Global and Local Persistent Homology for the Shape and Classification of Biological Data

Bernadette J. Stolz-Pretzer

Lincoln College

University of Oxford

Thesis submitted for the degree of

Doctor of Philosophy

Trinity 2019

To my family,

who gave me

love,

courage,

strength,

and persistence.

Acknowledgements

This thesis would not exist without a long list of people, first and foremost my academic supervisors Heather Harrington and Jared Tanner. Thank you for your guidance, continuous support, feedback, patience, and kindness from initial explorations of project directions to the long years spent extracting or generating data, coding, debugging, facing what felt like every possible computational issue, to the thesis slowly taking its final form. Thank you for teaching me the importance of communicating my research and pushing me to seize opportunities to attend and speak at many national and international conferences and workshops from my first DPhil weeks onwards.

In my different subprojects I have been very fortunate to benefit from the vast expertise of Mason Porter (Chapter 4), Helen Byrne (Chapter 3), and Vidit Nanda (Chapter 5), whose input critically shaped my research. Mason, thank you for teaching me about networks, how to write a research paper, and the importance of paying attention to clarity and punctuation. Many thanks in particular for your support during the many stages of the schizophrenia project and also for initiating Brexit-related side projects. Helen, thank you for your invaluable support in understanding the importance of spatial characterisation of tumour vasculature, for your help in getting access to data, for the many opportunities you gave me to interact with researchers from the biological side, and for your encouragement and great enthusiasm about my results even in their very early stages.

Vidit, thank you for your patience while you were teaching me the Mayer-Vietoris sequence and helping me double check my coding. Thank you for letting me embark on the search for geometric anomalies and keeping me company while staring at plots of intersecting planes that were coloured in all sorts of ways, just not the one we were looking for.

I am also very thankful to my industrial supervisors, Florian Lipsmeier and Franziska Mech (Roche), who gave me the opportunity to view my research from a completely different perspective. Thank you for the many helpful discussions, interesting research visits, and the option to experience pharmaceutical research from the inside. Franziska, thank you in particular for your many hours spent getting me the ultra-microscopy data and relevant information about it. Florian, thank you also for all your help in organising and supervising my industrial internship.

In all parts of this thesis I rely on either experimental or synthetic data. I would therefore like to thank Ruth Muschel and in particular Jakob Kaeppler and Bostjan Markelc for generating and giving me access to the intravital microscopy data set (Chapter 3) as well as for many helpful discussions. The fMRI motor-learning data (Chapter 4) was collected originally by Nicholas F. Wymbs and Scott T. Grafton, who I would like to thank for access to the data. Further thanks goes to Danielle S. Bassett for help in providing the data and helpful discussions. I would also like to thank Alessandro Bertolino, Fabio Sambataro, and the Bari psychiatric neuroscience group for permission to study their schizophrenia fMRI data set (Chapter 4). I am further grateful to Shawn Martin for kindly providing the cyclo-octane and Henneberg data sets (Chapter 5) for the study of geometric anomalies and answering my questions about this data.

I am very grateful to the many people who answered my software-related queries, in particular Russell Bates, James Grogan, and Almut Koepke for their input on running UNET-CORE (Chapter 3); Henry Adams and Mikael Vejdemo-Johansson for their help with JAVAPLEX (Chapter 3 and Chapter 4); Pawel Dłotko for his patient support with the Persistence Landscapes toolbox (Chapter 4), for providing new versions of the software during my work, and for helpful discussions; and Uli Bauer for constructive comments on using RIPSER (Chapter 3 and Chapter 5). I would like to acknowledge the use of Advanced Research Computing (ARC, University of Oxford) for parts of this work (Chapter 4) as well as the wonderful IT team at the Mathematical Institute who helped me with many computational hurdles on my path of applying persistent homology to data.

I would further like to thank Tegan Emerson for an interesting and fruitful collaboration on analysing the output of persistent homology (Chapter 4). Many thanks go to Peter Bubenik, Carina Curto, Parker Edwards, Peter Grindrod, Florian Klimm, Philip Maini, Satu Nahkuri, Nicola Richmond, and Ulrike Tillmann for helpful discussions and comments on my work. I am also incredibly grateful to Florian Klimm, Almut Koepke, and Hannah Kunde for proofreading my thesis.

I am extremely grateful to have been able to spend my DPhil in the stimulating and very friendly work environment in the Mathematical Institute. In particular, I would like to thank Barbara Mahler and Nina Otter for accompanying me through the ups and downs of persistent homology including mathematical discussions at the Jericho Café and the memorable search for inflatable tori and space balls to explain persistent homology to non-experts. I would further like to thank the WCMB for giving me

an academic family, as well as the wonderful admin staff at the DTC and the Mathematical Institute, in particular Sandy Patel, for their support.

I am very thankful to Lincoln College for giving me a wonderful home during my past years in Oxford and in particular also to my College advisor Dominic Vella for his support and for giving me the opportunity to teach tutorials. Katie Allan, Leandra Bias, Joshua Bull, Annina Grädel, Florian Klimm & Sarah Griffin, Almut Koepke, Tammo Rukat, and Marcel Stolz, thank you for the much needed coffee breaks, walks, rowing, brunches, and pizza evenings. Colette Bichsel, Corinne Huck & Christophe Bornand, Kathrin Jakob, Igor Jurosevic, Alex & Jane Schindler, thank you for cheering me on from my initial application through to submission.

I would like to gratefully acknowledge the Engineering and Physical Sciences Research Council and the Medical Research Council (EP/G037280/1) and F. Hoffmann-La Roche AG for funding my DPhil. I am also grateful to the Swiss Study Foundation (Schweizerische Studienstiftung) for academic inspiration from 2008 – 2018.

Most importantly, I would like to thank my family. Thank you to my parents, Zuzana and Michael Stolz, for their unconditional love and support, for shaping me into who I am, and for teaching me to recognise and seize opportunities when they come my way. I will also be forever grateful to my late grandparents, Jarmila and Jaromír Hladký, who sparked and encouraged my earliest interests in mathematics and biology. Thank you further to my brothers- and parents-in-law for their invaluable support, in particular with childcare and the logistics of moving between several countries. Finally, thank you to Christoph and Jeremias Pretzer for your love and patience, and for showing me life in all its beauty.

Oxford, September 2019.

Abstract

Persistent homology (PH) is an algorithmic method that allows one to study shape and higher-order interactions in high-dimensional data. Over the last decade, PH has been used in a wide variety of applications, including biology. PH considers topological invariants, such as connected components, loops, and holes, and their changes across a filtration which one can imagine as observing the data through multiple scales or resolutions. The filtration determines the questions that can be answered about the data and, in many cases, needs to be developed specifically for the problem of interest. There are also many other practical challenges when applying PH such as computational complexity and interpretation of PH output. This thesis has two parts: In the first part, we showcase how PH can be applied to two types of biological data: tumour blood vessel networks and functional neuronal networks. For tumour blood vessel networks, we develop a novel filtration that spatially characterises their structural abnormality. We show that the number of vessel loops and their distribution in the networks change over time when tumours undergo treatment with vascular targeting agents and radiation therapy. In functional neuronal networks, we find that PH can provide insight into dynamical processes in motor-learning data as well as in working-memory data from healthy versus schizophrenic human subjects. We highlight what type of information we can gain by applying persistence landscapes and persistence images to analyse and interpret the output from PH. In the second part of this thesis, we develop novel methods that consider PH locally around data points. To address computational issues when applying PH to large and noisy data sets – both traits are commonly found in biological data – we develop a novel landmark selection technique for point clouds. In contrast to existing methods, our subsampling process is robust to outliers and is developed specifically for PH. We further introduce a novel method that can detect geometric anomalies, such as intersections or boundaries, in point cloud data sampled from intersecting surfaces. Our detection is based on the computation of PH in local annular neighbourhoods around points and is less sensitive to the size of the local neighbourhood and surface curvature than an existing method.

Contents

List of Figures

List of Algorithms

"des himmels stern und meres griß
und alls das rechenunge hiß,
das floß durch mines herzen rink:
nach mir nature buwet dink."[a]

From the speech by the personified art of arithmetic (Arismetica) in *Der meide kranz*, Heinrich von Mügeln, 14th century. Edition by [263].

[a]Translation: "The stars of heavens and the sands of the sea and everything that was said to be connected with calculation, all that flowed through the ring of my heart. Natura builds things in accordance with me." [263].

1 Introduction

As applied mathematicians, we consider phenomena in the world around us as being countable, measurable, and classifiable. This view is not a modern phenomenon, see, for example, the quote above by Heinrich von Mügeln, a German poet from the 14th century who wrote about the significance of the seven liberal arts. Over the last century, however, instruments for measuring have become more precise and more easily available, leading to an unprecedented volume of data that is produced every day. Just over 10 years ago, Microsoft Research identified the lack of methods for understanding the available quantities of data in all their complexity as one of the big challenges of the 21st century and established data science as the fourth paradigm of science [134]. In particular in biology, there is great potential for novel insights and new hypotheses based on data [134,171]. For example, with novel imaging techniques it has become possible for cancer researchers to visualise tumour blood vessels and the effects that cancer drugs have on them in fascinating detail over multiple days. These blood vessels form through a process known as *tumour-induced angiogenesis*, which is very different to the growth of vessels in a healthy organ. The result of tumour-

induced angiogenesis is a chaotic network of inefficient blood vessels characterised by many loops and twists, which is highly abnormal compared to the vessel network of a healthy organ. Even though these abnormal characteristics are obvious to the human eye, quantifying them has, so far, proved to be very difficult. To take full advantage of this complex and cutting edge data, it is crucial to use analysis methods that can quantify shape. Since a tumour depends on its vascular system to survive, grow, and spread in the body, its vascular network is an excellent target for cancer drugs. Understanding this spatial and high-resolution data could thus have important implications for treatment.

In many areas of biology, one can also gain new knowledge by applying novel analysis techniques to existing data sets [171]. One such area is neuroscience, where large quantities of data already exist. The human brain is a highly complex organ, whose organisational structure from the scale of single neurons to the scale of brain regions can be studied using networks. For example, one can consider different neurons or brain regions as nodes and physical connections between them as edges and study the resulting structural network [183]. If one is more interested in functional aspects of the brain, one can construe different neurons or brain regions as nodes and give edges between them a weight that represents a measure of functional similarity over time. Such a construction is called a functional network [183]. Scientists have been observing structural and functional aspects of the human brain by creating (possibly time-dependent) neuronal networks from data for over 10 years [22, 27, 42, 59, 61, 193, 194, 234]. Most commonly used methods to study networks, however, consider only pairwise connections [183]. Only recently, techniques that can capture higher-order interactions and highlight new patterns or shapes in networks have been applied in neuroscience [24, 124, 237] .

To understand the shape of data, such as a vascular network or a functional neuronal network, mathematicians have in the past 10 –15 years started turning to

theoretically well-established areas such as topology. Ideas from topology have since been successfully applied to many biological settings, for example, to find a new subtype of breast cancer [184], to detect interactions between spinal chord injury and traumatic brain injury [186], to understand synaptic connectivity [210], and to unravel the neural codes [80]. A particularly popular and successful topological method is *persistent homology*, which we now discuss in more detail.

1.1 Persistent homology

Persistent homology (PH) [66, 97–99] is a data analysis method based on the topological concept of homology that forms a bridge between pure and applied mathematics through computation. With PH one can study topological invariants (e.g. characteristics of shapes) in high-dimensional data over multiple scales. Examples of topological invariants in different dimensions are connected components (dimension 0), loops (dimension 1), or holes (dimensions ≥ 2). When applying PH, one must choose *a filtration* which associates sequences of vector spaces and maps to the data and is the input for PH. One can imagine the vector spaces in the sequence as representing the data at different scales or resolutions. The result of applying PH to a filtration is a collection of intervals, which can be visualised by so-called *barcodes* [123] (see middle panel of Fig. 1.1 for an example). For every dimension, the intervals in the barcode represent topological features in the data and their persistence across the sequence of vector spaces. The barcodes can be interpreted by a choice of methods, which allow them to be analysed statistically. We show the typical pipeline for applying PH to data in Fig. 1.1. The theoretical basis of PH is well understood [76, 123, 279] and there are many available software packages to compute PH, such as [32, 173, 247], and to analyse the output of PH computations, see, for example, [3, 89]. Together with accessible introductions such as [190, 191, 232, 268] this has led to a noticeably increased interest in PH over the last 5 years. However, the application of PH to data

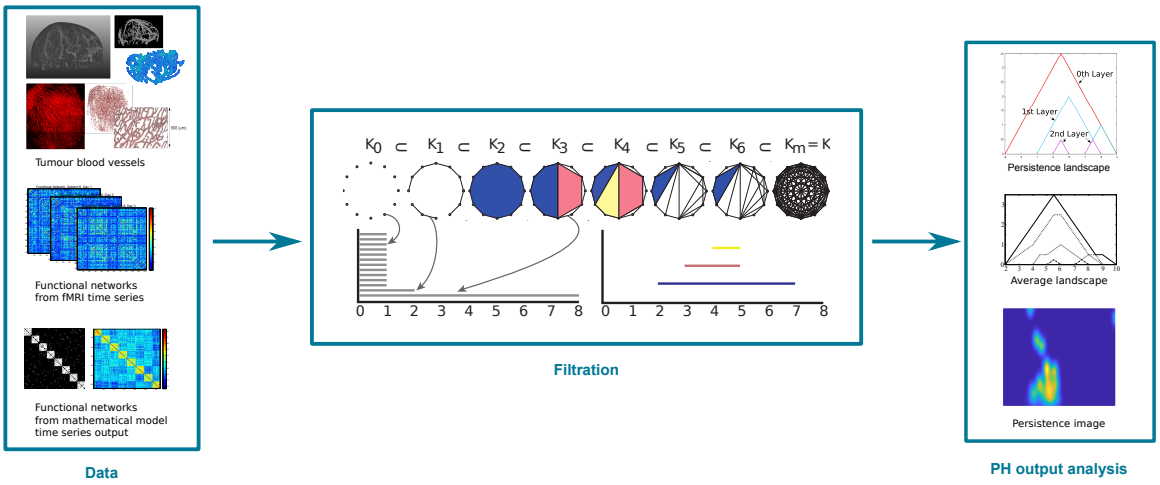


Figure 1.1: Schematic representation of the PH pipeline. Different types of data can be analysed by filtrations which assign barcodes to data. These barcodes are then analysed and interpreted to obtain information about topological features such as connectedness, or loops in the data. The figure contains modified versions of our images in [237, 238, 240] as well as images of experimental data provided by Russel Bates and Bostjan Markelc/Jakob Kaeppler (with permission).

still remains extremely challenging. We now discuss some of the difficulties that can arise in applications.

1.2 Practical challenges of persistent homology

PH works very well on synthetic examples such as distinguishing whether points were sampled from a torus or a sphere (see, for example, in [4]). However, PH comes with a unique set of challenges when applied to real-world data. Broadly speaking, the challenges fall into two categories: the output of PH on a given data set must be both meaningful and computable.

1.2.1 Meaningful filtrations

PH is often explained with point cloud data in mind, for which the notion of shape is somewhat intuitive and established filtrations exist such as a filtration via the Vietoris–Rips complex or the Čech complex (see, for example, [66, 123] for a description). Many software packages are designed to compute these very specific filtrations such as RIPSER [32], which computes the Vietoris–Rips filtration of a point cloud.

Applications of PH to other types of data such as networks, however, require one to either identify or develop an appropriate filtration for the data. Luckily, the software package javaPlex [247] allows the user to build their own filtration. For (biological) networks, several filtrations have been developed successfully [33, 34, 145, 202]. However, different types of networks are studied with very different types of research questions in mind, and thus finding a filtration that can provide insights into the data at hand is not straightforward. Before developing a filtration, one must consider whether topological invariants such as loops are an intrinsic feature of the data and in what way they are meaningful [232]. For example, tumour blood vessel networks contain loops as one of their characteristic features. In contrast, all-to-all connected weighted functional neuronal networks do not contain obvious loops on first inspection, but if one imagines thresholding the networks at different weights, loops can emerge at different thresholds. However, even if we identify the topological features such as loops that we want to capture, we can still choose different types of filtering parameters, which will determine what information we can gain about these loops. In the case of the tumour blood vessel data, we can filter the vessel networks by parameters such as thickness or tortuosity of the branches. The result of such a filtration would be a barcode that can be translated to a histogram of the number of loops versus the maximal tortuosity value or thickness value of a vessel segment in the loop. Alternatively, one can create a filtration based on a spatial parameter, which provides insight on the spatial distribution of loops in the network. Since the choice of filtration fully determines the type of information that can be gained from the data, it is crucial to closely consider the research question at hand. In addition, when working with collaborators from other disciplines, the output of a filtration needs to have a clear interpretation, i.e. barcode intervals or their persistence need to be connected with a physical phenomenon or a characteristic of the data.

1.2.2 Persistent homology on large data sets

After identifying or designing an appropriate filtration for the problem one needs to compute PH. Building a filtration on a data set with N points results in spaces of the size $\mathcal{O}(2^N)$, although in practice this can be reduced to $\mathcal{O}(N^{\hat{n}+1})$ by posing a limit $\hat{n}$ on the dimension of the topological features considered [190]. Following the construction of a filtration, the algorithm for computing PH in the worst case has a complexity of $\mathcal{O}(\mathtt{k}^3)$, where $\mathtt{k}$ is the number of points, edges, triangles, and higher-dimensional connections constructed from the data points by the filtration (the number of such connections can be up to $2^N - 1$), although the complexity is often linear in practice [99, 190]. Methods exist to approximate specific filtrations or to reduce the sizes of the vector spaces associated to the data by a filtration that have been implemented in software packages, see [190] for an overview. Despite such improvements, computation is still very challenging on large data sets and can pose a hard limit on the filtrations that can be applied to a particular data set. Even on Oxford's most powerful computers many of the calculations that we present here run on the order of months. In such cases it can become necessary to preprocess data before applying PH (see Fig. 1.2). For point cloud data one can, for example, use subsampling techniques to identify so-called *landmarks* of the data set and then define a filtration on the landmarks [86]. A filtration on well chosen landmarks retains topologically important global information about the full data set and one can even choose to include information from non-landmark points when constructing the filtration. An example for such a filtration is the *lazy witness filtration* which was first introduced by de Silva & Carlsson [86] and has been used to study noisy artificial data sets by Kovacev–Nikolic [151], primary visual cortex cell populations by Singh *et al.* [228], and cancer gene expression data by Lockwood & Krishnamoorthy [165]. Roughly, the lazy witness filtration consists of the following steps[1]:

[1]For the full definition see [86].

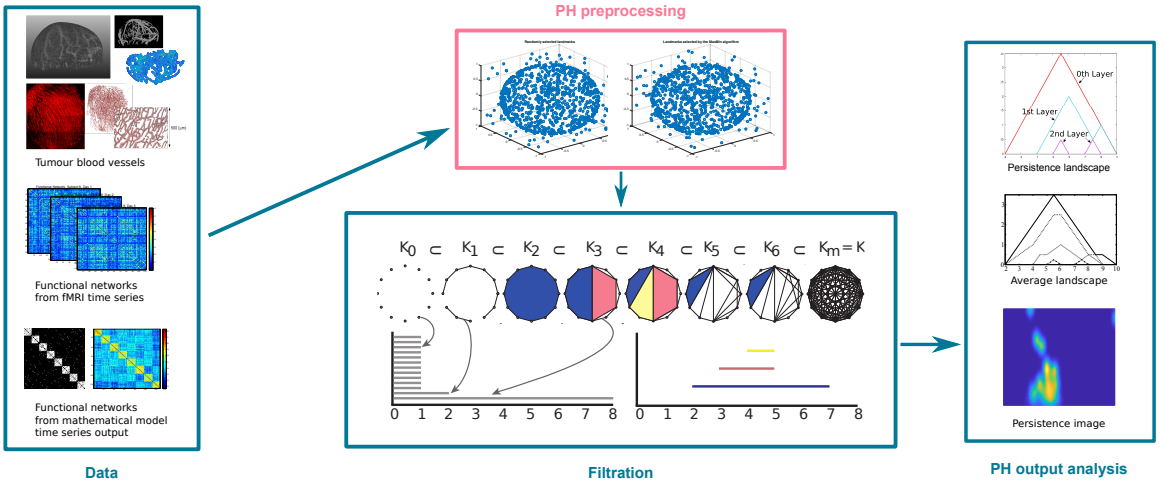


Figure 1.2: Schematic representation of the PH pipeline including data preprocessing. For large data sets a reduction of the number of data points can be necessary to obtain results within reasonable timeframes. The figure contains modified versions of our images in [237, 238, 240] as well as images of experimental data provided by Russel Bates and Bostjan Markelc/Jakob Kaeppler (with permission).

1. Selection of a (small) subset[2] of *landmark points* L from the data set D.

2. Construction of a *lazy witness filtration* on the landmarks L where the landmarks are vertices and data points from the full data set D can serve as *witnesses* for higher order interactions between sets of landmarks in the filtration.

Even though the choice of landmarks from the data inevitably has a large influence on the results that can be obtained, currently there are only two standard approaches to select landmarks: uniform random selection and selection via the so-called *maxmin* algorithm. We give a full description of both procedures in Chapter 5. Neither is ideal and in particular the maxmin algorithm tends to include outliers [4, 86], which is a problem since large real-world data sets often include noise. In the biological application of the lazy witness filtration in [165], for example, the maxmin algorithm leads to the discovery of loops in the data set which we could not reproduce using uniform random landmark selection or when discarding a small proportion of the initially chosen maxmin landmarks from the data set and choosing a new set of

[2]Although there is no systematic lower bound for the number of landmark points, the authors of [86] suggest using $> 5\%$ of the data points as landmarks.

landmarks with the maxmin algorithm. Moreover, in that particular case we found that the authors' log_2 scaling of the data seems to cause the appearance of a persistent loop in the data. We also note that the results of the lazy witness filtration are difficult to interpret. In addition to the mentioned disadvantages of the *maxmin* algorithm, neither of the proposed landmark selection methods were designed specifically for PH. While there are other methods that address subsampling for PH such as [76, 96, 187] these do not explicitly consider noisy data.

Because existing approaches are not ideal, it is desirable to develop new methods that lead to a reduction of large and noisy data while retaining interesting topological features. Ideally the reduced data set can be used as input for PH directly without additional preprocessing steps.

1.2.3 Barcode interpretation and analysis

The output from the computation of PH for every dimension of topological features consideres is a collection of intervals that correspond to topological features of that dimension and their persistence in the filtration. One possible visualisation of the PH output intervals is a barcode or, for data sets consisting, for example, of multiple networks, a collection of barcodes, i.e. one barcode for each network. In order to obtain information from barcodes, one needs to identify which barcode features are meaningful either by visual inspection, which is difficult for real-world data, or by applying a method that can create a summary of the barcodes.

1.2.3.1 Interpretation of bar length

A common interpretation of PH output is that short bars in a barcode, which represent topological features that are present across a small number of vector spaces associated to the data, represent noise [99, 123]. The reason for this interpretation is that barcodes are stable with respect to small perturbation of the input data [76]. In particular, such a perturbation could cause short bars to disappear completely

as a perturbation of data points by ϵ_p with respect to the original point cloud has the potential to remove bars in the barcode that are of length smaller than ϵ_p (in radius-based barcodes).

In some applications this interpretation of short bars as noise is very reasonable and has indeed been crucial to their success, see, for example, Gaimero *et al.* where all bars below a set threshold length had to be removed in order to get a good fit between predicted and experimental protein compressibility [118]. However, the stability of the barcode with respect to data perturbation was proven for points that are sampled from hypersurfaces in $\mathbb{R}^n$ [76]. In the context of (possibly weighted) networks, i.e. non-Euclidean space, it is much less clear what a small perturbation of the data is and whether it depends on the filtration how such a perturbation manifests itself in the barcode. The relationship between low persistence of a feature and it representing noise in the data rather than signal in such a context has not yet been verified statistically. Indeed, as Carlsson [66] points out, what is labeled as noise and what as signal can be dependent on the problem at hand.

We first observed that short bars appearing early in the filtration of functional neuronal networks created from functional magnetic resonance imaging (fMRI) data of a motor-learning experiment seemed to be a feature of the data in [237]. Meanwhile there have been other applications where short [56, 145] and medium-sized [34, 145] intervals in barcodes represented important features of the data.

The unknown significance of bar length for a specific data set adds additional complexity to the interpretation of PH on real-world data.

1.2.3.2 Analysis of collections of barcodes

There are several methods that allow the analysis of a collection of barcodes. The most widely used ones are persistence landscapes [89] and persistence images [3]. Both methods require a series of decisions, for example, whether to include or exclude topological features that persist across the entire filtration (and thus dominate

barcode comparison) in persistence landscapes, or whether to apply particular weights to longer bar lengths in persistence images. These decisions require intuition as to which aspects of the barcode are relevant for the interpretation of the data which is challenging, in particular for large collections of barcodes. Some aspects of performance of persistence images have been compared to persistence landscapes in [3] and later in [278] where both methods were compared together with alternative methods to a novel vectorised representation of PH output. However, in both cases the comparison was performed on synthetic data. No exemplary comparison of using the two methods next to each other has been conducted on real-world data sets.

1.3 Contributions

In this thesis, we showcase and improve the applicability of PH to biological data. The thesis has two distinct parts with different aims. In the first part we aim to answer biological questions using PH. We are in particular interested in studying the following two problems:

1. Quantification of abnormality of tumour blood vessels.

2. Understanding aspects of functional neuronal networks both in healthy human subjects and in subjects with schizophrenia.

Using PH for the study of tumour blood vessels and schizophrenia data is a novel approach. For tumour blood vessel data, there are to date no existing approaches to quantify abnormality that reveal detailed spatial information or go beyond summary statistics. Our work to generate novel quantification techniques that provide useful insights from a biological perspective is in close collaboration with researchers at the *Oxford Radiation Oncology Institute* in the *Department of Oncology*.

Due to the challenges of PH that we outlined in Section 1.2, innovation was necessary in all steps of the PH pipeline from developing adequate filtrations to the

analysis and interpretation of results. In addition, we were confronted with data formats that required non-trivial adaptations to make them suitable for the study with PH.

In both biological applications, we were at the edge of computational feasibility. This motivated the second part of the thesis, in which we contribute two novel methods that are built on computing PH locally around data points in point clouds. In the first method, we compute PH in a small neighbourhood around data points and use the output of this computation to rank the data points with respect to their suitability as landmarks for the computation of PH. We explicitly develop the method for the computation of PH on noisy data sets and also adapt an outlier-robust version of the k-means algorithm for the selection of landmarks for comparison. In our second method, we modify the shape of the neighbourhood that we consider for local PH such that we can use the number of persistent features in the local barcodes of points to detect whether they are close to geometric anomalies such as intersections of boundaries in data sampled from intersecting surfaces.

As the topics addressed in the present thesis are very diverse, each of the results chapters was written such that it can be understood independently after reading the introduction to the methodology in Chapter 2, which is relevant to all chapters.
The remainder of this thesis is structured as follows:

- **Chapter 2: Methodological Background.** We introduce the topological concepts behind PH in a way that is accessible and oriented towards applications. We include short descriptions and pointers to the mathematical framework that we consider important for the interested data scientist and provide our own examples to develop intuition. We show existing methods for the analysis of point cloud and network data with PH and describe which techniques one can use to analyse the output of PH computations.

- **Chapter 3: Persistent Homology Applied to Tumour Blood Vessel Networks**. We provide a short description of why blood vessels of tumours are of biological relevance and why their quantification could lead to many interesting biological insights. We develop a novel filtration that can spatially characterise two defining features of tumour blood vessels: tortuosity (in dimension 0) and the presence of loops (in dimension 1). We apply our filtration to two data sets from different imaging modalities and different types of tumours, whose blood vessel networks have distinct growth behaviour. The first data set allows observation of partial vessel networks over multiple time points of tumour growth, the second data set contains only one time point per tumour but shows the full vessel network in fascinating spatial detail. We illustrate how the number of loops in a tumour blood vessel network changes spatially under treatment with radiotherapy and/or vascular targeting agents. For the first data set, we also investigate the number of short intervals in dimension 0 barcodes and find that for the different treatment groups this changes in a similar way to the number of loops. The description of the filtration that we develop is currently in press [64].

- **Chapter 4: Persistent Homology Applied to Functional Neuronal Networks.** We illustrate what insights can be gained from applying the weight rank clique filtration, a filtration for weighted networks, to study loops in functional neuronal networks created from two different functional magnetic resonance imaging (fMRI) data sets. In the first part of Chapter 4, we study motor-learning data. Our work on this data set builds on our observation in [237] that the barcodes for these functional networks contain short bars at the beginning of the filtration which correspond to loops between highly synchronised brain regions. Using persistence landscapes on the PH output that we obtained in [237] from both the data set and a mathematical model, we here provide

evidence that the short bars really are features of these networks that reflect underlying community structure. Construing short bars as features rather than noise is contrary to what common wisdom in PH would suggest and was later also suggested for other applications by [56, 145]. Our findings suggest that in general, when using PH to analyse a data set it is useful to consider the full PH output for interpretation rather than just persistent features. Our results have been published in [240], which we reproduce in the first part of Chapter 4. In the second part of Chapter 4, we investigate fMRI data from schizophrenia patients, healthy siblings of schizophrenia patients, and healthy controls. To analyse the PH output from this data set, we again use persistence landscapes. In addition, we initiated a collaboration with Tegan Emerson (at the time a graduate student at Colorado State University) who provided an analysis using persistence images which we summarise. We provide a comparison of the different types of insights that the two techniques provide on this data set. The results are available as a preprint in [238], on which this part of Chapter 4 was built, and have been submitted for publication. To our knowledge, we are the first authors to analyse a real-world data set using both persistence landscapes and persistence images and to provide a detailed comparison of our insights.

- **Chapter 5: Applications of Local Persistent Homology.** In the first part of Chapter 5, we address the problem of computational complexity of PH by proposing two novel approaches for the selection of landmarks from large and noisy data sets. Note that we explicitly consider noisy data in contrast to other methods that address subsampling such as [76,96,187]. For the first approach we adapt an outlier-robust version of the k-means algorithm for landmark selection. We develop a second landmark selection method specifically for PH and base it on the computation of local PH in a small neighbourhood around each point in the data set. We apply both landmark selection techniques to simple artificially

created data sets to test their suitability for the task. Both methods outperform the existing standard landmark selection techniques with respect to preserving the topological properties of the data set. In the remaining part of Chapter 5, we introduce a novel method that also builds on the computation of local PH and detects geometric anomalies in data sampled from intersecting surfaces. We show that the method has practical advantages over an existing method that can find intersections. The results from our geometric anomaly detection method gave rise to a preprint [242] which connects data science, topology, and geometry.

- **Chapter 6: Discussion and Outlook.** We discuss our observations from this thesis in the wider context of PH and present possible future directions.

- **Chapter 7: Conclusions.** We summarise our main findings from this thesis.

In summary, the key contributions of this thesis are the following:

- We lay the groundwork to spatially characterise the unique features of tumour blood vessels quantitatively using PH and show that the descriptors that we find are meaningful for this type of data. Our work enables further investigations of the data using tools from statistics and machine learning.

- We provide evidence that short bars in a PH barcode do not necessarily represent noise, an observation that we first made in [237].

- We observe what type of information one can obtain from functional neuronal networks using PH for the detection of loops in motor-learning data and schizophrenia data. We compare persistence landscapes and persistence images and highlight the different insights that we make on the schizophrenia data set.

- We present a novel method to select landmarks for PH computations on large and noisy data sets based on on local PH.

- We develop a technique that can detect geometric anomalies in data using local PH.

1.4 Publications and preprints arising from this thesis

The following publications, preprints, and blog posts arose directly from the work on this thesis:

1. Helen M. Byrne, Heather A. Harrington, Ruth Muschel, Gesine Reinert, Bernadette J. Stolz, and Ulrike Tillmann. Topology characterises tumour vasculature. *Mathematics Today*, 55(5):206 – 210, 2019 (in press)[3]. Based on filtration presented in Chapter 3.

2. Bernadette J. Stolz, Heather A. Harrington, and Mason A. Porter. Persistent homology of time-dependent functional networks constructed from coupled time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(4):047410, 2017. Based on Chapter 4, Section 4.3[4].

3. Bernadette J. Stolz, Jared Tanner, Heather A. Harrington, and Vidit Nanda. Geometric anomaly detection in data. arXiv: 1908.09397 – to be submitted. Based on Chapter 5, Section 5.3.

4. Bernadette J. Stolz, Tegan Emerson, Satu Nahkuri, Mason A. Porter, and Heather A. Harrington. Topological data analysis of task-based fMRI data from experiments on schizophrenia. arXiv:1809.08504, 2018 – to be resubmitted. Based on Chapter 4, Section 4.4.

[3]Author list arranged in alphabetical order.

[4]Note that the initial work for this publication was carried out in Stolz 2014 [237] and was then extended significantly to include analysis using persistence landscapes.

5. Bernadette J. Stolz, Heather A. Harrington, and Mason A. Porter. The topological "shape" of Brexit. arXiv:1610.00752, 2016 – to be resubmitted.

6. Bernadette J. Stolz and Barbara I. Mahler. H is for homology. `https://www.maths.ox.ac.uk/about-us/life-oxford-mathematics/oxford-mathematics-alphabet/h-homology`, 2016. Inspired by Chapter 2.

The following manuscripts are planned or currently in preparation:

7. Persistent homology for the study of vascular networks in tumours. Results paper based on Chapter 3.

8. Outlier-robust subsampling techniques for persistent homology. Based on Chapter 5, Section 5.2.

"ein iglich kunst die treit min kleit,
als ichs ir nach genaden sneit"[a]

From the speech by the personified art of arithmetic (Arismetica) in *Der meide kranz*, Heinrich von Mügeln, 14th century. Edition by [263].

[a]Translation: "Every art wears my cloth, as I cut it for her according to my mercy" [263].

2 Methodological Background

Persistent homology (PH) [66, 97–99] is a method from computational topology that quantifies global topological structures (e.g., connectedness and loops) in high-dimensional data. One can think of PH as looking at the 'shape' of data in a given dimension using a set of different lenses. Each lens conveys topological features inside the data at a different resolution, disregarding any changes made to the shape by stretching or bending. One then interprets structures that persist over a range of different lenses to represent a significant feature of the data. Structures that are observed only through a small number of lenses are commonly construed as noise [66, 123], especially in settings where the data are sampled from a manifold, although we will see in Chapters 3 and 4 that short-lived structures can represent important features and possibly genuine geometrical (not just topological) features of data. PH has led to insights in an increasingly large number of applications [190] in diverse topics, ranging from granular materials [153] to contagions on networks [168, 248], path planning [44], cosmology [77], collective behavior in animals [253], the structure of brain arteries [34], and medical image analysis [74]. PH is one of several methods that are often referred

to by the umbrella term *topological data analysis*.

We introduce the topological concepts behind homology in Section 2.1. We then describe PH in Section 2.2. In Section 2.3, we illustrate how PH can be used to understand two types of data that are common in biology: network data and point cloud data. We focus on introducing existing methods which we will use in subsequent chapters. We proceed to describe methods that allow the interpretation and statistical analysis of the output of PH computations in Section 2.4. Finally, in Section 2.5 we give a brief pointer to other methods from topological data analysis. The above introduction, as well as the following sections, are based on Stolz *et al.* 2017 [240] and Stolz *et al.* 2018 [238] with modifications. Our approach to describe (persistent) homology is in the spirit of introductions such as [98, 190, 232], we provide our own examples which allow the reader to develop intuition.

2.1 Homology

PH is based on the topological concept of *homology* (for intuitive introductions, see, for example, [232, 241, 252]; for more formal introductions see [133, 150, 177]). We motivate the use of homology by considering different types of cheese and how they differ: Homology can differentiate between the shape of a stereotypical Swiss cheese (of the Emmentaler sort) with holes and the shape of a Mozzarella cheese by giving us information on the presence or absence of holes in the cheeses (see Fig. 2.1). The method thereby considers the space surrounding the holes, the so-called *loops*. The space is simplified by using scaffolds that contain the same number of holes as the space, so-called *simplicial complexes*, a depiction of which can be seen in Fig. 2.1. Homology does not give information on the geometry of the cheeses, e.g., it does not 'see' that the Swiss cheese is cut in a cube and the mozzarella is a sphere (unless it happens to be hollow), it only detects the differences in the number of holes. Similarly, if we were to stretch or bend the Emmental cheese (without breaking it), homology

would still detect the same number of holes and we could conclude that despite the deformations the cheese is still the same. This property means that homology is a *topological invariant*. Homology can be studied in any dimension n and formally associates an algebraic object to a topological space[1] for every dimension n. We

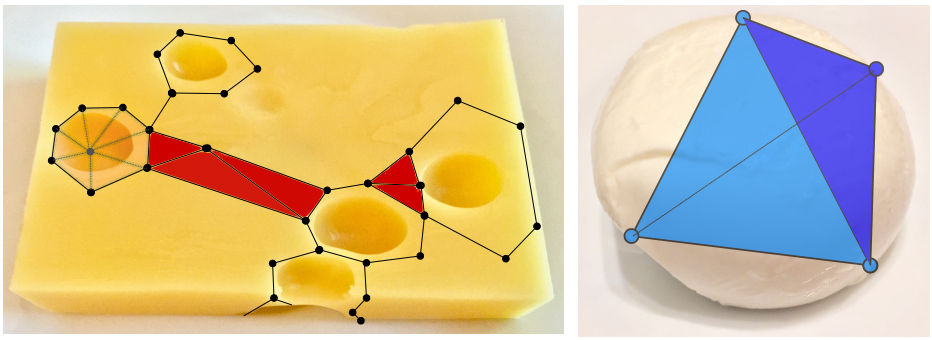


Figure 2.1: An example of two topologically different objects. Homology detects the topological differences by counting the number of holes in the cheeses. The simplicial complexes drawn on top of the cheeses approximate topological spaces and capture their properties. In order to capture the holes in the Emmental cheese, we glue together a collection of triangles and edges around the holes, enclosing the same number of holes as the original cheese. Note that we would only be able to capture the holes enclosed inside the cheese as the cut open holes that are visible on the surface can be deformed into a smooth surface of the cheese. For demonstrative purposes we therefore assume that the Emmentaler cheese is a cross section of a larger cheese enclosing the holes visible on the image. Image source: [238].

now briefly introduce some of the mathematical concepts necessary to understand homology.

2.1.1 Simplicial complexes

One can study the properties of a topological space such as the Swiss cheese or the mozzarella cheese by partitioning it into smaller and topologically simpler pieces, which when reassembled include the same aggregate topological information as the original space. The most trivial topological space $X = \{\emptyset, x\}$ consists of the empty set $\emptyset$ and a single point x. If we want to simplify the description of the topological properties of X, we would simply choose a single node to represent it. However, a node or even a collection of nodes does not allow one to capture the topological

[1] See Def. A.1.2 in Appendix A.

properties of more complicated spaces, such as a 2-sphere or the surface of the earth. In such cases, one needs a simple object that carries the information that the space is connected but also encloses a hole. For example, one could use a collection of triangles glued together to form (an empty) tetrahedron, which is an example of a mathematical object called a *simplicial complex*.

The building blocks that one uses to approximate topological spaces are called *n-simplices*, where the parameter n indicates the dimension of the simplex. Every n-simplex contains $n+1$ independent nodes: a point is a 0-simplex, an edge is a 1-simplex, a triangle  is a 2-simplex, and a (filled) tetrahedron  is a 3-simplex. By using a numbering x_i of vertices, we can write a 0-simplex as $[x_0]$, a 1-simplex as $[x_0, x_1]$, a 2-simplex as $[x_0, x_1, x_2]$, and a 3-simplex as $[x_0, x_1, x_2, x_3]$. Observe that the lower-dimensional simplices are contained in the higher-dimensional simplices. This allows one to build higher-dimensional simplices using lower-dimensional ones. The lower-dimensional simplices form so-called *faces* of the associated higher-dimensional objects.

One combines different simplices into a *simplicial complex* X to capture all different aspects of a topological space. For every simplex that is part of a simplicial complex, we demand that all of its faces are also contained in the simplicial complex. Additionally, two simplices that are part of a simplicial complex are allowed to intersect only in common faces, which does not allow situations such as  . The *dimension* of a simplicial complex is defined to be the dimension of its highest-dimensional simplex. A subcollection of a simplicial complex X is called a *subcomplex* of X if it forms a simplicial complex itself. If one is interested in the nature of a simplicial complex of dimension n, one can either consider the full complex, which can be very large, or one can examine subcomplexes. One can, for example, learn some of the properties of a simplicial complex X by considering the subcomplexes $\chi^{(\mathcal{P})}$, called *$\mathcal{P}$-skeletons* of X, that consist of all simplices of X of dimension $\mathcal{P}$ and their faces. In Fig. 2.2,

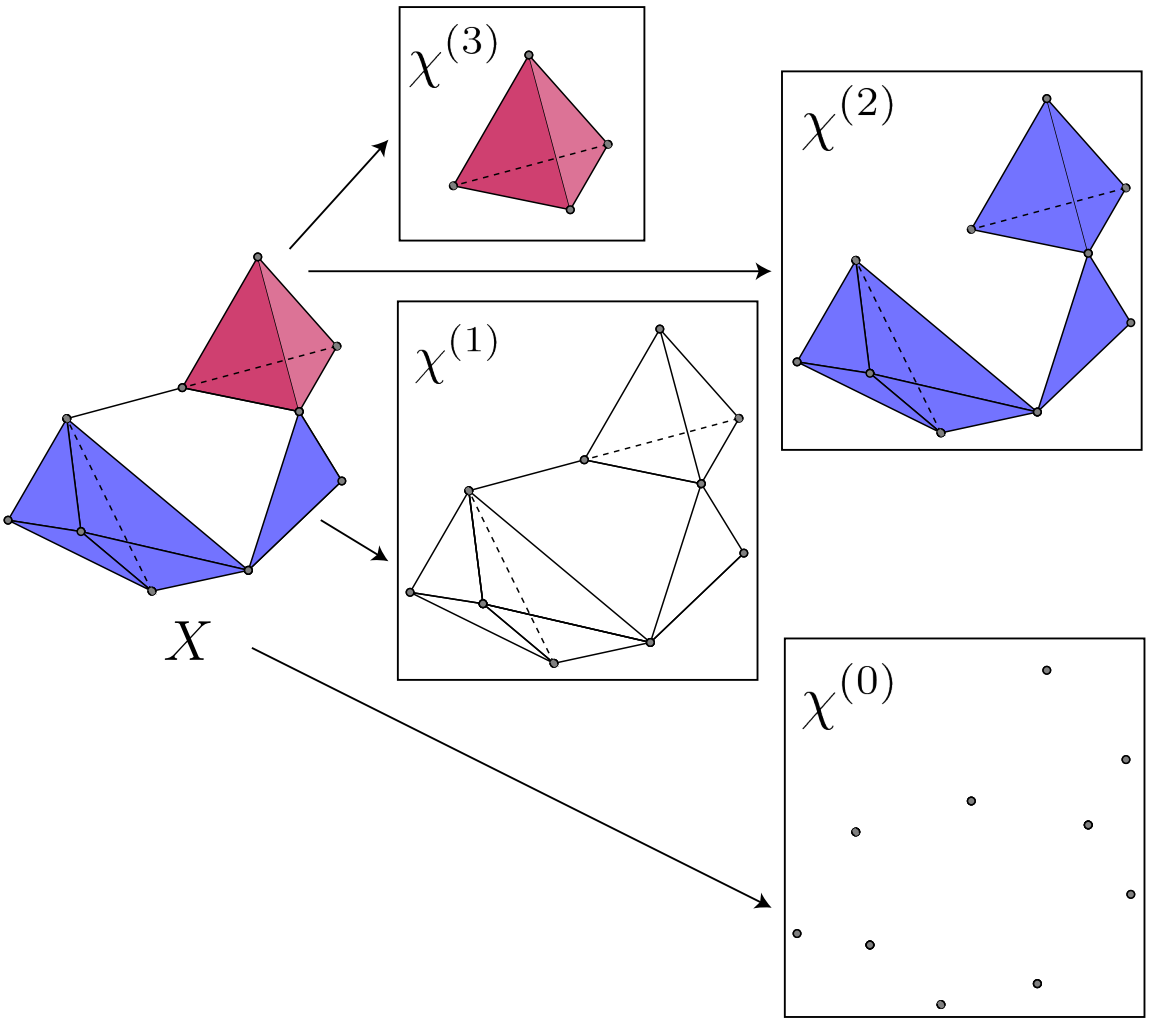


Figure 2.2: A simplicial complex X and its $\mathcal{P}$-skeletons $\chi^{(0)}, \chi^{(1)}, \chi^{(2)}, \chi^{(3)}$. We show 3-simplices in burgundy and 2-simplices in navy. Note that the 3-simplex in X appears as a 3-simplex in $\chi^{(3)}$ and is disassembled into its 2-simplices in $\chi^{(2)}$. Similarly, the 2-simplices in X appear as 2-simplices in $\chi^{(2)}$ and as 1-simplices in $\chi^{(1)}$.

we show an example of a simplicial complex and its $\mathcal{P}$-skeletons.

