SOM.

α should be set between 0 and 1 [56, p. 111], so the extreme values for $\alpha(0)$ of 0.1 and 1 are tested.

The number of training iterations should be at least 500 × the size of the map array in order to achieve statistical convergence [56]. For the largest map array under test this value is $500 \times 6 \times 5 = 15000$. All SOMs are built with 15000 iteration steps.

MySom takes an argument for random seed, so that the random values used for initiating training of the SOM can be set and therefore reproduced. Each of the parameter groups is analysed ten times with random seeds 1–10 to ensure that the output of training is consistent and therefore due to the parameters used, and not due to the random allocation of starting weights of the SOM nodes.

The parameters tested are grouped into families (1–8) of different parameter combinations, each with ten members (representing random seeds 1–10). Four errors are calculated for each SOM, based on quantisation and topographic errors (see Section 2.3.4). Minimum total error is desirable, but SOM learning dynamics cannot be expressed by a single energy function [28], meaning that minimisation of one error may cause inadvertent increase in the other. For this reason, the error difference is calculated to see if quantisation and topographic error are proportionate.

- Quantisation error: QE
- Topographic error: TE
- Total error: (QE + TE)
- Error difference: $|\text{QE} - \text{TE}|$

SOM Family	Map Dimensions	σ	α	QE	TE	(QE + TE)	|QE − TE|
1	7×4	0.5	0.1	0.00756	0.16667	0.17440	0.15894
2	7×4	0.5	1.0	0.00354	0.13333	0.13707	0.12990
3	**7×4**	**1.0**	**0.1**	**0.01572**	**0.00000**	**0.01618**	**0.01618**
4	7×4	1.0	1.0	0.01577	0.03333	0.05014	0.01841
5	6×5	0.5	0.1	0.00749	0.16667	0.17550	0.15757
6	6×5	0.5	1.0	0.00288	0.16667	0.16955	0.16379
7	6×5	1.0	0.1	0.01525	0.03333	0.04877	0.01790
8	6×5	1.0	1.0	0.01501	0.03333	0.04803	0.01909

Table 7.2: Parameters used to investigate SOMs built with the Known Prostate Cell Line dataset. The values for family 3 are shown in bold as these have the lowest total error. QE = quantisation error. TE = topographic error. All error values are median values of the ten family members.

The median error of each family is used to compare parameter groups. These data are summarised in Table 7.2.

7.5 Results and Discussion

Of the eight SOM families, family 6 has the lowest QE and family 3 has the lowest TE. It is not a surprise that the smallest errors do not belong to the same maps, as each is measuring a different aspect of the SOM, and no single metric can fully describe the characteristics of a SOM [28].

Family 3 is selected as the best representation of the data due to its low QE (0.01572) and TE (0.00000), and a low total error (0.01627). The low error difference shows that the errors are close together, so quantisation and topographic errors contribute to the total error similarly, and these SOMs are a good fit to both the distribution and topology of the dataset. The SOM in Figure 7.2 was built with a rectagular array configuration and the parameters from family 3:

- Map dimensions: 7×4
- $\sigma(0)$: 1.0
- $\alpha(0)$: 0.1
- Random seed: 1

Each family member showed a similar pattern and distribution of the input data, so the SOM produced with random seed 1 is selected arbitrarily for display.

This SOM in Figure 7.2 was trained unsupervised, and the input data were labeled once the SOM analysis was complete. The SOM shows good separation of the data into two clusters, with a stripe of low density nodes across the centre of the map from upper left to lower right dividing the PNT2-C2 and LNCaP observations. The PNT2-C2 cluster is contiguous, and the LNCaP cluster is separated from the PNT2-C2 cluster, suggesting that the map is not overfitted, which could result in breaking apart of the clusters, as seen with small σ values.

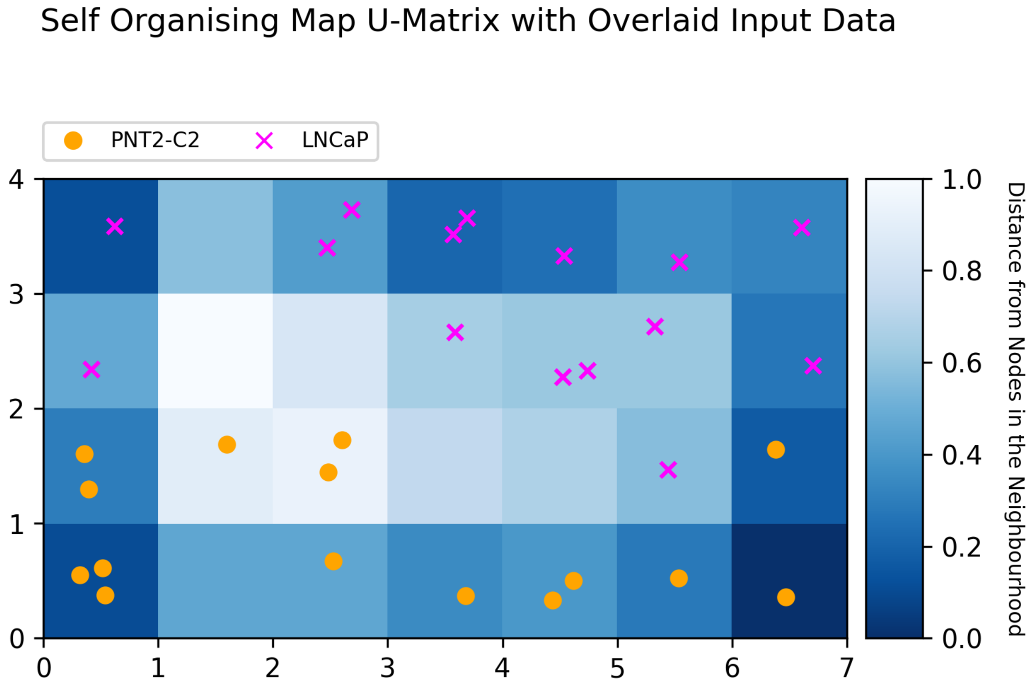


Figure 7.2: Som u-matrix trained unsupervised on the Known Prostate Cell Line dataset. There is a stripe of low density nodes across the centre of the map from (1, 3) to (4, 1), which separates the normal PNT2-C2 (lower) and malignant LNCaP (upper) clusters. The PNT2-C2 cluster is contiguous and the LNCaP cluster is segregated fully from the PNT2-C2 cluster.

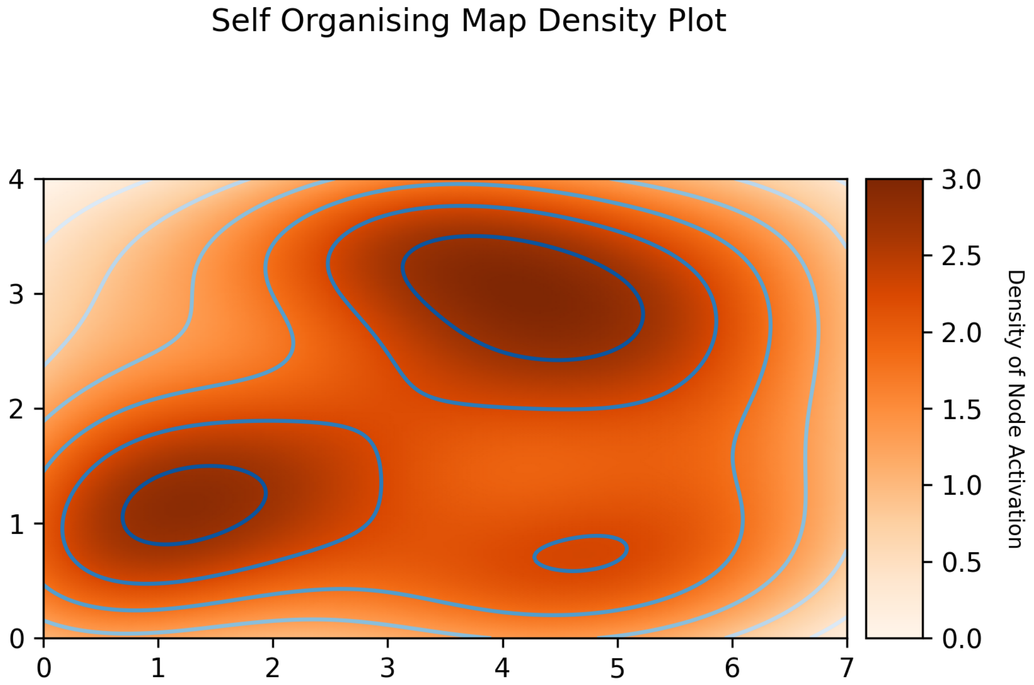


Figure 7.3: Density plot of the SOM in Figure 7.2 showing two dense regions around (1, 1) and (4, 3) representing PNT2-C2 and LNCaP, respectively.

The density plot in Figure 7.3 shows two dense foci of data at the lower left and upper centre of the plot. The lower left focus corresponds to nine PNT2-C2 observations and the upper centre focus corresponds to eight LNCaP observations situated in these positions in Figure 7.2.

Changing the parameters used to build SOMs alters how the training process occurs and ultimately the architecture of the resultant maps. Some of the errors seen when changing the parameters for this dataset are quite large, such as the median TE for family 5 of 0.15000 (Table 7.2). It is likely that such large errors are present due to the small number of observations in the dataset. TE measures the proportion of observations where the best matching node and the second best matching node are not adjacent. With a sample size of only thirty, each individual data point where this posit does not hold contributes a large proportionate error; with a larger dataset, each individual error would contribute less to the total error. The sample size thirty is approximate to $5\sqrt{n}$ in this case, meaning there is an average of 1 observation per node on the map—it is conceivable that the SOM may not cluster data if the neighbourhood radius does not reach far enough to adjacent neighbours, leading to distribution of data across the map rather than clustering around best matching nodes. Kohonen comments that the SOM is primarily a visualisation technique and is not the best tool for analysing small datasets [57].

7.6 Summary

MySom is able to cluster data from this dataset as expected into normal (PNT2-C2) and cancer (LNCaP) groups, given optimum parameters. This result is promising, as the input spectra are not clearly separable by eye, meaning that the SOM is finding patterns which segregate data by disease status rather than another source of sample heterogeneity. The next stage of experimentation involves an optimum parameter sweep and analysis of a larger, blinded prostate cell line dataset to test the SOM's ability to cluster data and discussion of how best to analyse the output.

Chapter 8

Blinded Prostate Cell Line Experiment

This chapter covers the experimental use of the MySom module to cluster Raman spectra gathered from prostate cell lines (derived from normal prostate and prostate cancer) in Dr Hancock's lab, Department of Physics, University of York. These data are blinded, and test the ability of MySom to cluster data and the methods by which this clustering can be interpreted.

Parameter choices impact the architecture of the SOM significantly, and values are selected for a parameter sweep. An index case of parameters is run with MySOM to check performance, and the resultant SOM (Figure 8.1) shows one observation in the dataset to be segregated far from the others on the map, indicating a very different pattern to the rest of the data in the dataset. Analysis of this spectrum shows it to be an outlying signal (Figure 8.2) with a grossly different shape to the rest of the spectra in the dataset and it is removed from subsequent analyses.

The Cleaned Prostate Cell Line dataset is used to perform a parameter sweep to determine the optimum parameter sets which give a good mapping to the input data. The resultant SOMs are assessed based on quantisation and topographic errors, and visual interpretation of cluster borders and data density foci. The selected optimum SOM parameters show delineation of the data into three clusters (Figures 8.5 and 8.6), which is unexpected given the two classes of input data.

Data are then unblinded and the three clusters shown to be one containing data from the normal (PNT2-C2) class and two containing data from the cancer (LNCaP) class. Investigation of the Raman spectra for the sample as a whole and for each cluster shows some potentially distinguishing features of the input spectra, which correspond to a difference in the amounts of saturated and unsaturated lipids between the two cancer clusters.

2760.53 (cm^{-1})	2760.91 (cm^{-1})	2761.29 (cm^{-1})	2761.66 (cm^{-1})	···
8.26E-05	6.84E-05	5.21E-05	4.16E-05	···
4.16E-05	4.66E-05	5.47E-05	6.34E-05	···
8.63E-05	8.99E-05	9.21E-05	9.3E-05	···
6.22E-05	6.72E-05	7.72E-05	8.72E-05	···

Table 8.1: Subset of the Blinded Prostate Cell Line dataset with wavenumber for reference.

8.1 Structure of the Dataset

The dataset for SOM analysis contains spectral data from an unknown (blinded) number of PNT2-C2 (normal prostate) and LNCaP (prostate cancer) cell lines, total 285 observations. The unlabelled dataset is stored as two files, the first containing a one-dimensional array of length 1056, each column containing a value for the wavenumber (cm^{-1}). The second file contains a two-dimensional array containing measured arbitrary intensity values that correspond to each wavenumber.

The data for the samples are a 285 × 1056 array, a subset of which is shown in Table 8.1, along with the wavenumbers for reference. These data are gathered from Dr Hancock's group, Department of Physics, University of York, and they have an instrumental accuracy of ± 3 cm^{-1}.

8.2 Parameter Selection Rationale

8.2.1 Parameter Selection

The parameters which affect training and output of the SOM are the configuration of the map array, the neighbourhood function, the learning rate, and the number of iterations of training, as discussed in Section 7.4.1. Training is stochastic, and a random seed can be used to set the starting point of this process. Further discussion of these parameters and their selection can be found in Section 2.3.3.

Map Array Configuration

The neural network array used for SOM training must be of appropriate topology, lattice configuration, and size, and should have optimum dimensions. SOM arrays are most often displayed as a two-dimensional regular lattice of nodes with either rectangular or hexagonal configuration, although irregular lattices or arrays with higher dimensions can be used [57]. A two-dimensional lattice with rectangular topology is used with MySom as it is computationally simpler and easier for non-experts to interpret.

The appropriate size of a SOM array cannot be calculated *a priori*, but can be investigated by trial and error to see how data cluster [56]. Vesanto investigated

Map Dimensions	Nodes in Map	Dimension Ratio	Comments
19 × 5	95	3.8	Optimum number of map nodes and map dimension ratio
14 × 6	84	2.33	Optimum number of map nodes
17 × 5	85	3.4	Optimum number of map nodes
15 × 4	60	3.75	Optimum map dimension ratio
23 × 6	138	3.83	Optimum map dimension ratio

Table 8.2: Table showing the selected map dimensions for a parameter sweep to optimise SOM parameters.

the computational complexity of clustering with SOMs, and combining partitive and agglomerative clustering [100]. In his discussion he uses the example of a SOM lattice with $5\sqrt{n}$ nodes (where n is the number of observations in the dataset), which reduces overall computational complexity. The figure of $5\sqrt{n}$ nodes has become widely used as a starting point for deciding on optimum SOM size [79, 95].

Kohonen comments that the first benefit of the SOM is in visualisation of the input dataset, whereby finer resolutions can be used to discern smaller subgroups of data, or coarser resolution where fewer groups are expected [57], so optimum size is determined by visual interpretation of the output SOM. The second use of the SOM is as a histogram of input data, whereby the nodes are coloured according to how many input data points map to them, requiring approximately fifty data points on average per node for statistical accuracy [57]. MySom uses overlaid scatter data of the input datset to show how many data points map to each node, so Kohonen's first visualisation interpretation of the SOM is used.

For a rectangular regular lattice, Kohonen suggests that the x and y dimensions should be in the ratio of the two highest eigenvalues of the input data autocorrelation matrix [57]. This configuration makes convergence in learning faster.

Reviewing the above points, the map dimensions summarised in Table 8.2 are used for a parameter sweep. The dataset contains 285 observations, for which $5\sqrt{n}$ is approximately 84.4. The two highest eigenvalues of the autocorrelation matrix of the input dataset are 512.692 and 133.623 giving on optimum ratio of 3.837. A map dimension of 19 × 5 is a good approximation of this number of nodes and dimension ratio; 14 × 6 and 17 × 5 give a good approximation of the size but not the dimension ratio; and 15 × 4 and 23 × 6 give a good approximation of the dimension ratio but not the size.

Neighbourhood Function

The neighbourhood function defines the region of nodes surrounding the BMU whose weights are affected in the updating step of the algorithm, and has a shape and starting radius, $\sigma(0)$. The function decreases monotonically with each iteration step.

Several neighbourhood function forms can be used, with the key features that the spread of the function is symmetrical about the BMU, and $\sigma(t) \to 0$ as $t \to \infty$. As long as $\sigma(0)$ is of an appropriate size, the choice of neighbourhood function is inconsequential [56, 79]. A Gaussian neighbourhood function is used with MySom.

The initial size of the neighbourhood function needs to be large enough that it reaches nodes surrounding the BMU, otherwise the map becomes discontinuously ordered in local pockets without map-wide organisation [55]. Kohonen suggests that $\sigma(0)$ can appropriately be more than half the smallest map dimension [55, 56].

Taking the above points into consideration, tested values for $\sigma(0)$ are 0.5, 1.0, 2.0, 3.0, and 4.0. $\sigma = 1.0$ is chosen as its neighbourhood should impact at least one surrounding node in each direction; $\sigma = 0.5$ is chosen to test if this value less than one gives a globally disordered map with locally organised clusters as Kohonen suggests [55]. The value $\sigma = 4.0$ is chosen as this is larger than half the map radius for all the map dimensions given in Table 8.2, and $\sigma = 2.0$ and $\sigma = 3.0.$ are chosen as reasonable values between 1.0 and 4.0.

Learning Rate

The learning rate, $\alpha(t)$, defines how much the weights of nodes in the neighbourhood are affected with iteration number, t. The learning rate should be between 0 and 1, and initially be close to 1 [55]. It should decrease monotonically throughout each iteration, although whether the decay is exponential, linear, or inversely proportional to t is inconsequential [55, 56, 79]. MySom uses an asymptotic decay shown in Equation 8.1.

$$\alpha(t) = \frac{\alpha(t-1)}{1 + \frac{t}{0.5 \cdot t_{max}}} \tag{8.1}$$

The values for initial learning rate, $\alpha(0)$, selected for the parameter sweep are 0.5, 0.75, 0.9, 0.95, and 0.99. These values are selected following Kohonen's advice of using values close to 1, particularly with random initiation of SOM weights [56, p. 112] as used in MySOM, and to get an idea of how decreasing the value of $\alpha(0)$ affects map organisation.

Iteration Number

The SOM learning method is stochastic in nature, and so requires many iteration steps to reach good statistical accuracy [56]. There is no way to guarantee that "enough" iteration steps have been used, but Kohonen suggests a rule of thumb of at least 500 times the number of nodes in the network [56, p. 112]. The dimensionality of the input data does not impact the number of iteration steps required [56].