One can use simplicial complexes to represent topological spaces if and only if there exists a continuous deformation that can stretch and bend the simplicial complex into the topological space, and only then are topological properties of the topological space preserved by the simplicial complex.

2.1.2 Chains, cycles, and boundaries

Given a simplicial complex X, we might want to learn how the different layers of simplices are connected. For example, for the set of all 1-simplices that consists of a collection of edges (and their end points), we would want to know whether they are simply connected in long lines or whether they also form loops. For every simplicial complex X we can define a vector space $C_n(X)$ that is spanned by its n-simplices with coefficients in the field[2] $\mathbb{Z}/2\mathbb{Z}$. The elements of the vector space $C_n(X)$ are called *n-chains*. We can now define a linear map, the so-called *boundary operator*, between vector spaces $C_n(X)$ and $C_{n-1}(X)$ which takes every n-simplex x to the (alternating) sum of its faces, i.e. its boundary:

$$\begin{aligned} \partial_n : C_n(X) &\to C_{n-1}(X), \\ x &\mapsto \sum_{j=0}^{n} (-1)^j [x_0, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n], \end{aligned} \tag{2.1}$$

i.e. in the j-th summand we omit the vertex x_j from the vertices spanning the $(n-1)$-simplex. Note that the sum in Equation 2.1 is again over the field $\mathbb{Z}/2\mathbb{Z}$. We can use the boundary operator to connect all n-chains of a simplicial complex X in a sequence, the so-called *chain complex* $\mathcal{C} = \{C_n, \partial_n\}$:

$$\begin{gathered} \ldots \xrightarrow{\partial_{n+2}} C_{n+1} \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \ldots \xrightarrow{\partial_1} C_0 \\ c \longmapsto \partial_n c. \end{gathered}$$

We can represent a collection of edges that are connected to form a loop in a simplicial complex as a 1-chain, for example, $[x_0, x_1] + [x_1, x_2] + \cdots + [x_j, x_0]$. If we apply the boundary operator to this 1-chain, we obtain $\partial([x_0, x_1] + [x_1, x_2] + \cdots + [x_j, x_0]) = [x_1] - [x_0] + [x_2] - [x_1] + \cdots + [x_0] - [x_j] = 0$. In contrast, for a collection of edges that does not form a loop this is not the case, e.g., $\partial([x_0, x_1] + [x_1, x_2] + \cdots + [x_{j-1}, x_j]) =$

[2]Note that chains can be defined much more generally, for example over the ring $\mathbb{Z}$, see [79] or [133]. We focus our explanation on the field of coefficients $\mathbb{Z}/2\mathbb{Z}$ since this is most commonly used in the algorithms for PH and it is more intuitive to understand.

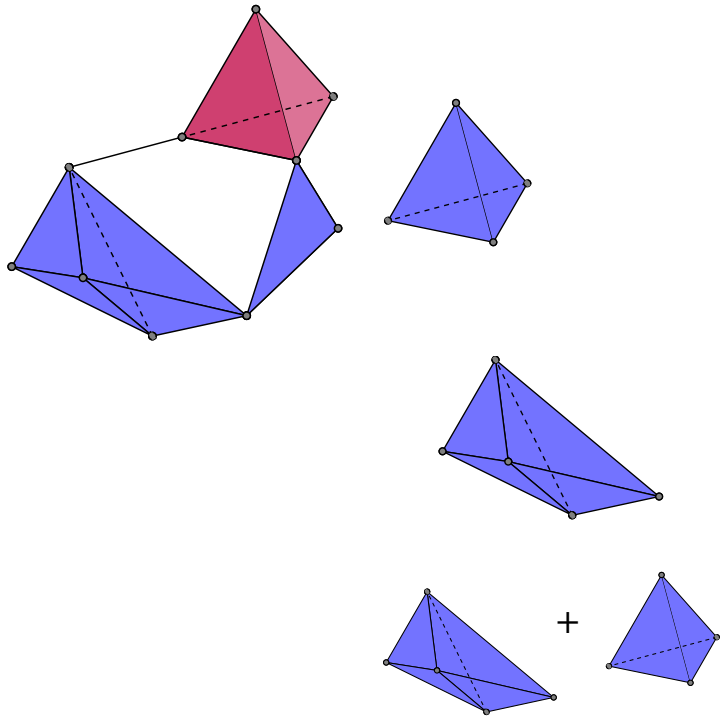
Simplicial complex	Examples of 2-cycles	Examples of 2- boundaries
		

Table 2.1: Examples of a simplicial complex and some of its 2-cycles and 2-boundaries. Note that the depictions of the 2-cycles and 2-boundaries represent elements in the vector space over the field $\mathbb{Z}/2\mathbb{Z}$ rather than simplicial complexes, i.e. they should be interpreted as sums of 2-simplices.

$[x_j]-[x_0] = [x_0]+[x_j]$ (for coefficients from $\mathbb{Z}/2\mathbb{Z}$). Chains that are in the kernel of ∂_n, i.e. their boundary is zero, are called *n-cycles*. One can compute that the composition of two boundary maps yields zero[3], i.e. $\partial_n \partial_{n+1} c = 0$ since the boundary of a boundary is empty. The image $\operatorname{im} \partial_{n+1}$ of the boundary operator is therefore a subspace of the kernel $\ker \partial_n$ and its elements are called *n-boundaries*. We show examples of 2-cycles and 2-boundaries of a simplicial complex in Table 2.1 and examples of 1-cycles and 1-boundaries of a simplicial complex in Table 2.2.

2.1.3 Homology groups and Betti numbers

One can associate a family of vector spaces known as *homology groups* to a simplicial complex X based on its cycles and boundaries. For every dimension $n \geq 0$ one defines

[3]For a proof, see, for example, [237]

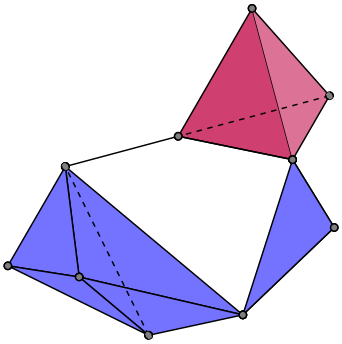
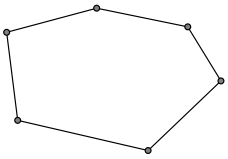
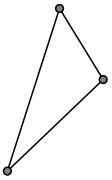
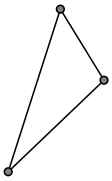
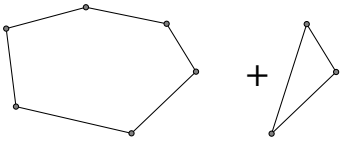
Simplicial complex	Examples of 1-cycles	Examples of 1- boundaries
		
		
		

Table 2.2: Examples of a simplicial complex and some of its 1-cycles and 1-boundaries. Note that the depictions of the 1-cycles and 1-boundaries represent elements in the vector space over the field $\mathbb{Z}/2\mathbb{Z}$ rather than simplicial complexes, i.e. they should be interpreted as sums of 1-simplices.

the nth *homology group* as:

$$H_n(X) = \frac{\ker \partial_n}{\operatorname{im} \partial_{n+1}}.$$

Intuitively, when we look at n-cycles ignoring n-boundaries, we are left with objects surrounding n-dimensional holes (see, for example, the 2-cycle in the second row of Table 2.1 or the 1-cycle in the first row of Table 2.2). In dimension 1, we therefore call the elements of the homology group H_1 *loops*; in dimension 0, we call the elements of the homology group H_0 *connected components*. Two elements in H_n are considered to be different, if they differ by more than a boundary, i.e. if they represent different n-dimensional holes. We then say that they belong to different *homology classes*. For example, in dimension 1, loops in the same homology class all surround the same 1-dimensional hole. We give an example of two loops that surround the same hole later in Fig. 2.4. If we want to measure the number of n-dimensional holes of a simplicial complex, we can consider its nth *Betti number* β_n:

$$\beta_n = \dim H_n(X) = \dim \ker \partial_n - \dim \operatorname{im} \partial_{n+1}.$$

One can interpret the first three Betti numbers, β_0, β_1, and β_2, to represent, respectively, the number of connected components, the number of 1-dimensional holes, and the number of 2-dimensional holes (i.e. voids) in a simplicial complex.

2.2 Persistent homology

While homology gives information about a single simplicial complex, PH allows one to study topological features across embedded sequences, so-called *filtrations*, of simplicial complexes. Constructing simplicial complexes from data involves many decisions that can, for example, require a choice of threshold (see Section 2.3). Being able to study a sequence of simplicial complexes created using multiple, or even all possible thresholds makes PH particularly suited for data analysis.

2.2.1 Filtrations and functoriality

A filtration [66,97,123] of a simplicial complex X is a sequence of embedded simplicial complexes,

$$\emptyset = X_0 \subseteq X_1 \subseteq X_2 \subseteq \cdots \subseteq X_{end} = X\,, \tag{2.2}$$

starting with the empty complex and ending with the entire simplicial complex X. The simplicial complexes in the filtration are connected by inclusion maps. One can now apply an important property of homology, *functoriality*: any map between simplicial complexes $f_{i,j} : X_i \to X_j$ induces a map between their n-chains $\tilde{f}^n_{i,j} : C_n(X_i) \to C_n(X_j)$ which induces a map between their homology groups $f^n_{i,j} : H_n(X_i) \to H_n(X_j)$ (see, for example, [190] for a more detailed description). In particular, this means that there exist maps between the homology groups of every simplicial complex in a filtration, e.g., there are maps that relate the loops or connected components in simplicial complexes across a filtration.

2.2.2 Barcodes

One can visualise the presence of topological features such as loops or connected components across a filtration in a summary diagram called *barcode* [68, 123]. For an appropriate choice of basis[4] of the homology groups H_n, a barcode represents the information carried by the homology groups and the maps $f^n_{i,j} : H_n(X_i) \to H_n(X_j)$. A topological feature of dimension n in $H_n(X_\eta)$ is *born* in $H_n(X_\eta)$, if it is not in the image of $f^n_{\eta-1,\eta}$. For example, intuitively, a loop is born in filtration step η, if the hole that it surrounds first appears bounded by the simplicial complex X_η. A topological feature from $H_n(X_i)$ *dies* in $H_n(X_\zeta)$, where $i < \zeta$, if ζ is the smallest

[4]Carlsson and Zomorodian [279] interpret the collection of homology groups of a filtered simplicial complex together with the induced maps between them as a *persistence module*, for which they show that there exists a choice of compatible bases such that the module can be uniquely decomposed (see Equation 5 in the paper). This decomposition can be represented by a collection of intervals, the barcode. The existence of this decomposition is sometimes referred to as *The fundamental theorem of PH* (see, for example, [190]) and is only possible when using field coefficients in the definition of homology, such as in the present thesis.

index such that the feature mapped to zero by $f^n_{i,\zeta}$. If the topological feature is a loop, intuitively it dies in the filtration step where it is first fully covered by triangles (or other higher-dimensional simplices). Note that some topological features never die in a filtration, for example, we always have one connected component in a non-empty simplicial complex that is never mapped to zero. In a barcode, topological features in the filtration of a simplicial complex are represented by half-open intervals $[\eta, \zeta)$. We show examples of barcodes in Fig. 2.3 and Fig. 2.5. The lifetime of a topological feature, the so-called *persistence* ϱ, is defined as

$$\varrho = \zeta - \eta \,.$$

For topological features that persist until the last filtration step (and beyond), we define the persistence to be infinite. Persistence was first used as a measure to rank topological features by their life time in a filtration in $\mathbb{R}^3$ [99].

2.2.3 Persistence diagrams

A *persistence diagram* [76] (PD) is an alternative visual representation of topological features in a filtration to a barcode. Instead of intervals $[\eta, \zeta)$, the topological features are represented by points (η, ζ) in a birth–death coordinate system. In addition to the (η, ζ) points, the points on the diagonal, i.e. points (η, ζ) where $\eta = \zeta$, are also considered to be part of the persistence diagram. The further away a point is from the diagonal line, the more persistent the corresponding feature is in the filtration. Alternatively, one can use a birth–persistence coordinate system, which can be the first step towards an analysis with persistence images (see Subsection 2.4.3). We show an example of a persistence diagram in Fig. 2.5.

Persistence diagrams (and barcodes) have been shown to be stable, i.e. small perturbations of the input data lead to small perturbations in the persistence diagram (or barcode) [76].

2.3 Persistent homology for different types of data

Typical types of data that lend themselves to be studied by PH are point cloud data (also referred to as finite metric spaces [190]) and network data. PH can however also be used for many other types of data, including digital images, and level-sets of real-valued functions [190]. One can define simplicial complexes and filtrations on point cloud or network data in many different ways. The choice of filtration tends to be motivated either by the type of questions to be answered about the data or by the consideration of computation time.

2.3.1 Persistent homology for point cloud data

Classically, PH is used to study data in the form of point clouds, i.e. the data is given by a finite number of discrete points, whose coordinates represent different measurements in a metric space. In order to obtain a filtration from the point cloud, one can, for example, apply a Vietoris–Rips filtration.

2.3.1.1 Vietoris–Rips filtration

The Vietoris–Rips complex is defined based on spatial proximity of data points: for a given distance ϵ two points are connected by an edge if their distance is at most ϵ. Higher-dimensional simplices are constructed as follows: if three points are pairwise connected by edges, they are interpreted as a 2-simplex, if four points are pairwise connected by edges, they are interpreted as a 3-simplex etc. The resulting simplicial complex and its properties are determined by the choice of ϵ: if ϵ is very small, no points are connected and we do not obtain any interesting topological information. If ϵ is very large, all points are connected pairwise and we simply obtain a high-dimensional simplex. A filtration of Vietoris–Rips complexes is therefore constructed by not fixing but varying ϵ [66, 123]:

1. Choose a sequence of increasing distances $\epsilon = \{\epsilon_1, \ldots, \epsilon_{\text{end}}\}$.

2. In the i-th filtration step define n-simplices by unordered $(n+1)$ - tuples of pairwise distance at most ϵ_i.

We show an example of Vietoris–Rips complex filtration with corresponding barcodes in Fig. 2.3. Among many other applications, the Vietoris–Rips complex has been

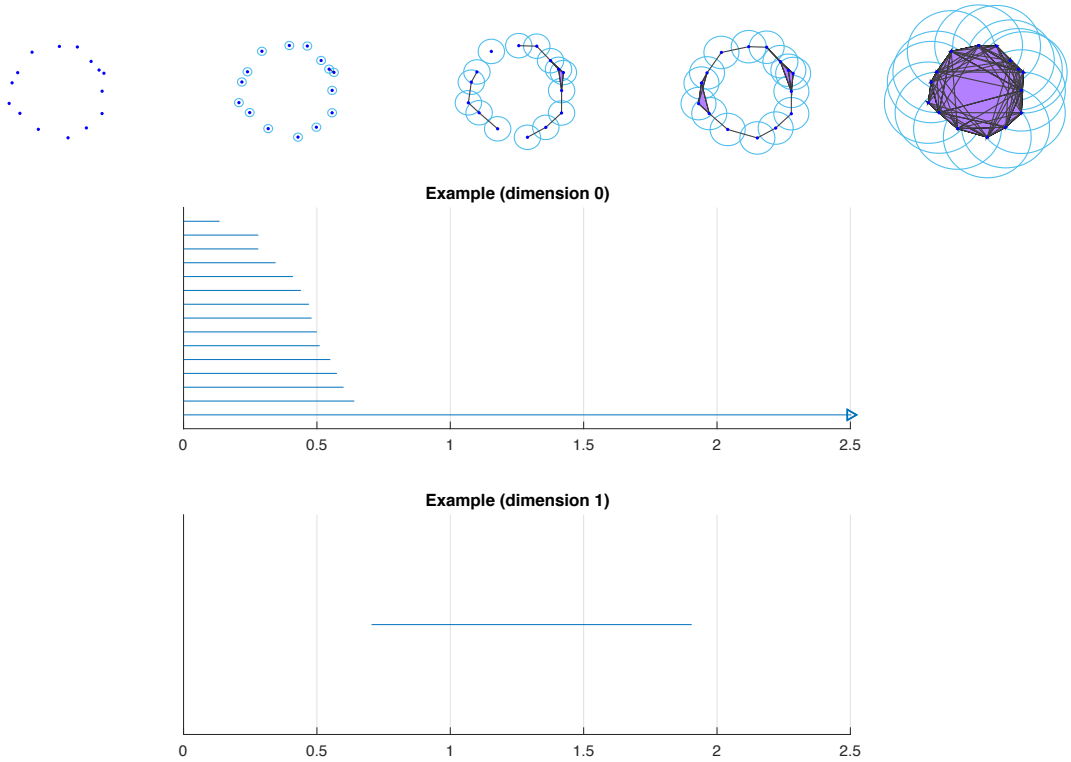


Figure 2.3: An example of a Vietoris–Rips filtration with simplicial complexes shown for $\epsilon = 0, 0.1, 0.55, 1, 2$ (top row) and the corresponding barcode (bottom row). The radius ϵ increases over the filtration steps and is shown on the horizontal axis of the barcode. Every line in a barcode represents a specific topological feature. The 0-dimensional barcode consists of 15 bars – one for each connected component – which all begin at $\epsilon = 0$. As ϵ increases, more vertices are connected by edges and connected components die in the filtration until only one large component is left. Similarly, for the 1-dimensional barcode, we observe a 1-loop that is born at $\epsilon = 0.7$. It dies once the data points are connected to form higher dimensional simplices that cover the hole. Image source: [64]

successfully used to study protein conformations [152], nanoporous materials [161], biological aggregation models [253], and different types contagion processes on networks [168, 248]. We apply the Vietoris–Rips filtration in Chapter 5 locally around data points for landmark selection in large and noisy data sets and for the detection of geometric anomalies in data sampled from intersecting surfaces.

2.3.2 Persistent homology for network data

From a topological point of view, a network [183] is a one-dimensional object that consists of points (the vertices) and lines connecting these points (the edges). In real-world applications networks are often weighted, i.e. every edge is assigned a specific weight. The standard methods from network theory are based on pairwise connections, which one can use to study micro-scale, meso-scale, and macro-scale structures [183]. PH gives us an alternative approach for studying networks, which explicitly incorporates 'higher-order' structures beyond pairwise connections and allows us to study topological invariants such as connectedness, loops, or holes. Although one can also represent higher-order structures using formalisms such as hypergraphs [47] (see, e.g., a recent paper [24] by Bassett *et al.*), those other approaches may not be the most convenient means for optimally conveying information about the shape or scale of mesoscale structures in a network. Other recent work concerns clustering in networks using higher-order structures [36].

PH has been successfully used on networks in many applications, ranging from granular materials (see, e.g., [153]) to neuronal networks, leading to several promising insights [14, 16, 18, 75, 80–82, 90, 122, 124, 125, 138, 157, 158, 201, 202, 236, 238, 240]. In this thesis we study functional neuronal networks using PH in Chapter 4 (see also [238, 240]), and tumour blood vessel networks in Chapter 3 (see also [64]). In both cases we use different filtrations on the networks. We now introduce the weight rank clique filtration which we will use for the functional neuronal networks. For the tumour blood vessel networks we develop our own filtration in Chapter 3.

2.3.2.1 Weight rank clique filtration for functional networks

We use PH for functional neuronal networks because it can detect loops in these networks which are invisible to other methods. The presence (or absence) of loops carries information on how a network is connected [157]. A loop in a graph is a set of

at least four edges that are connected in a way that forms a topological circle[5]. We give an example of such a loop in a network in Fig. 2.4.

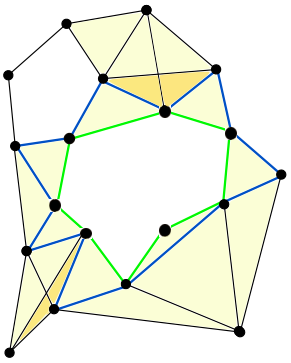


Figure 2.4: Example of a 1-dimensional loop in a simplicial complex. The green and the blue loop both surround the same hole and are therefore considered to be representatives of the same homology class. Image source: [240]

The simplest way to create a filtration from a weighted network is to filter the edges by their weights [159]. One first creates a sequence of embedded (binary) graphs by ranking all edge weights ν_i in descending order. In filtration step i, one retains an edge if and only if its weight is at least ν_i. To construct the filtration, one repeats this procedure until the graph is complete in the last step. Using this method, only 0-simplices (i.e., nodes) and 1-simplices (i.e., edges) are present in the filtration and we cannot distinguish a loop from three edges that form a triangle. The *weight rank clique filtration* (WRCF) [202], which we will use and which has been applied previously for examining weighted neuronal networks (see, for example, [125, 201, 202]), extends this definition to include higher-dimensional simplices. One constructs a WRCF as follows:

1. Define filtration step 0 as the set of all nodes.

2. Rank all edge weights $\{\nu_1, \ldots, \nu_{\text{end}}\}$, with $\nu_1 = \nu_{\text{max}}$ and $\nu_{\text{end}} = \nu_{\text{min}}$.

3. In filtration step i, threshold the graph at weight ν_i to create a binary graph.

[5]We sometimes also refer to these loops as 1-dimensional loops.

4. Find all maximal n-cliques for $n \in \mathbb{N}$, and define them to be n-simplices.

In every filtration step we obtain a valid simplicial complex: every $(n+1)$-clique in the graph guarantees the existence of an n-face on that clique, because cliques are closed under both intersection and taking subsets. Consequently, cliques satisfy the requirements for a simplicial complex. This type of simplicial complex on a graph is called a *clique complex*.

In Fig. 2.5, we show an example of a WRCF based on a neuronal network and its corresponding barcode.

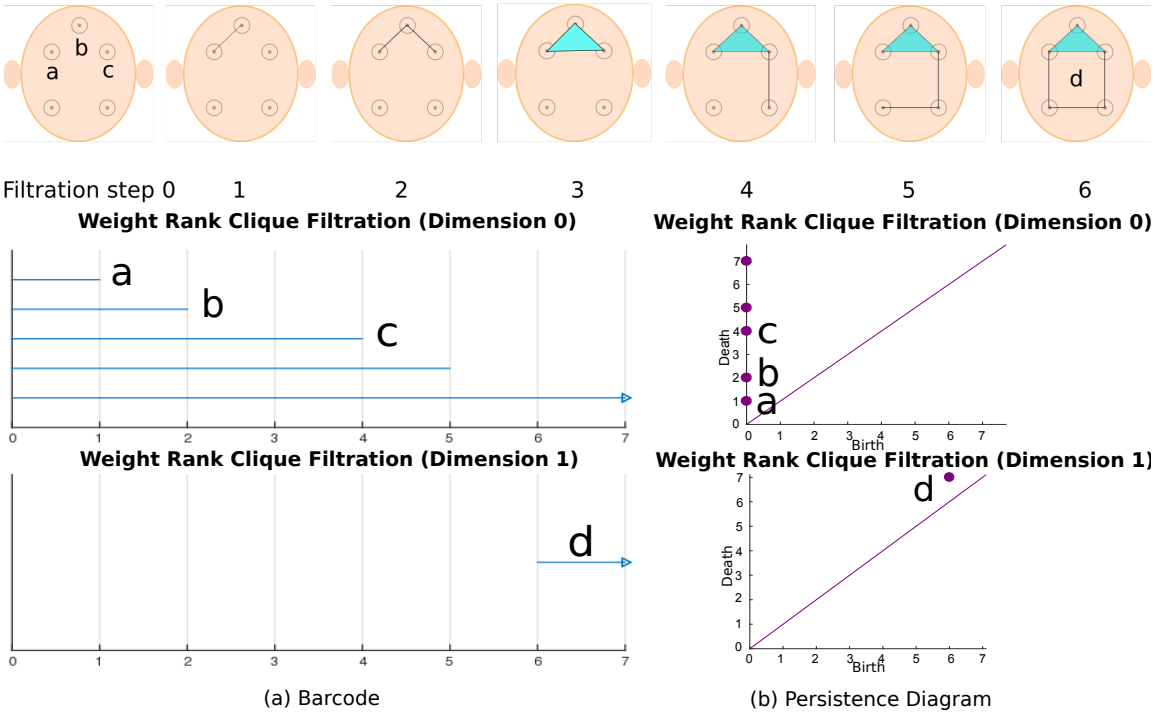


Figure 2.5: Example of a weight rank clique filtration (WRCF) of a neuronal network and the corresponding (a) barcodes and (b) persistence diagrams (PDs) in dimension 0 and 1. The neuronal network consists of different brain regions (indicated by circles), which we interpret as the nodes (indicated by dots) of a network, and weighted edges between the nodes. To construct the filtration, we add the nodes in step 0, followed by the edge with the largest weight in step 1, the edge with the second-largest weight in step 2, and so on. As soon as three nodes are all connected pairwise by edges, we cover the resulting region with a triangle. When four nodes are all connected pairwise, we fill in a tetrahedron. In a 0-dimensional barcode, each connected component is represented by a bar starting when the component is born and ending when it dies (e.g., when two components combine with each other). In a 1-dimensional barcode, each bar represents a loop, which consists of 4 or more edges and starts and ends at the same node. In persistence diagrams, one represents topological features by points rather than by bars. The distance of a point to the diagonal (the purple line) indicates the persistence of the corresponding feature in the filtration. Image source: [238].

2.4 Statistical analysis of persistent homology

In order to interpret the output of PH calculations on large data sets it is necessary to have a framework for statistical analysis. One of the first attempts of developing such a framework that, in particular, includes the calculation of a mean was conducted by Mileyko *et al.* 2011 [172] on the space of persistence diagrams. Persistence diagrams can be treated as elements of the space of persistence diagrams, on which one can define a metric. Mileyko *et al.* 2011 [172] show that it is possible to define probability measures in the space of persistence diagrams and that one can define statistical measures such as the mean and variance of a set of persistence diagrams. However, this mean is not well-defined as its definition includes a minimisation. An example for such a situation is given in Munch *et al.* 2015 [176]: consider two overlaid persistence diagrams with two points each located at opposite corners of a square (see also Fig, 5 in [176]). Intuitively, there are now two possibilities to place mean points on this square: either in the middle of the horizontal sides of the square or in the middle of the vertical sides. These intuitive positions correspond to the ones that arise from the formal definition of the mean given in [172]. The non-uniqueness of the mean was addressed in [176] where the authors defined a mean persistence diagram as a mixture or distribution of possible diagrams. Even though an algorithm was developed to find a local minimum for a version of the mean [257], there are to date no implementations. We therefore will only briefly introduce distances for barcodes and persistence diagrams which are implemented in standard packages for the computation of PH (see also Subsection 2.4.1).

Instead of relying on persistence diagrams or barcodes, one can use alternative ways of representing the output of PH calculations that were developed to allow statistical interpretation. We describe two approaches that have been used successfully in many applications and for which software packages are available: persistence landscapes and persistence images. Both of these methods can be used to generate

vector representations of PH output. Note that there are also a range of other summary techniques for PH output. These include kernel-based approaches which have a strong theoretical foundation (see, for example, [69, 155, 156, 211]) and the recently developed persistence codebooks [278] which combine aspects of kernel-based methods with vector representation. Persistence diagrams have also been combined with deep learning to create task-specific representations of topological features [135]. This supervised approach learns which topological features are of particular relevance to the task at hand, but requires a large training set.

2.4.1 Distances for barcodes or persistence diagrams

As described in Subsubsection 2.2.3, one can treat persistence diagrams as mathematical objects, and one can endow the space of persistence diagrams with a distance (see, for example, [190]). For two persistence diagrams Π_1 and Π_2, the $\mathtt{p}$*th Wasserstein distance* for $\mathtt{p} \in [1, \infty)$ is given by

$$d_{W_{\mathtt{p}}}(\Pi_1, \Pi_2) = \inf_{\mathcal{F}:\Pi_1 \to \Pi_2} \left\{ \sum_{\wp \in \Pi_1} \mathtt{d}(\wp, \mathcal{F}(\wp))^{\mathtt{p}} \right\}^{\frac{1}{\mathtt{p}}},$$

and for $\mathtt{p} = \infty$ by

$$d_{W_\infty}(\Pi_1, \Pi_2) = \inf_{\mathcal{F}:\Pi_1 \to \Pi_2} \sup_{\wp \in \Pi_1} \mathtt{d}(\wp, \mathcal{F}(\wp)),$$

where $\mathcal{F}$ denotes all possible bijections between Π_1 and Π_2 and $\mathtt{d}$ is a metric on $\mathbb{R}^2$. A common choice of metric is $\mathtt{d} = L^{\mathtt{q}}$ with $\mathtt{q} \in [1, \infty]$.

A special case of the Wasserstein distance is the *Bottleneck distance*, where $\mathtt{d} = L^\infty$ and $\mathtt{p} = \infty$. We use the Wasserstein distance in Chapter 5, Subsection 5.3.1.

2.4.2 Persistence landscapes

Persistence landscapes [53,55] can be constructed from barcodes and consist of piecewise-linear functions in a separable Banach space. For a given barcode interval $[\eta, \zeta)$, one

defines the function

$$\hat{f}_{(\eta,\zeta)}(x) = \begin{cases} 0\,, & \text{if } x \notin (\eta,\zeta)\,, \\ x-\eta\,, & \text{if } x \in (\eta, \frac{\eta+\zeta}{2}]\,, \\ -x+\zeta\,, & \text{if } x \in (\frac{\eta+\zeta}{2}, \zeta)\,. \end{cases} \tag{2.3}$$

For a barcode $\{\eta_i, \zeta_i\}_{i=1}^{I}$ and $i \geq 0$, the *ξth persistence landscape* is given by the set of functions

$$\begin{aligned} &\lambda_\xi : \mathbb{R} \to \mathbb{R}\,, \\ &\lambda_\xi(x) = \xi\text{th-largest value of } \{\hat{f}_{(\eta_\xi,\zeta_\xi)}(x)\}_{\xi=1}^{I}\,. \end{aligned} \tag{2.4}$$

If the ξth-largest value does not exist, then $\lambda_\xi(x) = 0$. One can think of the 0th persistence landscape as being the outline of the collection of peaks created by the images of the collection of functions $\hat{f}$ associated to a barcode. To obtain the 1st persistence landscape, one peels away this topmost 'layer' of peaks and then considers the outline of the remaining collection of peaks. This gives the 1st persistence landscape and one continues in this manner to obtain subsequent persistence landscapes. The *persistence landscape* λ of the barcode $\{\eta_i, \zeta_i\}_{i=1}^{I}$ is then defined as the sequence $\{\lambda_\xi\}$ of functions λ_ξ. We illustrate how to obtain a persistence landscape from a barcode in the first steps of Fig. 2.6. Even though persistence landscapes visualise the same information as barcodes and one can construct a bijective correspondence between the two objects, the former have distinct advantages over the latter. For example, one can calculate a unique 'average landscape' for a set of persistence landscapes by taking the mean over the function values for every landscape layer. This is not possible for barcodes, as they are not elements of a Banach space. For an average landscape, it is thus not possible to find a corresponding average barcode. We show a schematic illustration on how to obtain an average persistence landscape from two landscapes in Fig. 2.6 . One can also define $L^{\mathtt{q}}$ distances with $\mathtt{q} \in [1, \infty]$ between two (average) landscapes and thereby use a variety of statistical tools [53]. This allows one to compare multiple groups of barcodes by calculating a measure of pairwise similarity between

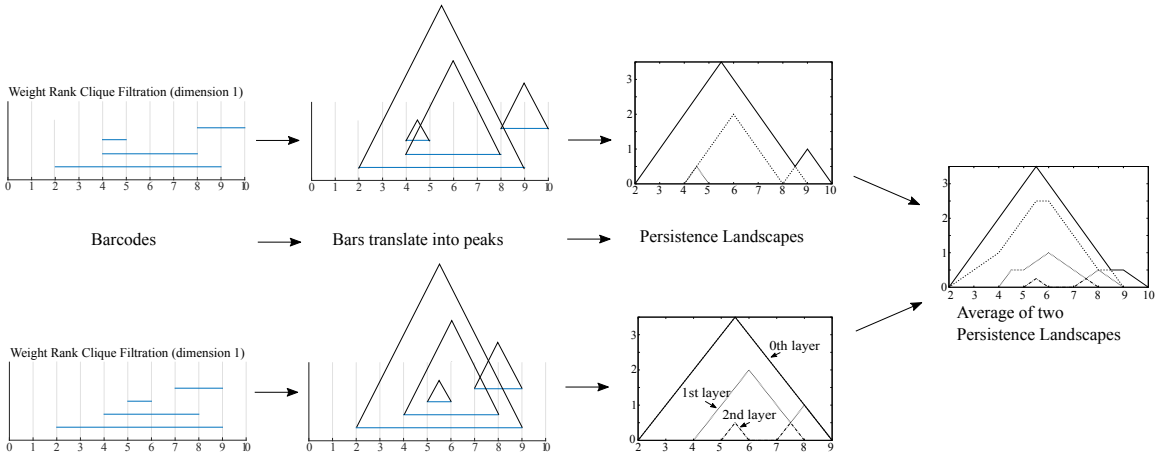


Figure 2.6: Visualisation of the relationship between barcodes and an (average) persistence landscape. The example is based on a weight rank clique filtration in dimension 1. To obtain a landscape from a barcode, one replaces every bar of the barcode by a peak, whose height is proportional the persistence of the bar. In the landscape all peaks are translated to touch the horizontal axis. The persistence landscape consists of different layers, where the ξth layer corresponds to the ξth-largest function value across the collection of peak functions. One creates an average of two landscapes by taking the mean over the function values in every layer. Image source: [240]

them. Moreover, one can use simple statistical methods such as permutation tests to determine whether the distance measured between two persistence landscapes is statistically significant. Like barcodes, persistence landscapes are stable with respect to small perturbation of the input data [53]. Persistence landscapes have been used to study conformational changes in protein binding sites [152], the origin of seizures in electroencephalographic (EEG) data from epileptic patients [265], phase separation in binary metal alloys [91], brain geometry in neurodegenerative diseases [120], and music audio signals [163]. We apply persistence landscapes to the study of functional neuronal networks in Chapter 4.

2.4.3 Persistence images

Persistence images [3] are a more recently developed way of representing the information captured in a barcode or persistence diagram. The distinct advantage of persistence images over persistence landscapes is that they are real valued vectors that can be used in combination with machine learning. Persistence images are stable with respect to input noise and they maintain an interpretable connection to the

persistence diagrams that they were obtained from even when averaged.

One can obtain a persistence image from a persistence diagram Π by first converting the coordinates of the persistence diagram from birth-death coordinates (η, ζ) to birth-persistence coordinates $(\eta, \zeta - \eta)$. This new diagram is denoted by $\hat{\mathcal{F}}(\Pi)$ and is then mapped to a persistence surface $\Upsilon_\Pi : \mathbb{R}^2 \to \mathbb{R}$, which is defined as follows:

$$\Upsilon_\Pi(\wp) = \sum_{\upsilon \in \hat{\mathcal{F}}(\Pi)} \Theta(\upsilon)\Gamma_\upsilon(\wp), \tag{2.5}$$

where Θ is a weighting function and Γ_υ is the normalised symmetric Gaussian with mean υ and variance $\hat{\sigma}^2$. The weighting function Θ needs to be 0 along the birth axis and can be chosen to emphasise features that are born or persist in a particular range of interest. The persistence surface is then reduced to a finite dimensional vector by choosing a pixel grid and assigning every pixel Ω the value $\Lambda(\Upsilon_\Pi)_\Omega = \int \int_\Omega \Upsilon_\Pi \mathrm{d}y\mathrm{d}x$. We show an illustration of how to obtain a persistence image from a persistence diagram in Fig. 2.7.

Persistence images have, for example, been used for the classification of different subtypes of neurons [145,146]. We apply persistence images to the study of functional neuronal networks in Chapter 4.

2.5 Other types of methods from topological data analysis

One can also study filtrations using *Betti curves*, which were introduced in [125]. Betti curves show the Betti numbers in each filtration step. For filtrations considering only simplicial complexes with 0- and 1-simplices, for example, a filtration of a network by edge weights [159], one can calculate Betti curves via the Euler characteristic omitting the expensive computation of PH. For an example where this approach enables the computation of Betti numbers across a complete filtration of – fully connected –

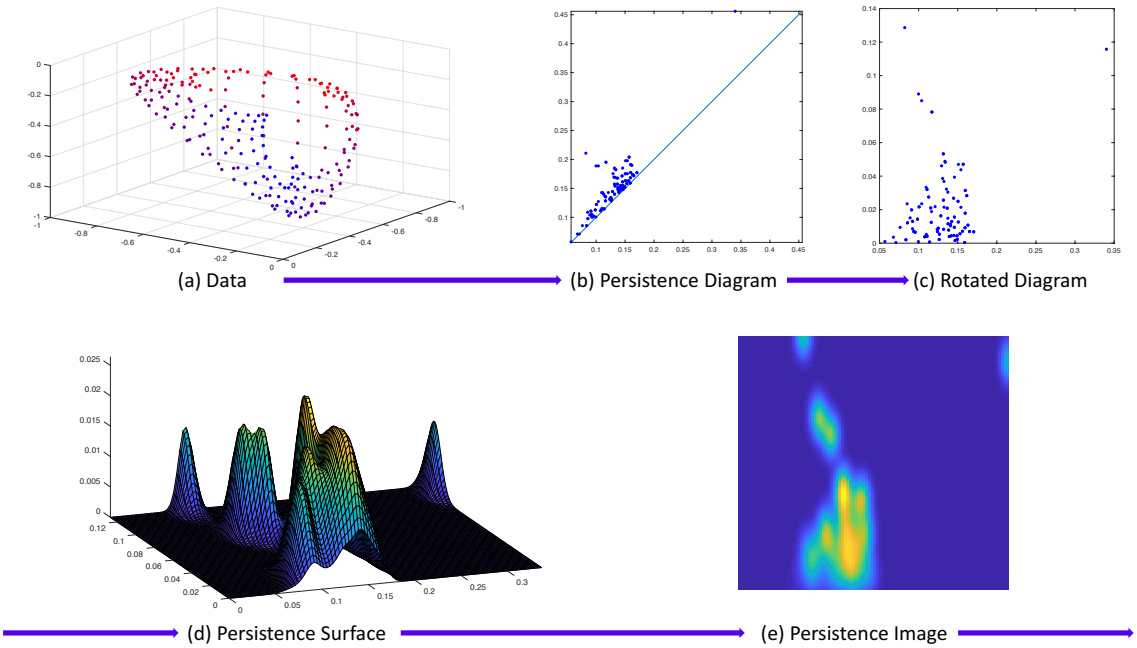


Figure 2.7: Schematic illustrating the primary steps for converting a persistence diagram (PD) to a persistence image (PI). (a) Sample point cloud in $\mathbb{R}^4$ plotted in $\mathbb{R}^3$ where the colouring corresponds to the fourth coordinate value. (b) PD in birth–death coordinates (i.e., the standard choice), with the diagonal identity line in blue. (c) PD in birth–persistence coordinates. (d) The process of generating a surface by centering 2D Gaussian distributions at each point in panel (c). (e) One generates a PI by summing the volume under 2D Gaussian distributions over the area of a pixel (i.e., the area of a square) in a uniformly-spaced grid overlay. Image source: [238] with permission from Tegan Emerson.

functional networks, see [75]. We apply Betti curves to functional neuronal networks in Chapter 4.

There are many other applications of topology in the study of data, for example, one can gain insights by monitoring the number of n-simplices in networks [210]. Another example of a prominent method from topological data analysis is the *Mapper* algorithm [184, 186, 227]. For examples of several other methods that are based on topological ideas, see, for example, [266].

" des himmels sterne, speren, kreiß,
wie hoch sie sint, min zirkel weiß;
wie tif, wie hoch, wie wit, wie lank,
in formen sie min zirkel twank."[a]

From the speech by the personified art of geometry (Geometria) in *Der meide kranz*, Heinrich von Mügeln, 14th century. Edition by [263].

[a]Translation: "The stars, sphere and orbits of heaven - my compasses know how high they are. However deep, however high, however wide, however long - my compasses forced them into shape." [263].

3

Persistent Homology Applied to Tumour Blood Vessel Networks

Globally, cancer is one of the leading causes of death [188]. The term refers to a large group of diseases that share common characteristics and can affect any part of the body. The world health organisation (WHO) estimates that the risk of developing cancer before the age of 75 worldwide is around 20% and the risk of dying from cancer before 75 is 10% [108]. While incidence and mortality rates can vary significantly between cancer types (see Fig. 3.1), the development of cancer from a single cell to an aggressive tumour follows common patterns, whose understanding is critical to developing new therapies.