The largest map network outlined in Table 8.2 contains 138 nodes, requiring a maximum number of iteration steps, t_{max}, of at least $138 \times 500 = 6.9{\times}10^4$.

All subsequent training with MySOM uses $10{\times}10^5$ iteration steps, as this number ensures the minimum value has been surpassed, and corresponds to values Kohonen used in his original SOM simulation experiments [55]. Higher values would increase the time taken to run experiments and would not be expected to make results more accurate.

Random Seed

Initial SOM weights can be random, or selected from a hyperplane spanned by the principal components of the input data associated with the two highest eigenvalues [56]. The former method gives evidence that the SOM can be used to cluster data from any arbitrary starting point, and the latter method reduces the number of iterations needed to reach convergence.

A random starting weight is used with MySom, and the random seed used to initialise these weights can be chosen. Each parameter set is run ten times with random seeds 1–10, for three key reasons. Firstly, this allows comparison between SOMs built with the same parameters to ensure their architecture results from the parameters used in training and not from the random starting point. Secondly, this allows comparison of SOMs built from the same starting points and trained with different parameters, ensuring that differences in their architecture result from different training parameters and not from the random starting point. Thirdly, this allows reproducibility of results.

8.2.2 Index Case SOM

As a first test of the SOM analysis to investigate if these parameters produce an interpretable map, the SOM in Figure 8.1 is built using the following parameters:

- Map dimensions: 19 × 5
- $\sigma(0)$: 2.0
- $\alpha(0)$: 0.9
- Random seed: 1

These parameters are selected because the map dimensions optimise computational complexity and eigenvalue ratio, and σ and α are in the middle of the ranges to be used for the parameter sweep.

The SOM in Figure 8.1 obtained from the Blinded Prostate Cell Line dataset shows a good spread of data across most of the map, and stripes of low nodal density around (2, 11) and (2, 16). There is a single observation which maps to (2, 4), and this node is surrounded by nodes of very low density, signifying that it is situated far from the other nodes in the map. For a single observation to be separated by a large region of map space where the rest of the map shows an even distribution of data, it must have a pattern very different from the other input data vectors. Such

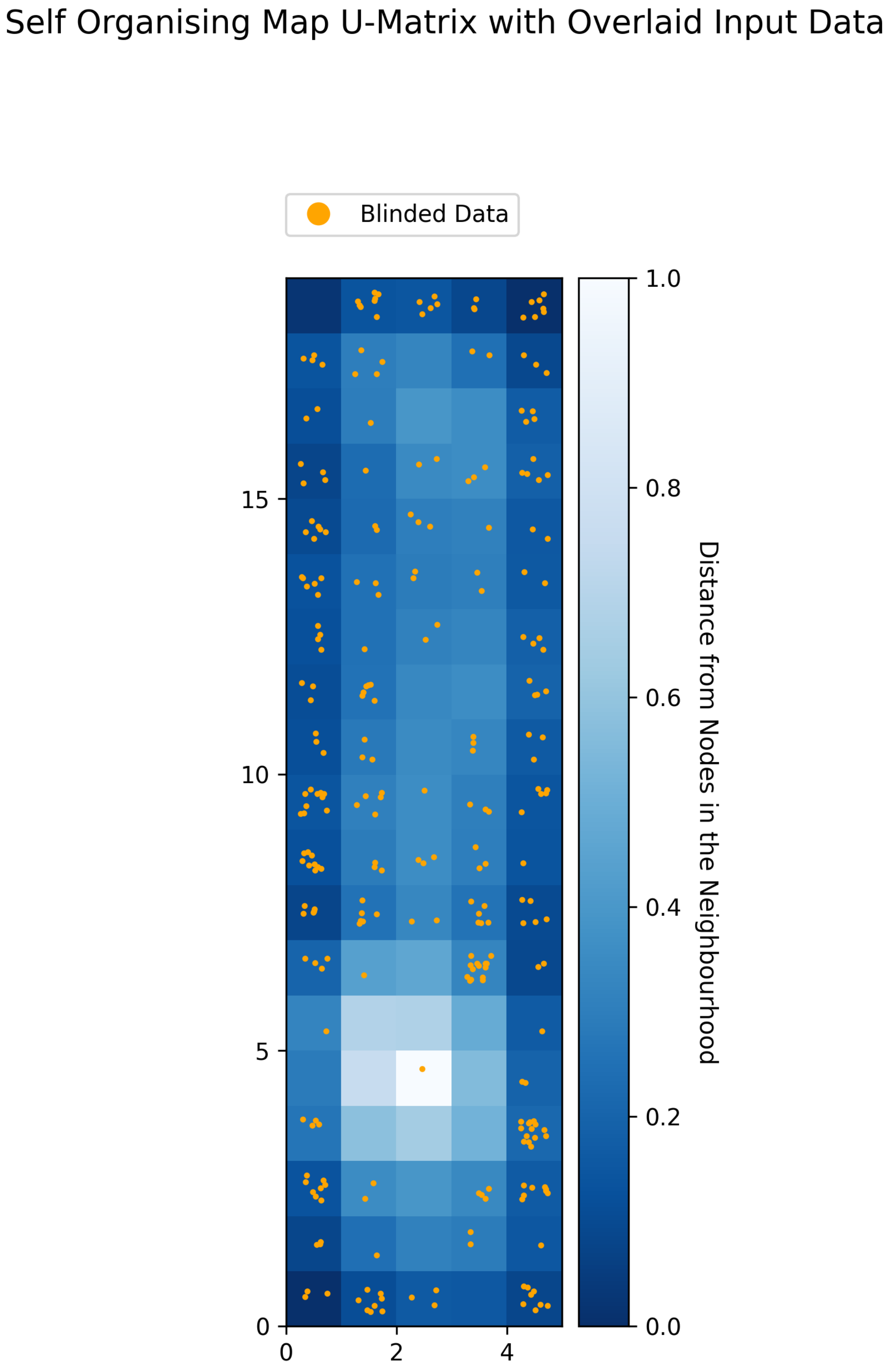


Figure 8.1: SOM u-matrix trained on the entire Blinded Prostate Cell Line dataset. There is a good distribution of data points across most of the map with stripes of low density around (2, 11) and (2, 16). There is a single datapoint mapping to (2, 4) surrounded by a region of low nodal density. This observation is likely an outlier and is investigated further.

a large difference could indicate this observation is a true outlier, or an abnormal signal, and it is investigated further.

8.3 Investigation of the Outlier

Three steps are taken to investigate the outlying observation at (2, 4) in Figure 8.1. Firstly, individual spectra for the entire dataset are plotted together for visual inspection. Secondly, the effect of removing the outlying observation on the average spectrum is investigated. Thirdly, the effect of removing the outlying spectrum on SOM training and output is investigated.

8.3.1 Individual Sample Spectra

Figure 8.2 shows plots for all the spectra within the dataset. The broad shape of the spectra is a rise from baseline to a peak at 2850 cm^{-1}, a small dip, then a shallow increase followed by a steep increase to a second peak at 2940 cm^{-1}, before falling to baseline around 3030 cm^{-1} with a small peak around 3060 cm^{-1}. Some of the spectra also show a slight shoulder to the main peak at 3010 cm^{-1}.

The outlier spectrum in Figure 8.2 has a different shape, with its first peak at 2890 cm^{-1} and a second bifid peak with zeniths at 2950 cm^{-1} and 2970 cm^{-1}. This shape is very different from the others, and may represent sampling error or data corruption.

The mean $\pm$ 3 $\times$ the standard deviation of the sample is also plotted on Figure 8.2, and these bounds encase the majority of spectra entirely or almost entirely. The outlying spectrum does not fit within these bounds as its main peaks fall outside of this region.

8.3.2 Average Sample Spectra

Given the shape of the outlying spectrum, how much it is different from the pattern of the other spectra, and that a large portion of it lies outside of the region bound by three standard deviations from the mean, it is very unlikely to be a true outlier and looks like a sampling error or data corruption.

Figure 8.3 shows the average spectrum for the entire dataset with and without the outlying spectrum. The shape of these spectra is the same, so removal of the outlying observation has little effect on the average spectrum.

The standard error of the mean is a measurement of the precision with which the sample mean reflects the population mean—the smaller this error, the smaller the uncertainty in the calculated value of the mean [4]. The standard error of the mean for the Blinded Prostate Cell Line dataset ranges from 2.4270 $\times 10^{-6}$ to 7.4999 $\times 10^{-5}$, indicating that the samples have converged and there is confidence that the

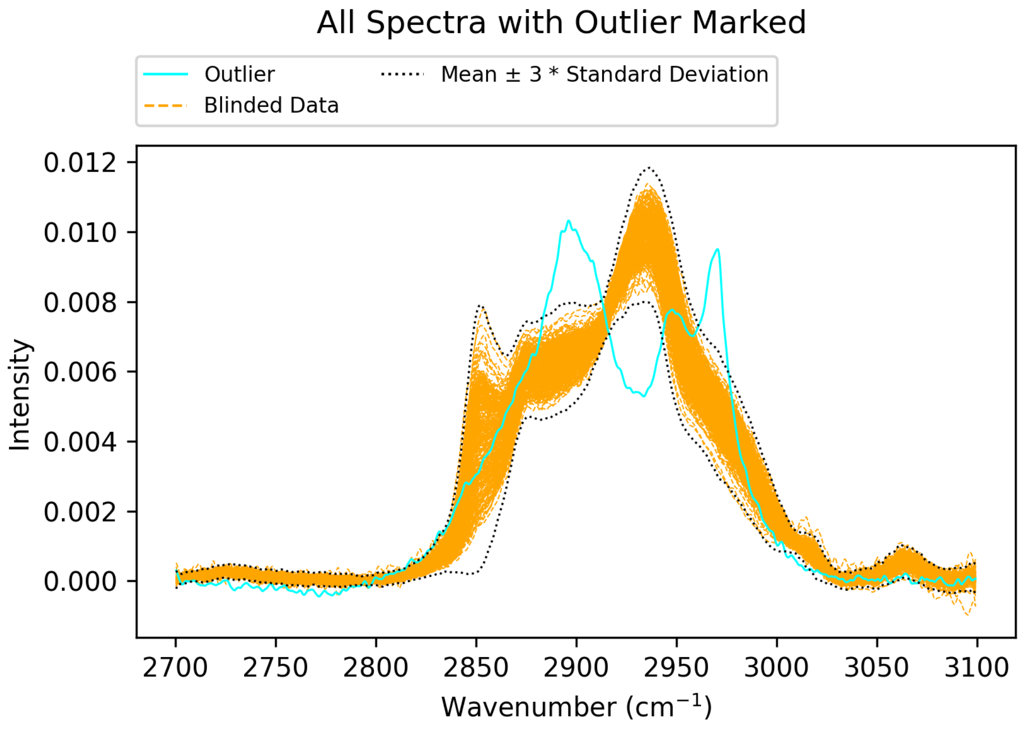


Figure 8.2: Plot of all spectra from the blinded prostate cell line dataset. The cyan spectrum drawn with a solid line is that of the outlying observation mapping to node (2, 4) in Figure 8.1. The orange spectra drawn with broken lines are the other spectra in the dataset.

The majority of the spectra follow the same shape, with a sharp increase from baseline to a peak at 2850 cm^{-1}, a small dip to 2860 cm^{-1}, slowly rising to 2920 cm^{-1}, rapidly rising to 2940 cm^{-1}, steeply descending to baseline at 3030 cm^{-1}, with a small peak at 3060-3070 cm^{-1}. Some spectra display a small shoulder at 3010 cm^{-1}.

The outlier looks spectrally very different to the rest of the spectra, with a steep increase from baseline to a peak at 2890 cm^{-1}, steep decline to 2930 cm^{-1}, then steep increase to 2950 cm^{-1}, small dip to 2960 cm^{-1}, sharp increase to 2970 cm^{-1}, then rapid descent to 3000 cm^{-1} and a slow descent to baseline at 3030 cm^{-1}.

The dotted black line marks the mean $\pm 3 \times$ the standard deviation of the sample. The majority of the spectra fit entirely or almost entirely within these bounds, except for the outlier spectrum whose peaks lie outside of this region.

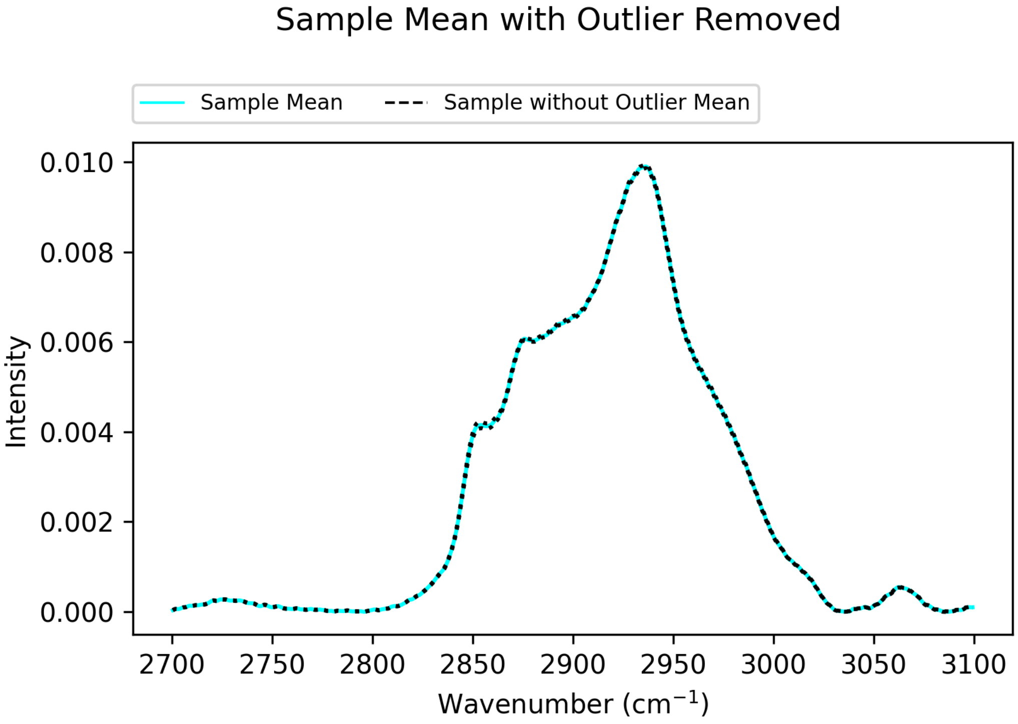


Figure 8.3: Average spectra for the entire Blinded Prostate Cell Line dataset (cyan, complete line) and the dataset with the outlier removed (black, dashed line). The shape of these average spectra is identical, so removal of the outlier has minimal effect on the average spectrum. The standard error envelope has been omitted from this image as it is too narrow to be discernible.

sample mean approximates the population mean. The statistical envelope of the standard error has been omitted from Figure 8.3 as it lies so close to the mean spectra that is is not discernible.

8.3.3 Removal of Outlier Spectrum

The outlying observation exerts little effect over the shape of the average spectrum within the dataset, as shown in Figure 8.3. However, this outlier does have a marked effect on SOM training, as evidenced by the 12 empty nodes surrounding it in the SOM shown in Figure 8.1. No other region of the map has more than four contiguous empty nodes.

Removal of this outlier from the Blinded Prostate Cell Line dataset yields the Cleaned Prostate Cell Line dataset. The SOM in Figure 8.4 was trained on the Cleaned Prostate Cell Line dataset with the same parameters as the SOM in Figure 8.1. This new SOM has a more even spread of data across the map space, with three regions of low nodal density at the lower right, centre, and upper centre. There are no single observations taking up large regions of map space, suggesting that the outlying spectrum was exerting a large effect on SOM training.

As the outlier appears to be of a very different spectral shape to the other spectra, it exerts little effect on the average spectrum, and it grossly impacts SOM training, it is removed for parameter optimisation. All subsequent SOM analyses are performed with the Cleaned Prostate Cell Line dataset.

8.4 Parameter Optimisation

Combination of the parameters selected in Section 8.2.1 yields 125 parameter sets (5 map arrays × 5 $\sigma(0)$ values × 5 $\alpha(0)$ values = 125 sets). Each of these parameter sets is used to train a SOM ten times with random seeds 1–10, respectively, giving a total of 1250 training runs. Each group of ten SOMs trained with the same parameters is denoted as a family, with each family numbered from 1–125.

In order to investigate which SOMs are the best representation of the data, a similar rationale to that laid out in Section 7.4 is used to select the SOMs with the lowest combined quantisation and topographic errors. The median QE, TE, total error, and error difference for each SOM family is calculated (Table 8.3). These errors are compared with the SOM outputs to decide on optimum parameters.

The optimum SOM family is chosen as being the one with the lowest total error, where borders between clusters on the u-matrix are well defined, and where data foci on the density plot are discrete. Of the SOM families summarised in Table 8.3, Family 117 is selected as optimum as it is the family with the lowest total error value for which the SOM (Figure 8.5) shows well demarcated regions of low nodal density and the density plot (Figure 8.6) shows discrete foci of data. For the families with

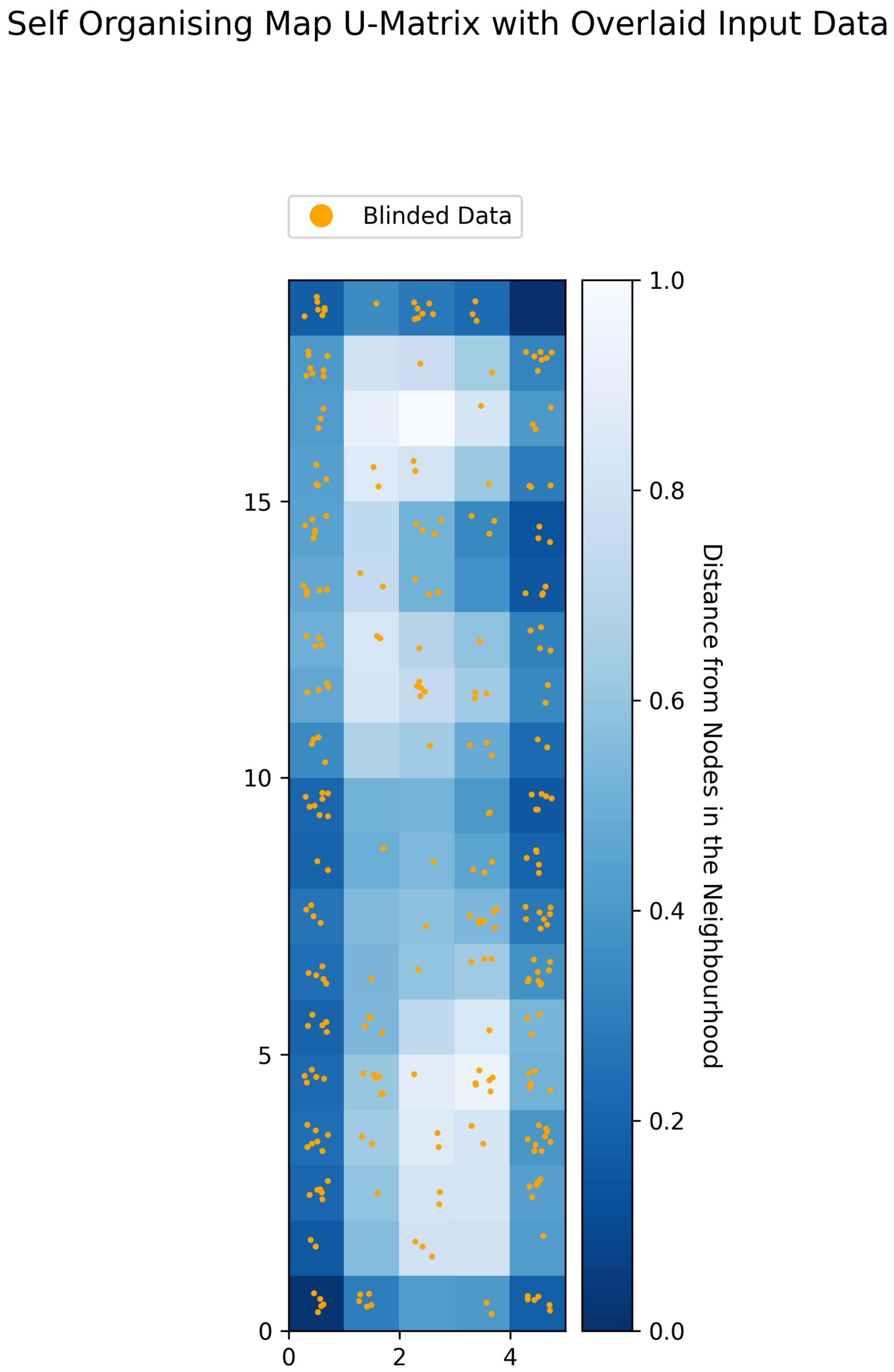


Figure 8.4: SOM built with the same training parameters as that in Figure 8.1 with the outlier removed from the dataset. There is a good spread of data across the entire map space, with stripes of low density around (1, 10), (1, 16), and (3, 1).

Family	Array	σ	α	QE	TE	(QE + TE)	|QE − TE|
81	15 × 4	1.0	0.5	0.019604	0.0	0.019604	0.019604
82	15 × 4	1.0	0.75	0.019908	0.0	0.019908	0.019908
83	15 × 4	1.0	0.9	0.020134	0.0	0.020134	0.020134
111	23 × 6	2.0	0.5	0.020138	0.0	0.020138	0.020138
61	17 × 5	2.0	0.5	0.021557	0.0	0.021557	0.021557
62	17 × 5	2.0	0.75	0.021627	0.0	0.021627	0.021627
64	17 × 5	2.0	0.95	0.021776	0.0	0.021776	0.021776
86	15 × 4	2.0	0.5	0.022923	0.0	0.022923	0.022923
106	23 × 6	1.0	0.5	0.016016	0.007042	0.023058	0.008973
87	15 × 4	2.0	0.75	0.023493	0.0	0.023493	0.023493
88	15 × 4	2.0	0.9	0.023742	0.0	0.023742	0.023742
89	15 × 4	2.0	0.95	0.023809	0.0	0.023809	0.023809
90	15 × 4	2.0	0.99	0.023850	0.0	0.023850	0.023850
12	19 × 5	2.0	0.75	0.021298	0.003521	0.024819	0.017777
63	17 × 5	2.0	0.9	0.021612	0.003521	0.025133	0.018091
117	**23 × 6**	**3.0**	**0.75**	**0.022025**	**0.003521**	**0.025546**	**0.018504**
91	15 × 4	3.0	0.5	0.026191	0.0	0.026191	0.026191
93	15 × 4	3.0	0.9	0.026643	0.0	0.026643	0.026643
48	14 × 6	4.0	0.9	0.026652	0.0	0.026652	0.026652
66	17 × 5	3.0	0.5	0.023308	0.003521	0.026829	0.019787

Table 8.3: Table of the twenty SOM families in the parameter sweep with the lowest total error. The values for family 117 are shown in bold as these are the optimum SOM training parameters based on errors and visual interpretation of the output map and density plots. QE = quantisation error. TE = topographic error. All errors given are median values for the ten family members.

lower total error values, at least one of these visual requirements does not hold. See Appendix B for these plots.

Family 117 is the first family in the list with a $\sigma(0)$ value of at least half the map dimension and an $\alpha(0)$ value close to 1, supporting Kohonen's suggestion of optimum values [55, 56]. The lower values for $\sigma(0)$ and $\alpha(0)$ in the few maps with lower errors may explain why they produce suboptimal maps, as values which are too small create maps with pockets of local organisation between which the direction of ordering changes [55]. This analaysis further supports the conclusion by Erwin that no single metric can fully describe a SOM [28], and emphasises the importance of visually inspecting the SOM rather than relying on error metrics alone.

8.5 Unblinding the Data

8.5.1 SOM Trained with Optimum Parameters

The SOM in Figure 8.5 is built with the optimum parameters selected from the parameter sweep:

- Map dimensions: 23 × 6
- $\sigma(0)$: 3.0
- $\alpha(0)$: 0.75
- Random seed: 1

The SOM (Figure 8.5) shows a large stripe of low nodal density down the centre of the map and a smaller stripe of low nodal density at the lower left around (1, 7). The density plot (Figure 8.6) shows three data density foci at the upper left, lower left, and right of the map. These results imply the presence of three clusters in the dataset.

8.5.2 Hypotheses

The SOM in Figure 8.5 and the density plot in Figure 8.6 contain data from two diseases classes and display separation into three clusters. There are three distinct possibilities for why three clusters are present.

- Single subgroup hypothesis
  - Cluster A contains data from class 1
  - Cluster B contains data from class 2
  - Cluster C contains data from class 1
- Mixed subgroups hypothesis
  - Cluster A contains data from class 1
  - Cluster B contains data from class 2 or classes 1 and 2

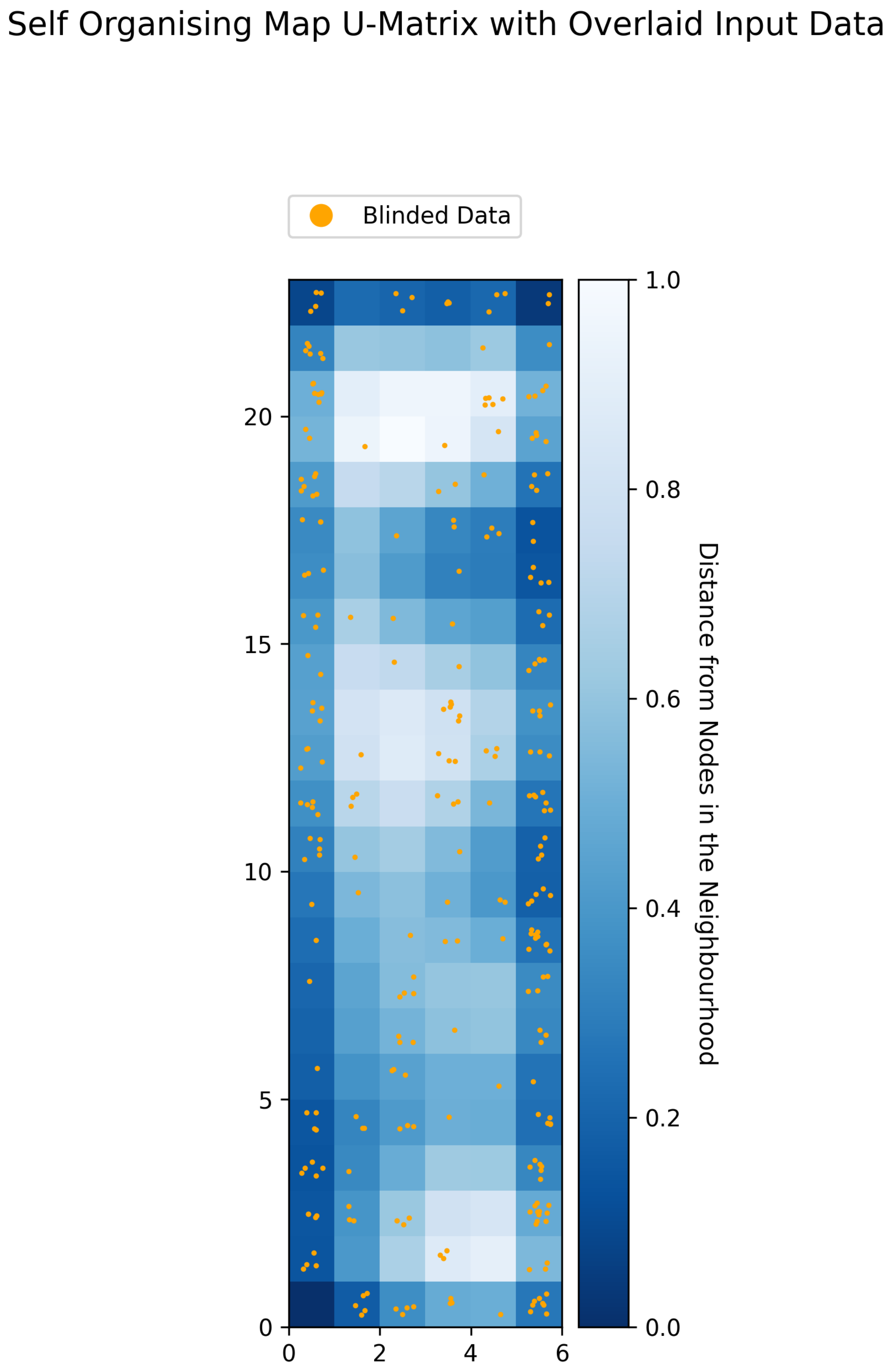


Figure 8.5: SOM u-matrix trained unsupervised on the Cleaned Prostate Cell Line dataset with the optimum parameters selected from the parameter sweep. There is a good spread of data across the map space, with stripes of very low density around (2, 11), (2, 20), and (4, 3), and areas of low nodal density around (1, 7), (1, 17), and (4, 7).

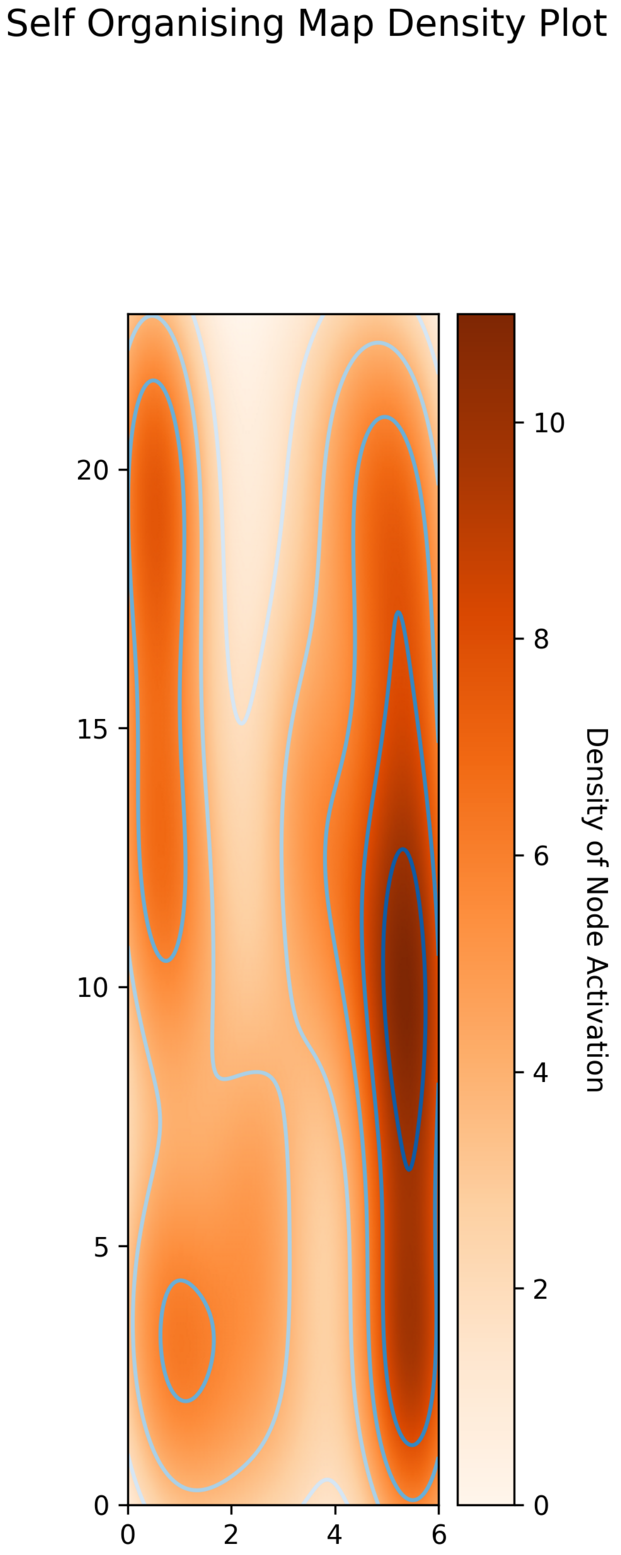


Figure 8.6: Density plot of the SOM shown in Figute 8.5. There appear to be three foci of data, one dense focus centred around (5, 9), and two less dense foci centred around (1, 3) and (0, 16). The borders between these foci are regions of low data density, which correpsond to the low nodal density stripes observed in the SOM in Figure 8.5.

    - Cluster C contains data from classes 1 and 2

- External Heterogeneity hypothesis
    - Cluster A contains data from classes 1 and 2
    - Cluster B contains data from classes 1 and 2
    - Cluster C contains data from classes 1 and 2

Under the single subgroup hypothesis, two clusters relate to the disease status classes encoded within the dataset, and the third cluster is a subgroup of one of these. In this scenario either the cancer class divides into two clusters, such as different grade or prognosis, or the normal class divides into two clusters, such as non-malignant overgrowth and true normal.

Under the mixed subgroup hypothesis, two clusters relate to the disease status classes encoded within the dataset, and the third cluster is a mixture of these two, or one cluster relates to disease status class and two are a mixture of both classes. In this scenario the unique clusters relate to one aspect which is different in each class, such as components of oncogenic protein expression levels in different cell types, and the mixed clusters relate to another aspect which is not different between classes, such as mitochondria producing cellular energy.

Under the external heterogeneity hypothesis, all three clusters contain a mixture of the disease status classes encoded within the dataset. In this scenario, an aspect unrelated to disease status is the source of heterogeneity, such as the location within the cell which is being sampled, the stage at which the cell resides within the cell cycle, or different culture techniques.

To fully ensure that the SOM is clustering data into groups of PNT2-C2 and LNCaP as expected, and to understand why three groups are being shown, the data from the optimised SOM in Section 8.4 are unblinded to produce the SOM in Figure 8.7. This SOM is built with an unsupervised training method and the labels PNT2-C2 (normal prostate) and LNCaP (prostate cancer) are added only after data unblinding.

8.5.3 Review of Unblinded Data

The SOM in Figure 8.7 shows the data for the two classes, PNT2-C2 (normal prostate) and LNCaP (prostate cancer) are divided by a line of low density nodes spanning across the map from upper centre to lower centre. There is a stripe of low nodal density centred around (1, 7), which may represent a cluster border—in the density plot in Figure 8.6, there are two lower density data foci around (1, 3) and (0, 16), separated by this low density region. These two lower density data foci contain data from the LNCaP class, and there is a high density focus around (5, 9) which corresponds to the PNT2-C2 cluster.

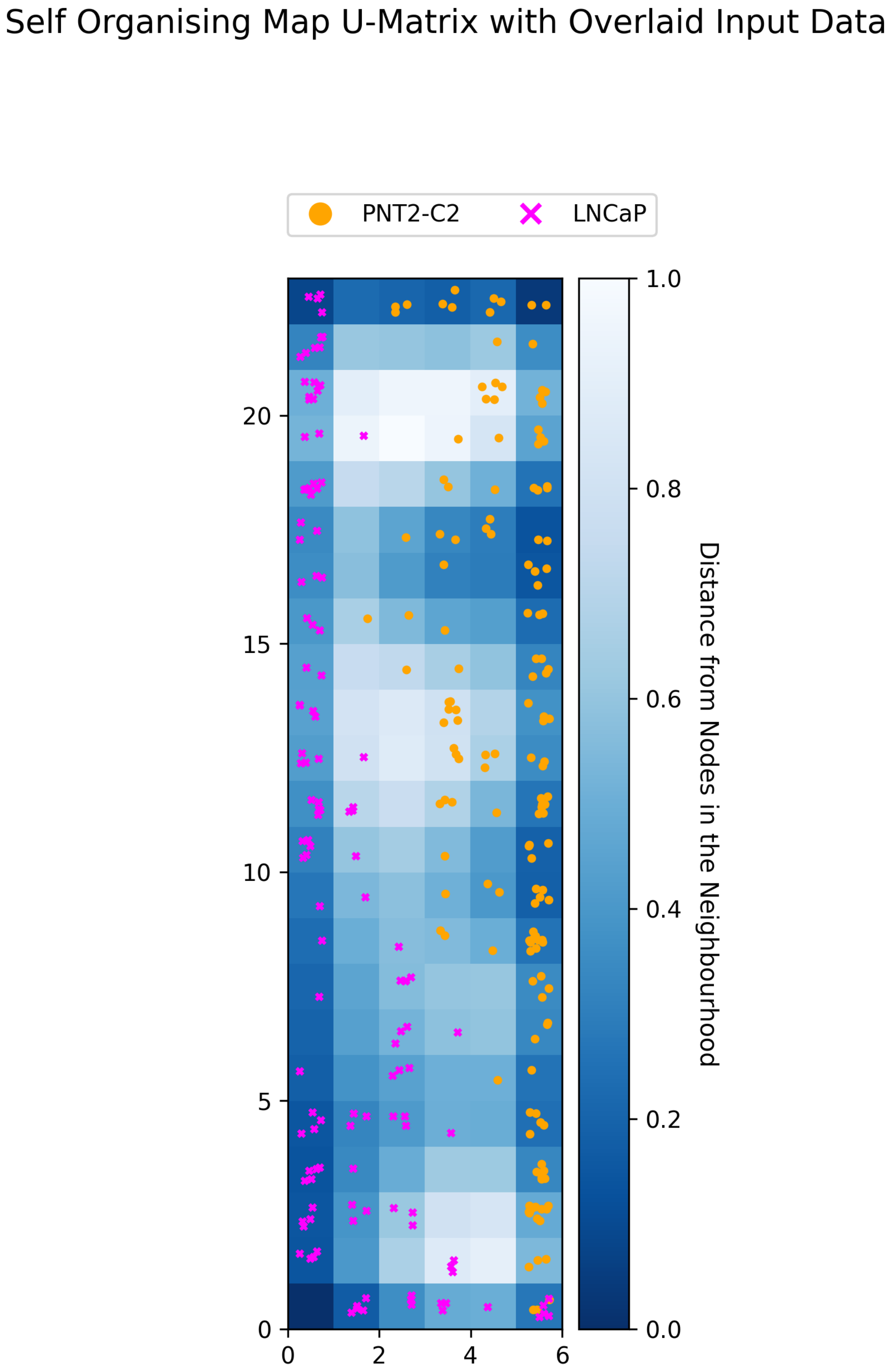


Figure 8.7: SOM u-matrix trained unsupervised on the Cleaned Prostate Cell Line dataset with the optimum parameters defined in Section 8.4. Regions of low nodal density around (2, 20), (1, 17), (2, 11), (4, 7), and (4, 3) divide the PNT2-C2 (normal prostate) and LNCaP (prostate cancer) classes. A stripe of low nodal density around (1, 7) within the LNCaP region may represent another border between clusters. Both PNT2-C2 and LNCaP spectra map to node (5, 0).