The formation of blood vessels in tumours, *tumour-induced angiogenesis*, has long been recognised as one of the hallmarks of cancer, e.g. a capability that a cell must acquire to become malignant [100,110,132,262]. Tumours not only rely on vasculature for the delivery of nutrients, cells, and oxygen, but blood vessels also provide physical support, serve as a starting point for metastasis, form a niche for tumour stem cells and even ensure immune suppression around the tumour [126,189]. For several cancer

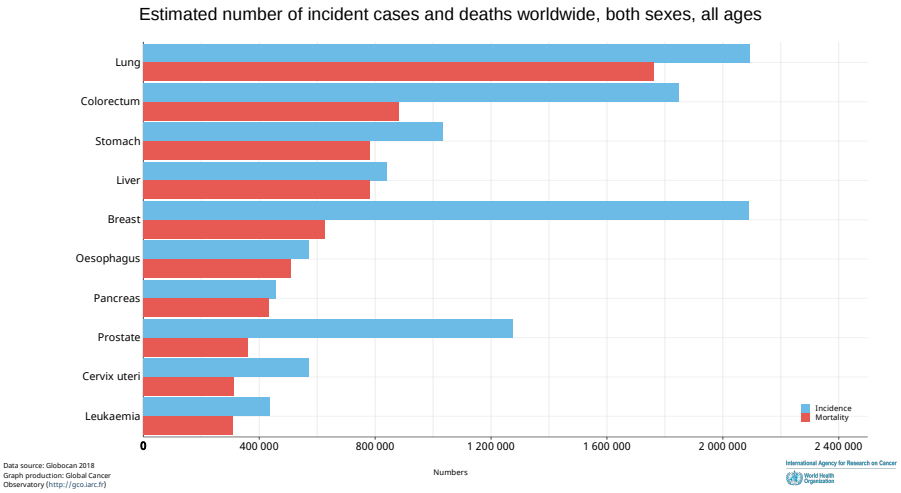


Figure 3.1: Incidence and mortality rates by cancer types estimated by the Global Cancer Observatory, 2018. Image source: [108].

types it has been observed that the degree of microvascular proliferation is an indicator of the level of the aggressiveness of a tumour [189].

Structurally, tumour blood vessels form a chaotic network of highly inefficient blood vessels characterised by many loops and twists. Even though these features are obvious to the human eye, to characterise them in a quantitative and informative manner has so far been very difficult. We will summarise existing approaches to quantification in Section 3.2. Here, we investigate persistent homology (PH) as a quantitative tool to characterise structural features of tumour blood vessels. While a large number of mathematical models of tumour-induced angiogenesis exist (see for example [9,63,87,192,262]), the focus here is to develop a method that enables analysis of tumour blood vessel data. We show that our method can spatially capture loops in tumour blood vessel networks and their changes over time in response to treatment in two different data sets. We further find features in the PH dimension 0 output which follow a similar trend as the loops and are likely to be connected to tortuosity of the blood vessels.

We now introduce the biological background of tumour-induced angiogenesis before illustrating our novel method for analysis. Our biological introduction is intended

as a brief overview and not an exhaustive review of the biological literature (for an excellent review of the biology and mathematical models of tumour-induced angiogenesis see [262]). Note that parts of the text of Subsubsections 3.2.2.1 and 3.2.2.2 were also used in our paper [64] with minor modifications.

3.1 Tumour-induced angiogenesis and characteristics of tumour blood vessel networks

Tumour-induced angiogenesis leads to the formation of blood vessels that are highly abnormal in structure and function. We first describe some of the biological processes regulating tumour-induced angiogenesis. We then outline the structure and function of tumour blood vessels.

3.1.1 Tumour-induced angiogenesis

As a tumour develops from a collection of cells into a solid mass, it initially receives nutrients and oxygen from its surroundings via diffusion. Once the tumour grows to a certain size, diffusion is no longer sufficient to meet the tumour's increasing demands. The tumour enters a dormant state until it acquires the ability to secure its own blood supply [262]. To acquire new blood vessels, the tumour uses existing blood vessels in its vicinity. This process is crucial in the further development of the tumour and is also often referred to as the *angiogenic switch* [262]. A tumour can induce the formation of new vessels using different mechanisms [126, 141, 262, 270]:

1. *Angiogenesis*: Sprouting of new vessels from existing vessels via molecular signals.

2. *Postnatal vasculogenesis*: Recruiting of endothelial progenitor cells from the bone marrow to differentiate into endothelial cells and form new vessels.

3. *Splitting angiogenesis*: Extension of a vessel membrane into its lumen such that the vessel is divided into two vessels.

4. *Vasculogenic mimicry*: Tumour cells line blood vessels mimicking the function of endothelial cells.

5. *Mosaic vessel formation*: Growth of tumour cells around vessel walls.

While a tumour may use a combination of these processes, it is typically assumed that a significant proportion of blood vessels is formed by angiogenesis [262]. As all of these mechanisms, angiogenesis is driven by the secretion of proangiogenic molecules, so-called *tumour angiogenic factors* (TAFs), by hypoxic tumour cells [126, 217, 262, 274]. The most prominent TAFs are *vascular endothelial growth factor*[1] (VEGF) and *angiopoietin 2* (ANG-2). TAFs activate endothelial cells of existing vessels, initiating sprouting of a finger-like tip which then grows into a solid strand of endothelial cells amongst the extracellular matrix. The formation of many sprouts in the same region on existing vessels is prevented via lateral inhibition by molecules such as the Delta-like ligand 4 (Dll4). After sprout formation, angiogenesis continues to be regulated by chemotaxis (i.e. vessels grow along a gradient of TAFs), as well as haptotaxis (i.e. vessels grow via interaction with the extracellular matrix), and mechanotaxis (i.e. vessels grow in response to mechanical cues such as fluid flow) [262]. Initially, newly formed blood vessels grow in parallel. Once the vessels reach a certain distance however, they start growing towards each other, which favours the formation of anastomoses, connections between different vessels. This process generates loops in the emerging blood vessel networks. New sprout tips grow from the loops, continuing the extension of the vessel network. As blood vessels grow towards the tumour, they exhibit a dramatic increase in branching before they penetrate the tumour.

[1]Note that there are different isoforms of VEGF, for a review see, for example, [261].

The resulting vasculature is highly abnormal, in both structure and function [126, 181,217,226]. The precise mechanisms of angiogenesis are however still not completely understood [258]. Mathematical modelling yielded many insights into the influence of different mechanisms involved in angiogenesis on the structure of the resulting vasculature (see for example [9, 63, 87, 192, 262]) and continues to do so.

3.1.2 Structure and function of tumour blood vessels

Abnormalities of tumour vasculature occur both at a macro- and microscopic level: while normal tissue contains evenly spaced, hierarchically ordered, well-differentiated vessels (i.e. large vessels branch into smaller vessels), tumour blood vessels are chaotic and characterised by a loss of hierarchy. Tumour blood vessels are larger and varied in diameter than normal vessels, show irregular branching patterns and can exhibit shunts (i.e. vessel connections) between arteries and veins. Moreover, tumours exhibit regions of very high vessel density and others with very low vessel density [126]. Tumour blood vessels are also more tortuous ('bendy') than normal vessels (i.e. they coil) due to a lack of space [181]. We present examples of images of typical tumour vasculature in Fig.3.2.

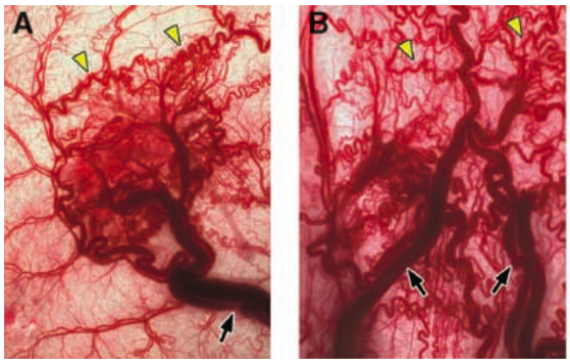


Figure 3.2: Images of typical tumour blood vessels. Tumour blood vessels are highly abnormal and exhibit large diameters (black arrows) and extreme tortuosity (yellow arrows). Image source: with permission from [181].

Tumour vasculature is highly dysfunctional as vessels tend to be very leaky and the direction of blood flow through them can vary over time [126]. Several math-

ematical models were developed specifically to study the influence of the structure of tumour blood vessels on their function. For example, in [17] Baish *et al.* used a mathematical model to show that the structural and functional characteristics of tumour blood vessels can have negative effects on nutrient supply and drug delivery. In particular, high tortuosity of vessels has been observed to reduce blood flow [199]. Mathematical modelling has further shown that the blood vessel structure determines oxygen availability in a tumour [128]. For radiotherapy on the other hand another mathematical model predicts the response of a tumour to be insensitive to different types of vessel network geometries in 3D [129]. The study compared different 3D vessel network geometries, both biologically and artificially derived, as well as 2D network representations and observed the changes in the viable fraction of cells in the surrounding tumour tissue during and after 5 rounds of simulated daily radiotherapy. In small tissue volumes, the cell growth and response to radiation therapy was very similar for the different vessel network geometries. However, the study assumed constant oxygen concentration in all parts of the vessel networks. Moreover, the model neglected the effects of vascular normalisation, induced for example by anti-angiogenesis drugs, even though they have been observed to improve the response to radiotherapy [88].

3.1.3 Anti-angiogenesis drugs and their effects on tumour vasculature

Anti-angiogenesis drugs block angiogenesis and inhibit vasculogenesis by targeting molecules such as VEGF [270]. While promising in pre-clinical studies, anti-VEGF therapy alone has shown little effect in the majority of cancers in clinical trials and prolonged the lives of patients only minimally [37, 141, 142, 262]. In many cases even after promising initial response to anti-angiogenic treatment, the cancer eventually progresses after a few months [37]. Cancers responsive to the therapy are well vascularised [270].

Interestingly, anti-VEGF therapy has been observed to improve the effect of classical chemotherapy when the two treatments are combined (see, for example, [137]). A possible explanation is that the vasculature undergoes a so-called *normalisation* process in response to anti-VEGF therapy, thereby losing many of its abnormal traits [126, 139, 140]. In particular, the vascular network appears to regain a clearer hierarchy of vessels, a more intact basal membrane and a more homogeneous blood flow. These changes are, in turn, thought to improve the delivery of oxygen and drugs to the tumour, and even to enhance the recognition of tumour cells by the immune system [141]. One of the current clinical challenges is, therefore, to identify anti-angiogenic drugs that cause permanent tumour vasculature normalisation [270]. There are, however, many less well understood effects of anti-VEGF drugs such as up-regulation of certain angiogenic pathways and even an increase in metastatic action [95].

3.1.4 Effects of radiation therapy on tumour vasculature

The effects of radiation therapy on the structure of tumour blood vessel networks are not well understood. Indeed, studies in the last 65 years are highly inconsistent [195]. In general, there seems to be a trend for fractionated irradiation therapy to result in a stabilisation of the vessel network structure in human tumours. For single-dose irradiation, a dose higher than 10 Gy has been observed to cause significant damage to vessel networks in human and mouse tumours [195]. Based on recent experiments however, it has been hypothesised that fractionated irradiation may cause more harm to the vessel network structure than previously thought [144]. Additionally, with novel technologies, it has become possible to target tumours with higher doses of radiation more specifically and thus researchers have become more interested in studying the effects of single-dose irradiation versus fractionated irradiation [144].

3.2 Spatial quantification of characteristics of tumour blood vessel networks

The quantification of structural characteristics of tumour blood vessels is of great interest for a number of reasons. It could help to determine whether cancer treatment can reach all parts of the tumour or whether vascular renormalisation has been achieved before classical chemotherapy is applied. As a longterm goal, this could enable prediction of a patients' response to chemotherapy or radiation treatment.

New imaging techniques, such as multispectral fluorescence ultramicroscopy (see, for example, Dobosz *et al.* 2014 [92]), intravital imaging (see, for example, Tozer *et al.* [255]), micro-computed tomography (see, for example, Ehling *et al.* 2014 [100]) or volumetric multispectral optoacoustic tomography (see, for example, Ron *et al.* [215]) allow detailed reconstruction of 3-dimensional vascular networks in tumours. From these reconstructions, parameters such as vessel length, vessel branching, vessel volume, vessel density, number of vessels, number of branching points, inter-vessel spacing, fractal dimension, and tortuosity can be extracted and summarised. Such structural parameters can reflect disease progress. For example, vessel size and vessel branching have been shown to correlate with tumour aggressiveness and angiogenesis [100]. Similarly, Bullitt *et al.* [58] find that tortuosity and aggressiveness of a tumour are closely related. Dobosz *et al.* [92] observe that vessel volume, the number of vessels, and vessel branching in the tumour periphery change significantly, when anti-angiogenic drugs are applied.

However, summary parameters are very coarse and do not exploit the 3-dimensional information available in the data. In particular, vascular abnormalities, such as loops and tortuosity, are not adequately resolved in a spatial manner. In order to use the full potential of the imaging techniques, a method is needed that can quantify these parameters mathematically while enabling spatial insights. Ideally, such a method

could be used to study the structural heterogeneity of blood vessels in a tumour and the effects of treatments in the clinic.

We now highlight existing approaches to quantify the structure of tumour blood vessels and then proceed to develop our own method based on PH.

3.2.1 Existing approaches for the quantification of tumour blood vessels

Tumour vascular networks are not tree-like structures; in particular, they are not rooted. This makes it impossible to apply metrics that were developed for the analysis of tree-like objects (see, for example, [107]). Existing methods for quantification rely on blood vessel metrics that can roughly be separated into two categories: direct measures such as averaged cell density, vessel lengths, diameters, tortuosity, and indirect measures such as oxygen distribution [262]. These metrics do not necessarily take global 3D structure of the vessel network into account. For example, although tortuosity measures were explicitly developed to obtain 3D information [57], these are currently used to provide summary statistics for the whole network. Moreover, these metrics can be sensitive to imaging methods. Vilanova *et al.* [262] emphasize the importance of vessel structure quantification for the validation of mathematical models with biological data and introduce a graph-based method: the graphs they construct consist of branching points as nodes and edges between them that represent capillaries weighted by distance. Considering graph descriptors such as the number of vertices, edges, anastomoses, loops, network efficiency, and connectivity the authors successfully quantify the structure and evolution of 2-dimensional tumour blood vessel simulations.

Other approaches that quantify blood vessel networks and account for shape but do not rely on a hierarchical network structure include fractal dimension [121], and multifractal analysis and lacunarity. For example, in [127], Gould *et al.* analyse 2D representations of 3D microvasculature networks from kidneys, the cortex, skin, and

thigh muscle to study how these networks occupy the space around them and how the gaps between them are distributed. While providing insight, these methods do not account for the 3D structure of the vessel networks.

3.2.2 Quantification of the characteristics of tumour blood vessel networks using persistent homology

Our aim is to use PH to characterise unique features of tumour blood vessels, in particular the occurrence of loops and their high degree of tortuosity. Since loops can be captured by PH in dimension 1 of any filtration[2] that uses the structure of the vessel network data, we focus on developing a filtration that will quantify tortuosity while providing an intuitive interpretation of the results in both dimension 0 and 1.

The data for our analysis consists of points sampled from networks of tumour blood vessels. We introduce the data sets in Section 3.3. We interpret the data as a spatial network in $\mathbb{R}^3$ (e.g. a collection of nodes with coordinates in $\mathbb{R}^3$ and edges between them). The nodes correspond to points sampled from tumour blood vessels in 3D, the edges correspond to physical connections between the points via branch segments. We sample multiple points from every branch (rather than just the end points) to capture the tortuosity of the vessels. In our analysis we are interested in structural aspects of these networks (in contrast to functional properties such as the quantification of blood flow or direction through a vessel). We refer to the networks as tumour blood vessel networks.

3.2.2.1 Filtrations for the study of spatial objects in biology

PH has been successfully used to study the structure of brain arteries [34], airways [33] and neurons [145]. The filtrations developed were able to quantify tortuosity and branching behaviour. PH has recently also been proposed to study the shape of individual blood vessels for diagnosis of vascular disease [185]. As we are interested

[2]See Chapter 2, Subsection 2.2.1 for definition.

in studying networks of vessels rather than individual vessels, we now review the methods employed by [33, 34, 145].

Bendich *et al.* [34] use PH in dimension 0 to study the tortuosity of brain arteries imaged by 3D Magnetic Resonance Angiography. They apply a 'sweeping plane' filtration that can be viewed as the stepwise sliding of a plane over a vessel network: In the first filtration step the full network is situated on one side of the plane. As the plane moves, it starts intersecting the network until eventually the whole network is located on the other side of the plane. The initially empty side of the plane thereby gives rise to a sequence of embedded objects that can be interpreted as a filtration of the network. The authors interpret the persistence of features in dimension 0 as a measure of the size of a bend in a brain artery: the larger a bend is, the longer it exists as an individual component in the filtration and the larger its persistence. They find that looking at vectors of medium-scale persistence values for every subject leads to a strong correlation of the vectors with the age of the subjects.

Belchi *et al.* [33] use the same filtration to study the geometric structure of airways. They slide a plane starting at the beginning of the trachea towards the bronchia. In this case, the dimension 0 PH is interpreted as a measure for branching behaviour of the airways. In a first approach, the authors use the number of bars in the dimension 0 barcode as an indicator of how often an airway branches 'upwards'. This simple measure outperforms other numerical markers in stratifying different subgroups of patients with chronic obstructive pulmonary disease, although classification based on this measure alone is not possible. In addition, using a version of the Wasserstein distance[3] on the 0-dimensional barcodes leads to a separation of inspiratory and expiratory scans of subjects. Although in this case the filtration was not used to quantify tortuosity, we note both these results since branching behaviour is also of interest for our tumour blood vessel network data.

[3]See Chapter 2, Subsubsection 2.4.1 for definition.

Finally, Kanari *et al.* [145] use a similar filtration to classify morphologically distinct groups of neurons. Instead of using a sweeping plane, they apply a filtration based on the radial distances of points on the neurons from the neuronal tree root. For trees with specified branch properties, this corresponds to shrinking a sphere with a decreasing radius around the tree root in every filtration step and including only nodes and edges that are outside the sphere in the growing graph. The filtration captures branching behaviour and spatial morphology in dimension 0. The authors base their classification on a distance measure that they define from the barcodes[4], as well as persistence images[5].

These examples motivate our approach.

3.2.2.2 Radial filtration for tumour blood vessel networks

In contrast to brain arteries or neurons, tumour blood vessels are not tree-like objects. They also do not have a natural orientation. Motivated by the fact that tumours are often viewed as spherical objects, we root our filtration in the tumour centre. Since we perform our analysis on the blood vessels rather than the tumour itself, we approximate the tumour centre by the centre of mass of the points sampled from the tumour blood vessels, e.g. the nodes of our networks. We then proceed in the following way. We search the neighbourhood of the tumour centre, increasing the radial distance stepwise to include all nodes within the radius. If two nodes that are connected by an edge are both within the given radius, we add the edge to our filtration. Fig. 3.3 shows a schematic representation of the radial filtration on blood vessel data. Based on this filtration we study the homology of the growing network at every filtration step, capturing tortuosity and branching behaviour in dimension 0, and loops in dimension 1.

[4]See Chapter 2, Subsection 2.2.2 for definition.
[5]See Chapter 2, Subsection 2.4.3 for definition.

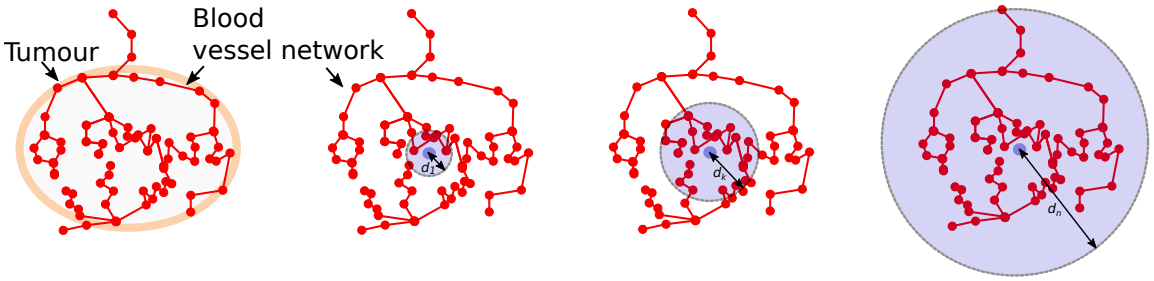


Figure 3.3: Schematic illustration of the radial filtration of a tumour blood vessel network. In the k-th filtration step we include all vessel nodes and edges that are fully contained in the ball of radius d_k around the centre of mass of the vessel points. Image source: [64].

The filtration generates topological information with respect to the tumour centre. We also obtain information about the heterogeneity of these characteristics within a single network, which is of particular interest since it has been observed that anti-angiogenic drugs affect the vessel network structure in the tumour periphery and the tumour core in different ways [92].

3.3 Data

For our analysis we use two different tumour blood vessel data sets: data obtained by multiphoton intravital 3D imaging [203] and data obtained by ultramicroscopy [93]. Both data sets consist of 3D stacks of images of tumour blood vessels subjected to different experimental conditions. We summarise the data in Table 3.1 and proceed to give a more detailed description of both data sets. We include only information which we consider relevant for the understanding of the data or our results. We also discuss the advantages and disadvantages of the two imaging methods. For a more detailed description of the experimental procedures we refer the reader to the references provided below.

3.3.1 Multiphoton intravital 3D imaging

Our first data set consists of tumour vasculature images that were obtained from multiphoton intravital 3D imaging [203] over several days while the animal was alive.

Data set	Type	Model	Experimental conditions
Multiphoton intravital 3D microscopy	Dynamic, over multiple days	Mouse colorectal cancer in mice	1. Control 2. Treated to increase sprouting 3. Treated to reduce sprouting (normalised vasculature) 4. Irradiated (single dose of 15 Gy) 5. Irradiated (fractionated 5×3 Gy)
Multispectral fluorescence ultramicroscopy	Static	Human breast cancer in mice	1. Control 2. Treated to slow growth of new blood vessels (normalised vasculature)

Table 3.1: Summary of data sets and experimental conditions.

We will refer to this data as *intravital data*. The experimental work was conducted by Bostjan Markelc and Jakob Kaeppler in the lab of Ruth Muschel at the *Oxford Radiation Oncology Institute, Department of Oncology, University of Oxford*. The data was obtained from transgenic mice that were injected with a murine colon adenocarcinoma MC38 tumour and surgically implanted a window chamber that enables intravital imaging of tumours [203, 213]. The mice were divided into groups that were subjected to different experimental conditions:

1. Controls (7 mice).

2. Anti-Dll4 treated tumours (3 mice): The mice were treated using anti-Dll4 antibodies which block Dll4 signalling, thus increasing vessel sprouting. The resulting networks are very dense and complex. The antibody was injected into the animal's belly every three days starting on the first day of imaging.

3. DC101 treated tumours (5 mice): The mice were treated using DC101 antibodies which block VEGFR-2 signalling and reduce vessel sprouting. The effect can

be interpreted as normalisation of the vasculature. The antibody was injected into the animal's belly every three days starting on the first day of imaging.

4. Single-dose irradiated tumours (7 mice): The mice were treated with a single dose of 15 Gy on the first day of imaging. The effects of irradiation therapy on vasculature are not well understood.

5. Dose-fractionated irradiated tumours (4 mice): The mice were treated with five doses of 3 Gy over 5 consecutive days followed by two days of rest starting on the first day of imaging. The effects of fractionated irradiation therapy on vasculature are not well understood.

The tumours were imaged from day 9 to day 14 and, in some cases, up to 20 days after tumour induction. The experimentalists began administering treatment once the tumours reached an approximate[6] size of 100 mm^3 and showed signs of vessel perfusion. In each case, we refer to the start of treatment as day 0 of treatment. We note that, by coincidence, the DC101 treated tumours tended to be larger in size on day 0 of treatment than the tumours of the other treatment groups. The vessels in all treatment groups were visualised using fluorescent staining on the endothelial cells and imaging was performed using a multiphoton microscope. For more details on experimental conditions we refer the reader to [28, 30, 31], where the data set has been used to develop an approach for feature extraction from vascular networks.

Multiphoton microscopy achieves high spatial resolution which allows detailed analysis of, for example, sprouting behaviour and vessel perfusion. In our case the resolution is 0.83 μm in the xy-plane with slices imaged every 5μm in the z-direction. The imaging method however has a small penetration depth. In practice this means

[6]The exact size of a tumour is impossible to determine during ongoing experiments, some variation of tumour sizes is therefore to be expected. The choice of size for the beginning of treatment was determined by the irradiation experiments: the effect of radiation treatment on tumours is very sensitive to initial tumour size, the experimentalists therefore aimed to start irradiation for tumours of size approximately 80 – 150 mm^3 [144].

that the imaging quality is very good close to the tumour surface but the dye fades when moving towards the centre of the tumour, enabling only a visible depth of around 300 μm. Since it is not possible to image the entire tumour through the window chamber, we only obtain the vessel network of a small segment of the tumour. It is also important to note that the presence of necrosis in the tumour has a negative effect on imaging quality. This was particularly evident for the anti-Dll4 treated tumours.

We obtained the data in the form of images generated by Bostjan Markelc and Jakob Kaeppler, and skeleton files produced by Russell Bates as part of his PhD thesis [28] at the *Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford*. The skeleton files were extracted by combining two segmentation models and taking their geometric average. The skeletons were then pruned (see p. 165 in [28] for a full description with references to the methodology). We note that the size of the images and skeletons in the xy-plane is considerably larger than in the z-axis [28].

3.3.2 Multispectral fluorescence ultramicroscopy data

Our second data set consists of multispectral fluorescence ultramicroscopy [93] images of blood vessels of human breast cancer tumours (cell line KPL-4, HER2 positive) that were implanted into 31 immunodeficient mice. We will refer to this data as *ultramicroscopy data*. The experiments were carried out by Michael Dobosz *et al.* [92], *Roche Diagnostics/Institute for Biological and Medical Imaging, Helmholz Zentrum, Munich*. The mice were divided into two groups that were subjected to different experimental conditions:

1. Controls (18 mice).

2. Anti-VEGF-A treated tumours (13 mice): The mice were treated with bevacizumab, an antibody which binds to VEGF-A and, thereby, induces normalisation of the vessel networks and reduces their permeability. Treatment was administered once the tumours reached a volume of approximately 60 mm^3.

To test the effect of the treatment on drug delivery at different time points, both controls and anti-VEGF-A treated mice were also treated with trastuzumab (anti-HER2 antibody) six hours before the tumour was extracted and prepared for imaging. Different subgroups of tumours were imaged on day 1 (5 controls, 5 treated), day 3 (5 controls, 4 treated), day 7 (5 controls, 2 treated), and day 14 (3 controls, 2 treated) after administration of bevacizumab. More details on experimental conditions can be found in [92] (note that the data set in [92] overlaps with the data used in this work, but they are not identical).

Multispectral fluorescence ultramicroscopy [92] is an *ex vivo* imaging technique that achieves high spatial resolution while maintaining high imaging depth. To avoid light absorption and scattering, tumour samples are treated chemically to make them optically transparent prior to imaging. The ultramicroscope sequentially illuminates different layers of the transparent tumour with a laser beam, the layers being perpendicular to the observation axis. It records images of the emitted fluorescence light from the tumour plane. In our data the resolution on the xy-plane is 5.1μm with an image taken every 5.1μm in the z-direction.

We obtain the data in image form as well as various outputs such as skeleton files and skeletonisations that were produced by Dobosz *et al.* [92] using a custom *Definiens* Developer script.

3.3.3 Synthetic hierarchical tree data

To evaluate whether there are PH features that represent characteristics of tumour vasculature, we compare these to features obtained from an extreme case of non-tumour vasculature: a hierarchical tree. Since we want to capture specific characteristics of the abnormality of tumour vasculature, i.e. tortuosity and loops, in a first step we need to understand which barcode features are simply a consequence of the filtration applied to a spatial network. This is in particular true for dimension 0 barcodes, where we expect tortuosity and potentially also branching behaviour to manifest itself. We choose a model that does not include tortuosity and shows regular branching to allow us to identify barcode features for the tumour vasculature that could be connected to these two properties. Our 'null-hypothesis' is that the barcodes that we obtain for tumour blood vessels are the same as the barcodes that we obtain from a synthetic hierarchical tree. The synthetic hierarchical tree data therefore serves as a type of initial 'null model'.

We use a simple model that captures regular branching based on the 2D fractal vasculature model introduced by Karshafian *et al.* [147]. This model includes information on vessel thickness and the computation terminates after a set diameter is reached by the daughter branches in the network. The model is based on very simple principles: Every branch of the network gives rise to two daughter branches, whose diameters D_i, $i = 1, 2$, and lengths L_{daughter} decrease with respect to the diameter D_0 and length L_{parent} of the parent vessel in the following way:

$$D_0^\gamma = D_1^\gamma + D_2^\gamma, \tag{3.1}$$

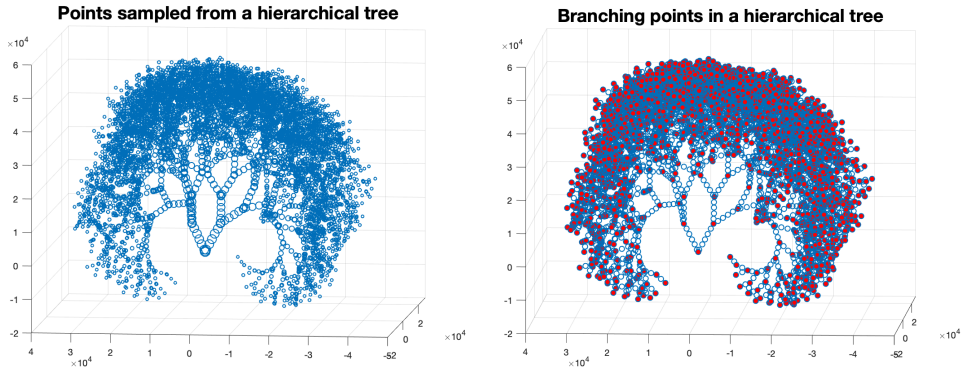
$$L_{\text{daughter}} = \kappa L_{\text{parent}}, \tag{3.2}$$

where γ is a constant that the authors set to be $\gamma = 3$ and the distance factor κ is given by $\kappa = 0.9$. The diameters of the daughter branches are further dependant on

the bifurcation index $0 < \beta < 1$:

$$\beta = \frac{D_1}{D_2}. \tag{3.3}$$

The daughter vessels branch away from the parent vessel at a branching angle θ, which is drawn uniformly at random from a prescribed interval. Karshafian *et al.* [147] modify the branching angles to obtain two different types of blood vessel networks: networks that share similarities with 'normal' vasculature such as found in kidneys, where $\theta \in [25.5, 28.5]$, and networks that share similarities with tumour vasculature, where $\theta \in [25, 140]$. We use Karshafian *et al.*'s 2D model for 'normal' vasculature to construct a 3D model by introducing a rotation angle ϕ for the daughter vessels around the parent vessel, which we draw uniformly at random from $[0, 360)$. In addition to the default parameters outlined above, we fix $\beta = 0.95$, an initial diameter of $D_0 = 500$ μm, an initial length of $L_0 = 1$ cm, and a termination diameter of $D_{\text{end}} = 40$ μm. We sample six points from every branch. A typical point cloud is presented in Fig. 3.4. Note that our definition of branching points includes vessel end points, as can be seen clearly in Fig. 3.4 (b).



(a) The size of the points is scaled according to the diameters determined by the model.

(b) We show branching points in red.

Figure 3.4: Points sampled from a 3D hierarchical tree based on Karshafian *et al.*'s 2D model of kidney vasculature [147]. Axes units are in μm.

3.4 Implementation

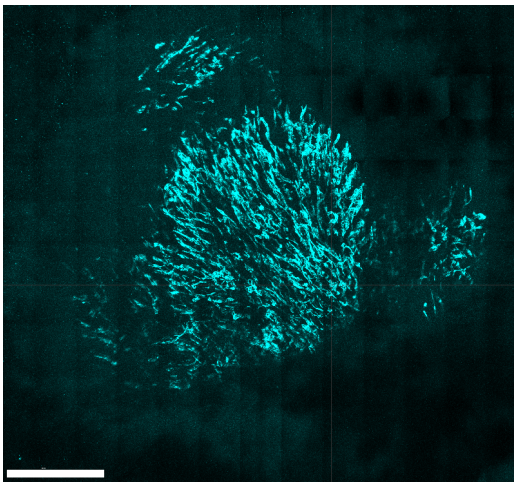
We give details on the implementation of the data preprocessing and the PH filtration.

3.4.1 Data preprocessing

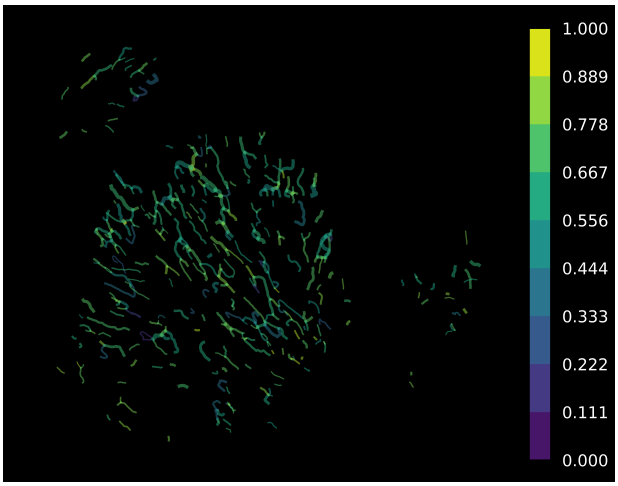
3.4.1.1 Intravital data

We write PYTHON scripts that use part of the PYTHON code package[7] UNET-CORE [29] developed by Russell Bates to extract blood vessel networks from skeleton files of microscopy images. The method we use is `VesselTree` from UNET_CORE.VESSEL_ANALYSIS. The networks consist of points on vessel branches (multiple points per vessel branch which also contain branching points) which represent the nodes, and the vessels between them that constitute the edges of the network. The method also enables one to extract network features such as diameter, length, and a measure of tortuosity (chord-length-ratio) for every branch point, although we are not using these features at the moment. We show a typical microscopy image and the different stages involved in data extraction in Fig. 3.5. As one can see in Fig. 3.5 (c), the different resolutions in the xy-plane and the z-axis do not appear to be accounted for in the extraction with UNET-CORE. Based on consultation with the experimentalists and further researchers who worked with the data, we account for this difference by rescaling the coordinates in the z-direction using the appropriate factor of resolutions before further analysis. For the following vessel networks it was necessary to reduce the point clouds for the computation of the radial filtration: control tumour 18_4E, day 17; control tumour 18_4E, day 18; control tumour 29_1B, day 15; control tumour 29_1B, day 16; control tumour 34_2A, day 14; control tumour 60_2A, day 14; DC101 treated tumour 51_2C, day 15; DC101 treated tumour 54_2D, day 17; anti-Dll4 treated tumour 24_2A, day

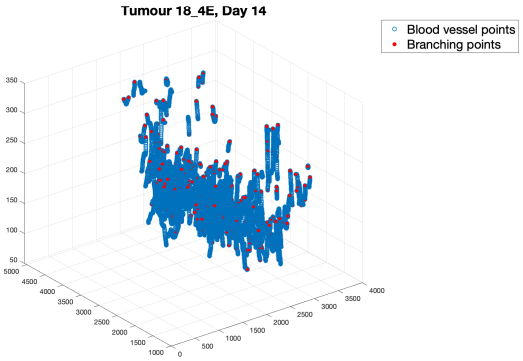
[7]Note that these codes have no documentation except for a list of required PYTHON packages. In addition to the required packages, we found that the codes only run with PYTHON 3 and require the networkx package to be version 1.9 (newer versions require extensive reprogramming of almost all functions and even then continue to cause issues that we were not able to resolve based on the error messages).



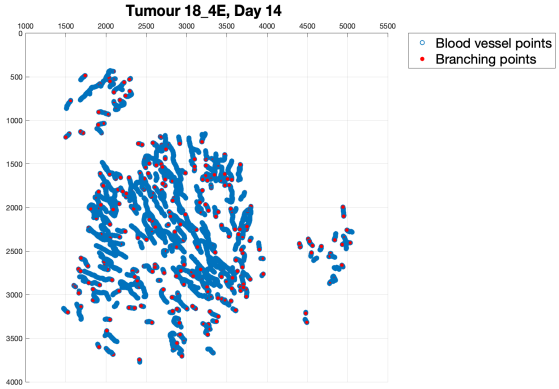
(a) Tumour blood vessels as seen under the microscope after endothelial staining. The grey bar corresponds to 1000μm. The image was taken by Bostjan Markelc and Jakob Kaeppler.



(b) Skeleton of tumour blood vessels coloured according to chord-length-ratio (clr) values. The skeleton and clr values were extracted by Russell Bates [28].



(c) Vessel points we extracted from the skeleton image using UNET-CORE [29]. Axis units are in μm.



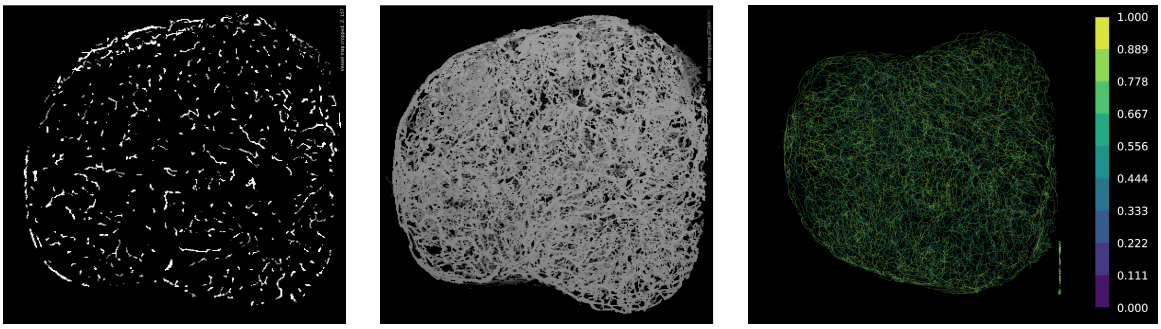
(d) Perspective corresponding to 2D images (a) and (b) on vessel points we extracted from the skeleton image using UNET-CORE [29]. Axis units are in μm.

Figure 3.5: Example images of control tumour 18_4E, day 0, multiphoton intravital 3D imaging data set. Image source: [64].

13; anti-Dll4 treated tumour 24_2A, day 14. The days listed here refer to the days after tumour induction rather than days after treatment. We reduced the number of points (i.e., the number of nodes of the networks) by including all branching points and only sampling every second point of every branch. We note that this can change the tortuosity in the reduced networks compared to the 'full' networks.

3.4.1.2 Ultramicroscopy data

We preprocess the grey scale skeletonisation files provided in the ultramicroscopy data set from individual `.tif` files (one for every xy-plane slice of the vessel network) to .tif stacks in `uint8` format using the software ImageJ [209]. We convert the `.tif` stacks to `.nii` format using the function `tiff2nii.m` from a Matlab toolbox [70] designed to aid preprocessing images for an image analysis software. We use the `.nii` files as image input for our python scripts for unet-core [29]. Even though unet-core was trained on multiphoton intravital 3D imaging we justify our approach by the fact that the skeletonisations are clear, high-contrast images. Any imaging specific effects were removed by the skeletonisation process which was developed specifically for this data set [92]. Visual inspection of the results that we obtained by applying unet-core to the skeletonisations suggest that the code extracts realistic vessel networks from the skeletonisations (see Fig. 3.6). We present examples of extracted



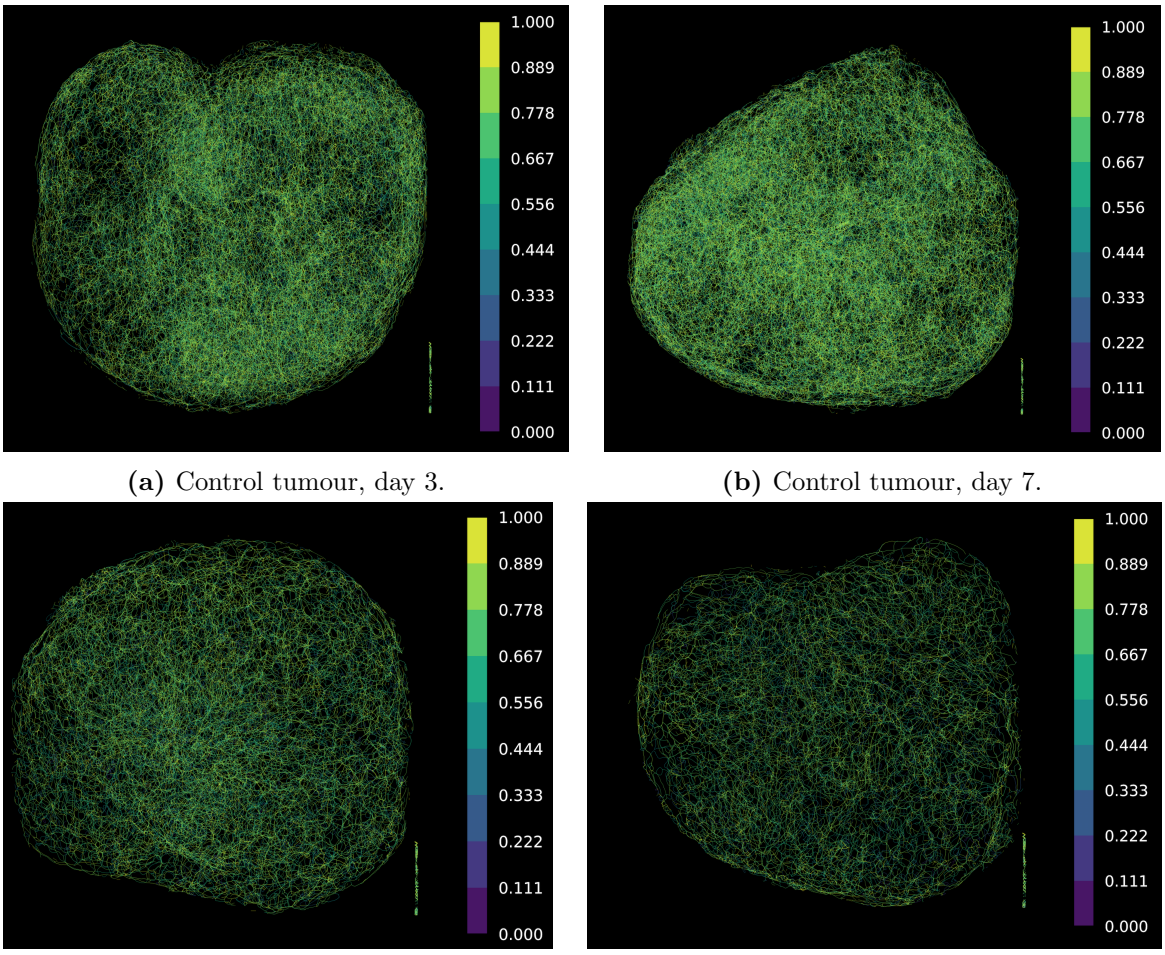
(a) Example of a slice of a skeletonisation.

(b) 3D projection of skeleton file in ImageJ.

(c) Example of extracted vessel network from skeleton file using unet-core.

Figure 3.6: Example images of the blood vessel network of anti-VEGF-A treated tumour 703, day 7, multispectral fluorescence ultramicroscopy data set. Note that the collection of lines in the bottom right corner of image (c) corresponds to text that was present in the skeleton images in the data set – visible in the upper righthand corner of images (a) and (b) – and was interpreted as a collection of vessels by unet-core. We removed these artefacts from our extracted point clouds manually. For technical reasons image (c) is axially rotated in comparison with images (a) and (b).

vessel networks in Fig. 3.7. We do not show the obtained point clouds in Matlab as these contain between around 0.5 to 2 million points and it is challenging to visualise these. Note that we do not know how unet-core determines the scaling of the data



(a) Control tumour, day 3.

(b) Control tumour, day 7.

(c) Anti-VEGF-A treated tumour, day 3.

(d) Anti-VEGF-A treated tumour, day 7.