Cluster	Location	PNT2-C2	LNCaP	Proportion
A	Upper left	0	46	0.162
B	Lower left	0	13	0.046
C	Right	33	0	0.116

Table 8.4: Table summarising the number of PNT2-C2 and LNCaP observations found within the centre of each of the three clusters in Figure 8.7. The final column gives the proportion of the total observations ($n = 284$) found within that cluster centre.

Figure 8.8 shows the density plot for the SOM in Figure 8.7 with the input data overlaid. This image shows a region of low data density down the centre of the map, which marks a divide between the majority of PNT2-C2 and LNCaP observations. The PNT2-C2 data have one focus centred around (5, 9), whereas the LNCaP data have two foci centred around (1, 3) and (0, 16) separated by a small region of low data density.

Analysis of Figures 8.7 and 8.8 favours the single subgroup hypothesis, as two clusters around (1, 3) and (0, 16) contain data from LNCaP, and one cluster around (5, 9) contains data from PNT2-C2.

8.5.4 Cluster Spectral Plots

Investigation of the three clusters is performed by comparing the spectra at their centre, these spectra being the most typical of the pattern in each cluster. The selected spectra are those of data points which map to nodes within the three data density foci in Figures 8.6 and 8.8, and are summarised in Table 8.4.

Sample Spectra

Figure 8.9 shows the average spectra for the entire Cleaned Prostate Cell Line dataset and for each cluster shown in the SOM in Figure 8.7.

The average spectrum (solid black line) shows a large peak around 2940 cm^{-1} with two shoulders on its upstroke (2850 cm^{-1} and 2870 cm^{-1}), and a small peak around 3060 cm^{-1}. The cluster A average spectrum (dashed magenta line) follows the same basic shape as the sample average spectrum; however, the intensity of the shoulder at 2850 cm^{-1} and the upstroke between 2870 cm^{-1} and 2910 cm^{-1} are lower. The cluster B average spectrum (blue dot and dashed line) follows the same basic shape as the sample average spectrum, but has a much higher intensity at the shoulder at 2850 cm^{-1}, a higher intensity at the shoulder at 2870 cm^{-1}, and a lower intensity at the main peak at 2940 cm^{-1}. The cluster C average spectrum (dotted orange line) follows the pattern of the sample average spectrum.

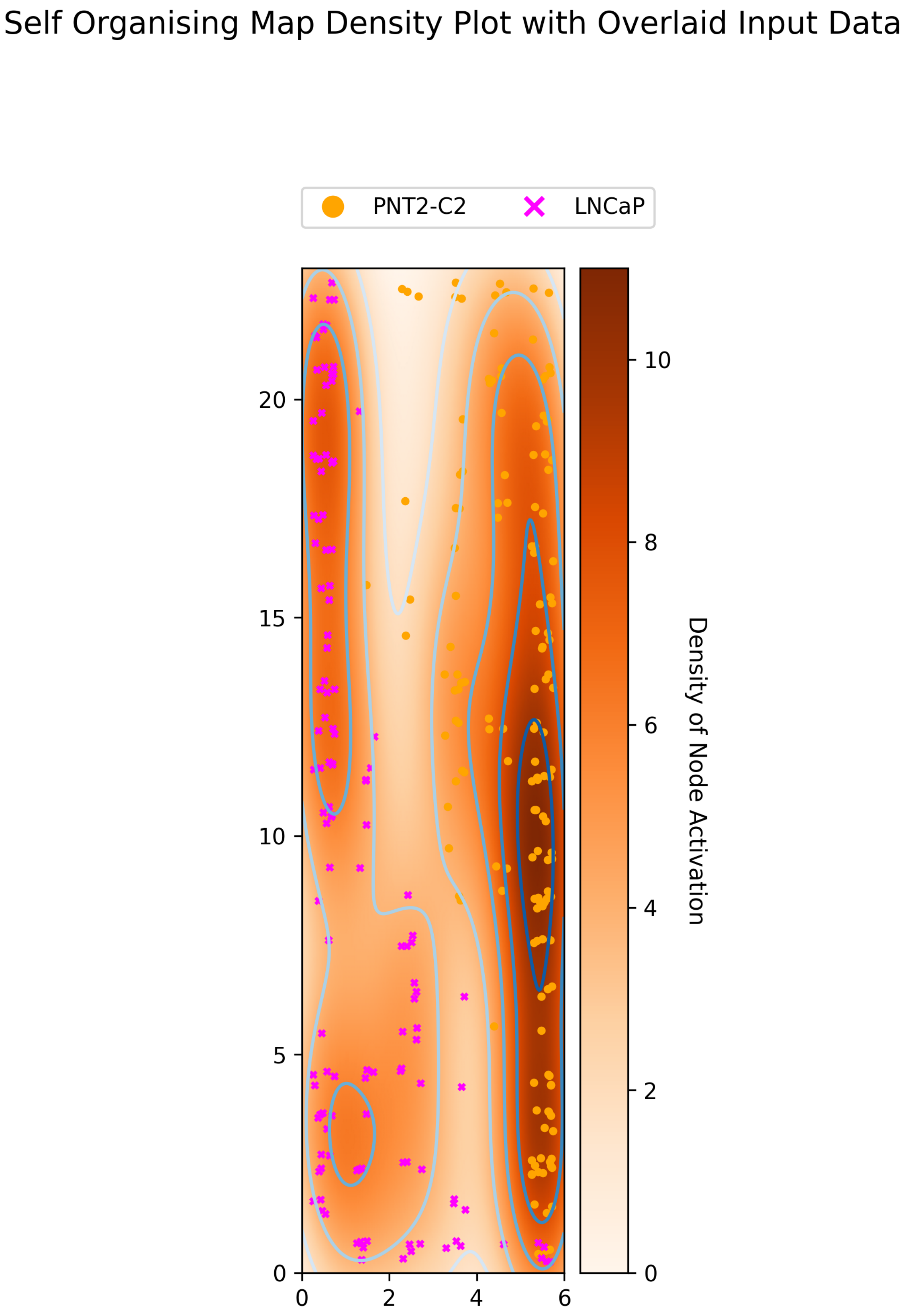


Figure 8.8: Density plot of the SOM in Figure 8.7 with the input data overlaid. There is a region of very low density spanning the plot from upper centre to lower centre, which divides the PNT2-C2 cluster around (5, 9) and the LNCaP clusters around (1, 3) and (0, 16). There is a second region of low density between the two LNCaP clusters around (1, 7).

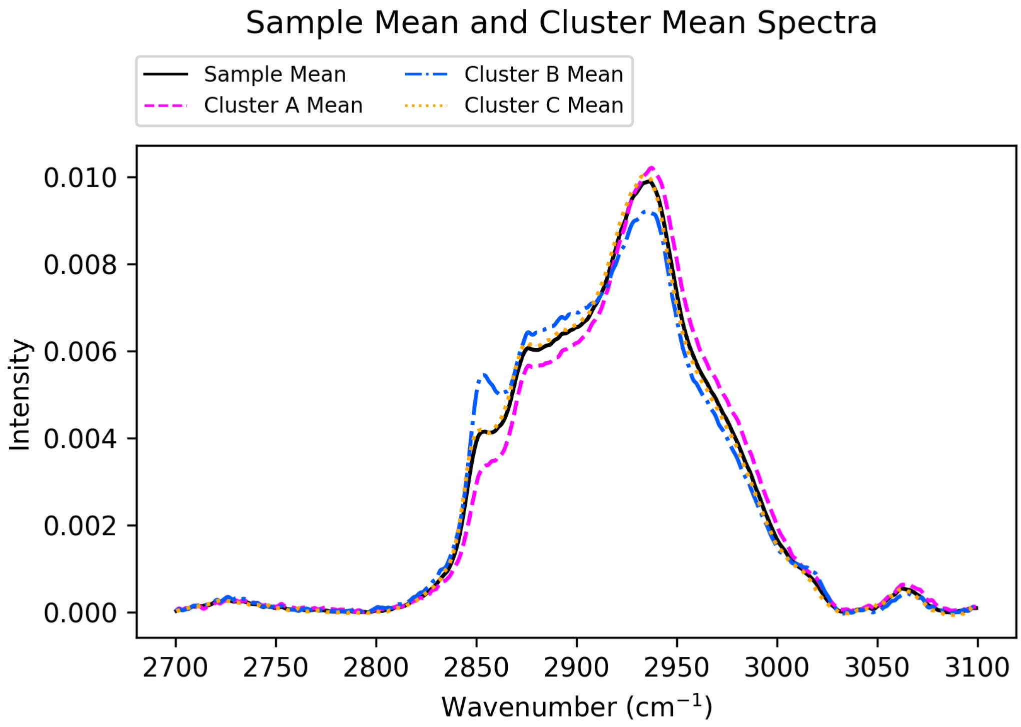


Figure 8.9: Average Raman spectra for the Cleaned Prostate Cell Line dataset.
The solid black line represents the average spectrum for the entire sample, with a large peak around 2940 cm^{-1} and a small peak around 3060 cm^{-1}. There are two shoulders to the upstroke of the large peak at 2850 cm^{-1} and 2870 cm^{-1}.
The magenta dashed line represents the average spectrum from cluster A at the upper left of the SOM in Figure 8.7, which has a lower intensity at the shoulder at 2850 cm^{-1} and at the upstroke between 2870 cm^{-1} and 2910 cm^{-1}, and otherwise follows the pattern of the sample average spectrum.
The blue dot and dashed line represents the average spectrum from cluster B at the lower left of the SOM in Figure 8.7, which has a higher intensity for the shoulders at 2850 cm^{-1} and 2870 cm^{-1}, and a lower intensity of the peak at 2940 cm^{-1} than the sample average spectrum.
The orange dotted line represents the average spectrum from cluster C at the right of the SOM in Figure 8.7, which follows the pattern of the sample average spectrum.

Observed Difference Discussion

There is a difference in the pattern observed for the spectra in clusters A and B, all of which are derived from LNCaP cell lines. The data are statistically converged (see Section 8.3.2), and so are representative of the constituent cell line populations across replicates. The different patterns seen could occur for a number of reasons.

Firstly, the laser providing the incident light for the analysis could be sampling a different part of the cell, such as the cytoplasm instead of the targeted nucleus. The laser used has a wavelength of 532 nm and the numerical aperture (NA) of the objective lens used is × 63 / 1.0. Calculation of the lateral spatial resolution of a laser spot, d, is shown in Equation 8.2 [30]. For the experimental setup used to gather data in these prostate cell line experiments, $d = 0.61 \times 532/1.0 \approx 0.3\mu$m. The hypothesis that different parts of the cell are being sampled is unlikely due to the fine spatial resolution of Raman spectroscopy being in the order of 1 μm, and the size of human nuclei being in the order of 10 μm [92].

$$d = \frac{0.61\lambda}{NA} \tag{8.2}$$

Secondly, being sure that sampling is occurring within the nuclei of cells, it could be that the area being sampled is somehow in a different state. When DNA is being replicated, the molecule relaxes and unwinds so it can be read and copied by cellular machinery, becoming 2 nm thick [103]. When not being replicated, DNA is wound around histone proteins to form 10 nm wide loops which coil to form 30 nm thick chromatin fibres [7]. When a cell prepares to divide, replicated DNA must be faithfully and equally partitioned between daughter cells; to do so, the 30 nm chromatin fibres are coiled tightly to form a 700nm wide chromatid, half of a 1400 nm wide chromosome [7]. The three orders of magnitude difference in scale of the width of DNA could potentially explain why Raman signals are different when analysing cells within the same disease state. The difference in the density of DNA and proteins may affect the vibrational energy levels of constituent molecules, thereby altering the proportion of them in a Raman activatable state and thus the resultant signal. However, this theory likely does not explain the cause of variation as all cells were synchronised to senescence via starvation before sampling, and so will have been in the same position within the cell cycle.

Thirdly, there could be diffraction artefacts affecting the measurement of the Raman signal. Due to the regular spacing of fibres which are in the order of 700nm width, it is possible that chromatids form a diffraction grating. If the regular spaces between fibres are parallel and slightly wider than the wavelength of incident light (in this case 532 nm), then light may be split into several beams travelling at different angles. Therefore, any interactions causing Raman shifts may be altered or more difficult to detect. However, this theory is unlikely to be physically plausible due to the small cross-section of cell sampled with the Raman laser.

Fourthly, the different patterns are seen for two groups of LNCaP cells, denoting that one may be a subgroup of the other. These cells are an immortalised cancer cell line, and it is biologically plausible that a mutation arose during passage of the cells at some point, which has been propagated through some of the progeny cells, giving two similar yet distinct groups. If this were true, it would be expected that a change could be traced to a single point in time, whereby all progeny from that point exhibit the new state. Dr Hancock has reviewed the data from each cluster and confirmed that there is no obvious pattern to which samples are found in each cluster, making this idea unlikely to explain the two subclusters found.

Observed Similarity Discussion

The data for PNT2-C2 and LNCaP were collected several months apart and by different people—they are distinct datasets. Nevertheless, there is very little visual difference between the average spectra for clusters B and C, despite cluster B being derived from LNCaP data and cluster C being derived from PNT2-C2 data. Given the similarity of these two average spectral shapes, the SOM must be detecting very slight differences in their patterns in order to be able to correctly segregate them into their respective clusters.

Node (5, 0) Spectral Plots

The node (5, 0) of the SOM in Figure 8.7 has both PNT2-C2 and LNCaP data mapped to it. The spectra of these observations are plotted in Figure 8.10, showing that each spectrum follows the general shape of the average spectrum shown in Figure 8.9, although there are no clearly discernible shoulders to the upstroke of the first peak, and the intensity of this peak is much lower than for the average spectrum. These spectra map far away from the main three clusters and are similar between classes.

8.5.5 Biological Plausibility of Cluster Differences

The data for LNCaP are statistically converged as evidenced by their low standard error of the mean of 2.4270 $\times 10^{-6}$ to 7.4999 $\times 10^{-5}$, and so they are representative of the LNCaP cell line population. The two clusters formed by the LNCaP data likely represent cancer heterogeneities.

Mean spectral subtraction is performed to highlight the differences between the mean spectra from clusters A and B. Each of these mean cluster spectra is subtracted from the mean spectrum of the PNT2-C2 cluster, this being from a normal prostate cell line and thus taken as a baseline value.

Figure 8.11 shows the mean subtracted spectrum for cluster A and Figure 8.12 shows the mean subtracted spectrum for cluster B. The large deflections seen for

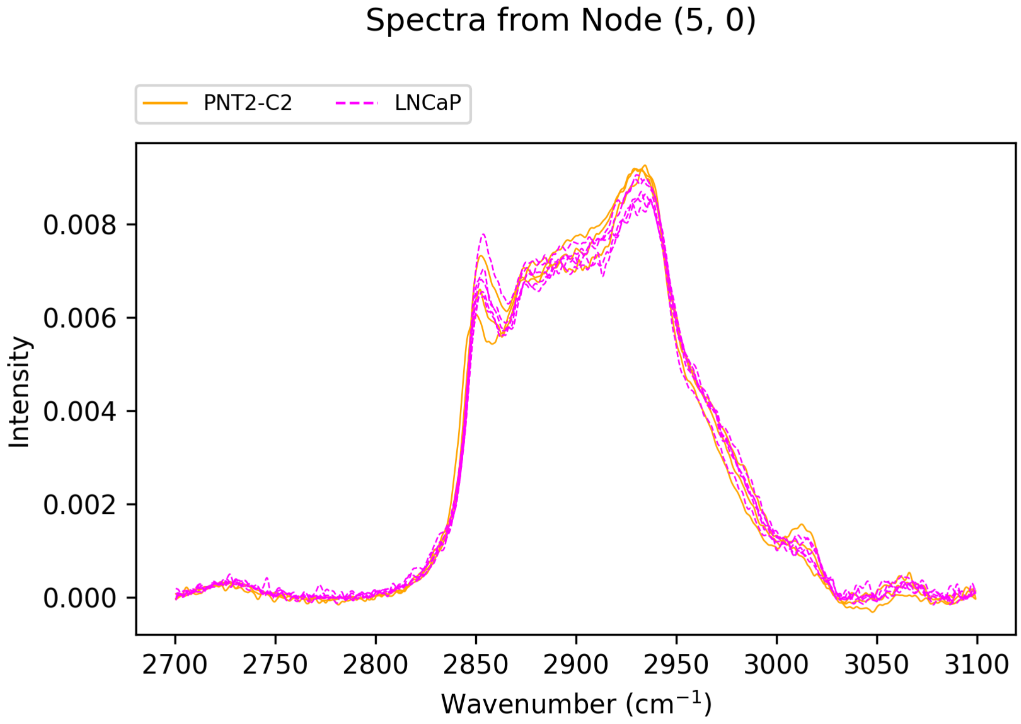


Figure 8.10: Spectra mapping to node (5, 0) in Figure 8.7. The solid orange lines belong to PNT2-C2 (normal prostate) spectra and the magenta dashed lines belong to LNCaP (prostate cancer) spectra. Each spectrum follows the general shape of two peaks at 2940 cm^{-1} and 3060 cm^{-1}. There are no clearly discernible shoulders to the upstroke of the first peak, and the intensity of the first peak is lower than for the average spectrum shown in Figure 8.9.

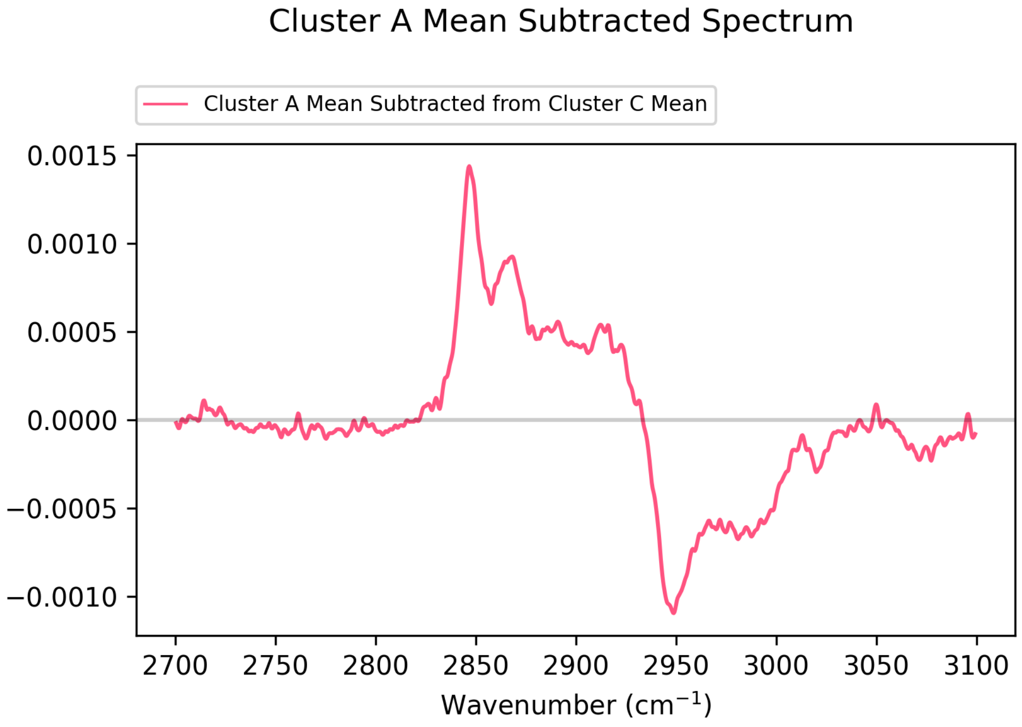


Figure 8.11: Resultant spectrum from subtracting the average cluster A (LNCaP) spectrum from the average cluster C (PNT2-C2) spectrum. As PNT2-C2 is taken as the baseline normal, positive deflections correspond to decreased intensity and negative deflections correspond to increased intensity in the average LNCaP signal compared to the average PNT2-C2 signal.
There is a broad positive deflection between 2840 cm^{-1} and 2930 cm^{-1}, with two large peaks at 2850 cm^{-1} and 2860 cm^{-1}. There is a broad negative deflection between 2935 cm^{-1} and 3010 cm^{-1}, with a deep trough at 2950 cm^{-1} and a broad trough at 2965–3010 cm^{-1}. There is a small negative deflection at 3020 cm^{-1}.

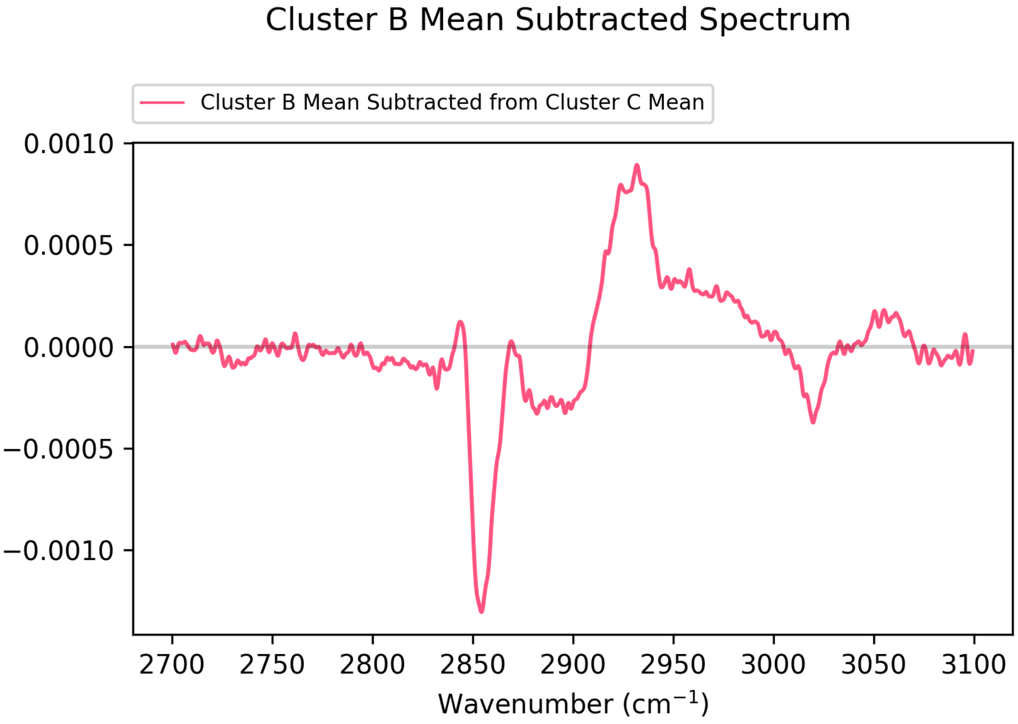


Figure 8.12: Resultant spectrum from subtracting the average cluster B (LNCaP) spectrum from the average cluster C (PNT2-C2) spectrum. As PNT2-C2 is taken as the baseline normal, positive deflections correspond to decreased intensity and negative deflections correspond to increased intensity in the average LNCaP signal compared to the average PNT2-C2 signal.
There is a deep negative deflection at 2850 cm^{-1}, a broad negative deflection at 2875-2905 cm^{-1}, a large positive deflection with a peak at 2920 cm^{-1} and a small negative deflection at 3010 cm^{-1}.

the two spectra are analysed with reference to a standard library of Raman signals generated from biomolecules [74, 93], to try to elucidate the underlying bimolecular changes responsible for the differences in signal.

- Cluster A
  - Positive deflection at 2840 cm^{-1} (saturated lipid)
  - Negative deflection at 2940 cm^{-1} (lipid chains)
- Cluster B
  - Negative deflection at 2850 cm^{-1} (saturated lipid)
  - Positive deflection at 2915 cm^{-1} (lipid chains)
- Clusters A and B
  - Negative deflection at 3015–3025 cm^{-1} (unsaturated lipids)

As the mean subtracted spectra are generated by removing the PNT2-C2 average spectrum from the LNCaP average spectrum, positive deflections correspond to decreased signal intensity and negative deflections correspond to increased signal intensity in the LNCaP compared to the PNT2-C2 signal. Cluster B appears to have more saturated lipids than normal, with fewer or shorter lipid chains; conversely, cluster A appears to have less saturated lipid than normal with more or longer lipid chains. Both have more unsaturated lipid than normal, and cluster A has more than cluster B.

8.6 Summary

This chapter has explored the use of SOM analysis of blinded high dimensionality spectral data. It has shown that the SOM is very sensitive to outlying data points, and that optimisation of SOM training parameters can yield reliable results. Two clusters were expected during SOM training with the cleaned Prostate Cell Line dataset, yet three clusters were found: spectral analysis shows that two clusters share the same input data class (LNCaP, prostate cancer), each with a different average spectral shape. Interestingly, the third cluster (PNT2-C2, normal prostate) has a spectral shape which, by eye, is the same as one of the LNCaP clusters, yet the SOM is still able to separate PNT2-C2 from LNCaP data.

The two clusters found for LNCaP display different Raman signal intensity at wavenumbers which correspond to lipids: cluster A has more unsaturated lipids and a higher total amount of lipid than cluster B, which appears to have more saturated lipids. These bimolecular differences could contribute to a phenotypic difference within the cells, as lipids are involved in cellular signalling and increased consumption of saturated fatty acids has been linked to more aggressive forms of prostate cancer [78].

These experiments and analyses show that the SOM method can be a useful tool for analysing spectral data given optimum parameters and highlight the feasibility of using this method in biomedical research.

Chapter 9

Conclusions and Discussion

9.1 Summary

The research question proposed in Chapter 3 is:

"Can self-organising maps distinguish between cancerous and non-cancerous prostate cells?"

The experiments laid out in this thesis form a feasibility study for the use of SOMs to analyse high dimensionality data from complex biological systems. The MiniSom module [101] has been studied and inherited to form the basic class of the MySom module written for this thesis. This module is used to analyse data to train and display SOMs, and its source code is freely available at `github.com/thenakedcellist/prostate/blob/master/mysom/mysom.py`.

The preliminary experiments in Chapters 5 and 6 demonstrate the use of the MySOM module to analyse low and high dimensionality data, respectively, and how to interpret the SOMs produced. Each of the experiments within these chapters uses an unblinded dataset with a known outcome, which is used to validate MySom's outputs.

The experiments in Chapter 7 show that the SOM is able to cluster even a small dataset ($n = 30$) of Raman spectra gathered from normal and cancer prostate cell lines appropriately by class. Visual interpretation of the SOMs and density plots to discern clusters is discussed.

The final experiments in Chapter 8 show that the SOM is able to cluster blinded Raman spectroscopic data gathered from normal prostate and prostate cancer cell lines ($n = 284$) by class. The method is shown to be sensitive to outlying data, and analysis of the raw data allows removal of an aberrant signal. A parameter sweep then allows a SOM to be built which best represents the dataset: three clusters are uncovered, the normal class in one cluster and the cancer class organised into two subclusters. Analysis of the spectra within the subclusters found by the SOM shows subtle differences associated with lipid levels which could explain their segregation.

The SOM method shows the potential for use as a diagnostic test as it can correctly cluster Raman spectral data from normal prostate and prostate cancer cells. It also manages to segregate two subclusters of cancer cell data based on patterns corresponding to different lipid compositions. These different lipid profiles show biological plausibility for a difference in disease aggressiveness [78], and therefore could represent a different disease stage or progression. The results of this SOM analysis could be used to direct further experimentation into mechanisms of oncogenic action driven by different cellular lipid profiles, which could ultimately inform prognosis.

9.2 Conclusion and Discussion

9.2.1 Contributions to Knowledge

This work is the first to use SOMs to analyse Raman spectra gathered from human-derived cell lines, and demonstrates the ability of SOMs to cluster data based on the cell type. Methods for selecting the optimum parameters for SOM training to allow good representation of the input data are discussed, and the SOMs are able to discern subclusters of a single group.

9.2.2 Software Engineering Practice

Good software engineering practices are followed to ensure that experiments are well structured and code is rigorous.

Experiments were theorised and designed using pseudocode, to enable a clear visualisation of the logic used, and to aid in translating from pseudocode into source code. The pseudocode outlining the MySom process can be found in Appendix A.

The MySom module is accompanied by a functional test suite built with pytest (version 6.1.1) [59]. Testing is performed each time code is refactored and new functions are added. Assert statements are used within the code to stop a code run should input data be in an incorrect format, rather than allowing code to execute and potentially misrepresent the data.

A version control system is used to ensure a full history of code changes is preserved. The source code for this project is available on GitHub at `github.com/thenakedcellist/prostate`.

9.3 Future Works

9.3.1 Feature Selection

Three clusters are revealed in the Cleaned Prostate Cell Line dataset by the SOM method, and visual inspection of these clusters to understand the differences between them is performed in this work. The next appropriate step is to perform statistical analyses on these clusters, to gain an idea of which spectral bands are the most significantly different among them and hence deterministic of cluster membership. PCA and k-means clustering could be re-performed with the expectation of finding three clusters.

Following from the methods of Rauber and Merkl [82], and Tan [94], it may be possible to design a tool such as LabelSOM which is able to highlight spectral bands of importance to cluster membership. This method will require much investigation and calibration with the data, as the original experiments were performed with binary data, so a cut-off value of 0.5 could easily be used to classify a feature as influential or not. As spectral data are continuous, appropriate cutoff values for classifying a feature as important must be thoroughly tested.

9.3.2 Statistical Analyses

The next logical step following the experiments discussed here is statistical analysis of the SOM as a diagnostic method.

Confusion Matrix

Depending on how coarsely or finely spread the bounds defining density regions on the density plot are, points near the edges of clusters may be classified as within the cluster or outside of it. If the data can be clearly denoted as belonging to either the normal or disease group (no “unknown” labelling) on visual interrogation, then it is a binary classifier and can be recorded in a confusion matrix and used to calculate sensitivity (true positive rate) and specificity (true negative rate) for the SOM as a diagnostic classifier. This method of appraising binary classifiers is used commonly throughout medical science.

Receiver Operating Characteristic Curves

As the granularity of the density plot can be changed, so too can the status of some points near the edges of clusters—with a fine resolution edge points may be classed as outside of the cluster, whereas with coarser resolution they may be included within the cluster. Changing the granularity of the density plot will therefore alter the sensitivity and specificity of the SOM at the given density threshold.

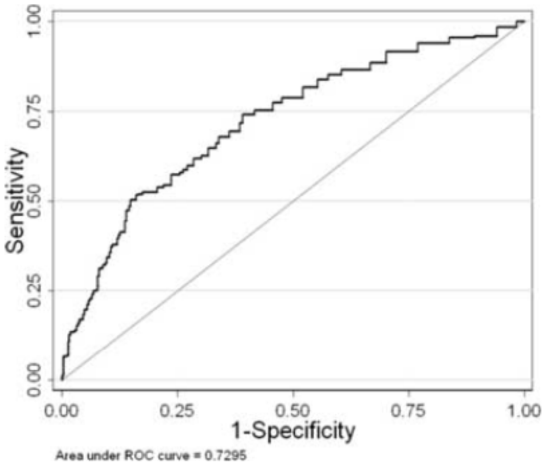


Figure 9.1: An example receiver operating characteristic curve (adapted from [2]). The area under the curve (AUC) is 0.7295, indicating that the diagnostic test under examination can correctly classify 72.95% of samples it is given.

A receiver operating characteristic (ROC) curve can be drawn plotting sensitivity against $1 -$ specificity (false positive rate) to find the threshold which gives the optimum values of sensitivity and specificity. An example ROC curve is given in Figure 9.1.

The nature of the disease under study, prostate cancer, is one which may be debilitating to many people who suffer from it, and fatal to an unfortunate minority. Therefore, maximising sensitivity while minimising error overall would be preferred, as the test would rarely misclassify an individual as healthy when they have the disease, although this preference may bias toward false positive results. A ROC curve can be used to decide on the optimum threshold to be used for the SOM density plots to achieve the desired levels of sensitivity and specificity.

Overfitting

One consideration of any simulation is overfitting of the resultant model. The Cleaned Prostate Cell Line dataset used in these experiments is not small, but given the vast heterogeneity of biological systems there is a risk of overfitting the SOM model to the random error intrinsic to the dataset, rendering it unable to be generalised to new data.

K-fold cross-validation can be used to assess for overfitting [39]. The dataset is split into K equal parts, and each Kth part is used to test the model trained on the other $K - 1$ parts. The errors of each of the K iterations of the experiment can then be combined to give an idea of the prediction error of the model.

9.3.3 Investigation of Raman Spectra

The average Raman spectra (Figure 8.9) from the two LNCaP (prostate cancer) clusters in the SOM (Figure 8.7) show very slight differences, which may be the cause for their segregation by the SOM. Analysis of mean subtracted spectra (Figures 8.11 and 8.12) suggests that these differences are due to differing lipid levels between the cells in each cluster.

Further analysis of these spectra and comparison to a standard library should reveal the biochemical species which are responsible for the observed Raman shift. Once defined, a biologically plausible reason for the differences in these species, such as increased production of a molecule involved in cancer signalling, or decreased production of a molecule involved in cellular repair, can be investigated.

9.3.4 Developing the SOM Method for Clinical Practice

This work outlines the first steps in an ongoing process of learning about the intracellular mechanisms of prostate cancer and diagnostic test development. It forms a feasibility study that has proven the SOM to be a useful analytic tool, and research groups are encouraged to investigate its use for high dimensionality data analysis within complex biological systems.

Appendix A

MySOM Pseudocode

Algorithm A.1 shows the pseudocode for the MySom module. It inherits the features of MiniSom [101], requires input of SOM parameters and the dataset under study, and produces consistent output plots.

Each function of the algorithm is performed in the order given: relevant functions may be omitted, such as data removal if the entire dataset is to be used, or data normalisation if the input data are already normalised.

Algorithm A.1 MySom Pseudocode

```
input: input data ($A_{m,n}$), SOM parameters

function SETUP
    MySom ← MiniSom                                    ▷ inherit from MiniSom [101]
    instantiate empty SOM object
    return SOM
end function

function INITIALISE SOM(SOM, input data, SOM parameters)
    SOM object ← SOM parameters
    SOM object ← input data
    if data are labelled as blinded data then
        RAISE MESSAGE(data are blinded)
    else if data are labelled then
        RAISE MESSAGE(all values are labelled)
    else if data are unlabelled then
        RAISE MESSAGE(all values are unlabelled)
    else if some data are unlabelled then
        RAISE MESSAGE(some values are unlabelled)
    end if
    return SOM
end function

function REMOVE OBSERVATIONS(SOM, list of indices for removal)
    remove selected index of input data from SOM
    return SOM
end function

function NORMALISE DATA(SOM)
    for $A_{m,n}$ in SOM do
        $x_{i,j} \leftarrow x_{i,j}/\|a_{i,*}\|_F$          ▷ divide each row by Frobenius norm
    end for
    return SOM
end function
```

```
function train SOM(SOM)
    random weight initiation
    for iteration $t$ do
        for row in input do
            calculate euclidean distance to each node
            BMU ← nearest node
            update BMU weights to better approximate input vector
            update weights of BMU neighbours
        end for
        update neighbourhood function radius, $\sigma(t)$
        update learning rate, $\alpha(t)$
    end for
    calculate quantisation error
    calculate topographic error
    return SOM, quantisation error, topographic error
end function

function plot SOM u-matrix
    create u-matrix of node density
    overlay input data on BMU
    add random jitter to input data
    return SOM u-matrix
end function

function plot SOM density function
    create u-matrix of node density
    overlay Gaussian kernel density estimate
    if overlay data required then
        overlay input data on BMU
        add random jitter to input data
    end if
    return SOM density plot
end function

output: SOM u-matrix, SOM density plot
```

Appendix B

Discarded SOMs from Blinded Prostate Cell Line Experiment

In Chapter 8 a sweep was performed to find the optimum parameters for a SOM trained on the Blinded Prostate Cell Line dataset. As discussed in Section 8.4, the outputs of training are judged on errors and visual interpretation of the SOM u-matrices and density plots.