Figure 3.7: Example images of extracted vessel networks from multispectral fluorescence ultramicroscopy data coloured according to chord-length-ratio (clr) values. We can see a clear difference between the vessel networks of the treated versus the untreated tumour on both day 3 and day 7 after treatment. Note that the collection of lines in the bottom right corner of the images corresponds to text that was present in the skeleton images in the data set. We removed these artefacts from our extracted point clouds manually.

points, but visual comparison with data points extracted by Franziska Mech, *Roche*, suggest that we obtain distances that scale linearly with the true distance in μm. Since we are only interested in features with respect to their relative distance to the tumour centre, this is sufficient. Due to the very high number of points, we reduce the point clouds for all tumours again by sampling all branching points but including only every second point from every branch (we refer to this as *half reduction*). For tumours where we are still unable to compute the full radial filtration, we run two separate

filtrations: one for the *tumour core*, which we define as the the region included in the sphere around the tumour centre with a radius corresponding to half of the maximal distance between the tumour centre and a node of the vessel network, and one for the *tumour periphery*[8], which we denote by the sphere around the tumour centre with a radius corresponding to the maximal distance between the tumour centre and a node of the vessel network minus the tumour core. Note that this approach affects the barcodes in dimension 0 and we lose up to 9% of the total number of loops[9]. In some cases, we further reduce the number of points in the networks by including all branching points but sampling only every fourth point from every branch (we refer to this as *quarter reduction*). Despite our reduction approaches, not all filtrations finish within reasonable time or are computable. We provide an overview of (to date) successful reduction approaches for each tumour in Table 3.2. For our analysis, wherever possible, we use the full filtrations rather than partial filtrations and consider results from the least reduced vessel networks.

3.4.2 Persistent homology

We implement the radial filtration in MATLAB and use the software package JAVAPLEX [247] to compute PH on our filtration. We divide the distance from the tumour centre (centre of mass) to the farthest away point in the blood vessel network into 500 steps to build the radial filtration.

[8]Note that our definitions of tumour core and periphery differ from Dobosz *et al.* [92]. The authors consider the radius of a fictional sphere with same volume as the tumour and define vasculature with a distance to the surface of the tumour smaller than 20% of this radius as 'periphery' and all other vasculature as being in the core.

[9]We draw this conclusion from observations on networks where we were able to obtain the full filtration.

Tumour	Full filtration	Tumour core	Tumour periphery
Veh 101.1d	quarter	half	
Veh 102.1d	quarter	half	half
Veh 103.1d		quarter	
Veh 104.1d	quarter	half	half
Veh 105.1d	quarter	half	quarter
Veh 301.3d	quarter	half	half
Veh 302.3d	quarter	half	half
Veh 303.3d		quarter	
Veh 304.3d			
Veh 305.3d		quarter	
Veh 301.7d		quarter	
Veh 302.7d	quarter	quarter	
Veh 303.7d		half	
Veh 304.7d			
Veh 305.7d		quarter	
Veh 403.14d		quarter	
Veh 404.14d			
Veh 405.14d		quarter	
Treat 201.1d	quarter	half	half
Treat 202.1d		half	half
Treat 203.1d		half	
Treat 204.1d		half	half
Treat 205.1d	half	half	half
Treat 401.3d		half	half
Treat 402.3d		half	half
Treat 403.3d		half	
Treat 404.3d	half	half	half
Treat 703.7d	half	half	half
Treat 705.7d	half	half	half
Treat 10-2.14d		quarter	
Treat 10-5.14d			

Table 3.2: Summary of reduction techniques that resulted in a finished computation of the radial filtration on the multispectral fluorescence ultramicroscopy data.

3.5 Results

We present our results from the radial filtration on both data sets. For the ultramicroscopy data, we are only able to obtain PH output for a subset of the tumours due to considerable computational complexity (see also Table 3.2 for included results).

Our approach is to identify interesting barcode features by visual inspection of the barcodes and then analyse some of these features in more detail. In contrast, in Chapter 4 we will use persistence landscapes and persistence images to enable analysis with a range of statistical and machine learning tools. We first examine the barcodes of the radial filtration for the synthetic hierarchical tree. We then show example barcodes from individual tumours from both data sets and determine those features which are likely to be associated with unique features of tumour blood vessels rather than a hierarchical tree. We then present an analysis of the number of loops on both data sets. Finally, we present results from dimension 0 PH output for the intravital data set.

3.5.1 Example barcodes for synthetic hierarchical tree

We show a barcode for the hierarchical synthetic tree data in Fig. 3.8. The scale on the horizontal axis is in μm. The distance from the centre of mass of the vessel points to the furthest point in the vessel tree data is approximately 5.25 cm. As expected, there are no features in dimension 1 as the tree does not contain any loops, so we only show the dimension 0 barcode. In dimension 0, we observe several persistent features that correspond to branches located less than 5000 μm, i.e. 5 mm, from the centre of mass. We also observe several short-lived components that seem to merge into one of the existing branches very quickly as the filtration radius increases. In almost regular steps of approximately 5 mm the more persistent branches merge into one another until after 4.5 cm we have one connected component.

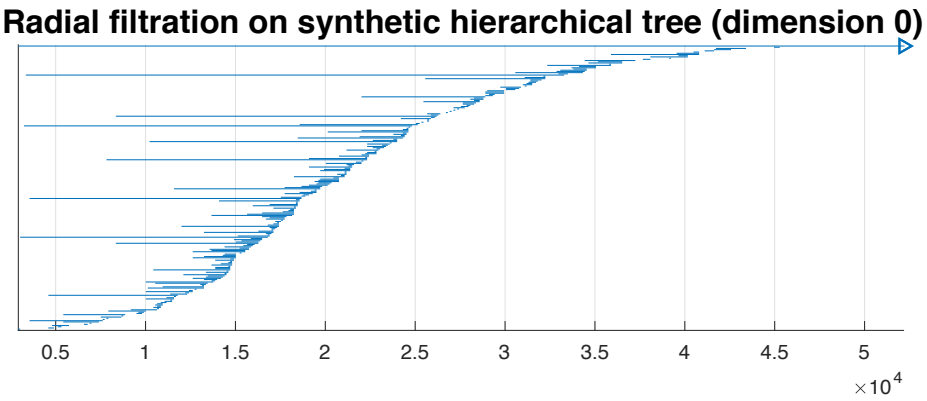
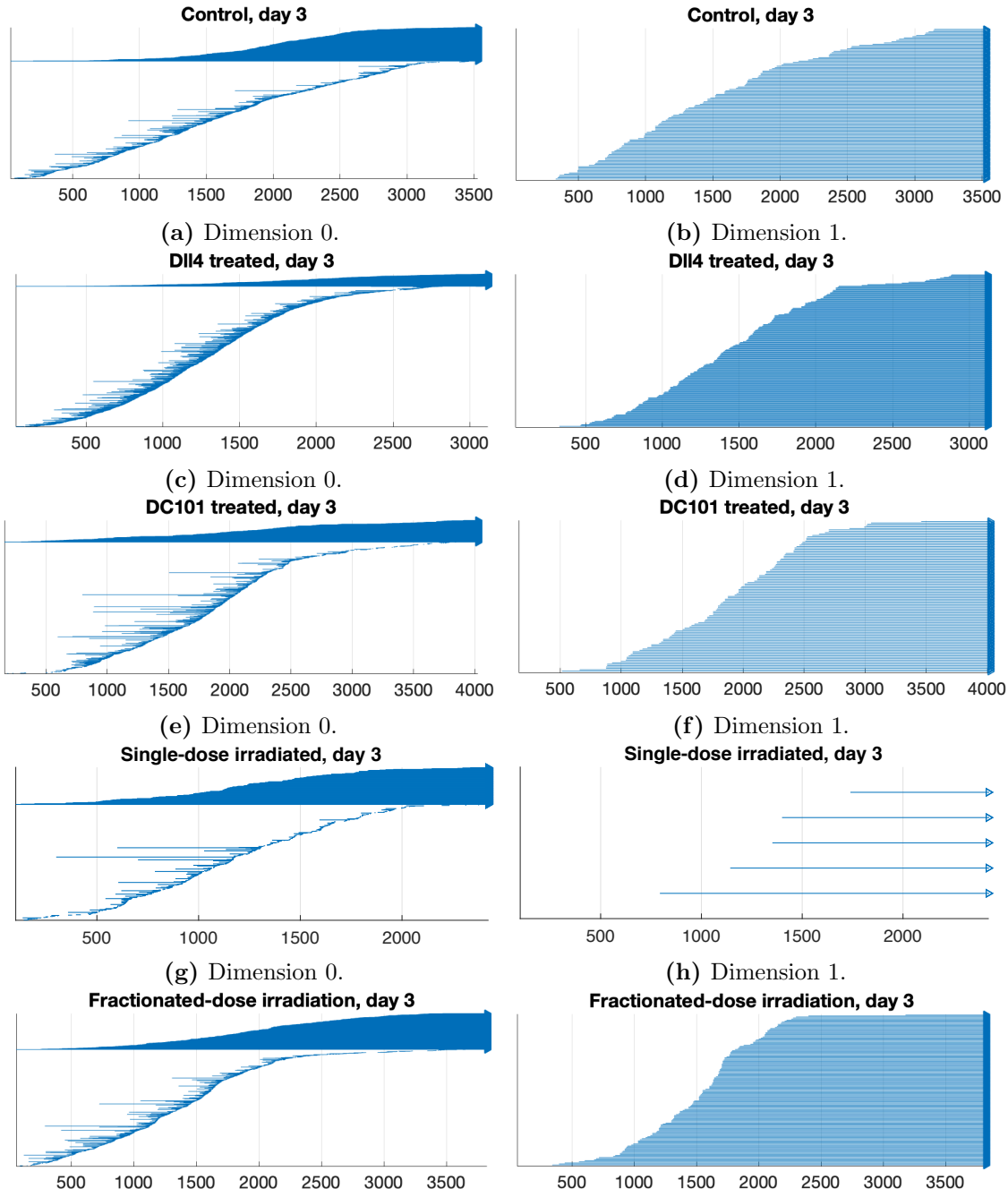


Figure 3.8: Dimension 0 barcode obtained from the radial filtration on a synthetic hierarchical tree. The horizontal axis shows the distance to the tumour centre in μm.

3.5.2 Example barcodes for intravital data

Fig. 3.9 shows example barcodes obtained from the radial filtration for five tumour blood vessel networks and different treatment regimes. For all tumours we show day 3 after treatment (day 3 of observation for the control tumour).

We note that the barcodes all end at different distances from the tumour centre, which reflects differences in tumour and/or vessel network sizes. In our description, we consider the persistence of features relative to the maximal filtration values. In all cases, the barcodes in dimension 0 differ from the barcodes for the synthetic hierarchical tree: When ignoring infinitely persisting bars, most barcodes show few (or no) persistent features that start early in the filtration. Only networks treated to decrease sprouting and treated with single-dose irradiation exhibit persistent bars, although these do not merge with existing features and/or die in such regularity as in the synthetic tree. DC101 treatment is known to normalise tumour blood vessel networks, so the presence of more persistent bars could indicate a less abnormal vessel network. The barcode for the vessel network treated with anti-Dll4 which increases vessel sprouting supports this hypothesis. Similarly, the control vessel network does not feature persistent bars. The fractionated-dose irradiated tumour seems to exhibit mesoscale persistent features, somewhere between the controls and the decreased



(a) Dimension 0. **(b)** Dimension 1.

(c) Dimension 0. **(d)** Dimension 1.

(e) Dimension 0. **(f)** Dimension 1.

(g) Dimension 0. **(h)** Dimension 1.

(i) Dimension 0. **(j)** Dimension 1.

Figure 3.9: Example barcodes for the radial filtration performed on each of the five treatment regimes: (a) Anti-Dll4 treated (increased sprouting), (b) DC101 treated (decreased sprouting), (c) single-dose irradiated, (d) fractionated-dose irradiated. The horizontal axis shows the distance to the tumour centre in μm. All vessel networks were imaged three days after treatment was administered. For all cases the blood vessel networks differ in size, which is reflected by the different maximal values on the horizontal axis.

sprouting case.

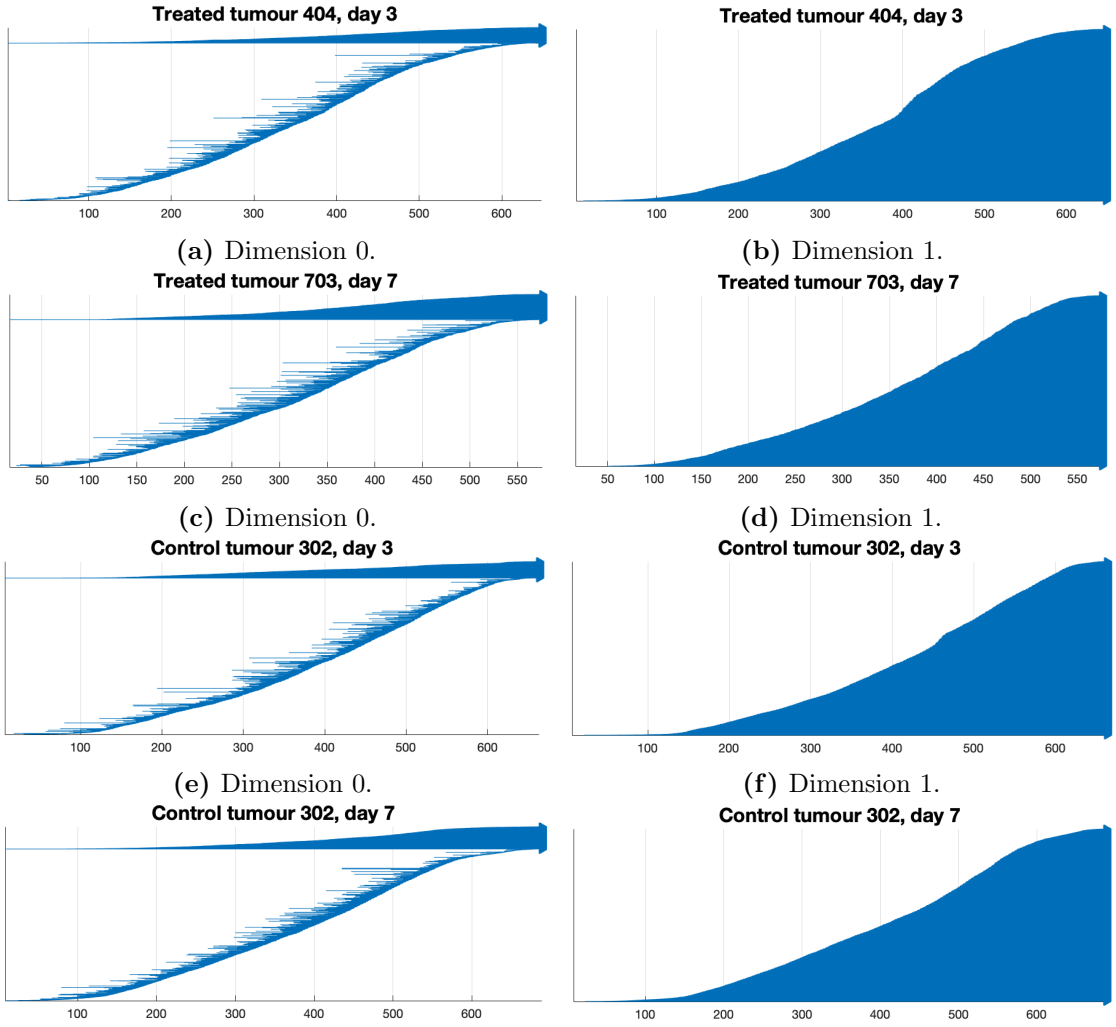
Even though there seem to be differences in the number of infinitely persisting bars in dimension 0, we do not consider these because the number of connected components of the vessel network can be influenced by image segmentation and can not be used as a definitive measure of the number of vessels growing into a tumour.

We also observe differences between the dimension 1 barcodes for the different treatment regimes. For example, the total number of vessel loops seems to differ, most prominently for the irradiated tumour where we find only 5 loops in the blood vessel network, while the other four tumours exhibit more than 100 loops. However, we find that the small number of loops of the irradiated tumour is not representative for the full group of irradiated tumours, in fact there is large variation in this group (see later in Fig. 3.11). There also seem to be slight differences in the distribution of vessel loops with respect to the tumour centre. In particular for the fractionated-dose irradiated tumour, we see an increase of loops at distances 1700 μm to 2000 μm from the tumour centre. These may be caused either by an increase in the number of loops within this particular annulus or a change in the shape of the tumour (for example a lump on an otherwise spheroid shape).

3.5.3 Example barcodes for ultramicroscopy data

We show example barcodes for the radial filtration performed on the ultramicroscopy data in Fig. 3.10. Again, the barcodes look very different compared to the barcode of the synthetic hierarchical tree in Fig. 3.8. For the treated tumours, we see only short to medium-scale bars in dimension 0, and no bars are particularly persistent. From day 3 after treatment to day 7 after treatment, the number of medium-scale persistent bars seems to increase. In contrast, for the control tumours, both on day 3 and on day 7 we observe less medium-scale bars than for both treated tumours. The overall persistence of bars in the controls seems to decrease from day 3 to day

7. In dimension 1, we can see an increase in the number of loops at a distance of



(a) Dimension 0. **(b)** Dimension 1.

(c) Dimension 0. **(d)** Dimension 1.

(e) Dimension 0. **(f)** Dimension 1.

(g) Dimension 0. **(h)** Dimension 1.

Figure 3.10: Example barcodes for the ultramicroscopy data. The horizontal axis reflects the distance to the tumour centre.

approximately 400 μm from the tumour centre three days after tumour treatment. Otherwise, the shape of the dimension 1 barcodes does not appear to be markedly different for the different tumours, although we note that there are substantial differences in the total number of bars, i.e. 6971 bars for the treated tumour, day 3; 4396 bars for the treated tumour, day 7; 13176 bars for the control tumour, day 3; and 19290 bars for the control tumour, day 7.

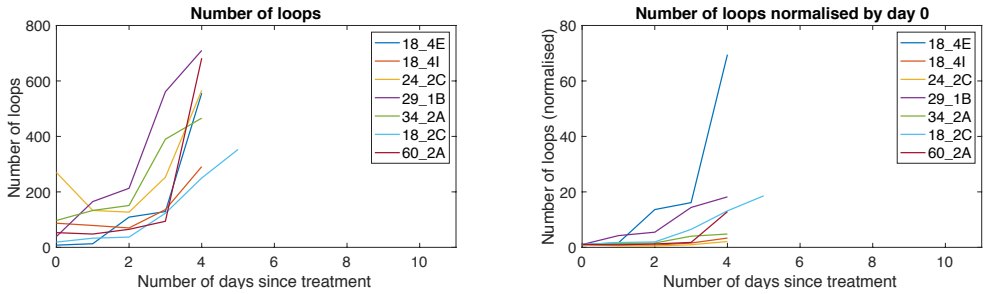
3.5.4 Number of loops and their distribution

We now investigate the dimension 1 barcodes from the radial filtration more closely for both data sets. These barcodes give information on the number of loops and their spatial distribution with respect to the tumour centre.

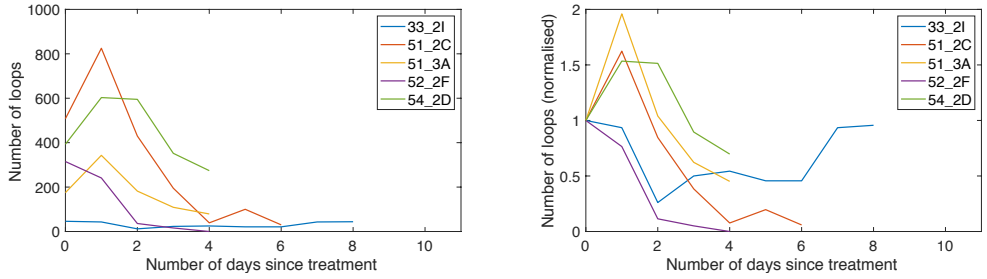
3.5.4.1 Intravital data

We first observe the total number of loops in the vessel networks and how it changes over time in every tumour, see Fig. 3.11 (left panel). To accommodate for different starting conditions of networks on day 0 of treatment/observation, we normalise[10] the average number of loops by day 0 in Fig. 3.11 (right panel). We observe considerable variation of the number of loops within one data category and between different data categories on all days of the experiments. Even after normalisation, the magnitude of the changes in the number of loops for the fractionated irradiation and DC101 treated groups is considerably lower than in the other data categories. Within the control group, the number of loops in the networks seems to increase similarly over time, with the exception of tumour 18_4E, whose number of loops increase very drastically from day 3 to day 4 as can be seen in the normalised plot in Fig. 3.11 (a). All anti-Dll4 treated tumours exhibit a noticeable increase in loops on day 3 or 4 after treatment, afterwards, the number of loops remains stable in two cases while it decreases in one case. After treatment with DC101 all tumours experience a decrease in the number of loops either on day 1 or on day 2 (in one case day 3). This could be caused by two subgroups of tumours - one fast responding and one slow responding – although there is not enough data to support this hypothesis. Fractionated- and single-dose irradiation seem to result in varied effects, in most cases stabilising the number of vessel loops. For fractionated irradiation this is visible with a time lag of four days.

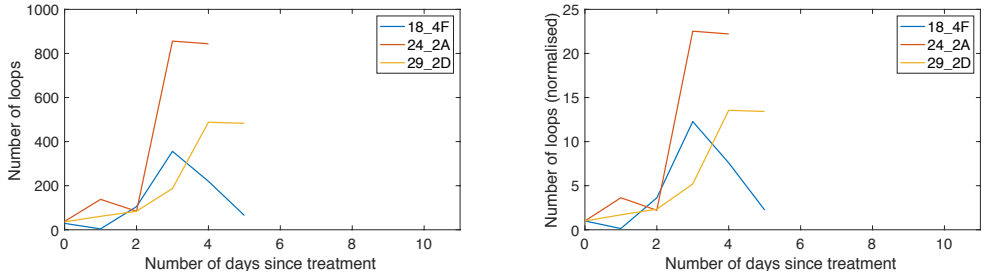
[10]This is a common strategy when working with time-dependent biological data and in particular [28,38] also applied normalisation to other vessel specific summary measures on the same data set.



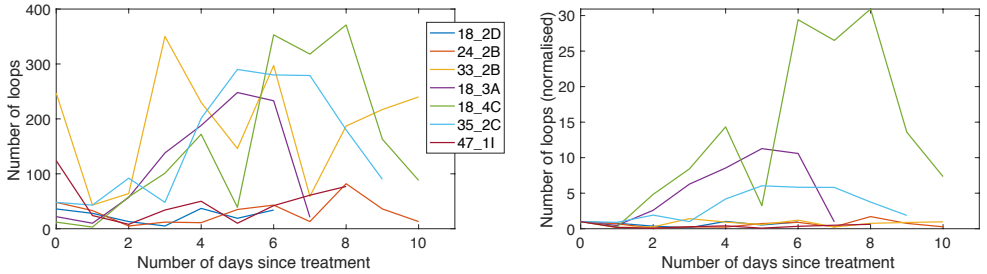
(a) Controls.



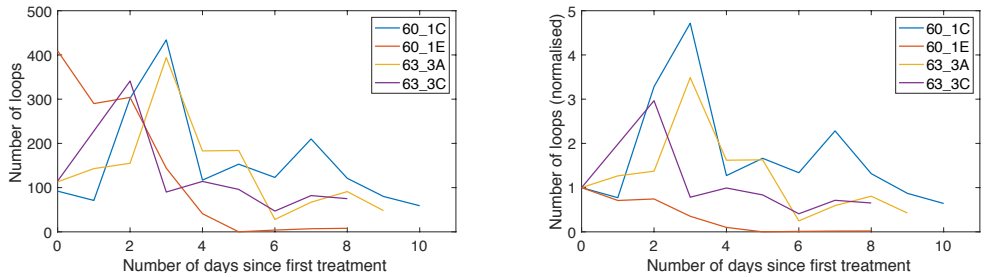
(b) DC101 treated.



(c) Dll4 treated.



(d) Single irradiation dose.



(e) Fractionated irradiation.

Figure 3.11: Number of loops for every observation day sorted by data category in the intravital data set. We show the total number of loops (left panel) and the number of loops normalised by the number of loops on day 0 (right panel).

After consultation with the experimentalists as well as inspection of the imaging data, we exclude the following data from further analysis: control tumour 24_2C day 4 (laser died, although we do not see this effect); anti-Dll4 treated tumour 18_4F day 5 (necrosis affects imaging quality); single-dose irradiated tumours 35_2C (strange growth pattern) and 47_1I (incomplete images due to technical problems); fractionated-dose irradiated tumour 60_1E day 5 onwards (bad quality of segmentation). In addition, we only include data up to day 4 for controls and anti-Dll4 treated tumour and data up to day 8 for all other categories to avoid artefacts in the trends that are caused by a rapid decline in the number of data points. For comparison, we present results using the full data set (without the above stated exclusions) in Appendix B, Subsubsection B.1.2.1.

We first study the total number of loops in the networks. Figures 3.12 (a) and (b) show how the average number of loops in the blood vessel networks of the different data categories change over time after treatment (or observation in the case of the controls). We present the same results using the median in Figures 3.12 (c) and (d). We first consider the results based on the raw data in Figures 3.12 (a) and (c). In particular in Figure 3.12 (c), we can see trends in the behaviour of the median for the different data categories, that match those we would expect from DC101 treatment (reduced sprouting) and anti-Dll4 treated (increased sprouting): on day 2 after treatment the number of loops in the networks treated with DC101 sprouting begins to decrease rapidly, while the number of loops in the networks treated with anti-Dll4 begins to increase rapidly on day 3. These effects continue over the remaining days of observation. The time lag between the beginning of treatment and its effects on the vessel networks being visible can be explained by the fact that the antibody treatment is administered to the animal's belly rather than directly to the tumour and first needs to spread in the system to reach the tumour blood vessels (also the vessels need to grow/die which takes time). For the anti-Dll4 treated networks, the response

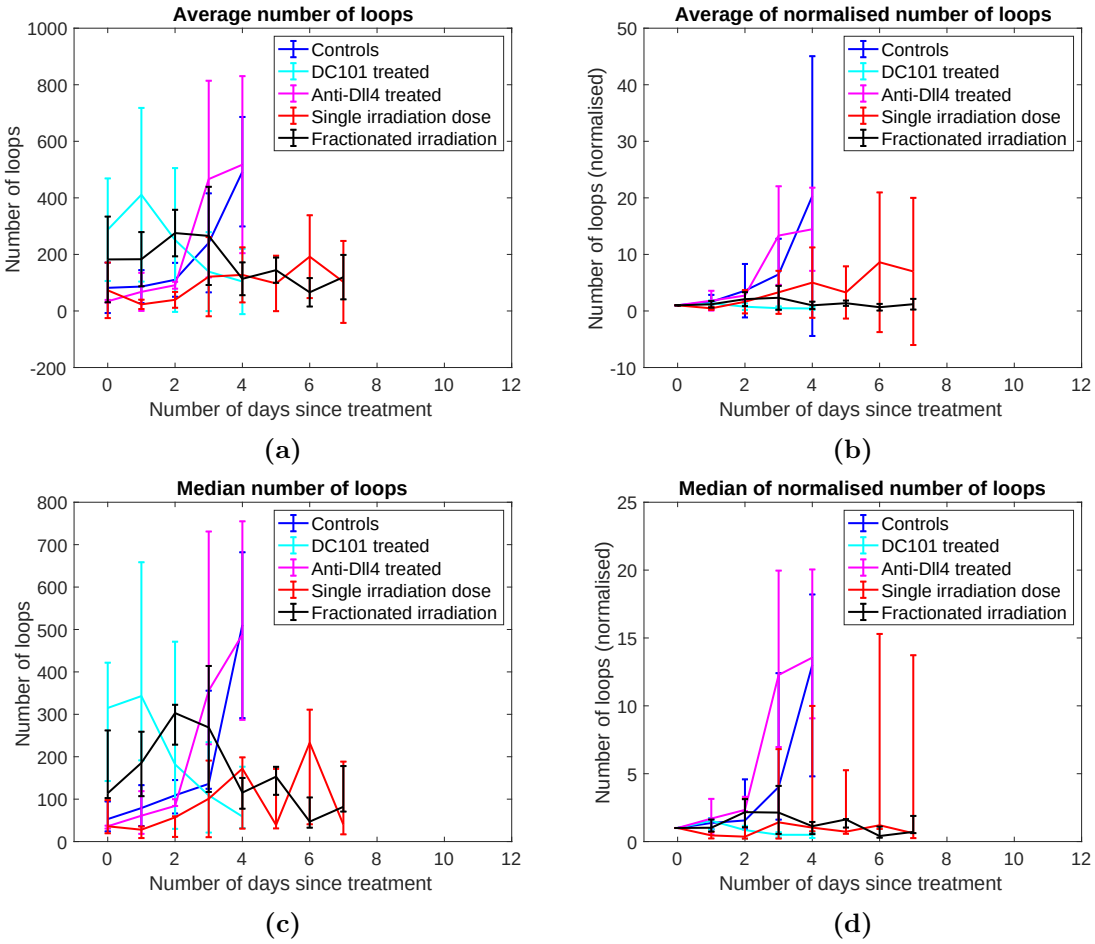


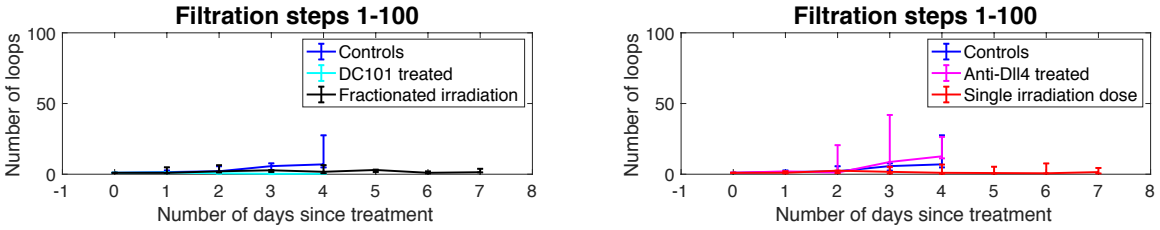
Figure 3.12: Total number of loops captured by the radial filtration in dimension 1 in the intravital data set. We show the average and standard deviation (top row) and the median and quartiles (bottom row) of the number of loops by treatment group. We show the total number of loops (left panel) and the number of loops normalised by the number of loops on day 0 (right panel).

time takes longer as the treatment first leads to the formation of vessel sprouts, which only start forming loops after an initial growth period. For the controls, we also observe an increase in the number of loops over time. However, we cannot completely differentiate this response from the increase in the anti-Dll4 treated group due to the large standard deviation and inter-quartile distance respectively. For single-dose irradiation the number of loops stabilises in the days immediately following treatment. At longer times (day 4 onwards), we see wide variation in the number of loops in the irradiated tumours. Our findings are consistent with a trend observed in experiments where single-dose irradiation strongly affects small, proliferating vessels [144], which

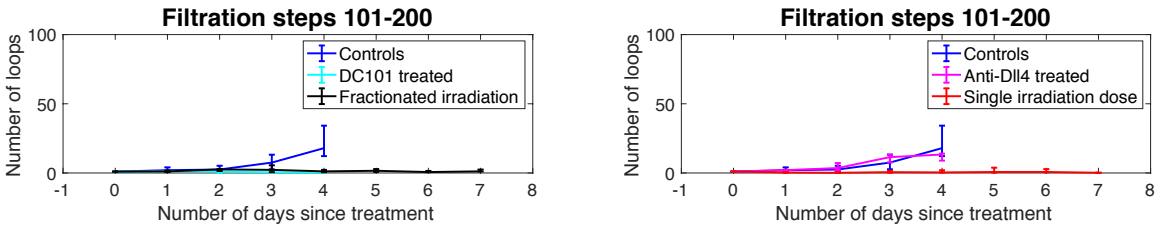
are also more prone to form loops, and results in a stable vessel structure. The effect of fractionated-dose irradiation on the number of loops is observable from day 4 of the treatment onwards when it manifests itself in the stabilisation of the number of loops.

To accommodate for different starting conditions of networks on day 0 of treatment, we consider the average and median of the normalised number of loops (by day 0), see Figures 3.12 (b) and (d). We observe the same trends as for the non-normalised cases, but we also see that the variability in the responses within the different treatment groups measured by both the standard deviation and the interquartile distance is very big, in particular for the controls, the anti-Dll4 treated tumours (note that the observations are based on only three tumours in this case), and the single-dose irradiated tumours. The effects of DC101 treatment and fractionated-dose irradiation over time seem to be more uniform within the groups. Although separation of the groups according to the number of loops in the vessel networks is not possible, our barcodes seem to capture the biological responses of the vessel networks to different tumour treatments.

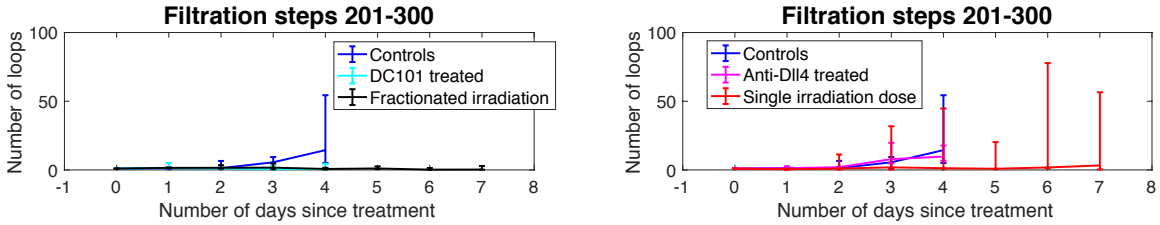
We now investigate how the different treatments affect the number of loops in different parts of the vessel network. In Fig. 3.13 we consider how the median of the normalised number of loops changes over time in different radial shells of the vessel networks with respect to the tumour centre. We divide the radial distance from the tumour centre to the maximal distance of a vessel point to the tumour centre into five consecutive intervals of equal length. In consequence, we normalise all tumours to the same size. We count the number of loops within each interval. We will refer to filtration steps 1–100, which are closest to the tumour centre as radial interval I, filtration steps 101–200 as radial interval II, filtration steps 201–300 as radial interval III, filtration steps 301–400 as radial interval IV, and filtration steps 401–500 as radial interval V. For the controls, we see an increase of loops that manifests itself the most



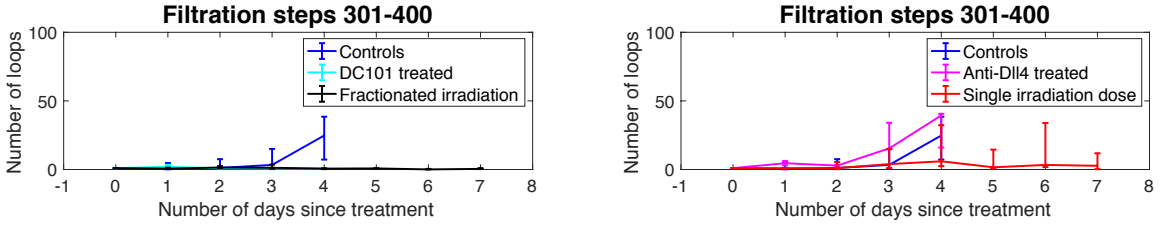
(a) Radial interval I.



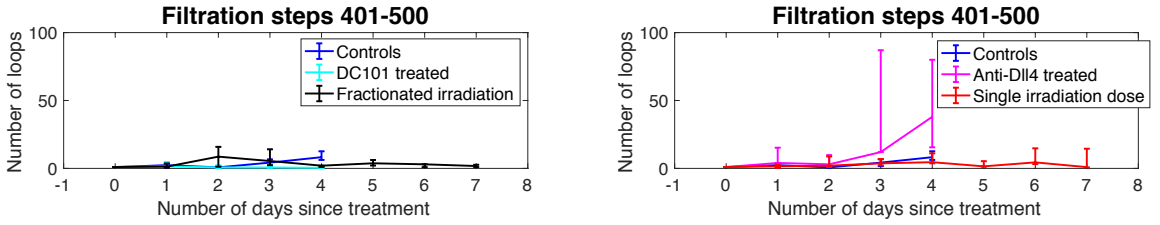
(b) Radial interval II.



(c) Radial interval III.



(d) Radial interval IV.



(e) Radial interval V.

Figure 3.13: Median of the normalised number of loops for different filtration intervals in the intravital data set. Filtration steps 1–100 correspond to the radial region closest to the tumour centre, while filtration steps 401–500 represent parts of the vessel network that are farthest away from the tumour centre. We perform normalisation with respect to the number of loops in the specified filtration interval on day 0. We separate the treatments into two groups to facilitate the distinction of the trends.

in radius intervals II-IV, in particular on day 4 (see Fig. 3.13 (b) – (d)). For anti-Dll4 treated tumours, we observe a significant increase in the number of loops on days 3 and 4 in radial interval V (i.e. closest to the outside of the tumour), particularly compared to the controls (see Fig. 3.13 (e)). Interestingly, the effect of single-dose irradiation seems to vary markedly between different tumours in radial intervals III and IV from day 3 onwards, whereas in comparison the number of loops remains more stable at the centre of the tumour and its outer rim (see Fig. 3.13 (c) and (d)). The DC101 treated tumours respond on day 3 in radial intervals I,II, and V, and in all intervals on day 4. Thus, for DC101 the effects on the vessel network seem to be more homogeneous than for the other treatments. Fractionated-dose irradiation significantly reduces the number of loops in all radial intervals from day 4 onwards, except for radial interval I. We also observe that fractionated-dose irradiation seems to induce variability close to the outside of the tumour (in radial interval V) on days 2 and 3 of treatment (see Fig. 3.13 (e)).

Overall, we find that the number of loops in the tumour blood vessel networks follows trends that align with the biological processes that underly the different treatments. Further, our analysis reveals that the different treatments have different effects on vessels depending on their location in the tumour.

3.5.4.2 Ultramicroscopy data

We now study the number of loops in the ultramicroscopy data set. Here, we only have one time point per tumour, thus a normalisation by day 0 of treatment is not informative. Due to the large number of points, i.e. 0.5 to 2 million points, in these vessel networks, we can only present partial results[11]. We show the median number of loops for different time points after treatment in Fig. 3.14. To compute the median we include all vessel networks for which we have either results from the full filtration,

[11]The results from the radial filtration that we show required a total of 2 months of parallel computation on a machine with 768 GB of RAM and 2×8 cores of 3.3 GHz.

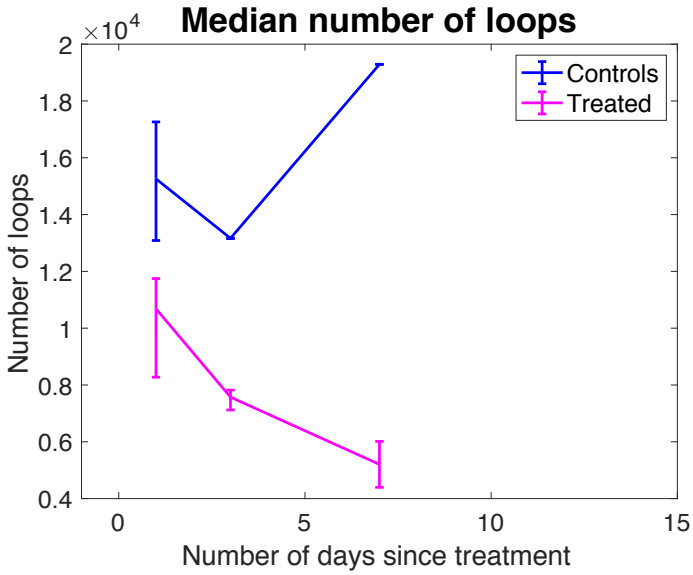


Figure 3.14: Total number of loops captured by the radial filtration in dimension 1 in the ultramicroscopy data set. We show the median and quartiles of the number of loops for controls versus bevacicumab treated tumours.

or results from both the tumour core filtration and the tumour periphery filtration. For the latter cases, we approximate the total number of loops by adding the number of loops from the tumour core and the tumour periphery filtration. We estimate that the difference between the approximated total number of loops and the total number of loops in the full filtration is at most 9% (based on the 10 cases where we are able to compute the full filtration and both partial filtrations).

We observe that the vessel networks in the ultramicroscopy data set are much more dense than in the intravital data set; this is reflected in the much higher number of loops (up to 20000 loops in contrast to under 900 loops in the intravital data) and the significantly longer computational time (on the order of months rather than weeks). Even though we only have partial results for three out of the four time points in the data set, we observe that for the treated tumours the median number of loops seems to steadily decline. We see a clear difference in the number of loops already one day after treatment. For both tumour groups there is only one computation missing for day 1 after treatment, so this difference is likely to remain once we obtain the full

results. Even after seven days, the treatment appears to still have effects on the number of loops in the vessel network. For the controls, we only have results for two data points (out of five) for day 3, versus three out of four for treated tumours, and one out of five for day 7, versus two out of two for treated tumours. But, thus far, it appears that the number of vessel loops in the controls for these days remains higher than for the treated tumours. The fact that we are able to compute the radial filtration on markedly more tumours in the treated group than in the control group also strongly suggests that the vessel networks in the controls are more dense than the treated networks at all stages of observation. We note that Dobosz *et al.* [92] did not see a significant change in the tumour volumes of treated tumours compared to the controls, the computational complexity is therefore directly influenced by properties of the tumour vasculature.

We consider the number of vessel loops in different filtration intervals in Fig. 3.15. Since, to enable computation, we divide the radial filtration into two parts (with few exceptions), the tumour core filtration and the tumour periphery filtration, we consider these two intervals and include all available results. In the tumour core, we see a decline in the number of loops during days 1, 3, and 7 for treated tumours. For these days, the results in the tumour core include all treated tumours in the data set. We observe a steep increase in the number of loops on day 14 after treatment in the tumour core for the treated tumour on which computations finished (one out of two). For the control tumours, a trend in the tumour core is less clear, but for the data points where we were able to obtain results it seems that there is at least a difference in magnitude of the number of loops in comparison to the treated tumours. Interestingly, there is hardly any change for the number of loops from day 7 to day 14 of observation in control tumours in the tumour core (for both time points only one computation is missing). For the tumour periphery, the trends in our available results seem markedly more clear with the treated tumours showing a decline of loops

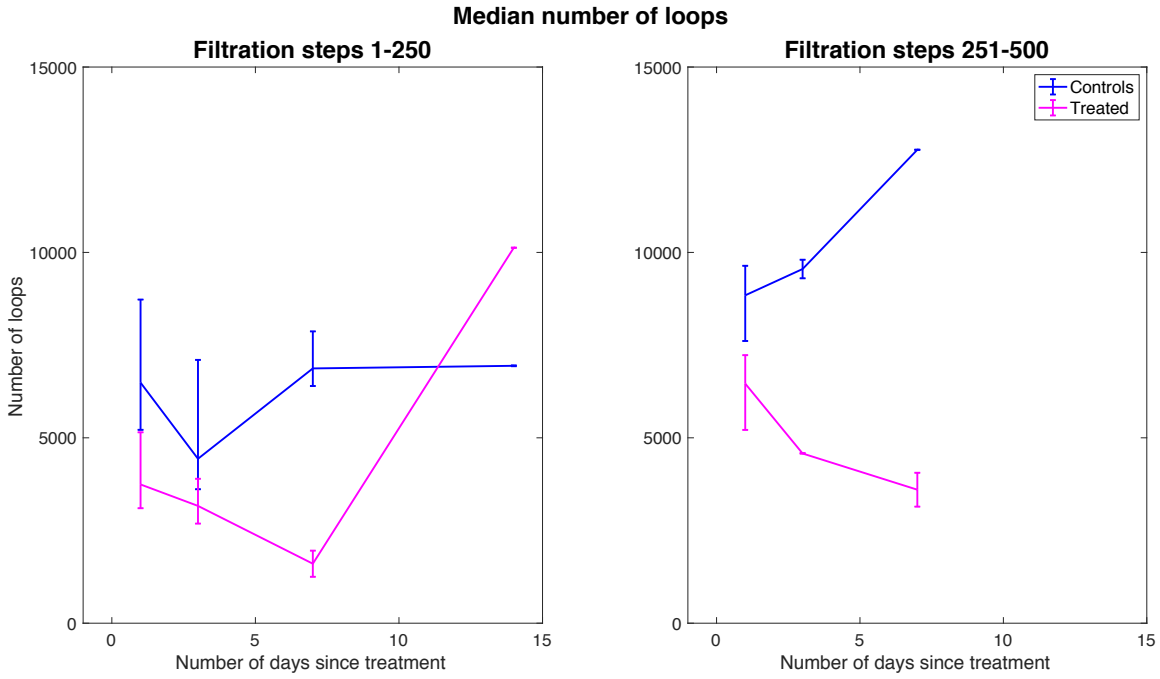


Figure 3.15: Median number of loops for different filtration intervals in the ultramicroscopy data set. We show the median and quartiles of the number of loops for controls versus bevacicumab treated tumours. Filtration steps 1–250 correspond to the radial region closest to the tumour centre (tumour core), filtration steps 251 – 500 represent the radial region closest to the surface of the tumour.

over time and the control tumours exhibiting an increase in the number of loops.

Finally, we consider the distribution of the loops in the tumours in Fig. 3.16. For all tumours, only a few loops are located in the tumour centre and in most cases the number of loops close to the outside of the vessel networks is also low. So far, the tumours for which we were able to obtain a full radial filtration[12] seem to suggest that the distribution of vessel loops with respect to the distance from the tumour centre seems to change over time for the treated tumours up to day 7 after treatment while for the untreated tumours a trend is more difficult to identify. The distribution of loops for the tumour on day 14 after treatment seems markedly different in comparison to earlier time points.

Overall, preliminary observations suggest that the number of loops is a good

[12]I.e., treated tumours 201, 205, 404, 703, and 705 and control tumours 101, 102, 104, 105, 301 (day 3), and 302 (days 3 & 7).

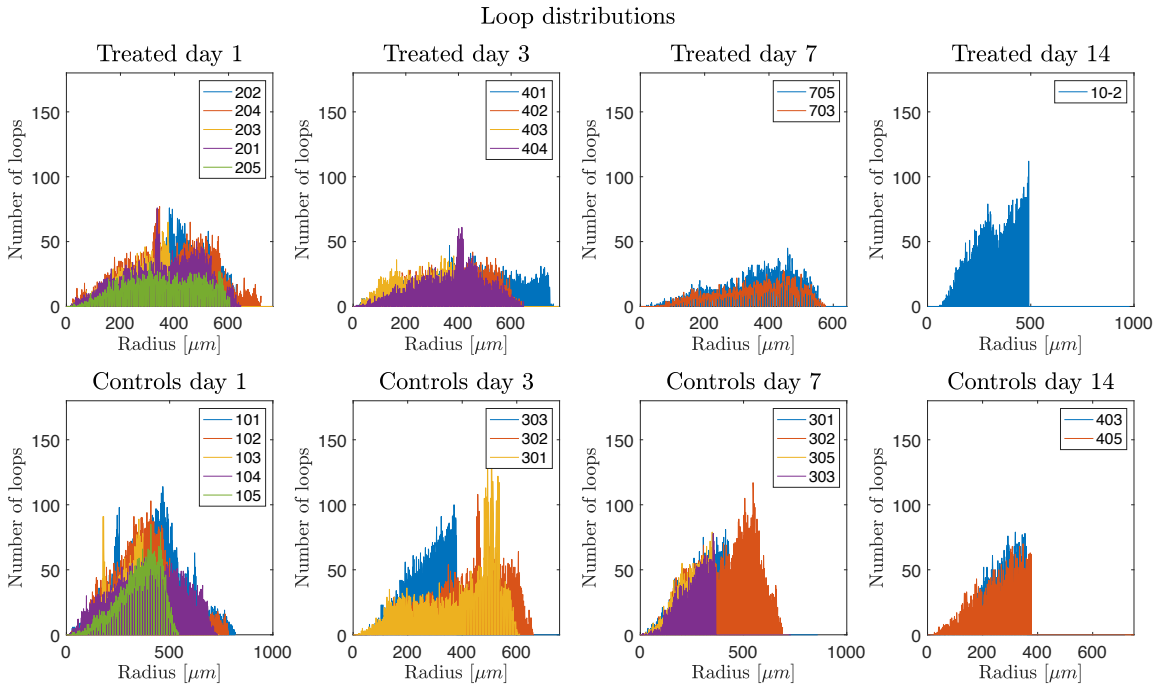


Figure 3.16: Distribution of loops for every observation from the ultramicroscopy data set. The horizontal axis represents the distance from the tumour centre, the vertical axis corresponds to the number of loops in the vessel networks. Note that, with exception of treated tumours 201, 205, 404, 703, and 705, and control tumours 101, 102, 104, 105, 301 (day 3) and 302 (days 3 & 7), the distributions are approximated by adding the distribution of the tumour core filtration and the tumour periphery filtration where possible. In cases where the distribution is constantly zero over either the tumour core or periphery or both, the computation of the radial filtration has not yet finished.

measure for structural changes in vessel networks treated with bevacicumab in the ultramicroscopy data set.

3.5.5 Dimension 0 features in the intravital data

We investigate whether the radial filtration can capture tortuosity in dimension 0 by looking at the number of finitely persistent bars with persistence up to 10% of the maximal filtration value. We present our results in Fig. 3.17. Interestingly, both the raw and the normalised number of short bars exhibit trends for the different treatment groups which strongly resemble those in the number of loops in Fig. 3.12. We consider the normalised data. Again, for the anti-Dll4-treated tumours there is a time-lag (day 3 onwards) before effects of the treatment are visible, even though they can not be

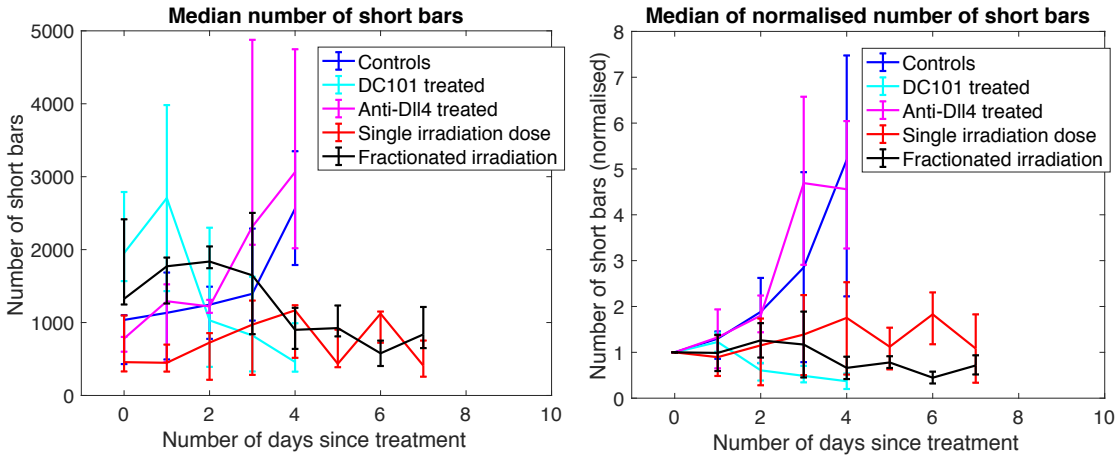


Figure 3.17: Total number of short bars in the radial filtration in dimension 0 in the intravital data set. We show the median and quartiles of the number of short bars by treatment group.

separated from the control tumours. For the DC101 treated tumours we see a very clear response from day 2 onwards. Single-dose irradiation and fractionated-dose irradiation both seem to have a stabilising effect on the number of short bars and, in comparison to the number of loops, we observe less variability in the response on individual days. The stabilising effect is particularly marked for fractionated-dose irradiation from day 4 onwards.

3.6 Summary and discussion

We developed the radial filtration for PH to characterise the abnormalities in the spatial structure of tumour blood vessel networks spatially. We applied the radial filtration to two data sets from different imaging sources. The first data set consisted of multiphoton intravital 3D imaging data, which allowed us to observe vessels from the same tumour over multiple days. The data set included groups of tumours subjected to five different treatment conditions with radiotherapy and vascular targeting agents. The second data set was imaged with multispectral fluorescence ultramicroscopy and allowed us to study large and highly detailed tumour blood vessel networks. We extracted the vessel networks from both data sets using the same approach. For both

data sets, we compared example barcodes to a synthetically created hierarchical tree and observed that these differed in their dimension 0 barcodes. We further investigated the number of loops and how it changes over time in both the full blood vessel networks and different radial intervals with respect to the tumour centre. Finally, we observed the number of intervals with persistence smaller or equal to 10% of the full filtration length in the intravital data.

The intravital imaging data showed that the number of loops increased over time for tumours treated with anti-Dll4 and decreased for tumours treated with DC101. In both cases, we observed a time lag of 2-3 days before we were able to see effects of treatment. For the two irradiation treatments, single-dose irradiation stabilised the number of loops in the first few days after treatment. At longer times (3 days plus), individual responses differed markedly. In contrast, fractionated-dose irradiation started to stabilise the number of loops four days after treatment. We found that treatments affected different parts of the vessel networks in different ways. For example, treatment with anti-Dll4 seemed to increase the number of loops mostly in the outer regions of the vessel network while treatment with DC101 acted similarly on all parts of the vessel network. For single-dose irradiation the stabilising effect on the number of loops deteriorated three days after treatment showing a diverse set of reactions in the medium region of the vessel network between the tumour surface and its centre, but not in other regions of the network. Similarly, following fractionated irradiation, we saw varied responses on the first three days after treatment begin, followed by a reduction of loops close to the tumour surface while the remaining vessel network does not show many changes. The results for the radiation therapy treatments are of particular biological interest, as the effect of radiation therapy on tumour vasculature is, so far, not well understood [144, 195]. We further found that the number of short bars in dimension 0 in the intravital data set appeared to follow the trends of the number of loops, with similar time lags for all treatments. Inter-

estingly, for the radiation treatments, in particular single-dose irradiation, there was less variation between different tumours in comparison with the other treatments than for the number of loops. For both dimension 0 and dimension 1, the trends that we observe in the data support the proposed mechanisms of action of the anti-Dll4 and DC101 treatments and are therefore very promising.

For the ultramicroscopy data, we found a trend in the total number of loops that matches what we would expect from treatment with bevacicumab. In contrast to the treatments in the intravital data set, bevacicumab seemed to have an immediate effect on the number of loops in the vessel networks that was already visible one day after treatment. This was also observed in [92], using measures of vessel density, and was interpreted as a sign of normalisation of the vessel network induced by bevacicumab. We also saw different developments in the number of loops in the tumour core and periphery: in the tumour periphery, the number of loops decreased following treatment and increased without treatment; in the tumour core the number of loops decreased on days 1, 3, and 7 after bevacicumab treatment, but stayed at a similar level for controls. Interestingly, we found a steep increase in the number of loops in the tumour core on day 14 after treatment, for which the total number of loops was in particular higher than for the two available control results. This could imply that the normalising effects of bevacicumab treatment on the vessel networks wear off in the second week after treatment. Dobosz *et al.* [92] only found effects of bevacicumab treatment on the structure of the outer regions of the vessel network, corresponding to vasculature with a distance to the surface of the tumour smaller than 20% of the radius of a fictional sphere with same volume as the tumour, but not on the inner regions. The study only includes results for days 1, 3, and 7 after treatment. For this data we are yet to examine features in dimension 0. It will be very interesting to see whether the trend in the number of short bars is similar to the number of loops and to the statistics that were examined in [92].

3.6.1 Advantages and limitations of the radial filtration

Our results suggest that the radial filtration is successful in capturing characteristics of tumour blood vessels in a spatial manner. Since we find that the number of loops is reduced in tumours treated with agents that promote vascular normalisation in the initial days after treatment and increases in tumours, whose treatment leads to the development of chaotic vasculature, we conclude that the number of loops is a relevant measure for quantifying abnormality in tumour vasculature. One could also study loops in these networks using other methods such as the Hodge Laplacian on graphs [157, 162]. The eigen-decomposition of the Hodge Laplacian can be used to give several geometrically 'nice' representatives of loops (so-called 'harmonic holes'). However, there is currently no available algorithm to study such loops and their persistence across a filtration. Indeed, a recent study that applies the Hodge Laplacian to brain networks [157] uses the PH algorithm to identify the birth and death filtration steps of a persistent loop before using the decomposition of the Hodge Laplacian of the corresponding birth and death graphs to estimate harmonic representatives of the loop. Using PH and, in particular, the radial filtration has the advantage that we obtain spatial information with respect to the tumour centre from the birth time in the filtration rather than just representatives of loops.

For dimension 0, our initial results on the intravital data set are very promising. We need to perform further analysis to investigate whether the short bars reflect underlying tortuosity or another characteristic of the vasculature. We note that using other simple measures such as the average or median persistence of features in dimension 0 did not lead to any significant differences between the different treatment groups. Nor could we see any trends in the corresponding Betti curves, even after excluding infinitely persisting features. In [28], Bates studied tortuosity in a subset of the intravital data using two different measures, chord-length-ratio (clr) and sum-of-angles-metric (SOAM). For controls, anti-Dll4 treated tumours, and DC101 treated

tumours, these measures of tortuosity did not appear to be discriminative (the study did not include data from radiation treatment). Our result in dimension 0 is therefore of interest, in particular if we can show that it is indeed a symptom of tortuosity.

As the number of loops and short bars in dimension 0 are structural properties of the network skeleton, they are more robust to imaging and segmentation induced errors than other statistics for vasculature, such as vessel lengths and vessel diameters. One could question whether the trends that we are observing in dimension 0 and in dimension 1 are a simple consequence of physical properties of the tumour, for example tumour volume in response to treatment. However, both treatment with DC101 and anti-Dll4 is known to slow tumour growth (see [206] and [212], respectively), i.e. in particular for DC101 we would not expect the tumours to reduce their volume. Similarly, bevacicumab was not observed to result in significant reduction of tumour volume in the ultramicroscopy data [92]. In Appendix B, Subsection B.1.3, we further show plots of estimated tumour volume for the different treatment groups and data sets. We approximated the tumour volume by calculating the volume of the sphere with a radius corresponding to the maximal filtration value of the radial filtration for every tumour. We do not find that the volumes fully reflect the trends from either the number of loops or the number of short bars. We will nevertheless, in the future, investigate normalisation of our results by tumour volume. We will also consider normalisation by the number of vessels.

There are, of course, limitations to our method. We constructed the radial filtration with spherical tumours in mind. In reality however, tumours do not necessarily grow in such an idealised way, as we can see in the example pictures in Fig. 3.5 and Fig. 3.6. Indeed, many of the tumours in our data sets exhibited multiple lumps. This can influence the results on the spatial distribution of loops in these tumours which depend on the position of the tumour centre. In future investigations, one could therefore divide the tumour into different spherical parts and define a centre

for every part separately. It would further be interesting to compare our approach to a sweeping plane filtration. It is also important to note that the experimental conditions and imaging modality can have an influence on the shape and structure of the vasculature. For example, in the intravital data set, the tumour grows directly under the window through which it is imaged. Consequently, vessels only grow into the tumour from below the window, as opposed to from all sides, which would be the case in a more biologically realistic scenario. The tumour growth itself is bounded by the window which flattens the upper part of the tumour and therefore also the vasculature. The limitations of the imaging to approximately 300 μm depth further mean that we only see a small sector of the tumour in the data and it is therefore difficult to define a tumour centre. We believe that using the centre of mass of the vessel points provides a good approximation, but further investigation on the influence of the location of the tumour centre on the results, in particular those based on spatial distribution, is necessary. We also found that for the intravital data set the vessel networks of the tumours are 'flat' when comparing the extent of the z-coordinates with the x- and y-coordinates even after correcting for different imaging resolutions. One could therefore attempt to use a filtration based on an ellipsoid around the tumour centre rather than a sphere. Ellipsoids instead of spherical neighbourhoods have been proposed for PH in the context of building simplicial complexes [50], for example the Vietoris–Rips[13] complex. Note that our use of ellipsoids would be different as we do not build our simplicial complexes based on pairwise distances but based on the distance to the tumour centre point.

3.6.2 Interpretability and further analysis of results

All our results have to be taken with caution: while both data sets are considered to be large for this type of data, we only have between two and seven vessel networks

[13]See Chapter 2, Subsubsection 2.3.1.1 for definition.

at each time point. Even for such small numbers, we see that individual tumours respond in different ways. This could be caused by physiological differences in the animals that determine how quickly treatment can travel from the place of injection to the tumour [144]. Another data specific complication is that for the intravital data set it was in some cases difficult for the experimentalists to determine the correct day for treatment administration. For example the DC101 treated tumours tended to be larger on day 0 of treatment than the tumours following other treatment regimes [144].

For both dimension 1 and dimension 0 features of the barcodes in both data sets, it would be beneficial to conduct a thorough analysis using persistence landscapes [53, 55] or persistence images [3]. We attempted an analysis with Betti curves, but did not find any obvious trends in either dimension 0 or dimension 1. Another option would be to create feature vectors for every vessel network based on the PH output, whose entries could, for example, include the birth times or persistence of the top 10 most persistent features in dimension 0, median of persistence in dimension 0, birth times of loops, number of loops etc. These vectors could then be used to train a machine learning classifier for vessel networks that could distinguish between highly abnormal vessel networks and more normalised vessel networks. One challenge in this approach, however, would be that, even though we have different treatment groups, all vessel networks in our data set come from tumours. Ideally, such an analysis would therefore include additional vasculature data, for example from healthy tissue and from experiments on wound healing. A further challenge in such a scenario would be to account for the fact that in the intravital data set the same tumour gives rise to several data points over time. This could lead to overtraining of the classifier on particular networks. The small number of data points in our data sets would be a further difficulty. Output from PH has successfully been used to create feature vectors and train a machine learning classifier recently by Bardin *et al.* [18]. The authors use four different approaches to create filtrations and use four measures

based on the Betti 0 and 1 curves from each filtration to populate their feature vectors. These features are different from what one can obtain using persistence landscapes, persistence images or other summaries for PH output.

Following a more sophisticated analysis of the output of the radial filtration, it would be interesting to compare our results to more conventional tumour blood vessel statistics, such as vessel density, vessel size, vessel length, and their prognostic power for the data. For the intravital data, our results for controls, anti-Dll4 treatment, and DC101 treatment follow similar patterns as average branch point density analysed in [28], but differ from average branch length, average branch diameter, and the number of large vessels. One could also investigate whether there is a link between the number of loops or number of short bars in dimension 0 and the function of a tumour blood vessel network.

3.6.3 Null model

It would be beneficial to use a more sophisticated null model than the synthetic hierarchical tree for both data sets, for example a mathematical model that allows the output of a tumour blood vessel-like network but with normalised features, and to incorporate this into a more thorough analysis of features that characterise tumour vasculature. The radial filtration on the synthetic hierarchical tree, however, also likely produces features related to hierarchical branching, it would therefore be interesting to identify these by additionally investigating the same model but using the branching angles that Karshafian *et al.* [147] suggest to be more representative of tumour vasculature. One could also use the model to test the robustness of the barcode features from the radial filtration to perturbation.

3.6.4 Future work

One could build on our spatial results from the radial filtration and see how these relate to regional differences in the tumour tissue such as the presence or absence

of necrotic tissue. Such information was studied by Dobosz *et al.* [92] and could therefore be available for the ultramicroscopy data. It would also be interesting to include anatomical information of the loops such as tortuosity, vessel length or vessel thickness and study whether there are spatial differences in the tumour. For the ultramicroscopy data, we found that by separately computing the filtration for the tumour core and periphery, we loose up to 9% of the total number of vessel loops in the network, which implies that up to 9% of loops start before and end after our boundary between tumour core and periphery. As this is a considerable number, it could imply that the size of loops in these networks is large. The combination of anatomical characteristics of the loops would require us to either output representatives of loops from our PH computations with JAVAPLEX or we could also use the Hodge Laplacian as suggested by [157, 162].

It would further be interesting to consider other filtrations on tumour vasculature. For example, the Vietoris–Rips filtration[14] could provide interesting insights on the vascularisation of a tumour reflected in features in dimension 1 or even 2. Unfortunately, all our attempts to compute the Vietoris–Rips filtration on the intravital data set were computationally infeasible due to the large number of vessel points and only resulted in partial results (see Appendix B, Subsection B.1.1). We were not able to improve computational feasibility despite several reductions of the number of points, for example by computing the Vietoris–Rips filtration only on branching points or vessel end points. Our attempts to use a reduced version of the algorithm available in the software package RIPSER [32] also failed. In future, it could be interesting to see whether the α-complex (see, for example, [190] for a description), which is particularly suited for 3-dimensional data, is computationally feasible.

It is also important to consider that different types of tumours can have different degrees of vascularisation and different structural characteristics in their blood

[14]See Chapter 2, Subsubsection 2.3.1.1 for definition.

vessel networks [149]. In the two data sets that we studied, we saw a very pronounced difference between the vascular networks: they were much more complex in the ultramicroscopy data than in the intravital data, even though the tumours in the ultramicroscopy data were smaller at the beginning of imaging. These differences are not just a consequence of the different imaging modalities. For the intravital data set, further tumour types have been imaged, whose characteristics are known in comparison to the MC38 tumours that we studied here. It would be interesting to compare results from the radial filtration between these different tumour types.

Finally, we recognise that the radial filtration is also of great interest for model selection for tumour-induced angiogenesis. This will be the focus of future research. PH has been used successfully for model selection in the past, for example in [259].

In summary, we have developed a filtration to spatially characterise the abnormal features of tumour blood vessel networks. Our results on two different data sets highlight that the number of vessel loops in tumour vasculature seems to reflect underlying biological processes in tumour development and treatment. We further observe that the number of short bars in the dimension 0 barcodes resulting from this filtration capture effects of treatment in one of the data sets.

" das mer und erd nach maßes trift
uß minem zirkel wart gerift. [...]
welch dink der maß ergibet sich,
das mak man teilen ewiklich:
ein ieglich lip, seit dir min list,
zu meßen und zu teilen ißt. "[a]

From the speech by the personified art of geometry (Geometria) in *Der meide kranz*, Heinrich von Mügeln, 14th century. Edition by [263].

[a]Translation: "The sea and earth emerged mature from my compasses in measurable form. [...] Anything which is subject to matter may be divided infinitely: every body, my art tells you, can be measured and divided." [263].

4

Persistent Homology Applied to Functional Neuronal Networks

The human brain consists of approximately 100 billion neurons, whose major task is to receive, conduct, and transmit signals. Analysis of neuronal networks is crucial for understanding the human brain [22,27,42,59,61,193,194,234]. Every neuron consists of a cell body and one long axon, which is responsible for propagating signals to other cells [6]. Neurons or (on a larger scale) different brain regions can be construed as nodes of a network, whose edges represent either structural or functional connections between those nodes. Examining neuronal data using a network-based approach allows one to use mathematical tools from subjects such as graph theory to better understand structural and functional aspects of neuronal interactions, identify key regions in the brain that are involved in physiological and pathological processes, and compare the structure of neuronal interactions to those of other complex systems. For example, data analysis using network theory has led to the insight that the brain has underlying modular structures, with small subunits that are able to carry out specific functions while minimally influencing other parts of the brain [61,62,194].

Here we present results from using persistent homology (PH) to analyse functional networks which we create from functional magnetic resonance imaging (fMRI) time series and the output of a mathematical model. We first describe how we obtain functional networks from time series data in Section 4.1 and how we apply PH to study these networks in Section 4.2. We then present our results from applications motivated by two different task-based fMRI data sets: the first was collected to understand motor-learning in healthy human subjects, see Section 4.3, the second was obtained to understand processes involved in Schizophrenia, see Section 4.4. In Section 4.3 we also analyse time series from a mathematical model. In all applications we use PH to understand loops (with four or more edges) in functional networks and investigate what we can learn from these network structures in the context of the biological problem studied.

The above introduction as well as the following sections are based on Stolz *et al.* 2017 [240] (mainly used in Section 4.3) and Stolz *et al.* 2018 [238] (mainly used in Section 4.4) with minor modifications[1]. For Stolz *et al.* 2018 [238] we only fully include work here that was conducted by the author of this thesis. We summarise Tegan Emerson's contributions to the paper in passive voice pointing the reader to the manuscript and Appendix B.2.2 for further details. We discuss and compare the results from all different methods, including those used by Tegan Emerson.

4.1 From fMRI to functional networks

Building a network based on experimental data involves several steps and decisions that determine the type of information that can be gained when studying the network. One approach to construct a neuronal network based on experimental data, for exam-

[1]Note that the initial work for Stolz *et al.* 2017 [240] was carried out in Stolz 2014 [237], which included preliminary results and observations from performing PH on both the mathematical model and the biological data, but no systematic analysis of the output from PH. In Stolz *et al.* 2017 [240] we extended the work significantly to include analysis using persistence landscapes, which we present here.

ple fMRI data, is to form a so-called *functional network* [59,61,200,235,240], in which each node represents a brain region, and the edges between them are weighted based on some measure of the similarity between their fMRI time series[2]. A functional network contrasts with a 'structural network,' which refers to underlying physical connections (e.g., anatomical connections) between nodes. For example, neurons are connected to each other in structural networks, but one can analyze the similarity in their firing patterns through functional networks. We use the term 'functional network' in a more general way: by constructing a matrix of similarities between coupled time series using some measure (and enforcing the diagonal entries to be 0), one obtains a functional network whose weighted adjacency matrix (sometimes also-called an 'association matrix') $\tilde{A} = (\tilde{a}_{ij})_{i,j=1}^{N}$ has elements that indicate the similarity between the time series of entities i and j. Studying functional networks is common in neuroscience, and they are also used in a wealth of other applications (e.g., finance [106], voting among legislators [267], and climate [94]). Importantly, the time series can come either from empirical data or from the output of a dynamical system (or stochastic process), and the latter is helpful for validating methods for network analysis [21]. Here, we consider time series either from a set of spatially distinct brain regions defined by a fixed anatomical atlas or from coupled oscillators (i.e., as the output of a dynamical system). In the context of functional brain networks, the adjacency-matrix element $\tilde{a}_{ij}$ arises as a measure of 'functional connectivity' (i.e., behavioral similarity) between the time series for nodes (i.e., brain regions) i and j. There are many different ways to measure similarity of time series [59, 233, 277], and that can be a major issue when it comes to interpreting results. Comparing the networks that arise from different similarity measures is beyond the scope of our work, so we will simply use common measures (pairwise synchrony, wavelet coherence

[2]Note that one can study coupled time series using a variety of different approaches [116, 182, 245, 251].

and Pearson correlation) of time series similarity. However, the methods that we employ can be applied to functional networks that are constructed using any measure of similarity between time series.

We construct functional networks using fMRI data from different data sources: (1) fMRI data that originates from a motor-learning study (see Section 4.3), and (2) fMRI data from experiments on schizophrenia patients, healthy siblings of schizophrenia patients, and healthy controls (see Section 4.4). In addition, for our investigations of data set (1) we also consider time series from coupled oscillators (i.e., as the output of a dynamical system). In Fig. 4.1, we show a pipeline of how to construct a functional network from fMRI time series data. Note that when interpreting fMRI studies, it is very important to consider the cautionary notes in [101].

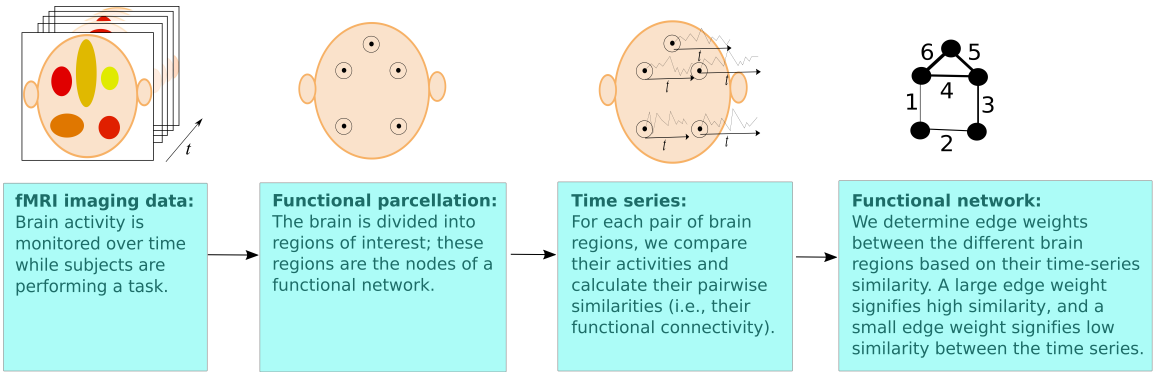


Figure 4.1: Pipeline for the construction of functional networks from imaging data (e.g., fMRI data). Image source: [238].

4.2 Persistent homology for functional networks

Even though PH [66, 97–99, 123, 232] has traditionally been applied to point-cloud data[3], it has become increasingly prominent in neuroscience in the last few years [80, 124]. Among other applications, it has been used to determine differences in brain networks of children with hyperactivity disorders and autism spectrum in comparison

[3]Note that one can also create point cloud data from time series, see for example a recent study where PH has been used to study point clouds of delay reconstruction of time series [219].

to normal situations [158], study the effect of the psychoactive component of 'magic mushrooms' (psilocybin mushrooms) on functional brain networks of humans [201], analyse covariates that influence neural spike-train data [236], and study structural and functional organisation of neural microcircuits [90]. Other neuronal applications have included consideration of place cells in the hippocampus of rats during spatial navigation [81, 82, 125], analysis of mathematical models of transient hippocampal networks [15], and a demonstration that topological features of networks of brain arteries in humans are correlated with their age [34]. Many other examples of applications of PH to neuronal networks can be found in [14, 16, 18, 75, 81, 122, 125, 138, 157, 202, 238, 240]. PH is not the only topological method that has been used to study the human brain or time series. More than fifty years ago, for example, Zeeman [276] used tolerance spaces and Vietoris homology theory to study aspects of visual perception. In the 1990s, Muldoon et al. [175] developed a method to study the topology of manifolds that underlie time-series data. Further, Sciamarella and Mindlin studied branched manifolds and trajectories in dynamical systems using homology [220, 221].

In contrast to standard methods of network analysis [183], employing PH allows one to explicitly go beyond pairwise connections; this is helpful for gaining global understanding of low-dimensional structures in networks. Although one can also use frameworks such as hypergraphs [47] to study higher-order network structures (see, e.g., [26]), such a formalism does not by itself give direct information about the shape or scale of mesoscale features in networks. By contrast, PH, allows one to explore the persistence of features, such as connectedness or loops, in data sets [190, 196].

In many studies based on experimental data, functional networks are used to construct binary graphs (i.e., unweighted graphs) [59]. To do this, one typically applies a global threshold $\xi \in \mathbb{R}^+$ to a weighted adjacency matrix to obtain a binary adjacency matrix $\tilde{B} = (\tilde{b}_{ij})_{i,j=1}^{N}$ associated with an unweighted graph. The adjacency-

matrix elements are then

$$\tilde{b}_{ij} = \begin{cases} 1\,, & \text{if } \tilde{a}_{ij} \geq \xi\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{4.1}$$

The choice of threshold has a strong influence on the resulting matrix, and it thereby exerts a major influence on the structure of the associated graph [59]. Some approaches to address this issue include determining a single 'optimal' threshold, thresholding the weighted adjacency matrix at different values [7, 224], examining the network properties as a function of threshold, or not thresholding at all and considering the weighted adjacency matrix itself [59, 194]. (One can also threshold a weighted adjacency matrix by setting sufficiently small entries to 0 but keeping the values of the other entries.) If one is thresholding and binarizing data, there is no guarantee that there exists an interval of thresholds that yield networks with qualitatively similar properties, and arbitrarily throwing away data can be problematic even when such intervals do exist. For example, parameters such as graph size (i.e., number of nodes) need to be taken into account when interpreting results on thresholded, binarised networks [103]. An advantage of using PH is that one can examine a graph 'filtration' (see Section 2.2.1) generated by multiple — ideally all — possible global thresholds and systematically analyze the persistence of topological features across these thresholds. Such a filtration can also be created using decreasing local thresholds.

For all functional networks that we consider, we create a nested sequence of networks in which we add edges, one by one, to the networks in order from largest edge weights to smallest. (In the unlikely case of two edges having the exact same weight, we add both edges simultaneously in one step.) We then construct a weight rank clique filtration (WRCF)[4] [202] by determining cliques and tracking their changes in every step of the network sequence. We compute PH and Betti numbers [79, 98] of the WRCF and examine the results by applying tools from statistics and machine learning, respectively, to output summaries such as persistence landscapes and persistence

[4]See Chapter 2, Subsubsection 2.3.2.1 for definition.

images that result from our computation of PH. We focus on loops (with four or more edges)[5] in the networks in our nested sequence, rather than on connected components, because one can also study the latter using more conventional approaches, such as by examining the spectrum of the combinatorial graph Laplacian [47, 183]. It has been demonstrated in other applications (e.g., contagions on networks [248]) that loops are important topological features of graphs, and a recent study [231] demonstrated the importance of loops (and related higher-dimensional objects) in structural neuronal networks. Because structural and functional neuronal networks are related and share many common network features [61], we expect loops to provide interesting insights.

4.3 Application to motor-learning data

In this application we use two sources of time series data to construct the functional networks that we study: time series output from networks of coupled Kuramoto oscillators, and fMRI data that was acquired from human subjects during a simple motor-learning task in which subjects were monitored on three days in a five-day period[6]. We use PH with a weight rank clique filtration to gain insights into these functional networks, and we use persistence landscapes to interpret our results. With these examples, we demonstrate that (1) using PH to study functional networks provides fascinating insights into their properties and (2) the position of the features in a filtration can sometimes play a more vital role than persistence in the interpretation of topological features, even though conventionally the latter is used to distinguish between signal and noise. We find that PH can detect differences in synchronisation patterns in our data sets over time, giving insight both on changes in community structure in the networks and on increased synchronisation between brain regions

[5]We use the term 'loop' to refer to at least four edges in a network that are connected in a way that forms a cycle. Conventionally, loops (other than self-loops) in undirected graphs must have at least 3 edges, and loops in directed graphs must have at least 2 edges. Here we adapt this terminology to represent the topological features that we detect in our simplicial complexes.

[6]Both data sets were also introduced and studied in Stolz 2014 [237]

that form loops in a functional network during motor learning. For the motor-learning data, persistence landscapes also reveal that on average the majority of changes in the network loops take place on the second of the three days of the learning process.

We first introduce the functional networks that we study and then present our results.

4.3.1 Data and construction of functional networks

The following subsections describe the time series data that we use. Note that while the descriptions of the mathematical model and data are taken from Stolz *et al.* 2017 [240], there is considerable overlap with the data description in Stolz 2014 [237], which is a precursor to the work presented in Stolz *et al.* 2017 [240].

4.3.1.1 The Kuramoto model

The Kuramoto model [11,131,154,214,244] is a well-studied model for a set of coupled phase oscillators with distinct natural frequencies that are drawn from a prescribed distribution. The model was developed in the 1970s to understand collective synchronisation in a large system of oscillators. It has subsequently been used as a toy model by many neuroscientists (as well as scholars in many other areas), as some of the characteristics of its synchronisation patterns resemble some of the ones in neuronal communities [13, 49, 160, 260]. The Kuramoto model and its generalisations have also been applied to numerous other applications in chemistry, biology, and other disciplines [11, 204, 214].

When all oscillators are coupled to each other, the Kuramoto model is most commonly written as [214, 244]

$$\frac{d\theta_i}{dt} = \omega_i + \frac{\mathcal{K}}{N} \sum_{j=1}^{N} \sin(\theta_j - \theta_i) \,, \quad i \in \{1, \ldots, N\} \,, \tag{4.2}$$

where θ_i denotes the phase of oscillator i, the parameter ω_i is its natural frequency, $\mathcal{K} \geq 0$ parametrises the coupling strength between different oscillators, and N is

the number of oscillators in the model. The normalisation factor $\frac{1}{N}$ ensures that the equations are bounded as $N \to \infty$. The distribution from which the frequencies ω_i are drawn is usually assumed to be unimodal and symmetric about its mean frequency, which can be set to 0 due to the rotational symmetry of the model (because Eq. (4.2) is invariant under translation of θ_i). The parameter ω_i then denotes the deviation from the mean frequency.

We also adapt Eq. (4.2) to create a network of N oscillators with uniform coupling between the oscillators [11, 12, 21, 214, 251]. We consider the following generalised version of Eq. (4.2):

$$\frac{d\theta_i}{dt} = \omega_i + \sum_{j=1}^{N} \kappa \mathcal{S}_{ij} \sin(\theta_j - \theta_i)\,, \quad i \in \{1, \ldots, N\}\,, \tag{4.3}$$

where $\kappa \geq 0$ denotes the normalised coupling strength and the entries of the coupling matrix $\mathcal{S} = (\mathcal{S}_{ij})_{i,j=1}^{N}$ indicate whether oscillators i and j are coupled. That is, $\mathcal{S}$ is an unweighted adjacency matrix, and $\mathcal{S}_{ij} = 1$ for coupled oscillators and $\mathcal{S}_{ij} = 0$ for uncoupled oscillators. The coupling matrix $\mathcal{S}$ thereby imposes a 'structural network' between the oscillators. One can further generalise Eq. (4.3) by using heterogeneous coupling strengths κ_{ij} or by considering functions other than sine on the right-hand side.

We divide the oscillators into 8 separate communities [7] of 16 distinct oscillators each, and we suppose that every oscillator has exactly 14 connections, 13 of which are with oscillators in the same community and 1 of which is to an oscillator outside the community. To enable comparison to known results about the dynamics and the community structure of coupled Kuramoto oscillators, we choose all parameters for the Kuramoto model as in Bassett *et al.* [21], i.e. we choose a coupling strength of $\kappa = 0.2$, consider a network with $N = 128$ oscillators, and suppose that the ith natural

[7] In this context, we use the term *community* to indicate a set of densely-connected nodes with sparse connections to other nodes outside of this set. There are also other uses of the term, and community structure is a popular subject in network science [113, 205].

frequency $\omega_i \sim \mathcal{N}(0,1)$. (That is, we draw natural frequencies from a Gaussian distribution with mean 0 and standard deviation 1.) Note however, that our network architecture differs somewhat from that in Bassett *et al.* [21], where every oscillator had at least 13 connections inside its community and at least 1 connection outside its community. Using the parameters above and the imposed structural communities, Bassett *et al.* [21] found that the resulting functional network of Kuramoto oscillators exhibits temporal changes in its community structure with increased synchronisation of oscillators in the same structural community over time.

We simulate the basic Kuramoto model using the Runge–Kutta MATLAB solver `ODE45` (with an integration time interval of $[0, T_{\max}]$, where $T_{\max} = 10$)[8]. We observe the system for $M = 500$ time steps in total (including the initial time step) and obtain time series $\mathcal{T}_i = (\theta_i(t_0), \ldots, \theta_i(t_{499}))$ as the output of the model for every oscillator θ_i. Kuramoto oscillators with a similar imposed community structure were demonstrated previously to initially synchronise rapidly within their communities, followed by a phase of global synchronisation in an entire network [21]. (There have also been other studies of community structure via synchronisation of Kuramoto oscillators [12, 243].) To study the dynamics of the coupled Kuramoto oscillators, we follow the work of Bassett *et al.* [21] and partition the time series into two time regimes, which we denote by $\hat{k} = 1$ and $\hat{k} = 2$ (we note that the authors in [21] used a different number of time steps). In our example, these time regimes each consist of 250 time steps.

To quantify the pairwise synchrony of two oscillators i and j, we use the local measure [12, 21]

$$\phi_{ij}^{\hat{k}} = \left\langle \left| \cos\left(\mathcal{T}_i^{\hat{k}} - \mathcal{T}_j^{\hat{k}} \right) \right| \right\rangle , \tag{4.4}$$

where the angular brackets indicate that we take a mean over 20 simulations. We use the absolute value both to facilitate comparison with Arenas *et al.* [12] and Bassett

[8]We used an input time step of $\Delta t = 0.02$, we however note that ODE45 internally uses an adaptive step size.

et al. [21] (by making the same choice that they made) and to avoid negative values, which can complicate interpretation and pose other difficulties in network analysis [106, 233, 256].

In each simulation, we choose the initial values for the phases θ_i from a uniform distribution on $[0, 2\pi)$ and draw the natural frequencies ω_i from $\mathcal{N}(0, 1)$. We apply the same underlying coupling matrix $\mathcal{S} = (\mathcal{S}_{ij})_{i,j=1}^{N}$ for all 20 simulations and then use the values ϕ_{ij} to define the edge weights in the fully connected, weighted network of Kuramoto oscillators for each time regime. We also study a network based on one full time regime that consists of 500 time steps. In analogy to neuronal networks, we call these networks 'functional networks.' In Fig. 4.2, we illustrate our pipeline for creating a functional network from the output of a simulation of the Kuramoto model.

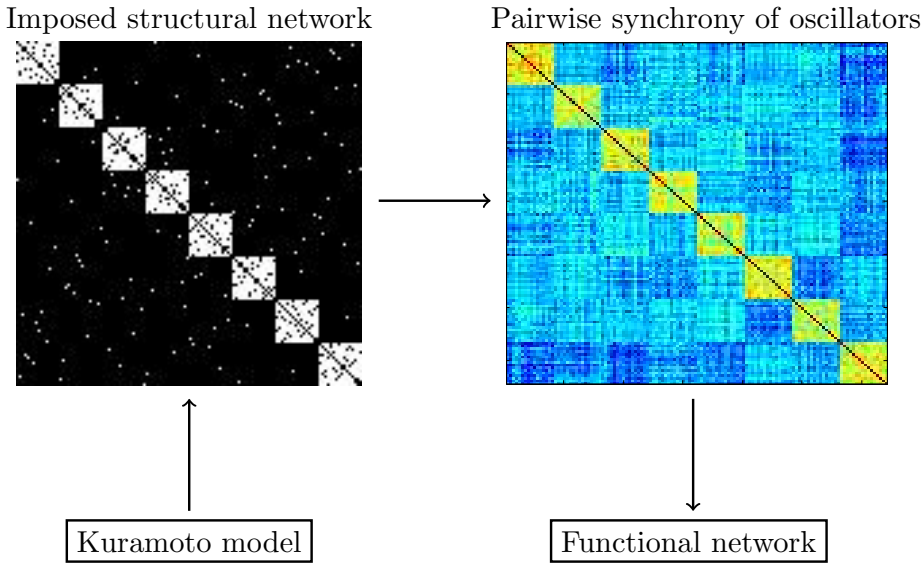


Figure 4.2: We construct a structural network for coupled Kuramoto oscillators by grouping the oscillators into 8 separate communities. Oscillators are coupled predominantly to other oscillators in their community, and they are coupled only very sparsely to oscillators outside their community. We use the time series output of a simulation of the Kuramoto model to create a functional network based on the similarity of the time series of individual oscillators. We use the measure of similarity in Eq. (4.4). Image source: [240].