The optimum parameter set is chosen as the one with the lowest total error, for which the SOM plots show distinct borders between clusters and the density plots show discrete data foci. Table B.1 is reproduced from Table 8.3 and shows the twenty SOM parameter families with the lowest total error. Family 117 is selected as the optimum parameter set, as those families with a lower total error do not show clear SOM cluster borders or discrete data foci. Figures B.1–B.30 show the SOM u-matrix and density plots for each of the families 81–63.

Family	Array	σ	α	QE	TE	(QE + TE)	|QE − TE|
81	15 × 4	1.0	0.5	0.019604	0.0	0.019604	0.019604
82	15 × 4	1.0	0.75	0.019908	0.0	0.019908	0.019908
83	15 × 4	1.0	0.9	0.020134	0.0	0.020134	0.020134
111	23 × 6	2.0	0.5	0.020138	0.0	0.020138	0.020138
61	17 × 5	2.0	0.5	0.021557	0.0	0.021557	0.021557
62	17 × 5	2.0	0.75	0.021627	0.0	0.021627	0.021627
64	17 × 5	2.0	0.95	0.021776	0.0	0.021776	0.021776
86	15 × 4	2.0	0.5	0.022923	0.0	0.022923	0.022923
106	23 × 6	1.0	0.5	0.016016	0.007042	0.023058	0.008973
87	15 × 4	2.0	0.75	0.023493	0.0	0.023493	0.023493
88	15 × 4	2.0	0.9	0.023742	0.0	0.023742	0.023742
89	15 × 4	2.0	0.95	0.023809	0.0	0.023809	0.023809
90	15 × 4	2.0	0.99	0.023850	0.0	0.023850	0.023850
12	19 × 5	2.0	0.75	0.021298	0.003521	0.024819	0.017777
63	17 × 5	2.0	0.9	0.021612	0.003521	0.025133	0.018091
117	**23 × 6**	**3.0**	**0.75**	**0.022025**	**0.003521**	**0.025546**	**0.018504**
91	15 × 4	3.0	0.5	0.026191	0.0	0.026191	0.026191
93	15 × 4	3.0	0.9	0.026643	0.0	0.026643	0.026643
48	14 × 6	4.0	0.9	0.026652	0.0	0.026652	0.026652
66	17 × 5	3.0	0.5	0.023308	0.003521	0.026829	0.019787

Table B.1: Table of the twenty SOM families in the parameter sweep with the lowest total error. The values for family 117 are shown in bold as these are the optimum SOM training parameters based on errors and visual interpretation of the output plots. QE = quantisation error. TE = topographic error. All errors given are median values for the ten family members.

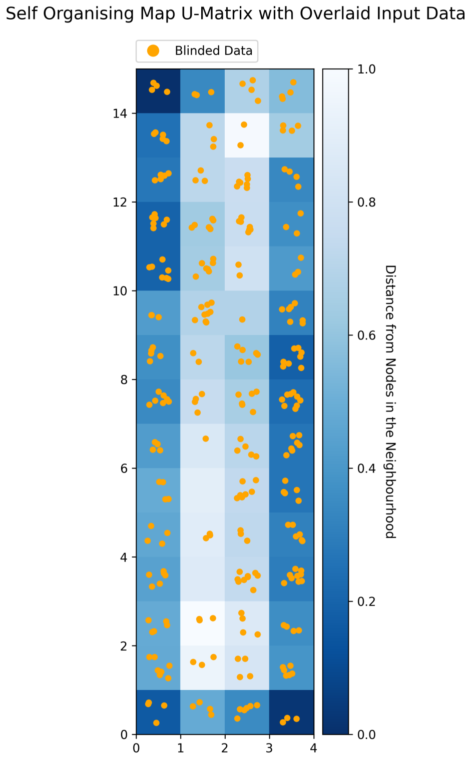


Figure B.1: Family 81 SOM u-matrix showing no distinct cluster borders.

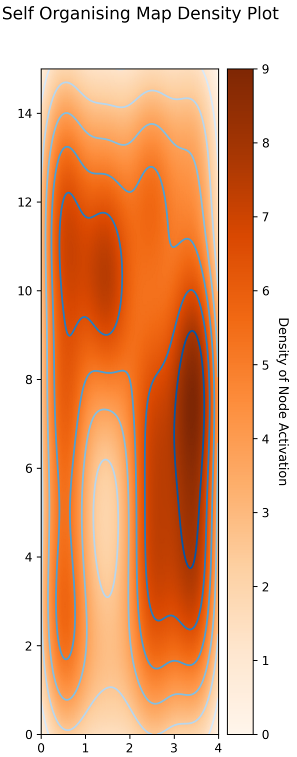


Figure B.2: Family 81 density plot showing no discrete data foci.

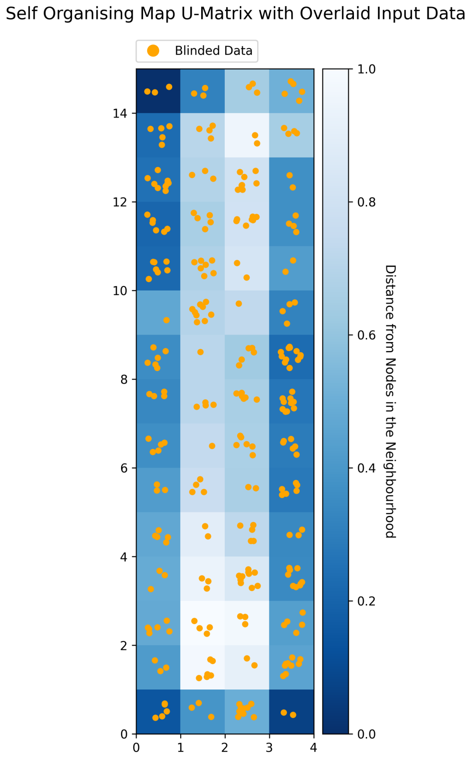


Figure B.3: Family 82 SOM u-matrix showing no distinct cluster borders.

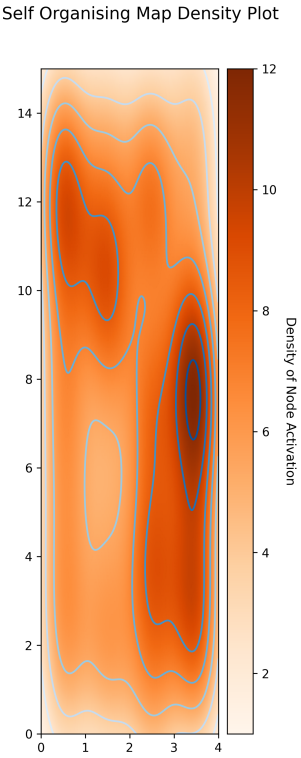
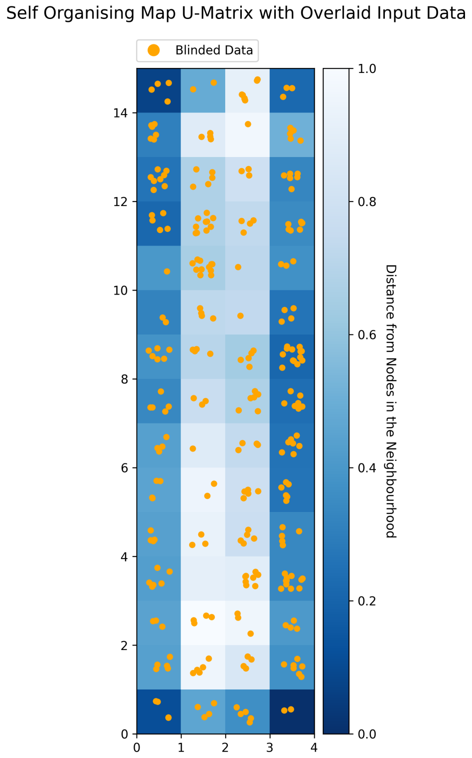


Figure B.4: Family 82 density plot showing no discrete data foci.



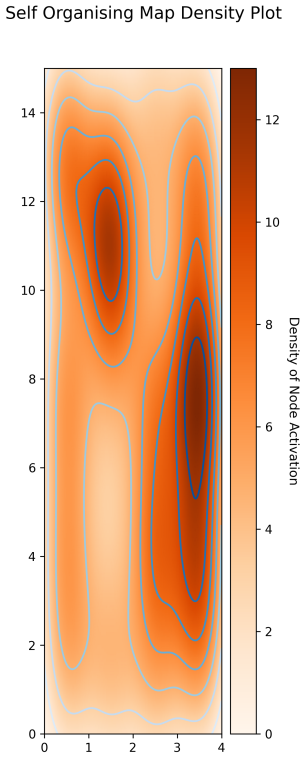
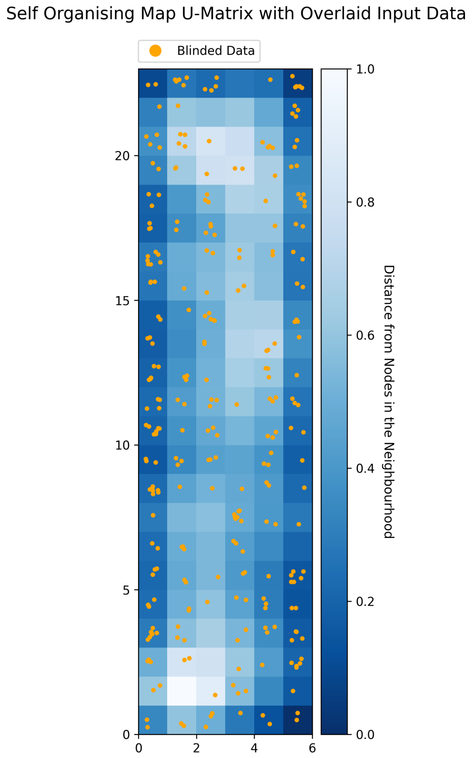


Figure B.5: Family 83 SOM u-matrix showing no distinct cluster borders.

Figure B.6: Family 83 density plot showing no discrete data foci.



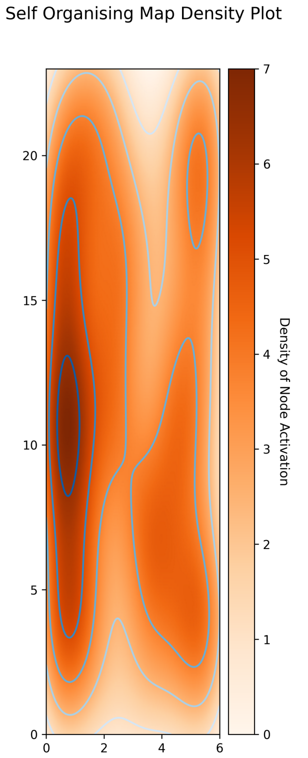
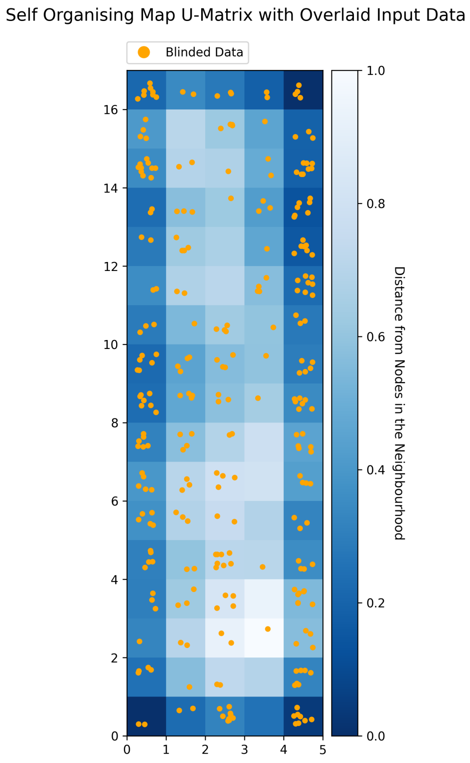


Figure B.7: Family 111 SOM u-matrix showing regions of empty nodes with no distinct cluster borders.

Figure B.8: Family 111 density plot showing some separation of data foci.



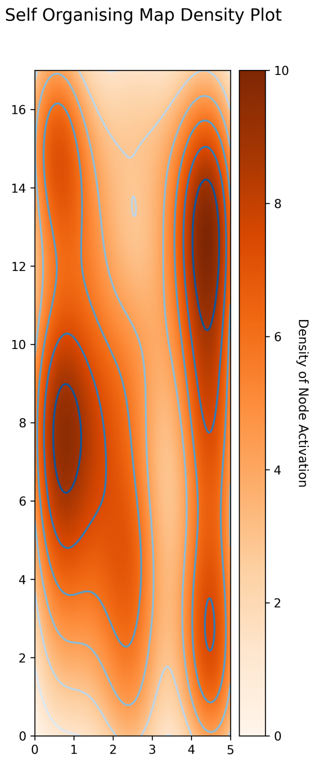
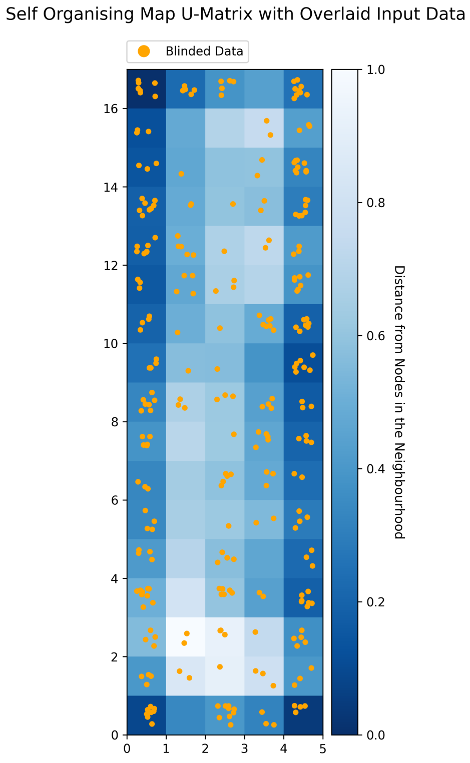


Figure B.9: Family 61 SOM u-matrix showing regions of empty nodes with no distinct cluster borders.

Figure B.10: Family 61 density plot showing some separation of data foci.



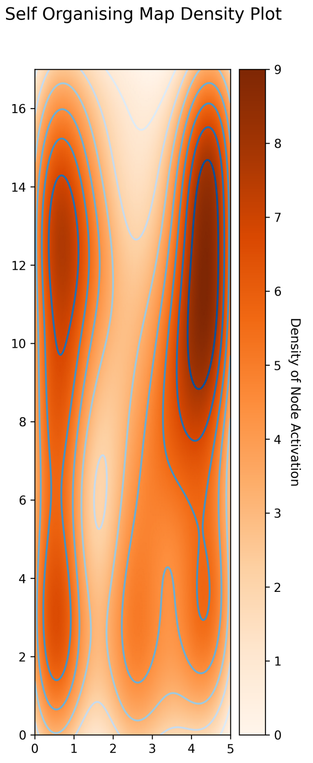


Figure B.11: Family 62 SOM u-matrix showing regions of empty nodes with no distinct cluster borders.

Figure B.12: Family 62 density plot showing some separation of data foci.

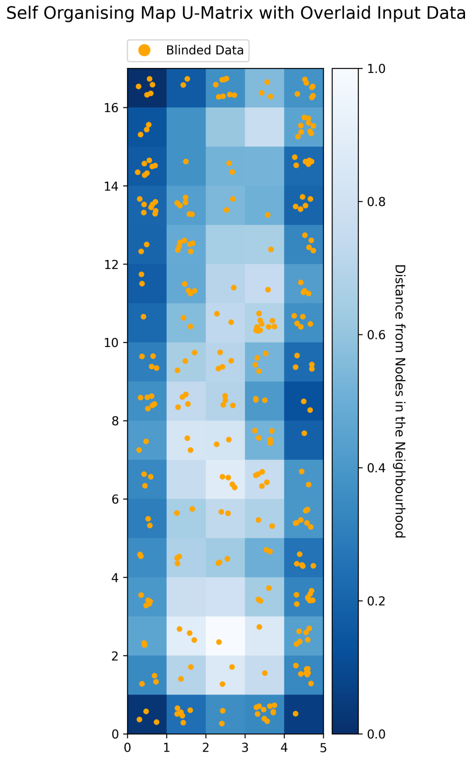


Figure B.13: Family 64 SOM u-matrix showing no distinct cluster borders.

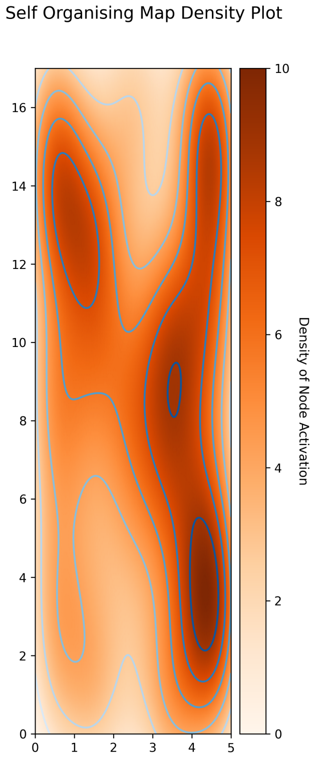


Figure B.14: Family 64 density plot showing no discrete data foci.

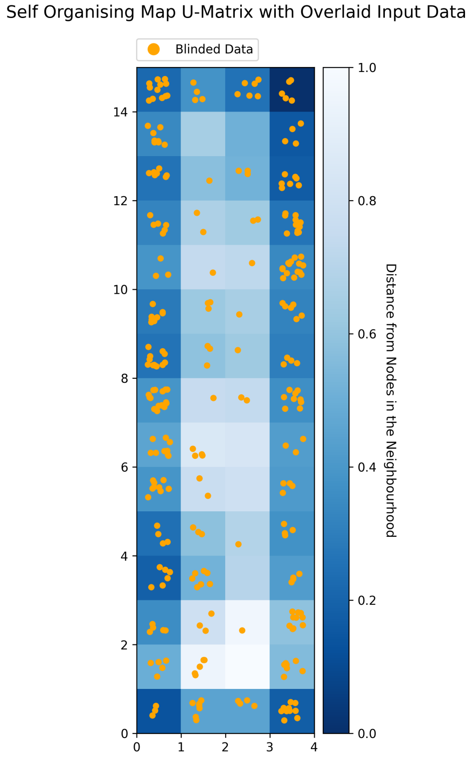


Figure B.15: Family 86 SOM u-matrix showing no distinct cluster borders.