4.3.1.2 Null models for the Kuramoto data

To assess whether our observations illustrate meaningful dynamics of the Kuramoto model or whether they can be explained by a random process, we consider two different null models based on the time series output. In the first null model, which we call the *simple null model*, we reassign the order of the time series for every oscillator according to a uniform distribution before computing the similarity measure with Eq. (4.4). The second null model, which we call the *Fourier null model*, is based on creating surrogate data using a discrete Fourier transformation. This approach [207] has the advantage of preserving not only the mean and the variance of the original time series but also the linear autocorrelations and cross correlations between the different time series.

To construct the Fourier null model, we start by taking the discrete Fourier transform

$$\hat{\mathcal{T}}_{\hat{n}} = \frac{1}{\sqrt{\hat{M}}} \sum_{\hat{m}=0}^{\hat{M}-1} \mathcal{T}_{\hat{m}} e^{\frac{2\pi i \hat{n}\hat{m}}{\hat{M}}} \tag{4.5}$$

of a time series vector $\hat{\mathcal{T}} = (\theta(t_0), \ldots, \theta(t_{\hat{M}}))$ of length $\hat{M}$. In our case, $\hat{M} = 250$ or $\hat{M} = 500$, depending on whether we are examining two different time regimes or just one. We then construct surrogate data by multiplying the Fourier transform $\hat{\mathcal{T}}_{\hat{n}}$ by phases $a_{\hat{n}}$ chosen uniformly at random from the interval $[0, 2\pi)$, aside from the constraint that they must satisfy the following symmetry property: for every $\hat{n} \leq \hat{M}$, there exists $\tilde{n}$ such that $a_{\hat{n}} = -a_{\tilde{n}}$. This symmetry ensures that the inverse Fourier transform yields real values. The surrogate data $\sigma = (\sigma_1, \ldots, \sigma_{\hat{M}})$ are thus given by

$$\sigma_{\hat{m}} = \frac{1}{\sqrt{\hat{M}}} \sum_{\hat{n}=0}^{\hat{M}-1} e^{\mathrm{i}a_{\hat{n}}} \hat{\mathcal{T}}_{\hat{n}} e^{-\frac{2\pi i \hat{n}\hat{m}}{\hat{M}}} . \tag{4.6}$$

Both the simple null model and the Fourier null model were used previously on time series output of coupled Kuramoto oscillators, and they exhibit different dynamics from those of the coupled Kuramoto oscillators [21, 24].

4.3.1.3 Human brain networks during learning of a simple motor task

We use a data set of functional brain networks from experiments that were first analysed by Bassett *et al.* [23]. The data set was collected to study human subjects during learning of a simple motor task, and a full description of the experiments conducted is available in [23]. We apply a WRCF to functional networks, and we compare our findings to previous studies on these and similar networks [23, 25, 26]. The functional networks are based on functional magnetic resonance imaging (fMRI) time series[9] from 20 healthy subjects who undertook a motor-learning task on three days (during a five-day period). During the imaging of the subjects, an 'atlas' of 112 brain areas was monitored while they were performing a simple motor-learning task (similar to a musical sequence), which they executed using four fingers of their non-dominant hand. For each subject and for each day of the study, the fMRI images are interpreted as 2000 time points for each monitored brain region. The brain regions and their time series were used subsequently to construct functional networks based on a functional connectivity measure known as the coherence of the wavelet scale-2 coefficients. This measure was applied to the time series to determine edge weights between every pair of brain regions in the network. The weighted adjacency matrices for the functional networks were then corrected for a false-discovery rate, as matrix elements under a certain threshold (which represents a correlation amount that one expects to occur at random) were set to 0. The other matrix elements were retained.

The functional networks that we just described were studied previously using community detection by Bassett *et al.* [23], whose results suggest that there is a significant segregation of the nodes in the functional networks into a small number of different communities with densely-weighted connections inside the communities and sparsely-weighted connections to nodes in other communities. Within these communities,

[9] See [101] for a recent discussion of fMRI inferences and potential perils in the statistical methods in use in neuroimaging.

certain nodes appeared to remain in the same community during the experiment, whereas others (the 'flexible' ones) often switched between different communities.

There have also been studies of networks from a similar experiment but with medium-term learning and including training sessions [25,26]. These networks have a noticeable core–periphery organisation, with the sensimotor and visual regions of the brain grouped into a temporally 'stiff' core of nodes, whose community memberships (in contrast to flexible, peripheral nodes) do not change much over the course of the learning task [25]. It was also shown subsequently that the interaction between primary and secondary sensorimotor regions and the primary visual cortex decreases as the regions (presumably) become more autonomous with task practice [26].

4.3.2 Implementation

For our PH calculations, we use MATLAB code that we construct using JAVAPLEX [4, 247], a software package for PH. For the WRCFs, we also use a maximal clique-finding algorithm from the Mathworks library [271] based on the Bron–Kerbosch algorithm, which is the most efficient algorithm known for this problem. For statistical analysis and interpretation of our barcodes, we apply the PERSISTENCE LANDSCAPES TOOLBOX [55].

4.3.3 Results

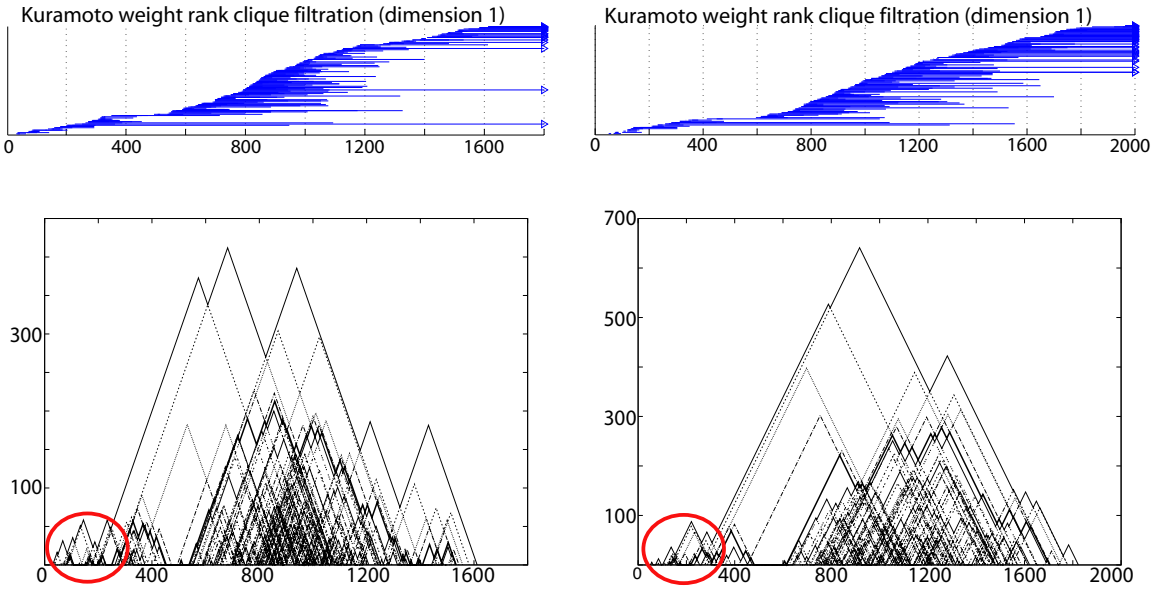
4.3.3.1 Persistent homology applied to the Kuramoto model and null models

We apply the WRCF to functional networks created from the output of two time regimes of the Kuramoto model, one time regime for the Kuramoto model, the simple null model, and the Fourier null model. We run the filtrations up to filtration step 1800 for the first time regime and up to 2000 for the second; we go up to filtration step 1100 for cases in which we only consider one time regime (note that using more filtration steps leads to very long computational times, see Subsection 1.2.2 for a

discussion of this type of problem). The total number of edges in the network, and thus the total number of possible filtration steps, is 8128. The number of filtration steps thereby correspond to respective edge densities of 0.22, 0.25, and 0.14 for the three examples above; in each case, this amounts to a threshold that is approximately in the middle of the range of the edge-weight values. In Stolz 2014 [237], which is a precursor to the work presented in this Section, we also applied PH to networks created from the Kuramoto model, and such an example was subsequently also studied using Betti curves by other authors [230].

As we described in Section 4.2, we focus our analysis on topological features in dimension 1, so examine loops in the network. In the first row of Fig. 4.3, we show the 1-dimensional barcodes for the networks constructed from time regime 1 (i.e., the first 250 time steps of the dynamics) and time regime 2 (i.e., time steps 251–500 of the dynamics) for the WRCF of the Kuramoto model. The barcode for each time regime includes several very short-lived bars between filtration steps 50 and 300. For the second time regime, we find more short bars for a longer filtration range at the beginning of the barcode. We extract representatives for the 1-loops that correspond to these short bars and find that these are all formed within the strongly synchronised communities. (See Fig. 2.4 for an illustration of different representatives of the same loop.) In fact, in time regime 1, the first 44 bars in the barcodes represent intra-community loops; in time regime. 2, only 2 of the first 28 bars represent intra-community loops. As strong intra-community edges are added to the simplicial complexes, they start to cover the 1-loops with triangles (i.e., 2-simplices), and the loops disappear from the filtration. Note that as we discuss in Subsubsection 4.3.3.2, one needs to be cautious when interpreting representatives given by the software Javaplex as these are not necessarily optimally chosen. Our findings do however align with what one would expect intuitively and from what was observed by Bassett *et al.* [21].

In the second row of Fig. 4.3, we show the persistence landscapes that we construct



(a) Time regime I. **(b)** Time regime II.

Figure 4.3: Dimension-1 barcodes and persistence landscapes for the WRCF for the two time regimes, (a) time steps 1–250 and (b) time steps 251–500, of time series output of the Kuramoto model. The horizontal axis represents the filtration steps in both the barcodes and the landscapes. The vertical axis in the persistence landscape captures the persistence of the features in the barcode. Note that for technical reasons the vertical axes in the two time regimes are different. In the first row, we show the barcodes for dimension 1. In the second row, we show persistence landscapes (although we ignore infinitely-persisting bars in the barcodes). The short peaks at the beginning of the filtration in the persistence landscapes that are indicated by the red ellipses represent loops formed within communities. The most prominent difference between the two landscapes is the occurrence of high peaks in the second time regime; these peaks correspond to persistent loops in the network that are formed between communities. Image source: [240].

from the 1-dimensional barcodes. Note that for technical reasons the vertical axes in the two time regimes are different. We ignore infinitely-persisting bars in the barcode. (We also studied persistence landscapes including the infinite bars as features with a death time that corresponds to the maximum filtration value but did not obtain any additional insights that way.) As expected, the landscapes have a group of small peaks early in the filtration for both time regimes. This feature occurs in a longer filtration range in the second time regime before more persistent loops appear. In the second time regime, some of the peaks that occur in the beginning of the filtration appear to almost double their heights to values of about 100. In contrast, in the first time regime, peaks at a similar location are about half as high (i.e., they are less

persistent).

The persistence landscapes reveal more persistent 1-loops in the second time regime (i.e., between time steps 251 and 500) than in the first (i.e., between time steps 1 and 250), and the second time regime also appears to reveal a clearer separation between the group of the very early short peaks and a second group of medium-sized peaks towards the end of the filtration. For this second group of medium-sized peaks, we observe a larger absolute increase in persistence in the second time regime than for the shorter peaks in the beginning of the filtration. These observations reflect the dynamics of the two time regimes in the Kuramoto model [21]. In time regime 1, there is strong synchronisation within the communities, and such dynamics are reflected by the appearance of short-lived intra-community 1-loops (corresponding to the short peaks in the persistence landscapes) at the beginning of the filtration. In the second time regime, the amount of global synchronisation is more prominent than in the first time regime. Moreover, in addition to intra-community loops, some of the peaks at the beginning of the filtration now represent inter-community loops, which are more persistent than the loops within communities. Additionally, as some of the peaks that correspond to inter-community loops have shifted to the beginning of the filtration, there is an increase in the gap between the initial group of peaks and the group of medium-sized peaks at the end of the filtration. In general, we observe an increase in the persistence of the peaks in the landscapes due to the stronger synchronisation between the communities. These observations are much easier to visualise using persistence landscapes than using barcodes.

We calculate pairwise L^2-distances between all dimension-1 persistence landscapes, and we note that L^2 distance has been used previously to compare persistence landscapes in an application to protein binding [152]. The L^2 distance between the two time regimes is 27078. Given the length of the support of the landscapes and the function values that they attain, this is a large distance, which captures the afore-

mentioned visible differences between the landscapes. What the L^2 distance can not capture is that in the first time regime the peaks that appear early in the filtration correspond to loops between nodes within one community, while in the second time regime they correspond to loops that form between nodes of different communities. This fact does therefore not contribute to the distance value.

We also compare the Kuramoto model to the two null models that we discussed in Section 4.3.1.2. To do this, we construct a functional network by considering a single time regime that consists of 500 time steps. In Fig. 4.4, we show the weighted adjacency matrices of the three functional networks, and we also show their corresponding persistence landscapes based on WRCFs of the functional networks. Note that the vertical axes for the Kuramoto model and the two null models are different. One can observe clearly that there is stronger intra-community synchronisation for the Kuramoto time series than for the null models, as there is a very distinct group of short peaks at the beginning of the filtration (which, as we discussed above, is also the case for the Kuramoto model when performing separate calculations in the two time regimes).

Again, the corresponding loops occur within communities. The peaks in the Kuramoto landscape appear to be separated from a second group of short peaks further along in the filtration. Between the two groups of peaks, there are two strikingly higher peaks that correspond to persistent loops, which appear to be formed by connections between different communities. For both null models, we also observe groups of short peaks at the beginning of the filtration, but these are less persistent and less clearly separated from other peaks than for the Kuramoto model. Indeed, we do not see any separation at all for the Fourier null model, which exhibits a much weaker intra-community synchronisation than the simple null model. Moreover, the persistence landscape for the Fourier null model appears to be 'noisier,' as the majority of the peaks in the landscape have similar persistences and appear in similar areas of

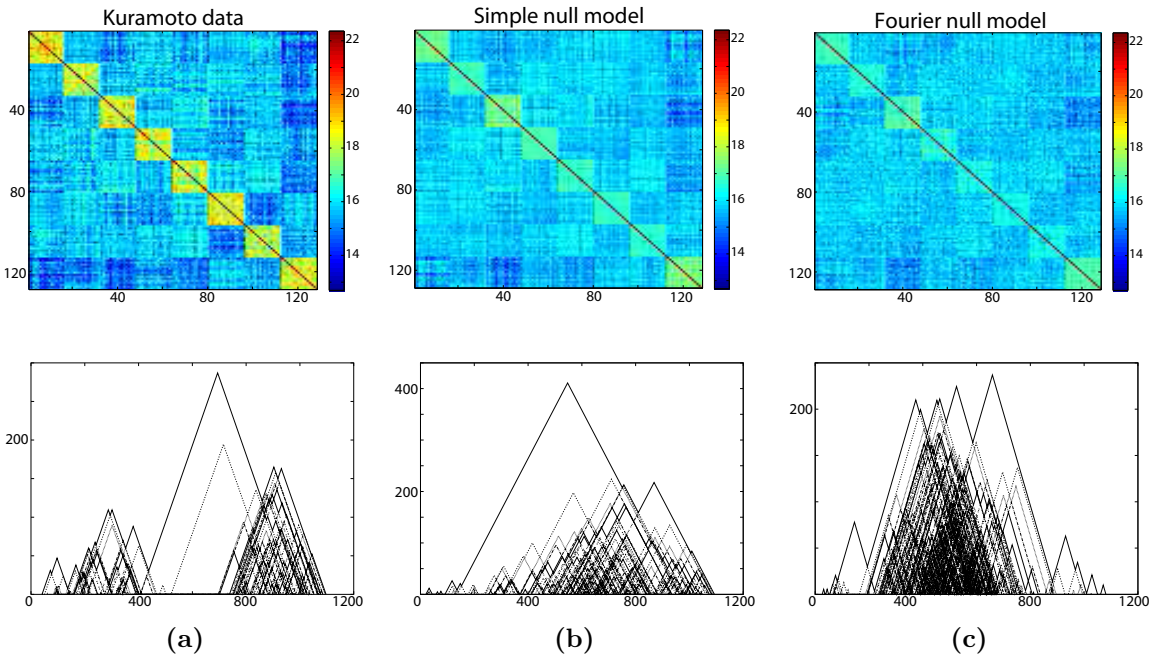


Figure 4.4: (Top row) Functional networks for (a) the Kuramoto model, (b) the simple null model, and (c) the Fourier null model. (Bottom row) Dimension-1 persistence landscapes for the WRCF of (a) the Kuramoto model, (b) the simple null model, and (c) the Fourier null model using one time regime and ignoring infinitely-persisting bars. Note that the vertical axes for the three persistent landscapes are different. The persistence landscapes illustrate differences in the occurrence of loops in the three different networks. Most prominently, these differences manifest in the height and distribution of the peaks in the landscapes, which appear to exhibit a stronger separation along the filtration between groups of peaks different heights for the Kuramoto model than in the two null models. Image source: [240].

the filtration.

The peaks in the landscapes of the null models appear to have a very different distribution along the filtration than in the Kuramoto model. They also possess more medium-sized and long persisting features than we observed in the Kuramoto data. These features occur in parts of the filtration in which the Kuramoto data has a smaller number of peaks. They consist of inter-community loops and are a symptom of the weaker intra-community and stronger inter-community synchronisation. The null models thus appear to have more topological features in the form of loops than the Kuramoto data, which is consistent with previous observations of null models in other studies [125, 202, 231]. The fact that there are fewer persistent loops in the Kuramoto model than in the null models implies that there are more high-dimensional simplices (e.g., triangles and tetrahedra) in the corresponding network than in the networks constructed from the null models.

To distinguish between the three landscapes, we calculate the L^2 distances between them. The L^2 distance between the Kuramoto landscape and the Fourier null-model landscape is 13540 the L^2 distance between the two null-model landscapes is 13263, and the L^2 distance between the Kuramoto landscape and the simple null-model landscape is 11703. Again considering the support of the landscapes and the attained function values, we see that three distances can be considered as large.

For the Kuramoto model, we find that PH can detect the dynamics of the system and that the persistence landscapes are rather different for the Kuramoto model and the null models. The L^2 distances between landscapes underscore these differences. We are also able to distinguish between the two null models using persistence landscapes. In contrast to conventional wisdom [66,123], we do not find for our examples that only the persistence of topological features distinguishes between signal and noise. In fact, the short bars at the beginning of the filtration of the Kuramoto model carry important information about the dynamics, and the medium-sized persistent

peaks in the Fourier null model are a symptom of the weaker intra-community and stronger inter-community synchronisation in that model. We therefore assert that the position of features in the barcode is as important as persistence length for their interpretation in our examples, and this provides an important point to consider for future studies. Note that persistence landscapes alone do not provide enough information to assess system dynamics. It is only by combining them with information about nodes that are forming loops (which are represented by certain groups of peaks) that we are able to obtain conclusions about intra-community and inter-community synchronisation.

4.3.3.2 Persistent homology applied to task-based fMRI data

We run the WRCF until filtration step 2600, which is when 42% of the edges are present in the network. (Note again that using more filtration steps leads to very long computational times, see Subsection 1.2.2 for a discussion of this type of problem.) We again focus our analysis on topological features in dimension 1. We construct persistence landscapes for dimension 1 (omitting infinitely persisting 1-loops). In Table 4.5, we summarise our results for one particular subject and for the whole data set. We use this subject to illustrate a representative example of the particular landscape features that we observe in the data.

Similar to the Kuramoto oscillators, we find a group of small peaks at the beginning of the filtration (between filtration steps 1 and 200). We can see this group very clearly both by magnifying either the landscape of individual subjects or the average landscape, where the height of the peaks is only slightly smaller than for the peaks in the individual landscape that we show. This feature of the heights indicates that a group of short peaks arises in the beginning of the filtration in the majority of the barcodes. We also consider the standard deviation from the average landscapes in the first 200 filtration steps. For all three days, it is very small: it is 127 for the first day, 167 for the second day, and 126 for the third day.

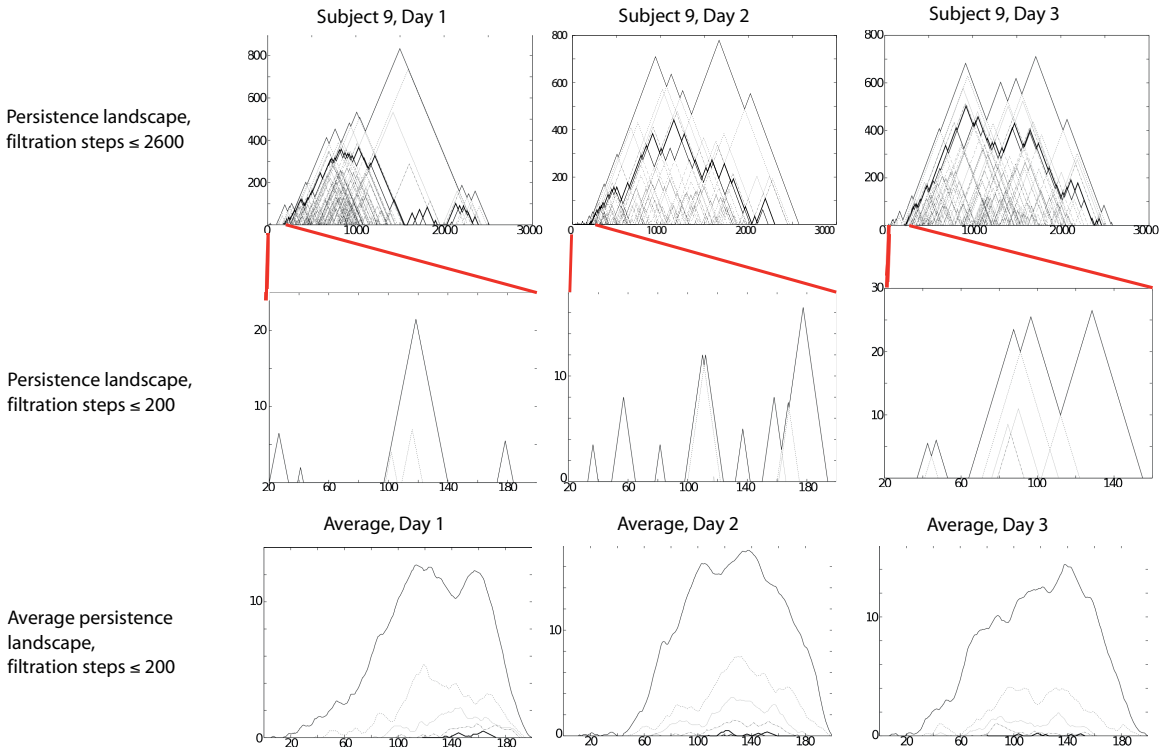


Figure 4.5: Persistence landscapes for dimension 1 of the WRCF applied to the human brain networks. (First row) Persistence landscapes for subject 9 based on filtration steps 1–2600 for days 1, 2, and 3. (Second row) Persistence landscapes for subject 9 based on filtration steps 1–200 for days 1, 2, and 3. (Third row) Average persistence landscapes over all subjects for days 1, 2, and 3. We observe on average that short peaks occur in the first 200 filtration steps of the landscapes. Image source: [240].

Based on our insights from the Kuramoto oscillators, we expect the observed short peaks in the beginning of the filtration to be associated with network communities, which have been observed previously in this data set using other methods [23]. We observe, in particular, that these short peaks undergo changes on day 2: during filtration steps 20 to 60, some of the peaks that are present in the landscapes for days 1 and 3 vanish, and more persistent peaks occur for day 3 than on the other two days between filtration step 80 and 200. This appears to suggest that there is a change in community structure that takes place on day 2, with either (1) very strong synchronisation in some of the communities, leading to very short-lived 1-loops; or (2) very strong individual differences between the subjects, leading to the vanishing of peaks in the average landscapes for the first 50 filtration steps. The particularly

	Cluster 1	Cluster 2	Cluster 3
Day 1	9	6	5
Day 2	5	4	11
Day 3	5	5	10

Table 4.1: Results for k-means clustering and average linkage clustering of pairwise L^2-distance vectors of persistence landscapes for $k = 3$.

persistent peaks on day 2 could represent either persistent loops between different communities or loops that occur due to sparse intra-community connections.

We calculate pairwise L^2-distances between all dimension-1 persistence landscapes. We create distance vectors, which we use as an input for k-means clustering and average linkage clustering for $k = 3$, and we obtain the same qualitative result for both methods. We find that 9 of the 20 distance vectors that correspond to persistence landscapes from day 1 are assigned to a common group (together with a small number of landscapes from days 2 and 3), whereas 11 and 10 landscapes from days 2 and 3, respectively, are assigned together to a separate group. We summarise our results in Table 4.1.

We also consider the average dimension-1 landscapes for WRCF steps 1–2600 and calculate the L^2-distances between them. We show the results of these calculations in Fig. 4.6. The distances between the average landscape for day 1 and the subsequent days of the experiment indicate that the WRCFs on average are able to detect changes in the functional networks across the filtration range. Based on the distances, we observe that most of these changes occur between the first and the second day. However, the standard deviations from the average landscapes are a factor of about 4 larger than the distances between the landscapes, and one therefore needs to be cautious about interpreting the results of these calculations. In a permutation test with 10000 regroupings of the landscapes, we do not find the distances to be statistically significant. We obtain p-values of about 0.4 for the distance between the average landscapes of day 1 and day 2, about 0.85 for the distance between the

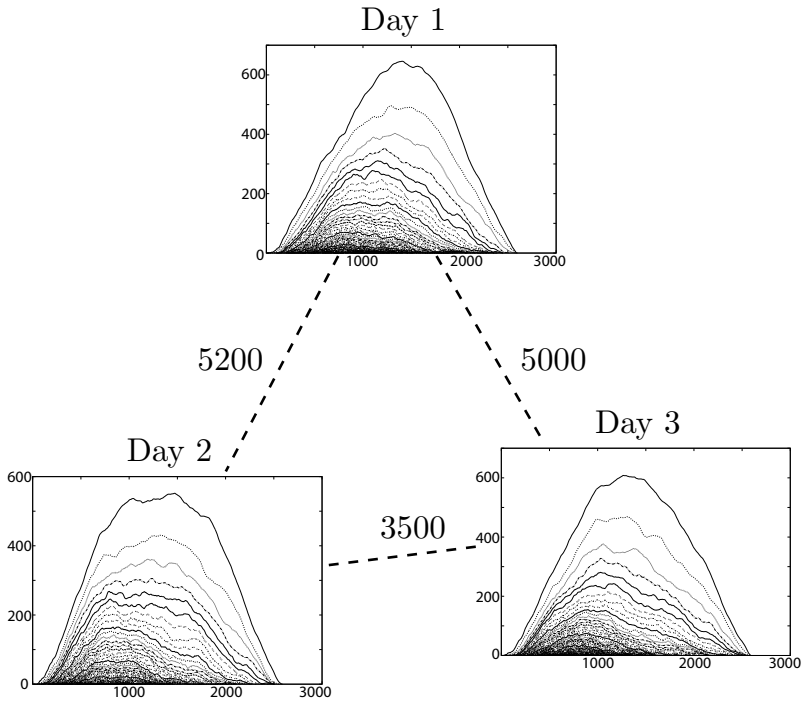


Figure 4.6: Visualisation of average persistence landscapes for days 1, 2, and 3 of task-based fMRI networks. The distance between the landscape for day 1 and the other two landscapes is larger than that between the landscapes for days 2 and 3. (The L^2 distances between them are 5200 between days 1 and 2, 5000 between days 1 and 3, and 3500 between days 2 and 3.) The standard deviations from the average landscapes are larger than the calculated distances, so these values need to be interpreted cautiously. We also observe a shift to the left of the landscape peak during the three days, indicating that the particularly persistent 1-loops in these networks arise earlier in the filtration for the later days. In other words, they are formed by edges with a higher edge weight, indicating that there is stronger synchronisation between the associated brain regions. Image source: [240].

average landscapes of day 2 and day 3, and about 0.6 for the distance between the average landscapes of day 1 and day 3.

For the average landscapes in Fig. 4.6, we also find that that the primary peak of the average landscapes shifts to the left over the course of the three days. This implies that the edge weights (between the brain regions) that give rise to persistent 1-loops increase on average over the three days (presumably due to stronger synchronisation). This can either mean that loops present on the first day synchronise more on the second and third day, or that new loops that appear on days 2 and 3 consist of more synchronised edges. Brain regions that synchronise in a 1-loop in a network may be an indication of an interesting neurobiological communication pattern that in this

case also becomes stronger over the course of the learning process. To analyze the most frequently occurring edges involved in these loops, we extract 'representatives' for all loops in dimension 1 across all subjects and days (see Fig. 2.4 for an illustration of two different representatives of a loop in a network). For each day, we construct a network, which we call the 'occurrence network,' using the same nodes (i.e., brain regions) that we used before and assign every edge an edge weight that is equal to the number of occurrences of that edge in 1-dimensional loops in the subjects on the given day. We then perform a WRCF on the three occurrence networks and study representative loops given by the algorithm. In Table B.1 in Appendix B.2.1.1, we list the brain regions that we find in loops that consist of edges that occur at least 50 times in functional networks in the subjects. We now examine loops in the occurrence networks. These particular loops may not exactly correspond to loops in the functional networks. However, it is very likely that they are also loops in the functional network. One also needs to consider that the representative loops given by the software JAVAPLEX are not necessarily optimally chosen or 'geometrically nice'[10] representatives of the loop [4]. (See Fig. 2.4 for an illustration of different representatives of the same loop.) Selecting a basis of homology generators that behaves in a biologically representative way corresponds mathematically to solving a problem known as the *optimal homology-basis problem*, which is not trivial and can be NP-hard [102]. We address the issue of the algorithm's choice of representatives to some extent by using persistent homolopgy on the occurrence network, but even then we cannot rule out possible artefacts. There exist loops in the occurrence networks that remain stable across the three days, although other loops occur on only one or two days. There also seem to be more loops that occur at least 50 times in the functional networks on days 2 and 3 than on day 1. It would be useful to study the brain regions involved in the listed loops (see Table B.1 in Appendix B.2.1.1) to

[10]For example, a loop may be represented by a double loop.

investigate their biological role in motor-learning tasks.

Finally, we also apply WRCF to the average networks for each of the three days. To create the average networks, we take the mean of the edge-weight values over all 20 subjects for each day separately and study the resulting network. We show the corresponding landscapes in Fig. 4.7.

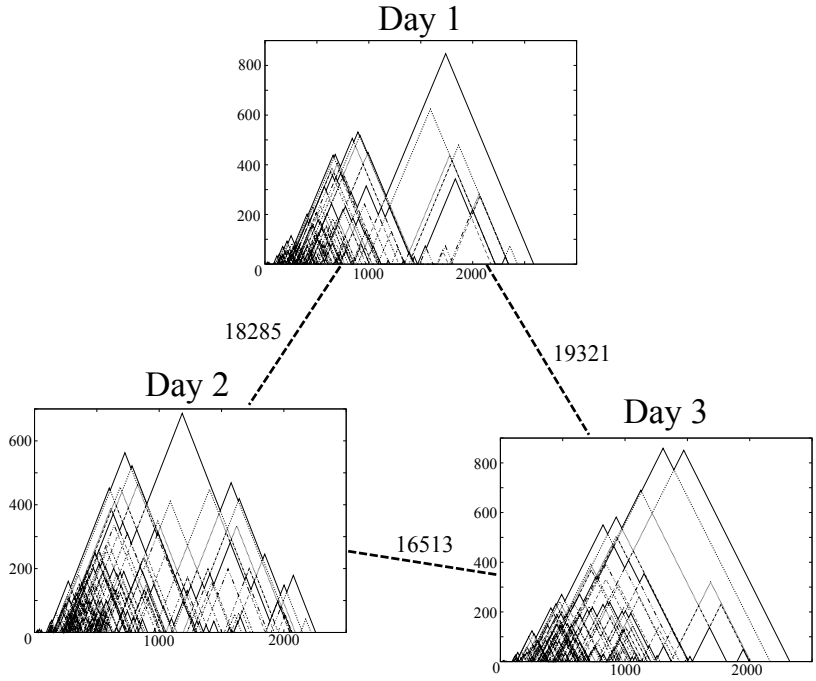


Figure 4.7: Visualisation of persistence landscapes based on average functional networks on days 1, 2, and 3 of the motor-learning task. The distance between the landscape for day 1 and the other two landscapes is larger than that between the landscapes for days 2 and 3. (The L^2 distances between them are 18285 between the first and second days, 16513 between the first and third days, and 19321 between the second and third days.) We find short peaks at the beginning of the filtration for all three landscapes, and larger peaks begin earlier in the filtration on day 3 than on day 1. Image source: [240].

As with the average landscapes, we find that the landscapes for the average networks have very short peaks in the beginning of the filtration. There are more persistent features (e.g., larger peaks) on day 1 and day 3 than on day 2, and we even find (as in the average landscapes) that the larger peaks appear earlier (at about filtration step 400) in the filtration on day 3 than on day 1 (where they appear at about step 900). Additionally, on day 2, we observe many short peaks, in particular in the later stages of the filtration. This is not the case for day 1 and day 3, so the

day-2 landscape is strikingly different visually from the other two landscapes. When calculating L^2 distances, we again find that the landscape distance between days 1 and 2 and that between days 1 and 3 are larger than the landscape distance between days 2 and 3. From visual inspection, we see that this arises from the fact that the day-1 landscape appears to have a clearer separation of short and high peaks than the landscapes for the later days. Taken together, the results for the landscapes of the average networks mirror our prior results for the average landscapes.

4.3.4 Summary and discussion

We have illustrated applications of PH to functional networks constructed from time series output of the Kuramoto model, null models constructed from the Kuramoto time series, and task-based fMRI data from human subjects. In all cases, we observed that non-persistent 1-loops occur at the beginning of the filtrations. Although such non-persistent features are commonly construed as noise in topological data analysis [66, 123], we observed that these features appear to be consistent with prior segregations of the studied networks into communities of densely-connected nodes. In one case (the Fourier null model), we even found that particularly persistent features appear to be linked to a network with a weak intra-community synchronisation. These very persistent features in the null model may thus represent noise. In other studies of PH in (different) null models [125, 202, 231] it was, however, also observed that the null models often exhibit a richer topology than data. Taking this into account one could perhaps interpret the persistent features in the Fourier null model as features of the null model rather than noise. Our results on the importance of non-persistent features match previous observations for synthetic examples with barcodes that consist of short intervals (which are commonly construed as noise), but where the differences between the corresponding persistence landscapes for the various spaces are nevertheless statistically significant [56]. Our results are also consistent with the

findings of a study on protein structure using PH for which bars of any length in the barcodes were equally important [273]. For weighted networks, we suggest that when using a filtration based on edge weights, one needs to consider the actual birth and death times of filtration features (such as 1-loops) in addition to their persistence to be able to determine whether they should be construed as part of noise or part of a signal. In particular, in the present paper, we observed that the early appearance of 1-loops in a filtration are important distinguishing features of these data. They may also yield important insights on the geometry [56] of data[11]. In general, when using PH to analyse a new data set it therefore seems beneficial to, at least initially, consider the full PH output for interpretation rather than just persistent features.

We also found — both by calculating average persistence landscapes and studying landscapes of average networks — that persistence landscapes for dimension 1 of the weight rank clique filtration (WRCF) are able to capture changes in the studied functional brain networks during the process of learning a simple motor task. Because we did not consider infinitely-persisting features and only included filtration steps 1–2600 when creating the landscapes, our results also suggest that the medium-lived (when compared to the the full filtration length) persistent 1-loops are able to capture changes in the network, so it is not always necessary to consider a full WRCF to study the dynamics of a system. This observation is similar to a finding in Bendich *et al.* [34], who observed in their study that medium-scale barcode features were able to distinguish human brain artery networks from different age groups. This again suggests that persistence length should not be the only measure of signal versus noise when applying PH. We also found that the persistent features that dominate the middle part of the filtrations appear in earlier filtration steps on days 2 and 3 of the experiment than they do on day 1, which suggests that interesting dynamics

[11]Note that we use the term *geometry* for properties that are called *shape* in other contexts (see, e.g., [167]) to avoid confusion with our previous usage of the term *shape*.

in synchronisation patterns are captured by medium-lived bars in the middle of a barcode.

As in other biological contexts, where PH has been applied successfully and has lead to new insights [34, 81, 82, 90, 125, 201], we find that PH can lead to fascinating insights about the dynamics of a system. We were able not only to detect symptoms of previously observed community segregation, but we also found notable differences between a setup with strong community structure (in the coupled Kuramoto oscillators) and weakly synchronised communities (in the associated null models). For the task-based fMRI data, we found that we can detect symptoms of community structure over the three days (in the short peaks at the beginning of the landscapes) of the data as well as changes in the 1-dimensional loops that appear on average in the functional networks. On average, most of these changes appear to take place on the second day of the learning task. In particular, brain regions that yield 1-loops in the functional networks on days 2 and 3 seem to exhibit stronger synchronisation on average than those that yield 1-loops on day 1. We obtained this observation both by calculating average persistence landscapes of the WRCF performed on individual functional networks and by calculating persistence landscapes based on the WRCF performed on average networks for each day. Although the landscape distances between the average landscapes are not statistically significant, our similar results in both of our approaches suggest that our findings indeed reflect the average dynamics of the system. Our obervations on 1-dimensional loops thereby provide novel insights that complement previous studies of synchronisation in functional brain networks. We note that it would be desirable to use our approach on larger data sets to draw clearer biological conclusions from the data.

There is a known relation between homology and graph Laplacians [73], and an interesting possible direction for future research would be to study possible connections between graph Laplacians (and, more generally, spectral graph theory) and our

results on barcodes and persistence landscapes.

Using methods from topological data analysis for studying networks has the important benefit of being both mathematically principled and generalizable. However, for biological interpretation, it is necessary to include information on the specific nodes that form part of the topological features such as loops. Moreover, the interpretation of the results and importance of persistence versus position of a topological feature in the barcode can differ depending on which type of filtration is employed. Different topological features can also have different levels of relevance for different dynamical systems. For example, the occurrence of many medium-sized persistent features in the persistence landscape for the Fourier null model is a symptom of the weak synchronisation in the communities, whereas the medium-sized persistent bars capture increasing synchronisation in 1-loops for the task-based fMRI data. It would be interesting to apply WRCF (and other types of filtrations) to different synthetic networks with underlying communities (e.g., using stochastic block models) to investigate such ideas further. Importantly, one should include both the persistence and the position of topological features in analysis of PH. It would also be beneficial to combine topological tools with additional methods, such as persistence images [3], to determine the exact topological features that are responsible for the detected differences between the persistence landscapes of the different networks.

In conclusion, in this application we have shown that PH and persistence landscapes can be applied successfully to functional networks (from either experimental data or time series output of models), and that they can lead to fascinating insights, such as segregation of a network into communities and changes of network structure over time.

4.4 Application to schizophrenia data

In this application we analyse functional networks constructed from task-based fMRI data from schizophrenia patients, healthy controls, and healthy siblings of schizophrenia patients using PH. We use persistence landscapes and Betti curves to create output summaries from our persistent-homology calculations, and we study the persistence landscapes and images using k-means clustering and community detection. Based on our analysis of persistence landscapes, we find that the members of the sibling cohort have topological features (specifically, their 1-dimensional loops) that are distinct from the other two cohorts. We first give a brief introduction to schizophrenia, then describe the data set and the construction of the functional networks from the data as well as the methods that we apply to analyse the PH output, and finally present our results.

4.4.1 Schizophrenia

Schizophrenia is a chronic psychiatric disorder that affects more than 21 million people worldwide [272]. Up to 80% of the risk factors appear to be genetic, although it has proven difficult to identify the specific genes that are involved in the disease [39]. The disease usually commences in early adulthood, and symptoms range from hallucinations and avolition to cognitive deficits (such as impaired working memory) [85, 272]. The cause of the cognitive deficits is thought to originate from compromised functional integration between neural subsystems [19, 60, 85, 198]. There can be significant differences in the properties of time series from imaging measurements of healthy versus schizophrenic individuals. Different studies have found seemingly contradictory results when comparing functional magnetic resonance imaging (fMRI) time series from two distinct brain regions in a schizophrenia patient and a healthy control. The majority of studies have concluded that schizophrenia patients have less-similar time

series across different brain regions [112]. Zalesky *et al.* [275] suggested that such reduced similarity may arise from an altered coupling between brain regions and local decoherence within brain regions in schizophrenia patients. However, some studies have observed that schizophrenia patients have more-similar series than controls across brain regions. For a detailed discussion of these seemingly contradictory findings, see [111]. In some cases, methodological steps in fMRI analyses seem to yield increases in these similarities, but abnormal neurodevelopment or drug treatment may play a role in increasing them in other cases [111].

Studies of functional networks of schizophrenia patients have revealed that such networks differ significantly from the functional networks of healthy controls [8, 19, 112, 164, 166, 216, 229]. For example, schizophrenia patients can have rather different community structure from controls [8, 109]. In one paper, Alexander-Bloch *et al.* [8] observed that a small subset of brain regions lead to significant differences in the community assignments in schizophrenia patients, whereas the communities for healthy subjects appear to be consistent with each other. Moreover, the maximum modularity of functional networks appears to be smaller for schizophrenia patients than in healthy controls [7, 8]. Two recent papers, Flanagan *et al.* [109] and Towlson *et al.* [254], compared the network structures of schizophrenia patients and healthy controls under the effects of different drugs and a placebo.

4.4.2 Data

We use a data set that consists of time series from blood oxygen level dependent (BOLD) functional magnetic resonance imaging (fMRI) data collected from 281 subjects (54 schizophrenia patients, 50 healthy siblings of schizophrenia patients, and 177 healthy controls) with 120 time steps (where the length of 1 time step corresponds to $\Delta t = 2\,\mathrm{s}$). The brain regions were determined according to the Montreal Neurological Institute template [246]. Prior to obtaining the time series, the fMRI

data were corrected for head motion, and they were normalised and smoothed with a Gaussian filter. The voxel-wise signal intensities were normalised to the whole-brain global mean. The data set was acquired by Bertolino, Blasi, and their collaborators as part of a larger fMRI data set over a period of approximately 10 years. Subsets of the data set have been studied previously [41, 208, 218]; these previous studies of the data did not include the data for siblings.

The experimentalists obtained fMRI images while subjects were performing a block paradigm of a so-called 'q-back task'. During a q-back task, subjects are presented with a sequence of numbers. In each step η of the sequence, they are first shown a number and then asked to recall the number from sequence step $\eta - q$. For example, during a 2-back task, subjects are shown a sequence $\{\dots, x_{i-1}, x_i, x_{i+1}, x_{i+2}, \dots\}$ and are asked to recall number x_{i-1} while being shown number x_{i+1}, recall number x_i while being shown number x_{i+2}, and so on. For the present data set, the stimuli consisted of alternating blocks of 30 seconds each of 0-back tasks and 2-back tasks.

We preprocess the data to remove signal noise, in particular noise contributions from brain white-matter [269] and cerebrospinal fluid [83, 269] (in these areas one does not expect a response related to neuronal processes), spontaneous global signal fluctuations [45, 114, 269], as well as signal mismatch between images caused through head motion of subjects [117]. For each subject and time step, we calculate the mean signal for white-matter brain regions, the mean signal for regions that consist of cerebrospinal fluid, and the mean of the global signal. In addition to these mean values, we also use the squares and cubes of the global signal means, as well as head-motion parameters (3 translation and 3 rotation parameters), to construct 11×120 subject-specific design matrices. We then perform linear regression for each time series using MATLAB's command for the Moore–Penrose pseudoinverse [12] PINV(); we

[12] As some of the matrices are ill-conditioned, there are variations in the resulting networks across different runs of the preprocessing code. However, in our observations, the matrices differ by only up to 0.2% of entries after two runs of preprocessing.

exclude brain regions without grey matter from our calculations. We then use the residuals from the regression as our time series for the 120 brain regions that we list in Tables B.7–B.11. Note that such preprocessing steps, while common when working with fMRI data, are not uncontroversial. In particular, the effects of global signal regression have been shown to alter correlation between time series (see, for example, [115, 180] and [111] in the context of schizophrenia).

4.4.3 Construction of functional networks

We obtain functional networks from the fMRI time series for each subject by using the 120 distinct brain regions (see Tables B.7–B.11) as the nodes of the networks and calculating Pearson correlations[13] (without a time lag) between the nodes' time series as a measure of pairwise functional connectivity. The values of the pairwise functional connectivity give the edge weights between the brain regions in the functional networks.

In all but one of our analyses, we consider four contiguous time regimes of 30 time points each, yielding four functional networks per subject. (The exception is Subsection 4.4.6.4, in which we use each subject's full time series, which consists of 120 time points, to construct a single functional network for each subject.) Although the four time regimes correspond temporally to one 0-back and one 2-back task each, our separation into time regimes is motivated by an interest in potential developments in the dynamics over time, rather than in relating the fMRI response to the task. We summarise each functional network in an adjacency matrix $\tilde{A} = \tilde{A}(\text{subject}, \text{time regime})$, whose entry $\tilde{a}_{ij}$ is given by the edge weight between node i and node j. We apply a statistical threshold, as described in [23], to the weighted adjacency matrices without modifying the remaining edge weights. To obtain the thresholded adjacency matrices, we estimate p-values for the correlations using the MATLAB function CORRCOEF and

[13] There are numerous ways to measure functional connectivity [59, 233, 277]. For a discussion in the context of schizophrenia research, see [112].

retain only those entries whose p-value is less than 0.05. Using this type of thresholding, we retain at most 44% of all edges in the network and on average 20–30% of the edges. We then separate each adjacency matrix into a positive and a negative part, $\tilde{A} = \tilde{A}^+ + \tilde{A}^-$, and study only the positive $\tilde{A}^+$ part of the adjacency matrix. By using this approach, we avoid the interpretation of negative correlations between time series. In Fig. 4.8 we show a diagram of the steps that we perform to construct our functional networks. (In the one case in which we study one functional network per subject instead of four, see Subsection 4.4.6.4, we skip the step in which we split time series; we perform all other steps in the same way.) In our computations in which we consider the four time regimes separately, we treat all subjects and all time regimes together as one data set.

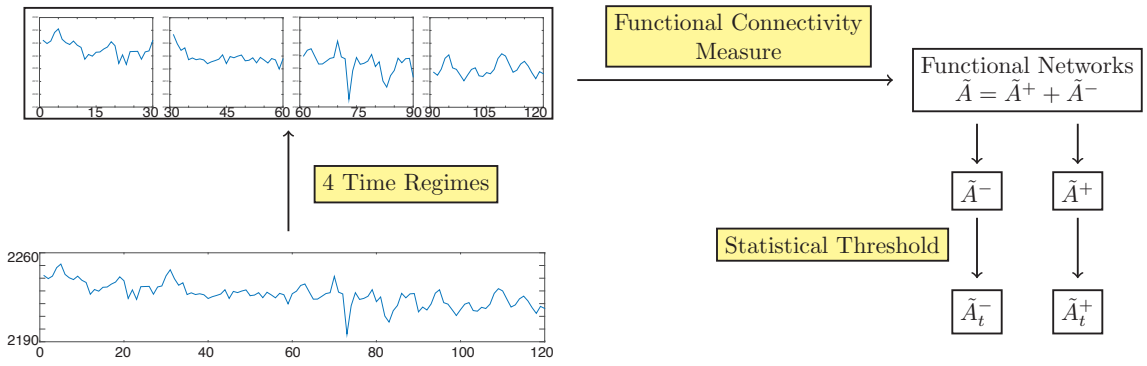


Figure 4.8: Steps that we perform on the preprocessed time series of each brain region to construct a functional network for each subject during each of four time regimes. We study the positive parts of the resulting networks using persistent homology. Modified from: [238].

4.4.4 Clustering methods from data mining and network analysis

We construct functional networks using fMRI data from schizophrenia patients, healthy controls, and siblings of schizophrenia patients. We construct a weight rank clique filtration (WRCF) [202] and compute PH and Betti numbers [79, 98] of the WRCF. We then construct persistence landscapes based on the PH output for dimension 1, i.e. loops in the networks, and examine the results by applying tools from statistics.

In [238] the PH output was also used to obtain persistent images and these were then analysed with tools from machine learning. We compare the findings from these two approaches and also study Betti numbers using Betti curves [125].

Given output of PH calculations, one can use clustering methods. There are myriad ways to proceed. We use a few different approaches: here, we apply the k-means clustering algorithm and community detection to examine whether we can separate the three subject groups based on the topological features of their functional networks. Further, in [238], linear sparse support vector machines (SSVMs) were applied to identify pixels in persistent images to discriminate between the subject groups and examine which brain regions are generators of loops that help discriminate between groups.

4.4.4.1 Employing k-means clustering for subject-group separation

The method of k-means clustering produces a partition of a metric space into k clusters of points [119]. In the initial step, the algorithm randomly selects k from the N data points to serve as initial cluster centers. All data points are then separated into clusters based on their closest cluster centers. Such a partition of the data is then given a 'score' which corresponds to the sum of the distances from each point to its nearest center. The algorithm then determines new cluster centers by taking the mean of the points assigned to every cluster, recalculates the clusters and partition scores. The process is usually repeated until the clustering score stabilises. One can apply k-means on either a distance matrix (which one can calculate for either persistence diagrams or persistence landscapes) or on a set of input vectors (such as those obtained from a persistence image).

4.4.4.2 Community detection for persistence-landscape classification

Community detection is a method from network analysis that attempts to partition a network into sets (called *communities*) of nodes that are more densely connected to

themselves than to other sets of nodes in the network [113, 183, 205]. One can detect communities in either weighted or unweighted networks. In a weighted network, one finds larger total edge weight within communities than between them.

One can also use community detection to partition data (e.g., for classification) by studying a given distance matrix of data objects such as (mean) persistence landscapes. One interprets the N persistence landscapes as N nodes of a network and converts the pairwise distances into edge weights, where a large edge weight signifies closeness in the distance matrix and a small edge weight signifies a long distance between two landscapes. We convert the distance $d(i,j)$ between landscapes i and j into an edge weight A_{ij} between nodes i and j with the following formula:

$$\tilde{A}_{ij} = 1.01 - \frac{d(i,j)}{\max_{i,j\in\{1,\dots,N\}}\{d(i,j)\}} \,. \tag{4.7}$$

This yields an adjacency matrix A with elements A_{ij}. Naturally, there are many choices for converting from pairwise distances to pairwise weights, and one has to be careful about how that influences community structure and other computations.

There are numerous methods that one can use for community detection in networks [113]. One approach for decomposing a network into communities (i.e., for performing a 'hard partitioning') is to seek a partition that maximises an objective function Q. The quality function that we use is modularity

$$Q = \sum_{i,j} [\tilde{A}_{ij} - \gamma P_{ij}] \delta(g_i, g_j) \,, \tag{4.8}$$

where P (with elements P_{ij}) is a null-model matrix (which specifies the expected edge weight between nodes i and j), the resolution parameter γ is a factor that determines how much weight one gives to the null model, and $\delta(g_i, g_j) = 1$ if nodes i and j are in the same community g (i.e., if $g_i = g_j$) and $\delta(g_i, g_j) = 0$ otherwise [113, 205].

For our computations, we use the GenLouvain package [143, 174], which maximises Q using a variant of the Louvain algorithm [46] to algorithmically detect com-

munities in our (mean) persistence landscapes. We vary the weighting factor γ (which is often called a *resolution parameter*) to compare results for different values of γ.

4.4.4.3 Linear sparse support vector machines for discriminatory feature selection

In [238] linear sparse support vector machines (SSVMs) were used to identify discriminatory features between different groups of vectorised persistence images. The vectorised persistence images were interpreted as data points that were separated into two groups with a hyperplane by SSVM using a *one-against-all* approach, i.e. three different hyperplanes were created to distinguish the persistence images from 1) controls versus those from siblings and patients, 2) siblings versus those from controls and patients, and 3) patients versus those from controls and siblings. In SSVM, each hyperplane is defined by a normal vector, whose entries are referred to as *SSVM weights*. The non-zero SSVM weights of each plane were used to determine discriminatory vector entries in the persistence images of the respective subject groups. Each discriminatory vector entry in a persistence image corresponds to a birth-persistence interval, a so-called *distinguishing pixel*. The distinguishing pixels were matched to their corresponding topological features.

4.4.5 Implementation

For our PH calculations, we implement MATLAB code constructed using JAVAPLEX [247]. For a given filtration of a simplicial complex, JAVAPLEX can output [birth, death) barcode intervals, representatives for each topological feature, and persistence diagrams. For the WRCF, we also use a maximal clique-finding algorithm (that is based on the Bron–Kerbosch algorithm [52]) from the Mathworks library [271]. For the analysis and interpretation of our barcodes, we use the PERSISTENCE LANDSCAPES TOOLBOX [55].

4.4.6 Results

We now present our results of our PH computations to examine loops in functional brain networks. We focus exclusively on topological features in dimension 1 and, except in Subsection 4.4.6.4, we perform our computations on all four time regimes as part of one data set rather than separating the data for each time regime. Aside from the aforementioned exception, we run our PH computations on four functional networks per subject. From the PH output, we create persistence landscapes. We then perform our computations either on (i) the full data set of persistence landscapes of 281 subjects and four time regimes (which gives 1124 landscapes or images, respectively, for the data set) or on (ii) the 12 subject-group means of the landscapes (from three subject groups with four time regimes each). We indicate which case we are examining in the relevant subsections. In Subsection 4.4.6.4, we consider one full time series for each subject; in other words, we study one functional network per subject.

In Stolz *et al.* 2018 [238] persistence images were further constructed from the PH output. For both persistence landscapes and persistence images, we find that there seem to be differences in the topological features of the functional networks between subject groups, although we only observe these for persistence landscapes when examining means across groups. To illustrate limitations of these methods, we also discuss results in which we were unable to find differences between subject groups.

4.4.6.1 Results of k-means clustering on persistence landscapes

Using k-means clustering on mean persistence landscapes, we are able to separate siblings of schizophrenia patients from controls and patients. For these calculations, recall that we use all four time regimes in each of the 12 mean landscapes.

We construct mean persistence landscapes from the dimension 1 barcodes (i.e., the barcodes that represent loops in the networks) for each time regime and each subject group. We obtain 12 mean landscapes and exclude infinitely-persisting bars, because all of our landscapes include (persistent) infinite features, and these tend to dominate the first several layers of the landscapes. Other researchers have excluded layers of landscapes (e.g., the first twenty) to filter out 'topological noise' [197]. Although we threshold our weighted networks prior to analysing them, this does not necessarily imply that we lose significant information by disregarding the infinite features. Additionally, infinitely persisting features do not necessarily correspond to the most persistent features in the barcodes, as even features that are born in the last filtration steps are infinitely persisting if they do not die over the course of the filtration. In our case, the presence of infinite features prevented us from discriminating between landscapes based on pairwise distances between them. When we considered infinite features separately, we did not observe any noticeable differences between the three subject groups (see Subsubsection 4.4.6.4).

We calculate a pairwise L^2 distance matrix of the mean landscapes, and we then perform k-means clustering on the distance matrix (which has 12×12 entries). For $k = 3$, we obtain the expected division of the mean landscapes into patients, controls, and siblings. Although the fact that one can separate the three cohorts based on fMRI data is not a new finding — see, for example, [8,19,112,164,166,216,229] for patients versus controls and [223] for patients versus siblings — the novelty of our work is that k-means clustering successfully distinguishes between the three different cohorts based on topological information (in the form of loops) in the functional networks.

We also perform k-means clustering for $k = 2$. Surprisingly, we find that the patients and controls are grouped in one cluster for all time regimes, whereas the siblings are in a separate cluster for all time regimes. We show the mean landscapes and clusters in Fig. 4.9.

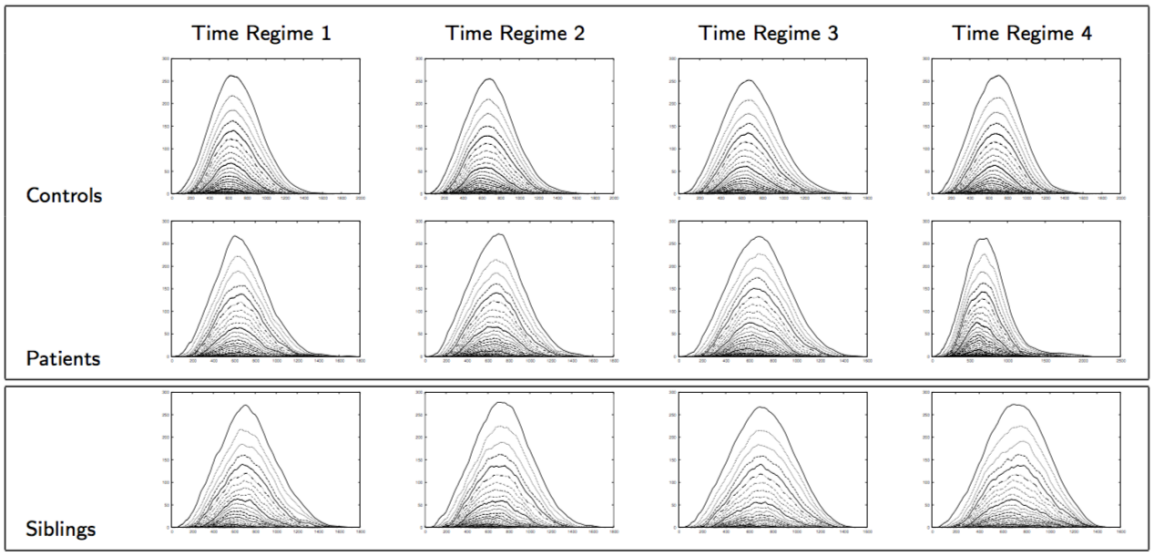


Figure 4.9: Mean persistence landscapes for each of the four time regimes and subject groups. Using k-means clustering with $k = 2$ on the set of 12 persistence landscapes (which consists of all subject-group means and time regimes as one data set) assigns patients and controls to one group. We show the mean persistence landscapes and their k-means-clustering grouping for the four time regimes separately. Image source: [238].

For larger values of k, we do not observe a clear subject-group separation. To compare our results with ones from other clustering methods, we also apply average linkage clustering to the distance matrix and perform community detection on networks that we construct from the distance matrices (as described in Subsection 4.4.4.2). We obtain the same qualitative result for these two methods. For community detection, we observe a clear separation for resolution-parameter values $\gamma = 0.82, 0.83, \ldots, 1.14$ into two communities (the siblings versus the patients and controls). Our results appear to indicate that the sibling cohort is particularly distinct from the other two cohorts, as compared to any other pairwise comparison among the three cohorts, with respect to their loop topology in the functional networks.

We also perform a permutation test on the mean persistence landscapes for each time regime to determine the significance of the landscape distances, as suggested in [55]. In the permutation test, we regroup the individual landscapes into three groups uniformly at random, create a new mean landscape for each newly assigned group, and calculate the pairwise L^2 distances between them. We then count how

many of the L^2 distances of the new groups are larger than the ones that we observe when using the mean landscapes of the three subject groups. We use 10000 permutations to obtain our results, which we summarise in Table 4.2.

Table 4.2: Using a permutation test, we calculate p-values for the pairwise distances between the mean landscapes of the three subject groups in each time regime.

p-values for	Controls vs Patients	Controls vs Siblings	Patients vs Siblings
time regime 1	0.302	0.200	0.051
time regime 2	0.460	0.009	0.052
time regime 3	0.477	0.102	0.270
time regime 4	0.736	0.110	0.229

Interestingly, for time regimes 1 and 2, we find significant distances between the patient and sibling mean landscapes, whereas the p-values for time regime 3 and 4 suggest that the distance is not significant (even though the p-values are comparably small). The distance between the mean landscapes of the controls and the siblings appears to be significant for time regime 2, but this does not appear to be the case for the other time regimes, although the p-values are again much smaller than for the distances between the mean landscape of the patients and controls. Thus, for the controls and the patients, there are many other divisions into two groups that lead to more extreme distances between the mean landscapes than what one obtains by simply assigning them to a control group and a patient group.

To see if we can further support our result from k-means clustering for $k = 2$, we artificially group the controls and patients into one group to create a mean landscape and again perform a permutation test to verify whether the distance between the mean landscapes for the two groups is significant. In Table 4.3, we show the p-values that we obtain with 10000 permutations.

For time regime 2, we obtain a significant distance, but the p-values for time regimes 1, 3, and 4 are approximately 0.1. Given the artificial grouping of the two subject groups, we construe these values as small, although they are not statistically significant.

Table 4.3: Using a permutation test, we calculate p-values for the controls-and-patients mean landscape versus the siblings mean landscape.

Time regime 1	Time regime 2	Time regime 3	Time regime 4
0.112	0.008	0.092	0.110

4.4.6.2 Results of community detection on a distance matrix from individual persistence landscapes

We construct persistence landscapes from each of the dimension 1 barcodes, which we calculate by examining each subject in each of the four time regimes, and we calculate the L^2 distance matrix for the resulting 1124 persistence landscapes. We again use the distance matrix to construct a network between the persistence landscapes, and we detect communities in this network by maximising modularity. For $\gamma = 0.92, 0.93, \ldots, 1$, we obtain a separation into two communities. The partition that is closest to what we observe with 2-means clustering for the mean landscape distance occurs for the resolution-parameter value $\gamma = 0.93$. We summarise our results in Table 4.4.

Table 4.4: Number of subjects from each subject group that are assigned to communities 1 and 2 by community detection using modularity maximisation.

Subject group	Nr. of subjects in community 1	Nr. of subjects in community 2
Patients	122	94
Controls	418	290
Siblings	93	107

We also apply k-means clustering and average linkage clustering to the distance matrix from the individual persistence landscapes (results not shown). Of all classification methods that we perform on these distance matrices, community detection appears to perform best at 'separating' the subject groups, although we do not observe a very clear separation.

4.4.6.3 Summary of results from analysis of persistence images

In Stolz *et al.* [238] persistence images were used to identify discriminatory topological features across the three subject groups considered. The persistence images were generated for each of the subjects and each of the four time regimes based on the dimension 1 persistence diagrams using the defaults in the persistence image code available from [2]. To compute persistence images one must in addition choose two parameters specific to the data: the maximum birth and persistence values, which determine the discretisation of the pixel boundaries in the images once one sets the resolution. Possibilities include taking the maximum birth and persistence values across all persistence diagrams or normalising each persistence diagram relative to its individual maximum. In the original paper on persistence images [3], the maximum values were chosen across all persistence diagrams under consideration although no theoretical rationale was provided for this choice. It was not possible to obtain clear results using either of these conventions on our data set. For more detailed analysis of the effect of using different versions of these data specific parameters by Tegan Emerson see Appendix B.2.2.1 or [238].

Using a priori knowledge of subject-group membership and fixing the maximum birth values separately for each subject group (based on the collection of persistence diagrams that were computed separately for each subject group), it was possible discriminate between the three subject groups. This provides a first interesting observation from persistence images: the maximum birth time which corresponds (or almost corresponds, in exceptional cases in which multiple edges have exactly the same weight) to the number of pairs of regions in the brain with positive functional connectivity, appears to contain non-trivial information. (Recall that we do not include edges that correspond to negative Pearson correlations.)

Surprisingly, despite the pronounced difference in persistence image performance when different maximum values were used for each class, the distributions of the

maximum birth times and persistences for each subject type are not statistically-significantly different from each other (see Fig. B.13 in Appendix B.2.2.1). The results that we discuss subsequently are based on the persistence images generated using a priori membership knowledge.

A linear SSVM was applied to the set of persistence images to identify distinguishing pixels that allow interpretation of classification results. Using a one-against-all SSVM with 5-fold cross validation, it was possible to obtain a 100% classification accuracy. See Fig. B.14 in Appendix B.2.2.1 for an illustration of the distinguishing pixels. Each distinguishing pixel corresponds to a bounded region in the birth–persistence plane. The following approach was taken to connect the distinguishing pixels with particular brain regions: for each subject, it is possible to determine whether or not a topological feature (in our case, a loop) in a filtration of a network exists in the bounded region of the birth–persistence plane that corresponds to a particular distinguishing pixel. If a loop does exist, one can identify a set of brain regions that comprise the loop (i.e., representatives of this loop). Brain regions that are consistently involved in the generation of particular loops across subjects are of particular interest. To find such brain regions, the set of nodes, called *top node(s)*[14], that are involved in the generation of loop(s) for each distinguishing pixel in each of the four of the time regimes for each subject was identified. Interestingly, for the patients there are only five distinguishing pixels for which top nodes were found. By contrast, for the siblings there are many distinguishing pixels for which top nodes were obtained. The control group lies between the other two in terms of its number of distinguishing pixels with top nodes, but there are still few top nodes, relative to the number of distinguishing pixels that have top nodes. We list the full list top nodes as well as further details in Appendix B.2.2.2.

[14]One can construe the calculation of top nodes in a similar spirit as calculations of node centralities [183].

4.4.6.4 Results from Betti curves

Finally, we also study *Betti curves*, introduced in [125], which describe Betti numbers and their changes across a filtration. We use the entire time period (i.e., one time regime, rather than four separate ones) of the experiment. In all other respects, we construct the functional networks as we described previously (see Subsection 4.4.3). We compute the mean and standard deviation across the Betti numbers for dimension 1 (i.e., the number of loops) for each cohort in each filtration step. We find that, apart from a slightly larger standard deviation in the patient cohort, the Betti curves for the three groups look essentially the same. We show our results in Fig. 4.10.

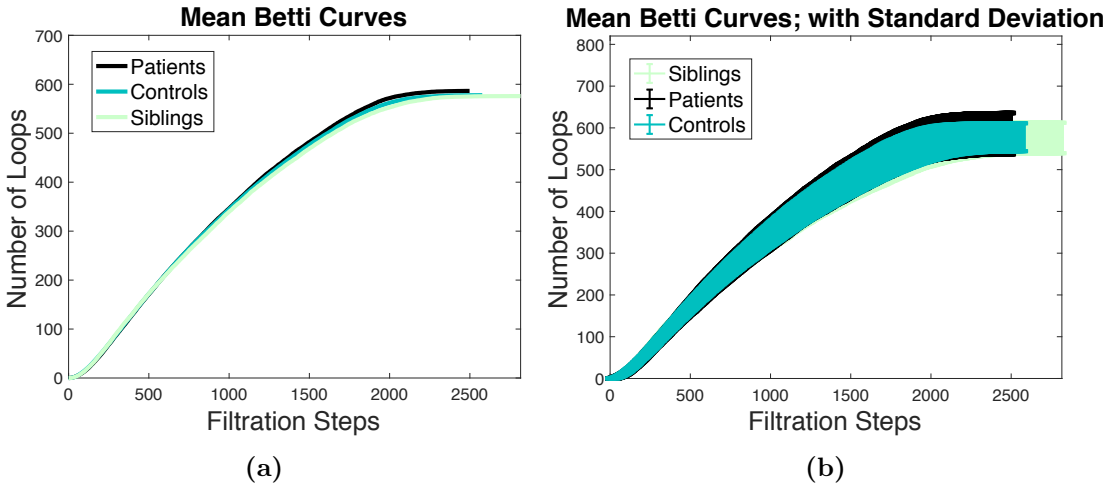


Figure 4.10: (a) Mean Betti curves for the patients, controls, and siblings. (b) Mean Betti curves and their standard deviations for patients, controls, and siblings. Image source: [238].

4.4.7 Summary and discussion

We applied methods from PH to analyse loops in functional brain networks of schizophrenia patients, siblings of schizophrenia patients, and healthy controls. We constructed persistence landscapes and we analysed them using several clustering techniques. In Stolz *et al.* [238] persistence images were constructed from the same PH output and also analysed using clustering techniques. We compare insights from both approaches.

We observed topological differences in the functional brain networks of schizophrenia patients, siblings of schizophrenia patients, and healthy controls with respect to the loops in their networks. We also found that persistence landscapes and persistence images have different practical advantages and disadvantages when applied to the same data set, and these insights may be useful for interpreting the results of PH computations in networks in diverse applications.

Computing persistence landscapes gave interesting results when comparing mean persistence landscapes of the cohorts but not when comparing individual landscapes of the subjects. Using mean persistence landscapes, we were able to separate the sibling cohort from the other two subject groups in each of the four time regimes. This is supported by the p-values that we obtained for the distances between the mean landscapes of the sibling cohort versus the controls and patient cohorts, though not all of our p-values are statistically significant. The shape of the mean persistence landscapes seems to suggest that loops that occur in the functional brain networks of siblings are on average more persistent than those in the functional networks of controls or patients. This could imply either that loops in the networks of siblings tend to be larger or that the third edge between three nodes has a small edge weight and thus that three brain regions with a large pairwise Pearson correlation between one region and two of the other regions do not necessarily imply that there is a large correlation between the other two brain regions; this facilitates the creation of a loop structure in the filtration. (Recall that we need at least four nodes for our loops.) To examine this issue further, it may be useful to analyse cross links in the functional networks, as in [24]. For the above computations and their interpretation, we need to take into account that we did not include infinitely-persisting loops (which persist until the end of a filtration). We also only include positive edge weights in our networks, so we only analysed loops that arise from brain regions with positive pairwise Pearson correlations.

Although we were able to obtain interesting insights about the data using mean persistence landscapes, we did not find interpretable results from comparing individual landscapes, and only being able to use the mean landscapes reduces the amount of information the we can obtain from this approach. By contrast, using individual persistence images and SSVMs allowed the separation of the entire set of subjects (with 100% accuracy) in each of the four time regimes. In previous work, Anderson and Cohen [10] obtained 65% accuracy for schizophrenia classification by applying machine-learning techniques to functional brain networks. A study on Alzheimer's disease using persistence landscapes [54] and machine learning attained a 73% separation of diseased and healthy subjects. It is important to note, however, that the results are based on using a priori knowledge of group membership, including specifically the maximum birth times of loops within subject groups. These birth times seem to include nontrivial information, which is important to pursue further in future studies. This a priori knowledge is tied closely to the choice of statistical thresholding when preprocessing fMRI data. Developing a statistical model that can classify a novel subject based on a persistence image representation thus also requires further exploration into how to choose such a threshold.

Computing persistence images also allowed the identification of brain regions with consistent involvement in loops in the functional networks within subject cohorts. Of the three cohorts, it was observed that siblings have the highest level of consistent brain-region involvement in the performance of the mental task in this study across the four time regimes. That is, regions that are involved in loops for siblings in one of the time regimes are more likely to also be involved in loops in other time regimes than is the case for patients or controls. It is particularly noteworthy that the number of brain regions that are consistently involved in the separation of the three cohorts is larger in the siblings of schizophrenia patients than in the healthy controls. We view variable involvement of brain regions in loops as a notion of neurological 'flexibility'. Various

works have studied concepts of brain flexibility using community structure [23, 48]. In those studies, flexibility was defined differently — based on how often a brain region changes its allegiance to a community of nodes over time, so it does not use loops directly — but it is noteworthy that Braun *et al.* [48] observed that relatives of schizophrenia patients have large flexibility than healthy controls. In Stolz *et al.* [238], it was found that a specific group of brain regions leads to the separation of the three subject groups when using persistence images and observed for the schizophrenia patients that the regions that lead to a separation consistently in each of the four time regimes are fewer in number than for the siblings and controls. Braun *et al.* [48] reported that there is larger node flexibility in network organisation of schizophrenia patients than in healthy controls. Additionally, Siebenhühner *et al.* [225] observed a greater variability in temporal networks constructed from Magnetoencephalography (MEG) data of schizophrenia patients than those from in healthy controls.

We did not observe any differences between the four time regimes, which each consist of responses during a 0-back task and a 2-back task, in any of our calculations. No significant changes seem to be occurring in the persistence or appearance of loops in the networks over the course of the data measurement. Additionally, when studying experiments as a single regime using Betti curves, we did not observe a clear difference between the cohorts.

Schizophrenia has a high genetic determinism, so siblings of schizophrenia patients have a significant genetic risk of developing the disease themselves [40], and it has been demonstrated that they have abnormalities in their structural neuronal networks [78]. Although our results that functional brain networks constructed from fMRI measurements of siblings differ both from patients and from healthy controls do not agree completely with the current standard in the literature, other studies have also reported that the features of fMRIs of siblings of schizophrenia patients differ from both schizophrenia patients and controls. For example, Callicott *et al.* [65] ob-

served in an fMRI study that there was no difference in task performance between healthy siblings of schizophrenia patients and healthy controls, yet they detected a physiological similarity between the sibling cohort and the schizophrenia patients in the corresponding fMRI data. Similarly, Sepede *et al.* [223] observed using fMRI data from a different data set that healthy siblings of schizophrenia patients exhibit differences in brain function to schizophrenia patients, although they did not differ significantly in task performance.

It was demonstrated recently that schizophrenia patients undergo a cortical normalisation process over the course of the disease [130], and a current study on blood samples of schizophrenia patients [222] has also observed that the measurements for patients who have had the disease for a long time are more similar to the measurements of healthy controls than to those of early stage patients. We would need further phenotypic information to assess whether any of the aforementioned studies can be connected more directly to our observations.

As our results are somewhat inconsistent with prior observations, it is also possible that our data set contains experimental noise that is beyond our control. Using standard network-analysis techniques, we do not observe any differences between the three subject groups. Nevertheless, we believe that our comparison of persistence landscapes to persistence images and the different types of results from these techniques provide a valuable example of a PH approach to functional brain networks.

As mentioned previously in Subsubsection 4.3.3.2, one needs to take into account that there are difficulties when interpreting the information about node participation in loops from computations of PH, as the software used for such computations (including, specifically, JAVAPLEX, which is what we used) only finds representatives of the loops. Despite these difficulties, the list of discriminating nodes in Stolz *et al.* [238] provides a useful starting point for further investigations into neuronal abnormalities in functional networks of schizophrenia patients.

Another important issue is that we preprocessed the data for our study. This is very common when working with fMRI data, but such steps are not uncontroversial [115, 180], indeed studies on functional connectivity in schizophrenia patients have found contradictory results depending on whether or not one performed global signal correction [111]. It is also relevant to keep in mind that the choice of functional connectivity measure can influence results [233]. We chose to use a Pearson correlation due to its simplicity and the fact that it is a widely used measure of functional connectivity [20, 264]. Many other choices are also available.

"kurz unde lank nach minem sin
die lingen von dem zentrum hin
ich leite zu dem ummesweif
uf aller speren zirkelreif
von punt zu punt in rechter saß
die ling uf alle winkelmaß."[a]

From the speech by the personified art of geometry (Geometria) in *Der meide kranz*, Heinrich von Mügeln, 14th century. Edition by [263].

[a]Translation: "Long and short as it pleases me, I draw lines from the centre to the outer periphery, through the orbits of all the spheres; [I draw] the lines through all angles, from point to point in proper order." [263].

5 Applications of Local Persistent Homology

One of the practical challenges of applying persistent homology (PH) is that it is computationally difficult for large data sets (see Chapter 1). For example, the Vietoris–Rips complex[1] can be infeasible for data sets with as little as 3000 points, even when considering PH only in the first two dimensions and performing computations on powerful computers with over 750 GB of RAM. Locally, in a small neighbourhood around a data point, these problems vanish. While one of the appeals of persistent homology is its ability to study multi-scale data sets globally, local information may also produce useful insights.

Here, we explore two applications of computing the Vietoris–Rips complex locally. In our first application, we illustrate how local PH[2] can be used to select landmarks from large and noisy data sets. We use very simple data sets that consist of signal

[1]See Chapter 2, Subsubsection 2.3.1.1 for definition.

[2]Note that the idea behind our notion of local persistent homology is similar to the one used, for example, in [5, 35, 104]. In contrast to these approaches, our definition does not use relative homology and can be computed in a more intuitive way by considering the Vietoris–Rips complex on points in a local neighbourhood.

points sampled from an object with a topologically interesting structure, such as a sphere, a torus, or a Klein bottle, as well as noise points. We specifically include noise in our data sets, as we are motivated by practical applications of PH to large and noisy real-world data sets. A consequence of the inclusion of noise is that a large data set cannot be reduced by applying subsampling techniques developed to infer the (persistent) homology of data, for example developed in [76, 96, 187]. We present an algorithm that computes PH in small neighbourhoods around data points and uses PH output to define a score for every data point which allows us to identify suitable candidates for landmarks. The (global) PH of landmarks selected in this way is close to the PH of signal points in the original data set. In comparison to existing landmark selection procedures, our landmarks based on local PH perform very well on our data sets, in particular for low sampling densities.

As our second application of local PH, we develop an algorithm that examines local PH around data points to detect non-manifold[3] like singular regions in data sampled from intersecting surfaces. The method can in particular be used to distinguish between points that are close to a boundary, points that are close to an intersection, and points that are neither close to an intersection nor to a boundary in such data, even when none of the points were explicitly sampled from singular regions. We showcase this method on complex, high-dimensional data, and notably find the intersection of two data surfaces in the 24-dimensional space of conformations of cyclo-octane. The method and results are also described in our preprint [242]. Here, we further compare our approach to local principal component analysis (PCA) which can also be applied to identify intersections. We find that our method has distinct advantages over local PCA.

For both of our applications of local PH, the local nature of the methods enables us to completely parallelise the computations.

[3]For a definition of a manifold, see Appendix A, Definition A.2.3.

5.1 Mathematical motivation

The use of our notion of local PH is inspired by the Mayer-Vietoris sequence, which can enable computation of the homology of a space X by considering subspaces, whose homology is easier to compute. The following definitions and explanations are based on [133, 177]. To understand the Mayer-Vietoris sequence, we first introduce *exact sequences*.

Definition 5.1.1 (Exact sequence)**.** Let $i \in \mathbb{Z}$ and

$$\cdots \to \mathcal{A}_i \xrightarrow{\Phi_i} \mathcal{A}_{i+1} \xrightarrow{\Phi_{i+1}} \mathcal{A}_{i+2} \xrightarrow{\Phi_{i+2}} \ldots, \tag{5.1}$$

be a sequence (finite or infinite) of abelian[4] groups and homomorphisms. We call the sequence *exact* at $\mathcal{A}_{i+1}$, if

$$\operatorname{im} \Phi_i = \ker \Phi_{i+1}. \tag{5.2}$$

If the sequence is exact at all $\mathcal{A}_i$, we say that it is an *exact sequence*.

Remark 1. If there exists a first and/or last group in the sequence, exactness is not defined at these points.

Two important types of exact sequences are the *long exact sequence* and the *short exact sequence*:

Definition 5.1.2 (Long exact sequence)**.** Let $i \in \mathbb{Z}$ and

$$\cdots \to \mathcal{A}_i \xrightarrow{\Phi_i} \mathcal{A}_{i+1} \xrightarrow{\Phi_{i+1}} \mathcal{A}_{i+2} \xrightarrow{\Phi_{i+2}} \ldots, \tag{5.3}$$

be an exact sequence. We call the sequence a *long exact sequence*.

Remark 2. Note that a long exact sequence can begin or end with an infinite string of trivial groups.

[4] One can define exact sequences for general groups, but we will only be using abelian groups in this thesis.

Definition 5.1.3 (Short exact sequence)**.** Let

$$0 \to \mathcal{A}_1 \xrightarrow{\Phi_1} \mathcal{A}_2 \xrightarrow{\Phi_2} \mathcal{A}_3 \to 0, \tag{5.4}$$

be an exact sequence. We call such a sequence *short exact sequence.*

Remark 3. It follows immediately by exactness that Φ_1 is an injective homomorphism and Φ_2 is a surjective homomorphism. Moreover, by the first isomorphism theorem (see page 44 of [136]) we have that $\mathcal{A}_3 \cong \mathcal{A}_2/\ker\Phi_2 = \mathcal{A}_2/\operatorname{im}\Phi_1$.

One can define a short exact sequence for chain complexes[5] in the following way:

Definition 5.1.4 (Short exact sequence of chain complexes)**.** Let $\mathcal{E}$,$\mathcal{F}$,$\mathcal{G}$ be chain complexes, e.g. $\mathcal{E} = \{E_n, \partial_E\}$, $\mathcal{F} = \{F_n, \partial_F\}$ and $\mathcal{G} = \{G_n, \partial_G\}$ with dimensions $n \in \mathbb{N}_0$. Let $\Phi : \mathcal{E} \to \mathcal{F}$, $\Psi : \mathcal{F} \to \mathcal{G}$ be chain maps and let 0 denote the trivial chain complex, whose groups are trivial in every dimension. We say the sequence

$$0 \to \mathcal{E} \xrightarrow{\Phi} \mathcal{F} \xrightarrow{\Psi} \mathcal{G} \to 0, \tag{5.5}$$

is exact if

$$0 \to E_n \xrightarrow{\Phi} F_n \xrightarrow{\Psi} G_n \to 0, \tag{5.6}$$

is an exact sequence for each dimension n.

We can now connect a short exact sequence of chain complexes to a long exact sequence of homology groups[6] via the zig-zag lemma.

Lemma 5.1.1 (Zig-zag Lemma)**.** *Let $\mathcal{E}, \mathcal{F}$,$\mathcal{G}$ be chain complexes with chain maps Φ and Ψ in a short exact sequence:*

$$0 \to \mathcal{E} \xrightarrow{\Phi} \mathcal{F} \xrightarrow{\Psi} \mathcal{G} \to 0. \tag{5.7}$$

[5]See Chapter 2, Subsection 2.1.2 for definition. For a definition of chain maps, see, for example, page 72 in [177].

[6]Note that in Chapter 2, Subsection 2.1.3 we define homology groups $H_n(X)$ of a simplicial complex X via the boundary operators of its chain complex $\mathcal{C}$. Here, instead of $H_n(X)$, we write $H_n(\mathcal{C})$ or $H_n(\mathcal{C}(X))$ when we want to emphasise which chain complex in the sequence we are considering. The definition of H_n, however, remains the same.

Then there exists a long exact sequence between the homology groups of the chain complexes:

$$\cdots \to H_n(\mathcal{E}) \xrightarrow{\Phi_*} H_n(\mathcal{F}) \xrightarrow{\Psi_*} H_n(\mathcal{G}) \xrightarrow{\partial_*} H_{n-1}(\mathcal{E}) \xrightarrow{\Phi_*} H_{n-1}(\mathcal{F}) \xrightarrow{\Psi_*} \ldots \xrightarrow{\Psi_*} H_0(G) \to 0, \tag{5.8}$$

where ∂_* is induced by the boundary operator in $\mathcal{F}$.

We refer the reader to [177] page 136 for a proof of the zig-zag lemma, in particular for the existence of the map ∂_*. The Mayer-Vietoris sequence for a topological space X is a special type of long exact sequence that follows directly from the zig-zag lemma:

Theorem 5.1.2 (Mayer-Vietoris sequence)**.** *Let X be a simplicial complex with subcomplexes $A, B \subset X$ such that $X = A \cup B$. Then there exists an exact sequence*

$$\cdots \to H_n(A\cap B) \xrightarrow{\Phi_*} H_n(A)\oplus H_n(B) \xrightarrow{\Psi_*} H_n(X) \xrightarrow{\partial_*} H_{n-1}(A\cap B) \to \cdots \to H_0(X) \to 0. \tag{5.9}$$

The sequence is called the Mayer-Vietoris sequence*.*

Proof. The proof follows [177], page 142. We start with the construction of a short exact sequence of chain complexes to which we can later apply the zig-zag lemma to obtain the Mayer-Vietoris sequence:

$$0 \to \mathcal{C}(A \cap B) \xrightarrow{\Phi} \mathcal{C}(A) \oplus \mathcal{C}(B) \xrightarrow{\Psi} \mathcal{C}(A \cup B) \to 0, \tag{5.10}$$

where $\mathcal{C}$ denotes that we consider the chain complexes of the simplicial complexes. Before showing that the sequence is indeed exact, we ensure that the definitions of all chain complexes and maps are clear. For a given dimension n we define the chain group

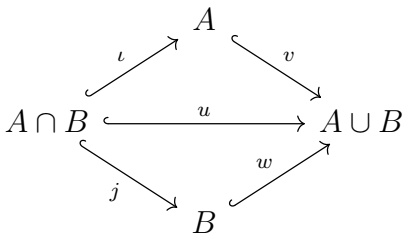
$$C_n(A) \oplus C_n(B), \tag{5.11}$$

with the boundary operator

$$\partial' : (a, b) \mapsto (\partial_A a, \partial_B b), \tag{5.12}$$

where ∂_A is the boundary operator in $\mathcal{C}(A)$ and ∂_B is the boundary operator in $\mathcal{C}(B)$.

Now consider the following commutative diagram of inclusion maps:



We can define the homomorphisms Φ and Ψ via the chain maps induced by the inclusion maps in the following way:

$$\Phi : c \mapsto (\iota_\#(c), -j_\#(c)), \tag{5.13}$$

$$\Psi : (a, b) \mapsto v_\#(a) + w_\#(b). \tag{5.14}$$

We now show that these maps and chain complexes give rise to an exact sequence.

1. Exactness at $\mathcal{C}(A \cap B)$: this is equivalent to showing that Φ is injective and follows immediately since $\iota_\#$ and $j_\#$ are inclusion maps of chains.

2. Exactness at $\mathcal{C}(A \cup B)$: we need to show that Ψ is surjective. Let $d \in \mathcal{C}(A \cup B)$. We can write d as $d = d_A + (d - d_A)$, where d_A is the part of d that is in A. We now have $\Psi(d_A, d - d_A) = d_A + d - d_A = d$.