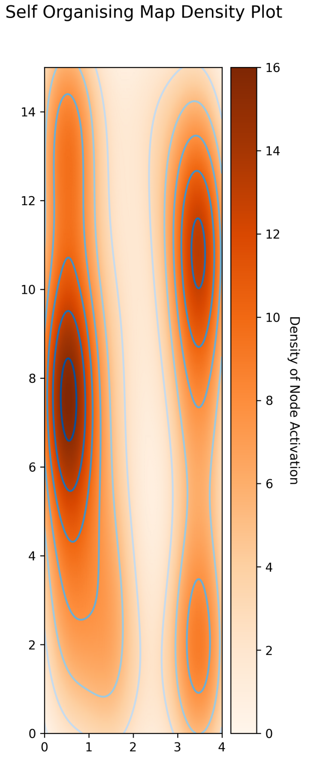


Figure B.16: Family 86 density plot showing separation into discrete data foci.

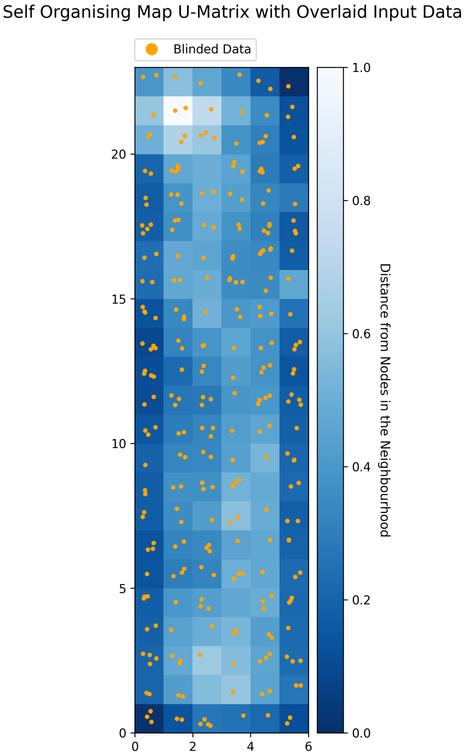


Figure B.17: Family 106 SOM u-matrix showing no distinct cluster borders.

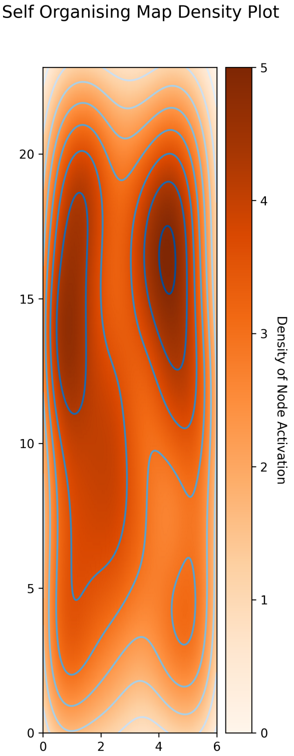


Figure B.18: Family 106 density plot showing no discrete data foci.

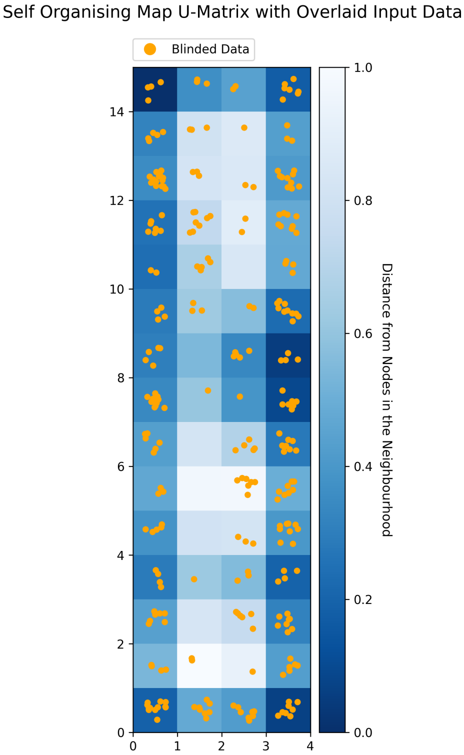


Figure B.19: Family 87 SOM u-matrix showing no distinct cluster borders.

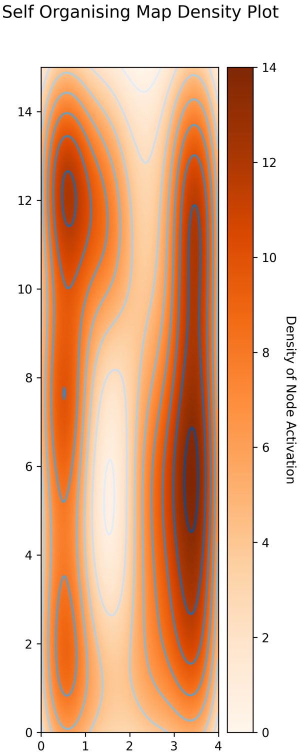


Figure B.20: Family 87 density plot showing separation into discrete data foci.

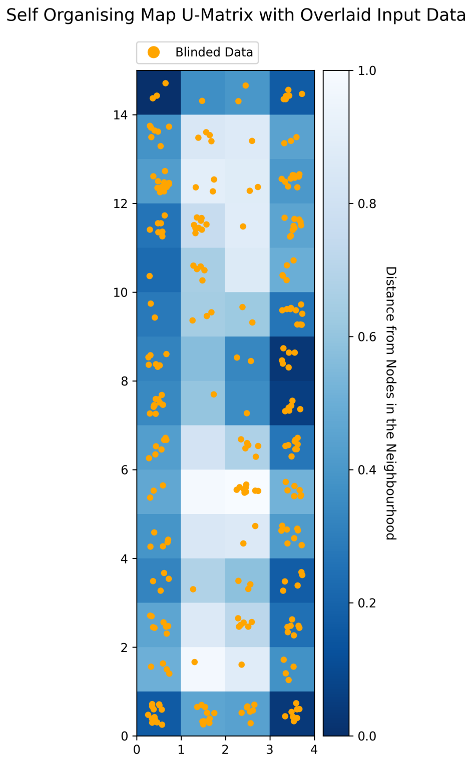


Figure B.21: Family 88 SOM u-matrix showing no distinct cluster borders.

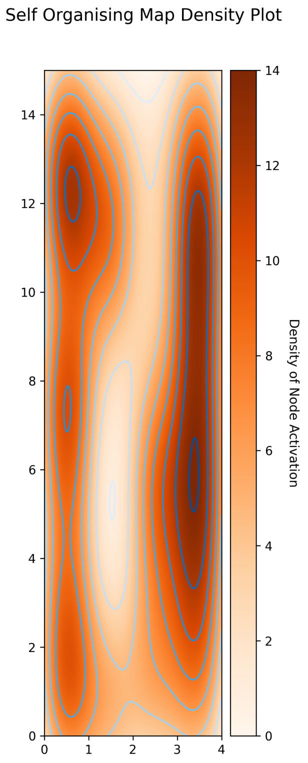


Figure B.22: Family 88 density plot showing separation into discrete data foci.

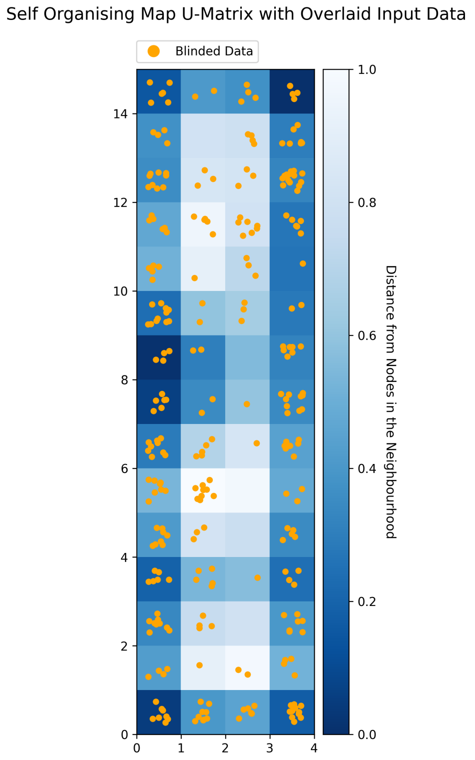


Figure B.23: Family 89 SOM u-matrix showing no distinct cluster borders.

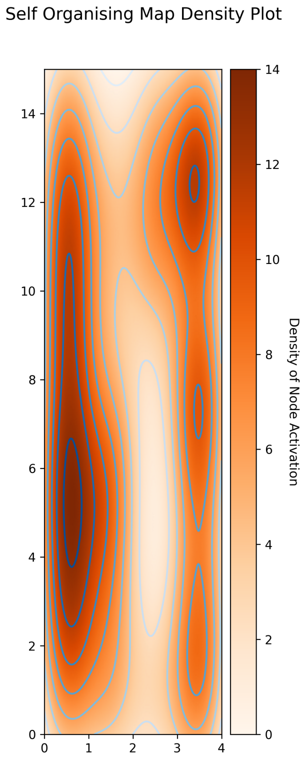


Figure B.24: Family 89 density plot showing separation into discrete data foci.

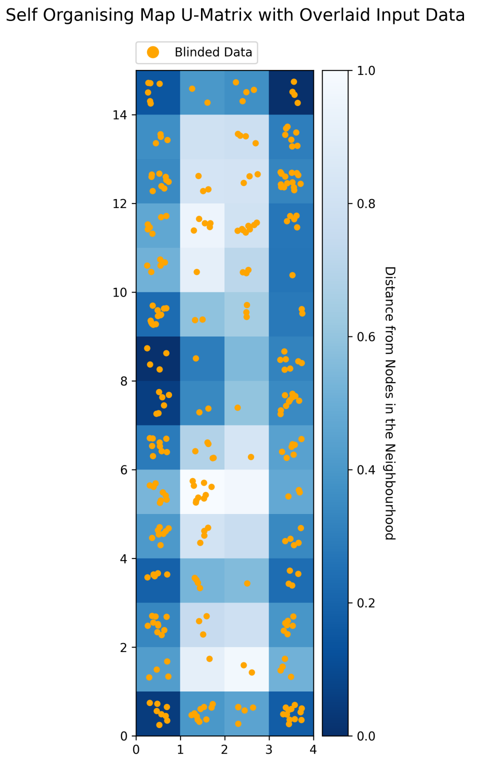


Figure B.25: Family 90 SOM u-matrix showing no distinct cluster borders.

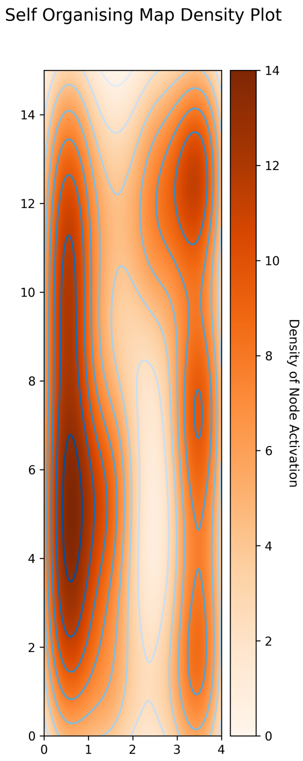


Figure B.26: Family 90 density plot showing separation into discrete data foci.

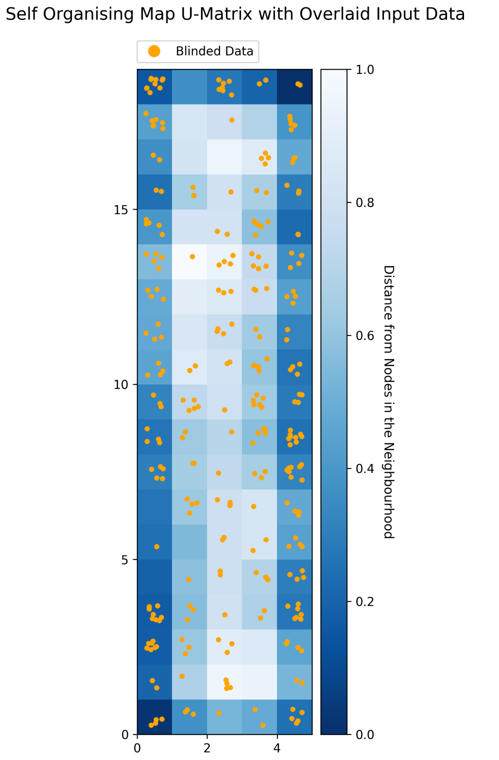


Figure B.27: Family 12 SOM u-matrix showing regions of empty nodes with no distinct cluster borders.

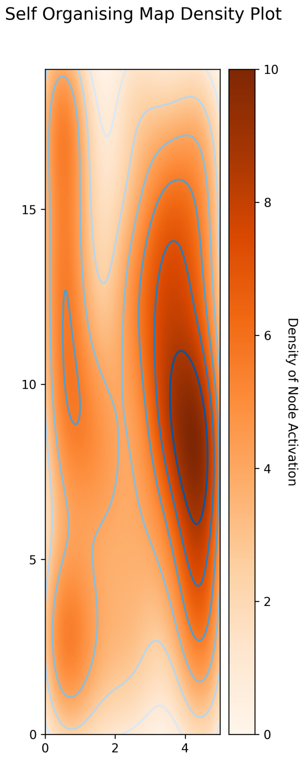


Figure B.28: Family 12 density plot showing no discrete data foci.

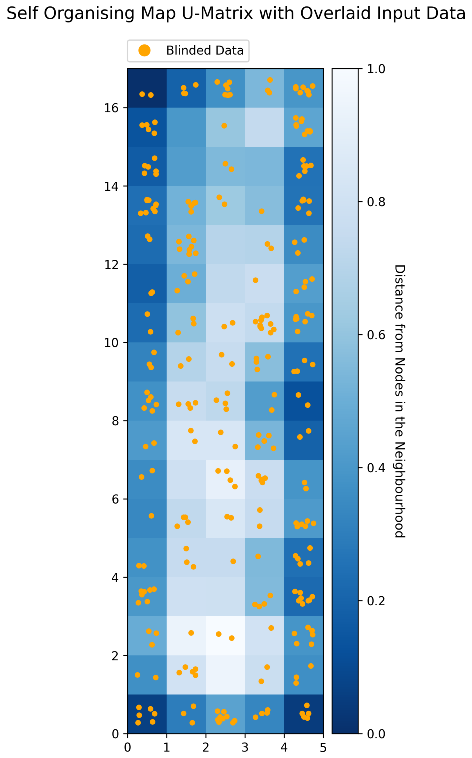


Figure B.29: Family 63 SOM u-matrix showing regions of empty nodes with no distinct cluster borders.

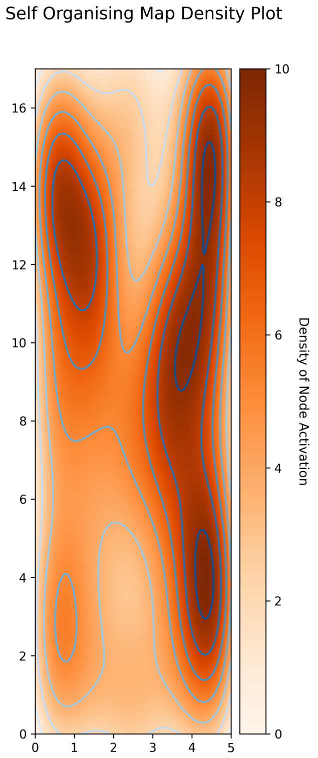


Figure B.30: Family 63 density plot showing no discrete data foci.

Bibliography

[1] H Abramczyk, A Imiela, and A Śliwińska. Novel strategies of Raman imaging for exploring cancer lipid reprogramming. *J. Mol. Liq.*, 274:52–59, January 2019.

[2] Anthony K Akobeng. Understanding diagnostic tests 3: Receiver operating characteristic curves. *Acta Paediatr.*, 96(5):644–647, May 2007.

[3] Bruce Alberts, Alexander Johnson, Julian Lewis, Keith Roberts, Martin Raff, and Peter Walter. *Molecular Biology of the Cell.* Garland Science, fifth edition, 2008.

[4] Douglas G Altman and J Martin Bland. Standard deviations and standard errors. *BMJ*, 331(7521):903, October 2005.

[5] Edgar Anderson. The irises of the Gaspé peninsula. *Bull. Am. Iris Soc.*, 59:2–5, 1935.

[6] Edgar Anderson. The species problem in iris. *Ann. Mo. Bot. Gard.*, 23(3):457–509, 1936.

[7] Anthony Annunziato. DNA Packaging: Nucleosomes and Chromatin. *Nature education*, 1(1):26, 2008.

[8] Kelly Aubertin, Vincent Quoc Trinh, Michael Jermyn, Paul Baksic, Andrée-Anne Grosset, Joannie Desroches, Karl St-Arnaud, Mirela Birlea, Maria Claudia Vladoiu, Mathieu Latour, Roula Albadine, Fred Saad, Frédéric Leblond, and Dominique Trudel. Mesoscopic characterization of prostate cancer using Raman spectroscopy: potential for diagnostics and therapeutics. *BJU Int.*, 122(2):326–336, August 2018.

[9] David A Bader and Sean E McGuire. Tumour metabolism and its unique properties in prostate adenocarcinoma. *Nat. Rev. Urol.*, 17(4):214–231, April 2020.

[10] Carl Banbury. An implementation of a Kohonen map in JavaScript extended to provide feature extraction and classification. `https://github.com/cbanbury/kohonen`, 2018 (accessed August 14, 2020).

[11] Carl Banbury, Richard Mason, Iain Styles, Neil Eisenstein, Michael Clancy, Antonio Belli, Ann Logan, and Pola Goldberg Oppenheimer. Development of the self optimising Kohonen index network (SKiNET) for Raman spectroscopy based detection of anatomical eye tissue. *Sci. Rep.*, 9(1):10812, July 2019.

[12] Yaneer Bar-Yam. General Features of Complex Systems. *Encyclopedia of Life Support Systems*, 2002.

[13] P Berthon, O Cussenot, L Hopwood, A Leduc, and N Maitland. Functional expression of sv40 in normal human prostatic epithelial and fibroblastic cells - differentiation pattern of nontumorigenic cell-lines. *Int. J. Oncol.*, 6(2):333–343, February 1995.

[14] Andrea G Bodnar, Michel Ouellette, Maria Frolkis, Shawn E Holt, Choy-Pik Chiu, Gregg B Morin, Calvin B Harley, Jerry W Shay, Serge Lichtsteiner, and Woodring E Wright. Extension of Life-Span by Introduction of Telomerase into Normal Human Cells. *Science*, 279(5349):349–352, January 1998.

[15] Eva Brauchle and Katja Schenke-Layland. Raman spectroscopy in biomedicine - non-invasive in vitro analysis of cells and extracellular matrix components in tissues. *Biotechnol. J.*, 8(3):288–297, March 2013.

[16] Nadezda A Brazhe, Marek Treiman, Alexey R Brazhe, Ninett L Find, Georgy V Maksimov, and Olga V Sosnovtseva. Mapping of Redox State of Mitochondrial Cytochromes in Live Cardiomyocytes Using Raman Microspectroscopy. *PLOS One*, 7(9):e41990, September 2012.

[17] Matthew Brush. Color names. `https://github.com/codebrainz/color-names/blob/master/output/colors.csv`, 2012 (accessed February 27, 2020).

[18] Anil Cheriyadat and Lori Mann Bruce. Why Principal Component Analysis is not an Appropriate Feature Extraction Method for Hyperspectral Data. In *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477)*, volume 6, pages 3420–3422, July 2003.

[19] Neandder A Correia, Lucas T A Batista, Roberto J M Nascimento, Maria C T Cangussú, Pedro J L Crugeira, Luiz G P Soares, Landulfo Silveira, Jr, and Antonio L B Pinheiro. Detection of prostate cancer by Raman spectroscopy: A multivariate study on patients with normal and altered PSA values. *J. Photochem. Photobiol. B*, 204:111801, March 2020.

[20] Sishan Cui, Shuo Zhang, and Shuhua Yue. Raman Spectroscopy and Imaging for Cancer Diagnosis. *J. Healthc. Eng.*, 2018:8619342, June 2018.

[21] O Cussenot, P Berthon, R Berger, I Mowszowicz, A Faille, F Hojman, P Teillac, A Le Duc, and F Calvo. Immortalization of human adult normal prostatic epithelial cells by liposomes containing large T-SV40 gene. *J. Urol.*, 146(3):881–886, September 1991.

[22] Richard Dawkins. *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe Without Design.* Penguin Books, 30th anniversary edition edition, 2016.

[23] Eric de Bodt, Marie Cottrell, and Michel Verleysen. Statistical tools to assess the reliability of self-organizing maps. *Neural Netw.*, 15(8-9):967–978, 2002.

[24] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[25] A Einstein. Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. *Annalen der Physik*, 322(6):132–148, 1905.

[26] Katherine J I Ember, Marieke A Hoeve, Sarah L McAughtrie, Mads S Bergholt, Benjamin J Dwyer, Molly M Stevens, Karen Faulds, Stuart J Forbes, and Colin J Campbell. Raman spectroscopy and regenerative medicine: a review. *NPJ Regen Med*, 2:12, May 2017.

[27] Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett Delahunt, John R Srigley, and Peter A Humphrey. The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *Am. J. Surg. Pathol.*, 40(2):244–252, 2016.

[28] E Erwin, K Obermayer, and K Schulten. Self-organizing maps: ordering, convergence properties and energy functions. *Biol. Cybern.*, 67(1):47–55, 1992.

[29] John R Ferraro. *Introductory Raman Spectroscopy.* Elsevier, January 2003.

[30] John R Ferraro and Kazuo Nakamoto. *Introductory Raman Spectroscopy.* Academic Press, December 2012.

[31] R A Fisher. The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 7(2):179–188, 1936.

[32] Rekha Gautam, Sandeep Vanga, Freek Ariese, and Siva Umapathy. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation*, 2(1):8, June 2015.

[33] R J Geraghty, A Capes-Davis, J M Davis, J Downward, R I Freshney, I Knezevic, R Lovell-Badge, J R W Masters, J Meredith, G N Stacey, P Thraves, M Vias, and Cancer Research UK. Guidelines for the use of cell lines in biomedical research. *Br. J. Cancer*, 111(6):1021–1046, September 2014.

[34] D F Gleason. Classification of prostatic carcinomas. *Cancer Chemother. Rep.*, 50(3):125–128, March 1966.

[35] Peter Godfrey-Smith. *Philosophy of Biology.* Princeton Foundations of Contemporary Philosophy. Princeton University Press, 2013.

[36] D Hanahan and R A Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, January 2000.

[37] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, March 2011.

[38] Andrew T Harris, Manjree Garg, Xuebin B Yang, Sheila E Fisher, Jennifer Kirkham, D Alastair Smith, Dominic P Martin-Hirsch, and Alec S High. Raman spectroscopy and advanced mathematical modelling in the discrimination of human thyroid cell lines. *Head Neck Oncol.*, 1:38, October 2009.

[39] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer, second edition, January 2017.

[40] Leonard Hayflick. The limited in vitro lifetime of human diploid cell strains. *Exp. Cell Res.*, 37:614–636, March 1965.

[41] Simon Haykin. *Neural Networks: A Comprehensive Foundation.* Pearson, second edition, 1999.

[42] G H Henry, A Malewska, D B Joseph, V S Malladi, J Lee, J Torrealba, R J Mauck, J C Gahan, G V Raj, C G Roehrborn, G C Hon, M P MacConmara, J C Reese, R C Hutchinson, C M Vezina, and D W Strand. A cellular anatomy of the normal adult human prostate and prostatic urethra. *Cell Rep.*, 25(12):3530–3542 e5, 2018.

[43] R Herrmann and C Onkelinx. Quantities and units in clinical chemistry: Nebulizer and flame properties in flame emission and absorption spectrometry (recommendations 1986). *Pure Appl. Chem.*, 58(12):1737–1742, January 1986.

[44] J S Horoszewicz, S S Leong, T M Chu, Z L Wajsman, M Friedman, L Papsidero, U Kim, L S Chai, S Kakati, S K Arya, and A A Sandberg. The LNCaP cell line–a new model for studies on human prostatic carcinoma. *Prog. Clin. Biol. Res.*, 37:115–132, 1980.

[45] A Jablonski. Efficiency of Anti-Stokes Fluorescence in Dyes. *Nature*, 131(3319):839–840, June 1933.

[46] Shuiwang Ji and Jieping Ye. Generalized Linear Discriminant Analysis: A Unified Framework and Efficient Model Selection. *IEEE Trans. Neural Netw.*, 19(10):1768–1782, October 2008.

[47] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philos. Trans. A Math. Phys. Eng. Sci.*, 374(2065):20150202, April 2016.

[48] Samuel Kaski and Krista Lagus. Comparing self-organizing maps. In *Artificial Neural Networks — ICANN 96*, pages 809–814. Springer Berlin Heidelberg, 1996.

[49] Rachel E Kast, Stephanie C Tucker, Kevin Killian, Micaela Trexler, Kenneth V Honn, and Gregory W Auner. Emerging technology: applications of Raman spectroscopy for prostate cancer. *Cancer Metastasis Rev.*, 33(2-3):673–693, September 2014.

[50] Gurvinder Kaur and Jannette M Dufour. Cell lines: Valuable tools or useless artifacts. *Spermatogenesis*, 2(1):1–5, January 2012.

[51] Toomas Kirt, Ene Vainik, and Leo Vohandu. A method for comparing self-organizing maps: Case studies of banking and linguistic data. In *Local Proceedings of ADBIS 2007*, volume 325, pages 107–115, January 2007.

[52] Hiroaki Kitano. Systems Biology: A Brief Overview. *Science*, 295(5560):1662–1664, March 2002.

[53] Teuvo Kohonen. *Construction of similarity diagrams for phonemes by a self-organizing algorithm.* Teknillinen korkeakoulu, 1981.

[54] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biol. Cybern.*, 43(1):59–69, January 1982.

[55] Teuvo Kohonen. The Self-Organizing Map. *Proc. IEEE*, 78(9):1464–1480, 1990.

[56] Teuvo Kohonen. *Self-Organizing Maps.* Springer-Verlag, third edition, 2001.

[57] Teuvo Kohonen. Essentials of the self-organizing map. *Neural Netw.*, 37:52–65, January 2013.

[58] Teuvo Kohonen. *MATLAB Implementations and Applications of the Self-Organizing Map*, volume 177. Unigrafia Oy, Helsinki, Finland, 2014.

[59] Holger Krekel et al. pytest. `docs.pytest.org/en/stable/index.html`, 2004 (accessed October 22, 2020).

[60] Anders Krogh. What are artificial neural networks? *Nat. Biotechnol.*, 26(2):195–197, February 2008.

[61] J Stephen Lansing. Complex Adaptive Systems. *Annu. Rev. Anthropol.*, 32(1):183–204, October 2003.

[62] Yongzeng Li, Wei Huang, Jianji Pan, Qing Ye, Shaojun Lin, Shangyuan Feng, Shusen Xie, Haishan Zeng, and Rong Chen. Rapid detection of nasopharyngeal cancer using Raman spectroscopy and multivariate statistical analysis. *Mol Clin Oncol*, 3(2):375–380, March 2015.

[63] Gavin R Lloyd, Kanet Wongravee, Christopher J L Silwood, Martin Grootveld, and Richard G Brereton. Self organising maps for variable selection: Application to human saliva analysed by nuclear magnetic resonance spectroscopy to investigate the effect of an oral healthcare product. *Chemometrics Intellig. Lab. Syst.*, 98(2):149–161, October 2009.

[64] Bartalanffy von Ludwig. *General System Theory: Foundations, Development, Applications*. George Braziller, 1968.

[65] Sayani Majumdar and Mary L Kraft. Exploring the maturation of a monocytic cell line using self-organizing maps of single-cell Raman spectra. *Biointerphases*, 15(4):041010, August 2020.

[66] Thaddeus Mann. *The Biochemistry of Semen*. Methuen, 1954.

[67] Muhammad Irfan Maqsood, Maryam M Matin, Ahmad Reza Bahrami, and Mohammad M Ghasroldasht. Immortality of cell lines: challenges and advantages of establishment. *Cell Biol. Int.*, 37(10):1038–1045, October 2013.

[68] Laura E Masson, Christine M O'Brien, Isaac J Pence, Jennifer L Herington, Jeff Reese, Ton G van Leeuwen, and Anita Mahadevan-Jansen. Dual excitation wavelength system for combined fingerprint and high wavenumber Raman spectroscopy. *Analyst*, 143(24):6049–6060, December 2018.

[69] Michael Mawhinney and Angelo Mariotti. Physiology, pathology and pharmacology of the male reproductive system. *Periodontol. 2000*, 61(1):232–251, February 2013.

[70] James Clerk Maxwell. VIII. a Dynamical Theory of the Electromagnetic Field. *Philosophical Transactions of the Royal Society of London*, 155:459–512, January 1865.

[71] Rudolf Mayer, Robert Neumayer, Doris Baum, and Andreas Rauber. Analytic comparison of self-organising maps. In *Advances in Self-Organizing Maps*, pages 182–190. Springer, 2009.

[72] J E McNeal. The zonal anatomy of the prostate. *Prostate*, 2(1):35–49, 1981.

[73] N Mottet, R C N Van Den Bergh, E Briers, L Bourke, P Cornford, M De Santis, and Others. EAU-ESTRO-ESUR-SIOG guidelines on prostate cancer 2018. Technical report, European Association of Urology, 2018.

[74] Zanyar Movasaghi, Shazza Rehman, and Ihtesham U Rehman. Raman Spectroscopy of Bological Tissues, 2007.

[75] Maria E Mycielska and Mustafa B A Djamgoz. Citrate transport in the human prostate epithelial PNT2-C2 cell line: electrophysiological analyses. *J. Physiol.*, 559(3):821–833, 2004.

[76] Y Pan, S Kytölä, F Farnebo, N Wang, W O Lui, N Nupponen, J Isola, T Visakorpi, U S Bergerheim, and C Larsson. Characterization of chromosomal abnormalities in prostate cancer cell lines by spectral karyotyping. *Cytogenet. Cell Genet.*, 87(3-4):225–232, 1999.

[77] P K Pandalai, M J Pilat, K Yamazaki, H Naik, and K J Pienta. The Effects of Omega-3 and Omega-6 Fatty Acids on in vitro Prostate Cancer Growth. *Anticancer Res.*, 16(2):815–820, March 1996.

[78] Colleen Pelser, Alison M Mondul, Albert R Hollenbeck, and Yikyung Park. Dietary Fat, Fatty Acids, and Risk of Prostate Cancer in the NIH-AARP Diet and Health Study. *Cancer Epidemiol. Biomarkers Prev.*, 22(4):697–707, April 2013.

[79] R Ponmalai and C Kamath. Self-Organizing Maps and Their Applications to Data Analysis. Technical Report LLNL-TR-791165, Lawrence Livermore National Laboratory, September 2019.

[80] Mariana C Potcoava, Gregory L Futia, Jessica Aughenbaugh, Isabel R Schlaepfer, and Emily A Gibson. Raman and coherent anti-Stokes Raman scattering microscopy studies of changes in lipid content and composition in hormone-treated breast and prostate cancer cells. *J. Biomed. Opt.*, 19(11):111605, 2014.

[81] C V Raman and K S Krishnan. A new class of spectra due to secondary radiation. part I. *Indian J Phys*, 1928.

[82] Andreas Rauber and Dieter Merkl. Automatic Labeling of Self-Organizing Maps: Making a Treasure-Map Reveal Its Secrets. In *Methodologies for Knowledge Discovery and Data Mining*, pages 228–237. Springer, 1999.

[83] Christine Rauh-Adelmann, Kin-Mang Lau, Nari Sabeti, John P Long, Samuel C Mok, and Shuk-Mei Ho. Altered Expression of BRCA1, BRCA2,

and a Newly Identified BRCA2 Exon 12 Deletion Variant in Malignant Human Ovarian, Prostate, and Breast Cancer Cell Lines. *Mol. Carcinog.*, 28(4):236–246, 2000.

[84] Prashanth Rawla. Epidemiology of prostate cancer. *World J. Oncol.*, 10(2):63–89, April 2019.

[85] Maciej Roman, Tomasz P Wrobel, Agnieszka Panek, Czeslawa Paluszkiewicz, and Wojciech M Kwiatek. Lipid droplets in prostate cancer cells and effect of irradiation studied by Raman microspectroscopy. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids*, 1865(9):158753, September 2020.

[86] K A Schafer. The Cell Cycle: A Review. *Vet. Pathol.*, 35(6):461–478, November 1998.

[87] Warren I Schaffer. Terminology associated with cell, tissue and organ culture, molecular biology and molecular genetics. *In Vitro Cell. Dev. Biol.*, 26:97–101, 1990.

[88] Punit Shah, Xiangchun Wang, Weiming Yang, Shadi Toghi Eshghi, Shisheng Sun, Naseruddin Hoti, Lijun Chen, Shuang Yang, Jered Pasay, Abby Rubin, and Hui Zhang. Integrated Proteomic and Glycoproteomic Analyses of Prostate Cancer Cells Reveal Glycoprotein Alteration in Protein Abundance and Glycosylation. *Mol. Cell. Proteomics*, 14(10):2753–2763, October 2015.

[89] Dustin W Shipp, Faris Sinjab, and Ioan Notingher. Raman spectroscopy: techniques and applications in the life sciences. *Adv. Opt. Photon., AOP*, 9(2):315–428, June 2017.

[90] Wiley Science Solutions. Knowitall spectroscopy edition software. `sciencesolutions.wiley.com/knowitall-spectroscopy-software/`, (accessed May 16, 2021).

[91] Lien Spans, Zeynep Kalender Atak, Filip Van Nieuwerburgh, Dieter Deforce, Evelyne Lerut, Stein Aerts, and Frank Claessens. Variations in the Exome of the LNCaP Prostate Cancer Cell Line. *The Prostate*, 72(12):1317–1327, September 2012.

[92] H B Sun, J Shen, and H Yokota. Size-Dependent Positioning of Human Chromosomes in Interphase Nuclei. *Biophys. J.*, 79(1):184–190, July 2000.

[93] Abdullah Chandra Sekhar Talari, Zanyar Movasaghi, Shazza Rehman, and Ihtesham ur Rehman. Raman Spectroscopy of Biological Tissues. *Appl. Spectrosc. Rev.*, 50(1):46–111, January 2015.

[94] Hiong Sen Tan. HLabelSOM: Automatic Labelling of Self Organising Maps toward Hierarchical Visualisation for Information Retrieval. In *AI 2003: Advances in Artificial Intelligence*, pages 532–543. Springer, 2003.

[95] Jing Tian, Michael H Azarian, and Michael Pecht. Anomaly Detection Using Self-Organizing Maps-Based K-Nearest Neighbor Algorithm. In *Proceedings of the European Conference of the Prognostics and Health Management Society*, pages 1–9, 2014.

[96] S Tsuchiya, M Yamabe, Y Yamaguchi, Y Kobayashi, T Konno, and K Tada. Establishment and characterization of a human acute monocytic leukemia cell line (THP-1). *Int. J. Cancer*, 26(2):171–176, August 1980.

[97] Cancer Research UK. What is prostate cancer? `cancerresearchuk.org/about-cancer/prostate-cancer/about`, (accessed February 17, 2021).

[98] Prostate Cancer UK. About prostate cancer. `prostatecanceruk.org/prostate-information/about-prostate-cancer`, (accessed February 17, 2021).

[99] Adrie van Bokhoven, Marileila Varella-Garcia, Christopher Korch, Widya U Johannes, E Erin Smith, Heidi L Miller, Steven K Nordeen, Gary J Miller, and M Scott Lucia. Molecular Characterization of Human Prostate Carcinoma Cell Lines. *The Prostate*, 57(3):205–225, 2003.

[100] J Vesanto and E Alhoniemi. Clustering of the Self-Organizing Map. *IEEE Trans. Neural Netw.*, 11(3):586–600, 2000.

[101] Giuseppe Vettigli. Minisom. `github.com/JustGlowing/minisom`, 2013 (accessed March 10, 2020).

[102] O Warburg. On the Origin of Cancer Cells. *Science*, 123(3191):309–314, February 1956.

[103] J D Watson and F H Crick. Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, April 1953.

[104] Li Yuan. Implementation of Self-Organizing maps with python. Master's thesis, University of Rhode Island, 2018.

[105] Hongjuan Zhao, Young Kim, Pei Wang, Jacques Lapointe, Rob Tibshirani, Jonathan R Pollack, and James D Brooks. Genome-Wide Characterization of Gene Expression Variations and DNA Copy Number Changes in Prostate Cancer Cell Lines. *The Prostate*, 63(2):187–197, May 2005.