3. Exactness at $\mathcal{C}(A) \oplus \mathcal{C}(B)$: we need to show that $\operatorname{im} \Phi = \ker \Psi$.

   - $\operatorname{im} \Phi \subseteq \ker \Psi$: Let $c \in \mathcal{C}(A \cap B)$, then

$$\Psi(\Phi(c)) = \Psi(\iota_\#(c), -j_\#(c)), \tag{5.15}$$

$$= v_\#(\iota_\#(c)) - w_\#(j_\#(c)), \tag{5.16}$$

$$= u_\#(c) - u_\#(c), \tag{5.17}$$

$$= 0. \tag{5.18}$$

- $\ker \Psi \subseteq \operatorname{im} \Phi$: Let $(a, b) \in \mathcal{C}(A) \oplus \mathcal{C}(B)$ such that $\Psi(a, b) = 0$. Then $\Psi(a, b) = v_{\#}(a) + w_{\#}(b) = 0$ and hence $v_{\#}(a) = -w_{\#}(b)$ and $a = -b$ when considered as chains in $\mathcal{C}(A \cup B)$. Since $a \in \mathcal{C}(A)$ and $b \in \mathcal{C}(B)$, we now know that $a, b \in \mathcal{C}(A \cap B)$ and $(a, b) = (a, -a) = \Phi(a)$.

For the chain complex $\mathcal{C}(A) \oplus \mathcal{C}(B)$, the n-th homology group is given by

$$\frac{\ker \partial'_n}{\operatorname{im} \partial'_{n+1}} = \frac{\ker \partial_{A,n} \oplus \ker \partial_{B,n}}{\operatorname{im} \partial_{A,n+1} \oplus \operatorname{im} \partial_{B,n+1}}, \tag{5.19}$$

$$\cong H_n(\mathcal{C}(A)) \oplus H_n(\mathcal{C}(B)), \tag{5.20}$$

$$= H_n(A) \oplus H_n(B). \tag{5.21}$$

The Mayer-Vietoris sequence now immediately follows from the zig-zag lemma. □

Remark 4. The Mayer-Vietoris sequence can be formulated more generally for a topological space X with subspaces $A, B \subset X$.

We can use the Mayer-Vietoris sequence to connect the homology of a simplicial complex X to the local homology around a vertex $\hat{x} \in X$. We do this via two simplicial subcomplexes connected to the vertex $\hat{x} \in X$: the link of the vertex $\hat{x} \in X$ and the closed star of the vertex $\hat{x} \in X$.

Definition 5.1.5 (Closed star of a vertex)**.** Let X be a simplicial complex and $\hat{x} \in X$ a vertex. Then the closed star of $\hat{x}$ in X, denoted by $\overline{\mathrm{St}}\ \hat{x}$, is the subcomplex of X that contains all the simplices which have $\hat{x}$ as one of their vertices.

Definition 5.1.6 (Link of a vertex)**.** Let X be a simplicial complex and $\hat{x} \in X$ a vertex. Then the link $\mathrm{Lk}\ \hat{x}$ is the union of all simplices of X lying in $\overline{\mathrm{St}}\ \hat{x}$ that are disjoint from $\hat{x}$.

We show a simplicial complex, the closed star of a vertex and the link of a vertex in Fig. 5.1 . Denoting the simplicial complex of all simplices in X that are disjoint from $\hat{x}$ as $X \setminus \hat{x}$, we make the following observations for $\hat{x} \in X$:

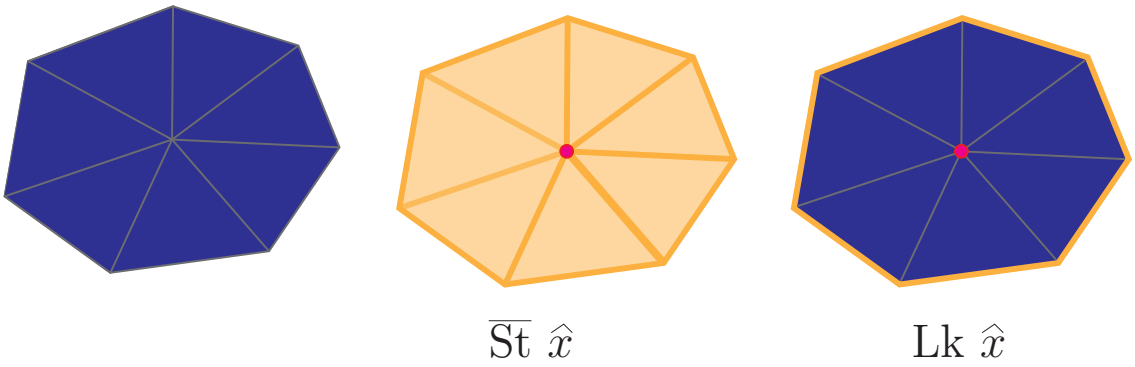


Figure 5.1: Examples of a simplicial complex, the closed star of a vertex $\hat{x}$, and the link of a vertex $\hat{x}$. We show the vertex $\hat{x}$ in red and highlight the closed star and the link in yellow.

Remarks 1.

1. $X = (X \setminus \hat{x}) \cup \overline{\text{St}}\ \hat{x}$.

2. $\text{Lk}\ \hat{x} = (X \setminus \hat{x}) \cap \overline{\text{St}}\ \hat{x}$.

Based on these definitions and observations we can now consider the following Mayer-Vietoris sequence:

$$\cdots \to H_n(\text{Lk}\ \hat{x}) \xrightarrow{\Phi} H_n(X \setminus \hat{x}) \oplus H_n(\overline{\text{St}}\ \hat{x}) \xrightarrow{\Psi} H_n(X) \xrightarrow{\partial} H_{n-1}(\text{Lk}\ \hat{x}) \to \cdots \to H_0(X) \to 0. \tag{5.22}$$

Now, $\overline{\text{St}}\ \hat{x}$ is contractible: every simplex that contains $\hat{x}$ is contractible and the intersection of simplices in $\overline{\text{St}}\ \hat{x}$ is either empty or a simplex that contains $\hat{x}$ and is hence also contractible. Thus $H_n(\overline{\text{St}}\ \hat{x}) = 0$ for $n > 0$. We observe that if we can ensure that $H_n(\text{Lk}\ \hat{x}) = H_{n-1}(\text{Lk}\ \hat{x}) = 0$, for $n > 0$ we obtain:

$$0 \xrightarrow{\Phi} H_n(X \setminus \hat{x}) \xrightarrow{\Psi} H_n(X) \xrightarrow{\partial} 0, \tag{5.23}$$

which gives us an isomorphism Ψ between $H_n(X \setminus \hat{x})$ and $H_n(X)$.

The Mayer-Vietoris sequence given by Equation 5.22 connects the homology of the large simplicial complexes X and $X \setminus \hat{x}$, which are both global and expensive to compute, to the completely local homology of the Link $\text{Lk}\ \hat{x}$, which is easy to compute.

In practice, we are however working with point cloud data D. We therefore consider the PH of a data point $y \in D$, rather than the homology, and apply the Mayer-Vietoris sequence 5.22 to the simplicial complexes that we obtain from the data by constructing a filtration. We consider a simplicial complex X in this filtration. A data point $y \in D$ is now a vertex $\hat{y}$ in X. Instead of $H_n(\text{Lk } \hat{y}) = 0$, we will demand $PH_n(\text{Lk } y) \approx$ small to quantify the failure of the isomorphism $PH_n(D \setminus y) \xrightarrow{\Psi} PH_n(D)$. To make computation easier, instead of looking just at the link of $\hat{y}$ in the simplicial complex X, we extend the link to a δ-neighbourhood of $\hat{y}$ in X, which we define to be the collection of simplices whose vertices are within a distance of at most δ from y in D. We now give the definition of a δ-Link of a data point $y \in D$:

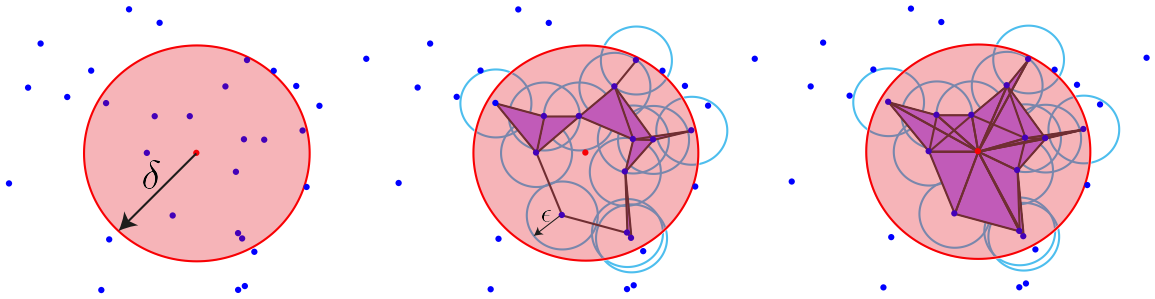
Definition 5.1.7 (δ-link of a data point y)**.** Let X be a simplicial complex in a metric space constructed from a data set D, $\hat{y} \in X$ a vertex, $\delta > 0$ a distance and $\delta(\hat{y})$ a δ-neighbourhood of $\hat{y}$ in X. Then the δ-link $\text{Lk}^{\delta}\ y$ is the union of all simplices in X that are disjoint from $\hat{y}$ and contained in $\delta(\hat{y})$.

Building on the δ-Link of a data point we can define the δ-Star of a data point:

Definition 5.1.8 (Closed δ-star of a data point y)**.** Let X be a simplicial complex in a metric space constructed from a data set D, $\hat{y} \in X$ a vertex, $\delta > 0$ a distance and $\delta(\hat{y})$ a δ-neighbourhood of $\hat{y}$ in X. Then the closed δ-star $\overline{\text{St}}^{\delta}\ \hat{y}$ is the union of the δ-link $\text{Lk}^{\delta}\ y$ with the vertex $\hat{y}$ and all simplices $[\hat{y}, \sigma]$ where $\sigma \in X$ and σ is fully contained in $\delta(\hat{y})$.

Remark 5. The closed δ-star of a data point is always contractible by construction.

We show an example of a data point, its δ-link and its closed δ-star in Fig. 5.2. From now on, we refer to computing the PH of the δ-Link of a point in a data set as computing the local PH of a data point. Our notion of local PH of a data point is motivated by the Mayer-Vietoris sequence and we will, in Subsection 5.2.2, use



(a) δ-neighbourhood of y in D. **(b)** Lk^{δ} y at scale ϵ. **(c)** $\overline{\mathrm{St}}^{\delta}$ y at scale ϵ.

Figure 5.2: Examples of a data point y and its δ-neighbourhood in a point cloud, the δ-link of y and the closed δ-star of y. We show the data point y in red and its δ-neighbourhood in light red highlighting the points within δ in blue. We use the data points within the δ-neighbourhood to build a Vietoris–Rips complex (see Chapter 2, Subsubsection 2.3.1.1 for definition) for a set filtration value $\epsilon > 0$, which represents the simplicial complex X in the definitions of the δ-link and the closed δ-star. We only show the subcomplexes of the Vietoris–Rips complex (or their extensions in the case of the closed δ-star) that are relevant to the illustrated definitions.

property 5.23 to define a new landmark selection method, in which we select points y in a data set D which are 'closest' to giving us the desired isomorphism between $PH_n(D \setminus y)$ and $PH_n(D)$ as landmarks. In Section 5.3 we demonstrate that local PH can also be used on its own to retrieve local geometric information about a data point on a surface.

5.2 Outlier-robust landmark selection in large and noisy data sets

5.2.1 Existing landmark selection methods

The two standard methods for selecting landmarks $L \subset D$ in a data set $D = \{y_1, \ldots, y_N\}$ are random landmark selection and the maxmin algorithm. Both are implemented as standard procedures for use in combination with the lazy witness filtration in javaPlex [247].

5.2.1.1 Random landmark selection

The simplest way to choose landmarks $L = \{l_1, l_2, \ldots, l_m\}$ from a point cloud D is to select m points from D uniformly at random. For data sets whose points are evenly

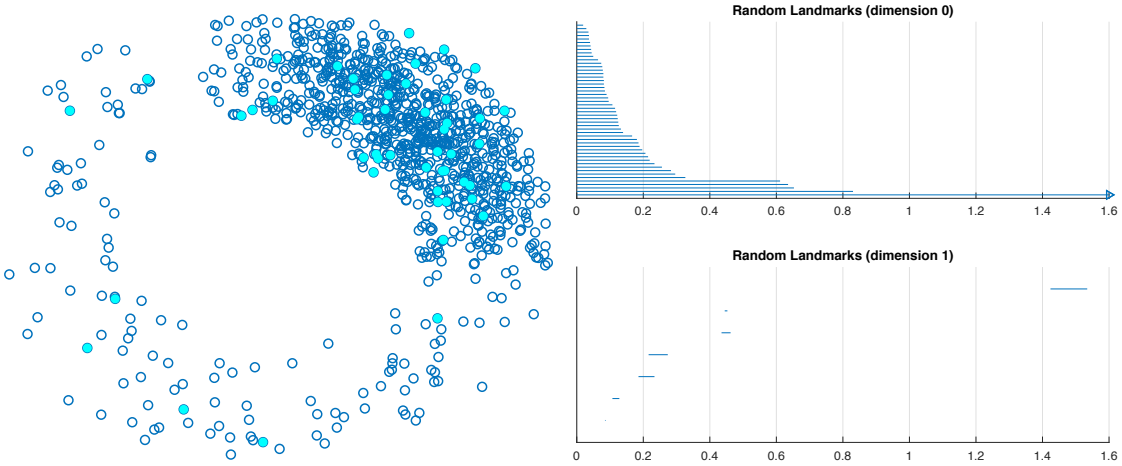


Figure 5.3: Example of a point cloud and landmarks selected at random (landmarks are shown in cyan). We observe that the dimension 1 barcode based on a Vietoris–Rips filtration on the selected landmarks does not capture the persistent homology of the point cloud correctly.

distributed, random selection achieves good coverage at a small computational cost. However, as soon as there are large differences in the density of the data, random selection will favour points from more dense regions, which can result in landmarks that do not represent the point cloud well. In extreme cases, the landmarks do not carry any topological similarity to the original point cloud. We show such an example in Fig. 5.3.

5.2.1.2 The maxmin algorithm

The sequential *maxmin* algorithm chooses the first landmark $l_1 \in D$ randomly. Inductively, for $i \geq 2$ and a landmark set $L_{i-1} = \{l_1, l_2, \ldots, l_{i-1}\}$, the algorithm selects the next landmark $l_i \in D \setminus L_{i-1}$ such that for a chosen metric d the function mapping

$$y \mapsto d(y, L_{i-1}),$$

is maximised, where $d(y, L) = \min_{l \in L} d(y, l)$ for a given a distance function $d : D \times D \to \mathbb{R}$ and $y \in D$.

We show pseudocode for the procedure in Algorithm 1. The method has been used successfully for image data in [1, 67]. Landmarks chosen this way tend to cover the data set well, are evenly spaced, and represent the underlying topological features

Algorithm 1 The maxmin algorithm (from [4])

Input: Data points $D = \{y_1, \ldots, y_N\}$,
a distance function $d : D \times D \to \mathbb{R}$
number of landmarks m.
Output: A set of m maxmin landmarks $L = \{l_1, \ldots, l_m\}$.
Select $y \in D$ at random
$l_1 \leftarrow y$
$L_1 \leftarrow \{l_1\}$
$D_1 \leftarrow D \setminus \{l_1\}$
for i = 2 **to** m **do**
  for all $y \in D_{i-1}$ **do**
    Calculate $d(y, L_{i-1})$
  end for
  Find l_i such that $d(l_i, L_{i-1}) = \max_{y \in D} d(y, L_{i-1})$
  $L_i \leftarrow L_{i-1} \cup \{l_i\}$
  $D_i \leftarrow D_{i-1} \setminus \{l_i\}$
end for

better than landmarks selected at random. The algorithm does, however, tend to include outliers [4, 86]. We show an example of a point cloud where the selected landmarks do not represent the underlying topology of the data correctly in Fig. 5.4.

De Silva and Carlsson [86] state that the maxmin algorithm is best suited to produce the qualities desired from landmarks. They do not recommend the use of clustering algorithms as an alternative due to the high computational cost and potential to accentuate accidental features. In our own investigations, we found that in the work of Lockwood and Krishnamoorthy [165], the maxmin algorithm leads to the discovery of loops in the data set. We did not obtain these loops using random landmark selection. When we discarded a small proportion of the landmarks chosen by maxmin from the data set and chose a new set of maxmin landmarks from the remaining data points, we also no longer observed the loops. These significant differences in the results have led us to believe that maxmin is far from ideal for landmark selection.

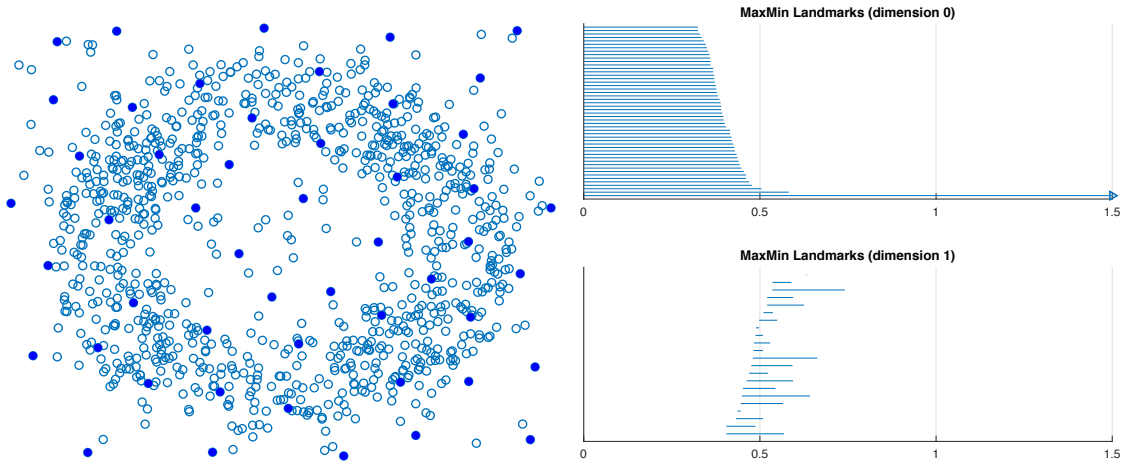


Figure 5.4: Example of a point cloud and landmarks selected by the maxmin algorithm (landmarks are shown in dark blue). We observe that the dimension 1 barcode based on a Vietoris–Rips filtration on the selected landmarks does not capture the PH of the point cloud correctly.

5.2.1.3 Dense core subsets

In the JAVAPLEX tutorial, Adams and Tausz [4] mention that using so-called dense core subsets before applying the maxmin algorithm can help overcome the selection of outliers as landmarks. This approach, however, is not considered as one of the standard approaches for landmark selection and we did not find examples where it was used for this purpose. De Silva and Carlsson [86] and Carlsson *et al.* [67] use dense core subsets to identify dense regions in their data sets. The authors subsequently use the maxmin landmarks to study the topology of these dense regions.

Dense core subsets are based on assigning density values to every point: for an integer K, the density value assigned to a point $y \in D$ is $\frac{1}{\rho_K(y)}$, where $\rho_K(y)$ is the distance to the K-th nearest neighbour of y. Large values of K provide a measure of the global density around the point in the data set, while smaller values of K give a more local perspective. Using the density values, one can select the m densest points in the data set as a dense core subset. Given a data set, it is not clear what values of K to use. As shown in [86], different values for K and m can produce markedly different subsets.

For our comparisons in Subsubsection 5.2.5.3, instead of selecting a dense subset

and then performing the maxmin algorithm to select landmarks, as proposed by Adams and Tausz [4], we choose the m densest points in the data as landmarks. This enables us to determine whether the information on which our own landmark selection technique (PH landmarks) is based differs from that given by the K-th nearest neighbour of a data point.

5.2.2 Proposed landmark selection methods

As we have seen, the currently existing methods for landmark selection from a point cloud that enable PH analysis on large point data sets are not ideal and have, in particular, not been designed with PH in mind. De Silva and Carlsson [86] state the most pertinent qualities for a landmark set to be good coverage of the data set and even spacing spacing of the landmarks. While these are certainly important properties for many data sets, we find that, regardless of whether the landmarks are intended for use in combination with the lazy witness complex or simply as a subset of the data set to apply PH to, the aim of a landmark selection method should be to represent the underlying topology of the data set. Since outliers can artificially introduce topological features such as loops, we in particular require outlier-robust landmark selection techniques. Based on these observations we formulate the following goals for landmark selection methods intended for TDA:

1. Good representation of the underlying topology of the data set, even at low sampling densities with small variance of the results between different landmark realisations.

2. Robustness to outliers, ideally including a measure for how much we consider a specific point to be an outlier.

We now introduce two landmark selection methods that we apply to achieve these goals: Persistent homology landmarks (PH landmarks) and $k--$ landmarks. We

design Persistent homology landmarks specifically for the application of PH, while $k--$ landmarks is a variant of the k-means algorithm, whose properties make it a promising candidate to overcome the downsides of both the random and the maxmin landmark selection.

5.2.2.1 Persistent homology landmarks

As outlined in Section 5.1, we mathematically motivate our landmark selection method by observation 5.23 which we can reformulate for point cloud data D and PH in the following way: if for a data point $y \in D$ and a neighbourhood radius $\delta > 0$ we can ensure that $PH_n(\mathrm{Lk}^\delta\ y) = PH_{n-1}(\mathrm{Lk}^\delta\ y) = 0$, then for $n > 0$ we obtain:

$$0 \xrightarrow{\Phi} PH_n(D \setminus y) \xrightarrow{\Psi} PH_n(D) \xrightarrow{\partial} 0, \tag{5.24}$$

where PH_n denotes the n-th homology groups associated with the different filtration steps. As we are working with data however, we are unlikely to achieve such strict conditions. For $PH_n(\mathrm{Lk}^\delta\ y) = PH_{n-1}(\mathrm{Lk}^\delta\ y) \approx$ small, however, we obtain something close to an isomorphism between $PH_n(D \setminus y)$ and $PH_n(D)$. As a measure for $PH_n(\mathrm{Lk}^\delta\ y)$, for every point $y \in D$ we use the maximal persistence[7] of a non-infinitely persisting feature across dimensions $n = 0, 1, 2$. We call this value the *PH outlierness* $out_{\mathrm{PH}}(y)$ of the point $y \in D$:

Definition 5.2.1 (PH outlierness of a point $y \in D$)**.** Let D be a point cloud, $y \in D$ a data point, $d : D \times D \to \mathbb{R}$ a distance function, $\Delta_y = \{\tilde{y} \in D \setminus \{y\} \mid d(\tilde{y}, y) \leq \delta\}$, $n = 0, 1, 2$ and $\mathcal{B}_n(y) = \{[\eta_i, \zeta_i)\}_{i=1}^{I(n)}$ the n-dimensional barcode of the Vietoris–Rips filtration performed on Δ_y excluding infinitely persisting features. Then

$$out_{\mathrm{PH}}(y) = \max_n \max_{i=1,\dots,I(n)} \{\zeta_i - \eta_i\},$$

is the PH outlierness of y.

[7] See Chapter 2, Subsection 2.2.2 for a definition.

The larger the PH outlierness of a point is, the further away we are from an isomorphism between $PH_n(D \setminus y)$ and $PH_n(D)$. Consequently, the inclusion or exclusion of the point changes the PH of the data set more than for a point with a small PH outlierness value. Under the assumption that each point has at least two neighbours within distance δ, the choice of δ determines by how much we allow the Bottleneck distance[8] between the persistence diagrams of the filtration on the point cloud containing y and a point cloud not containing y to differ. We can therefore think of it as a resolution parameter.

Interestingly, there are now two approaches one can take for the interpretation of the PH outlierness values for landmark selection:

1. Landmarks should be points with small PH outlierness values, since then their inclusion or exclusion in the full data set does not alter the PH of the full data set dramatically. Hence, the landmarks represent the data well and they do not introduce accidental topological features that can, for example, be caused by outlier points in the landmark set.

2. Landmarks should have large PH outlierness values, since then we are far away from an isomorphism between $PH_n(D \setminus \hat{y})$ and $PH_n(D)$. This means that the exclusion of such points from the data set would change the PH dramatically.

In practice, it is not immediately clear which approach works better. We therefore consider both. To avoid choosing points as landmarks that are very far away from other data points, we determine points $S = \{s_1, \ldots, s_o\}$ with fewer than two neighbours within their δ-neighbourhood to be *super outliers*. We include super outliers into the landmark set only once all other points have been chosen as landmarks. Note that the resolution parameter δ strongly influences the number of super outliers. As long as we have enough points in the data set that are not considered to be super

[8]See Chapter 2, Subsubsection 2.4.1 for definition.

outliers, we choose our landmarks to be the points $L = \{l_1, l_2, \ldots, l_m\} \subset D \setminus S$ such that $out_{\mathrm{PH}}(l_i) \leq out_{\mathrm{PH}}(y)$ for all $y \in D \setminus \{L \cup S\}$ and $i = 1, \ldots, m$ for approach 1 and $out_{\mathrm{PH}}(l_i) \geq out_{\mathrm{PH}}(y)$ for all $y \in D \setminus \{L \cup S\}$ and $i = 1, \ldots, m$ for approach 2. We call the set of landmarks obtained by this procedure *PH landmarks*. We show the pseudocode for PH landmarks using approach 1 for the interpretation of the PH outlierness values in Algorithm 2, an algorithm for approach 2 can be formulated accordingly.

Algorithm 2 The PH landmark algorithm

Input: Data points $D = \{y_1, \ldots, y_N\}$,
a distance function $d : D \times D \to \mathbb{R}$
number of landmarks m,
local neighbourhood radius $\delta > 0$.
Output: A set of m PH landmarks $L = \{l_1, \ldots, l_m\}$, a set of o super outliers $S = \{s_1, \ldots, s_o\}$.
for all $y \in D$ **do**
Find $\Delta_y = \{\tilde{y} \in D \setminus \{y\} \mid d(\tilde{y}, y) \leq \delta\}$
if $|\Delta_y| > 1$ **then**
Compute Vietoris–Rips filtration for Δ_y for $n = 0, 1, 2$.
Compute $out_{\mathrm{PH}}(y)$
else
$S \leftarrow S \cup \{y\}$
end if
end for
Re-order the points in $D \setminus S$ such that $out_{\mathrm{PH}}(y_1) \leq out_{\mathrm{PH}}(y_2) \leq \cdots \leq out_{\mathrm{PH}}(y_{N-o})$
$L \leftarrow \{y_1, \ldots, y_{\min\{m, N-o\}}\}$
if $N - o < m$ **then**
$L \leftarrow L \cup \{s_1, \ldots, s_{m-N+o}\}$
end if

In practice, we find that $out_{\mathrm{PH}}(y)$, when considering all PH dimensions, is usually determined by dimension 0, where we find the longest non-infinitely persisting features in our data sets. To avoid this, one can also apply Algorithm 2 with restriction to one particular dimension in the PH calculation, for example dimension 1. In this case, it is important to note, that one can obtain data points y with $out_{\mathrm{PH}}(y) = 0$. To avoid that the order of inclusion of such points in the landmark set is determined by the

ordering of the points in the original data set, which could, for example, favour noise points to be added before signal points or vice versa, we ensure that all points with $out_{\mathrm{PH}}(y) = 0$ are randomly permuted in the ordering of the data by PH outlierness.

5.2.2.2 $k--$ landmarks

The k-means$--$ algorithm was developed by Chawla and Aristides [71] to overcome the extreme sensitivity of the k-means algorithm[9] to outliers. The authors formulate their approach as a generalisation of the k-means algorithm: for an input data set $D = \{y_1, \ldots, y_N\}$ the algorithm provides a set of k cluster centres $\hat{L} = \{\hat{l}_1, \ldots, \hat{l}_k\}$ and a set of j outliers $O = \{o_1, \ldots, o_j\}$, $O \subset D$. For a given distance function $d : D \times D \to \mathbb{R}$ and $y \in D$ the authors use the following term in their algorithm:

$$c(y, \hat{L}) \coloneqq \arg\min_{\hat{l} \in \hat{L}} d(y, \hat{l}). \tag{5.25}$$

We show the pseudocode in Algorithm 3.

For our application of the algorithm to landmark selection, we further define:

$$\tilde{c}(D, \hat{l}) \coloneqq \arg\min_{y \in D} d(y, \hat{l}). \tag{5.26}$$

We show our modified version of the k-means$--$ algorithm in Algorithm 4.

5.2.3 Implementation

We implement PH landmark selection method in MATLAB using RIPSER [32] for the computation of the local Vietoris–Rips complexes. We also implement the $k--$ landmarks algorithm in MATLAB. For the calculation of maxmin landmarks, random landmarks, and dense core subsets we use the inbuilt functions in the JAVAPLEX package [247].

[9] For a description of the k-means algorithm see also Chapter 4, Subsubsection 4.4.4.1.

Algorithm 3 The k-means−− algorithm [71]

Input: Data points $D = \{y_1, \ldots, y_N\}$,
a distance function $d : D \times D \to \mathbb{R}$,
number of clusters k and number of outliers j.
Output: A set of k cluster centers $\hat{L} = \{\hat{l}_1, \ldots, \hat{l}_k\}$,
a set of j outliers $O = \{o_1, \ldots, o_j\}$, $O \subset D$.

$\hat{L}_0 \leftarrow \{k \text{ random points of } D\}$
$i \leftarrow 1$
while (No convergence achieved) **do**
  for all $y \in D$ **do**
    compute $d(y, \hat{L}_{i-1})$
  end for
  Re-order the points in D such that $d(y_1, \hat{L}_{i-1}) \geq d(y_2, \hat{L}_{i-1}) \geq \cdots \geq d(y_N, \hat{L}_{i-1})$
  $O_i \leftarrow \{y_1, \ldots, y_k\}$
  $D_i \leftarrow D \setminus O_i = \{y_{k+1}, \ldots, y_N\}$
  for r = 1 **to** k **do**
    $P_r \leftarrow \{y \in D_i \mid c(y, \hat{L}_{i-1}) = \hat{l}_{i-1,r}\}$
    $\hat{l}_{i,r} \leftarrow \text{mean}(P_r)$
  end for
  $\hat{L}_i \leftarrow \{\hat{l}_{i,1}, \ldots, \hat{l}_{i,k}\}$
  $i \leftarrow i + 1$
end while

Algorithm 4 The k-means−− algorithm [71] modified for landmark selection (our changes and additions are highlighted in blue)

Input: Data points $D = \{y_1, \ldots, y_N\}$, a distance function $d : D \times D \to \mathbb{R}$, number of clusters k and number of outliers j.

Output: A set of k cluster centers $L = \{l_1, \ldots, l_k\}$, $L \subset D$, a set of j outliers $O = \{o_1, \ldots, o_j\}$, $O \subset D$.

$\hat{L}_0 \leftarrow \{k \text{ random points of } D\}$
$e_0 = -1$
$i \leftarrow 1$
while (continuation_criterion $> 10^{-4}$ **and** $i < 100$) **do**
  for all $y \in D$ **do**
    compute $d(y, \hat{l}_{i-1})$
  end for
  Re-order the points in D such that $d(y_1, \hat{L}_{i-1}) \geq d(y_2, \hat{L}_{i-1}) \geq \cdots \geq d(y_N, \hat{L}_{i-1})$
  $O_i \leftarrow \{y_1, \ldots, y_k\}$
  $D_i \leftarrow D \setminus O_i = \{y_{k+1}, \ldots, y_N\}$
  for r = 1 **to** k **do**
    $P_r \leftarrow \{y \in D_i \mid c(y, \hat{L}_{i-1}) = \hat{l}_{i-1,r}\}$
    $\hat{l}_{i,r} \leftarrow \text{mean}(P_r)$
  end for
  $\hat{L}_i \leftarrow \{\hat{l}_{i,1}, \ldots, \hat{l}_{i,k}\}$
  for $y \in D_i$ **do**
    compute $d(y, \hat{L}_i)$
  end for
  $e_i \leftarrow \sum_{y \in D_i} d(y, \hat{L}_i)^2$
  continuation_criterion $\leftarrow |e_i - e_{i-1}|$
  $i \leftarrow i + 1$
end while
$L \leftarrow \emptyset$
$D_L \leftarrow D_{i-1}$
while $|L| < k$ **do**
  for $\hat{l} \in \hat{L}$ **do**
    $m_{\hat{l}} \leftarrow \min_{y \in D_L} d(y, \hat{l})$
  end for
  Re-order $\{\hat{l}_1, \ldots, \hat{l}_{\hat{k}}\}$ such that $m_{l_1} \leq m_{l_2} \leq \cdots \leq m_{l_{\hat{k}}}$
  $s \leftarrow 1$
  repeat
    $L \leftarrow L \cup \{y_s\}$, where $y_s = \tilde{c}(D_L, \hat{l}_s)$
    $s \leftarrow s + 1$
  until $y_s = y_t$ for some $t < s$
  $L \leftarrow L \setminus \{y_s\}$
  $D_L \leftarrow D_L \setminus L$
  $\hat{L} \leftarrow \hat{L} \setminus \{\hat{l}_1, \ldots, \hat{l}_{s-1}\}$
end while

5.2.4 Data sets

We introduce the data sets to which we apply the different landmark selection methods. The data sets are chosen to be simple to allow us to determine effects of the landmark selection. The data sets consist of signal points that are sampled from a topologically interesting structure – a sphere, a Klein bottle, or a torus – and noise points that we design to be topologically different from the signal.

5.2.4.1 3-dimensional data sets

Sphere-cube data set For a given number of points N and probability p we sample points uniformly at random from the surface of the unit sphere with probability p and points from the (filled) cube $[-1,1]^3 \subset \mathbb{R}^3$ with probability $1-p$.

Sphere-plane data set For a given number of points N we sample points uniformly at random from the surface of the unit sphere with probability p and points from the xy-plane $[-3,3]^2 \subset \mathbb{R}^2$ with probability $1-p$.

Sphere-line data set For a given number of points N we sample points uniformly at random from the surface of the unit sphere with probability p and points from $(\alpha,0,0)$, where $\alpha \in [-50,50] \subset \mathbb{R}$, with probability $1-p$.

Sphere-Laplace line data set For a given number of points N we sample points uniformly at random from the surface of the unit sphere with probability p and we sample points from $(\alpha,0,0)$ with probability $1-p$, where α is sampled from $[-50,50] \subset \mathbb{R}$ and Laplace distributed with $\mu = 4$ and $\sigma = 0.5$. We use the Laplacian random number generator code [72] to generate α.

5.2.4.2 4-dimensional data sets

Torus data set We use the following parametrisation of the torus $\mathcal{T}$:

$$(x, y, z, \omega) = (\cos(\gamma), \sin(\gamma), \cos(\varphi), \sin(\varphi)),$$

where $\gamma, \varphi \in (0, 2\pi)$. We add noise T_{noise} to the torus using the equation

$$(x_{\text{noise}}, y_{\text{noise}}, z_{\text{noise}}, \omega_{\text{noise}}) = (r * \cos(\gamma), r * \sin(\gamma), \hat{r} * \cos(\varphi), \hat{r} * \sin(\varphi)),$$

where $r, \hat{r} \in (0, 2)$.

For a given number of points N we sample points uniformly at random from $\mathcal{T}$ with probability p and points from $\mathcal{T}_{\text{noise}}$ with probability $1 - p$.

Klein bottle data set We use the following parametrisation of the Klein bottle $\mathcal{K}$:

$$\begin{aligned} x &= \cos(\gamma) * (r * \cos(\varphi) + C), \\ y &= \sin(\gamma) * (r * \cos(\varphi) + C), \\ z &= \cos(\gamma/2) * r * \sin(\varphi), \\ \omega &= \sin(\gamma/2) * \sin(\varphi), \end{aligned}$$

where $\gamma, \varphi \in (0, 2\pi)$, $r = 3$ and $C = 2$. We define noise $\mathcal{K}_{\text{noise}}$ for the Klein bottle using the equations:

$$\begin{aligned} x_{\text{noise}} &= \cos(\gamma) * (r_{\text{noise}} * \cos(\varphi) + C_{\text{noise}}), \\ y_{\text{noise}} &= \sin(\gamma) * (r_{\text{noise}} * \cos(\varphi) + C_{\text{noise}}), \\ z_{\text{noise}} &= \cos(\gamma/2) * r_{\text{noise}} * \sin(\varphi), \\ \omega_{\text{noise}} &= \sin(\gamma/2) * \sin(\varphi), \end{aligned}$$

where r_{noise} is sampled uniformly from the interval $[2, 4] \subset \mathbb{R}$ and C_{noise} is sampled uniformly from the interval $[1, 3] \subset \mathbb{R}$. We use and adapt the code from [190]. For a given number of points N we sample points uniformly at random from $\mathcal{K}$ with probability p and points from $\mathcal{K}_{\text{noise}}$ with probability $1 - p$.

5.2.5 Results

We present our results for the proposed landmark selection methods. For all methods we use the Euclidean distance as distance function and all data sets consist of 3000 points. We first study the PH landmark selection method in detail on the sphere-cube data set with $p = 0.6$, then proceed to showing our results in comparison to the current standard methods on the various data sets, and finally compare our methods to the dense core subsets also investigating the influence of the δ parameter on the performance of the method.

5.2.5.1 Persistent homology landmarks case study on the sphere-cube data set with $p = 0.6$

We apply PH landmark selection with $\delta = 0.2$ to the sphere-cube data set where a data point has a probability of 0.6 to be located on the surface of the unit sphere and 0.4 to be located in the unit cube. For all of our data sets, we find that the PH outlierness values, as defined in Def. 5.2.1, are determined by the maximal non-infinite bar in the dimension 0 barcode as this tends to be much longer than the persistence of any feature in the higher dimensional barcodes. We show example barcodes for dimension 0 for a noise point, a sphere point, and a super outlier in Fig. 5.5. In general, we expect a noise point to be located in a sparser region of this data set and thus to either be classified as a super outlier, or to exhibit a barcode with a small number of long bars and very few, or no, short bars. For a sphere point, we expect the dimension 0 barcode to have many short bars and occasionally some longer bars caused by noise points that lie within the δ-neighbourhood. Note that for the examples in Fig. 5.5, we choose the noise point with the highest outlierness score and the sphere point with the lowest outlierness score in the data set to illustrate ideal cases for the method. We find that for these example points, the dimension 0 barcodes behave as expected. To explore whether the outlierness scores reflect the properties of the different types of points as expected, we consider histograms of the outlierness

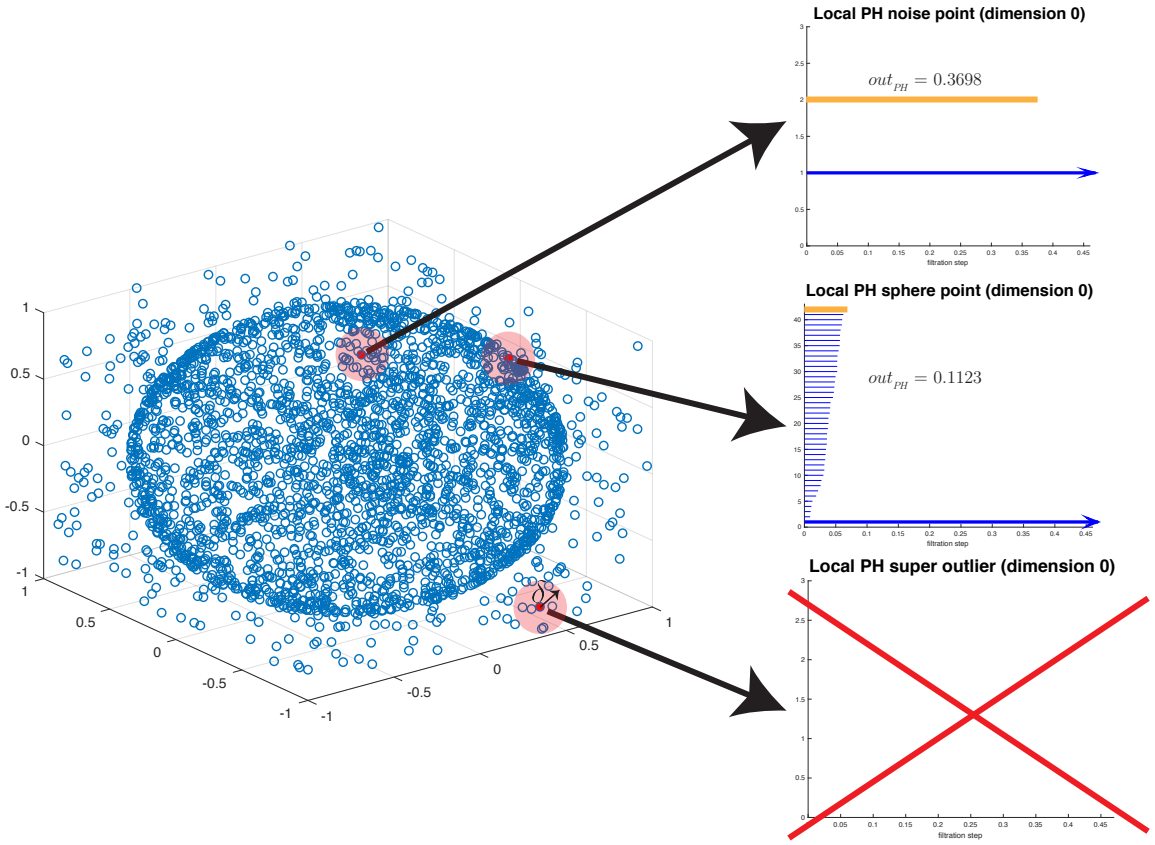


Figure 5.5: Schematic illustration of three different types of points y found in the sphere-cube data set, $p = 0.6$, and their local dimension 0 barcodes for $\delta = 0.2$: noise point, sphere point, and super outlier. The δ-neighbourhoods are shown in light red balls around the corresponding data points. We calculate the PH outlierness $out_{\mathrm{PH}}(y)$ for every point based on its local Vietoris–Rips barcode and ignore super outliers. We highlight the bars in the barcodes that are used to determine $out_{\mathrm{PH}}(y)$ in yellow.

scores (as per Def. 5.2.1) of all data points in Fig. 5.6. We find that we have 48 super outliers in the data set. For the noise and sphere points, we can see that the outlier scores are distributed differently and that by including points with low outlier scores as landmarks first, we should preferentially obtain sphere points rather than noise points. This corresponds to approach 1 described in Subsubsubsection 5.2.2.1.

As our landmark selection approach is motivated by Equation 5.24 which holds for PH in dimensions $n > 0$, it is not immediately clear that approach 1 is also preferable both for $n = 0$ and restriction to dimensions $n > 0$. We examine a variant of the method where we restrict ourselves to local PH in dimension 1. In this case the PH outlierness values correspond to the persistence of the most persistent feature

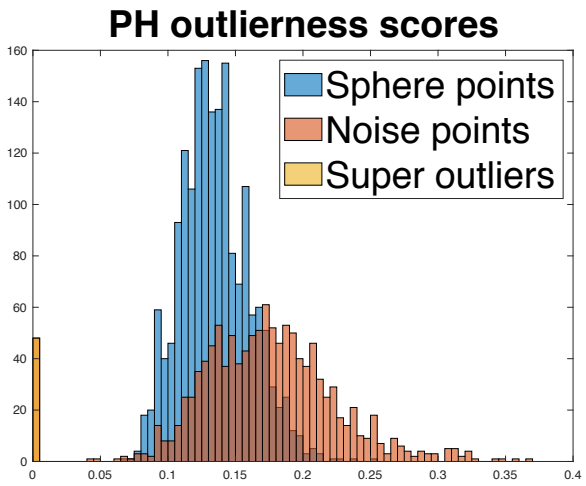


Figure 5.6: Histograms of the PH outlierness $out_{\mathrm{PH}}(y)$ values obtained on the sphere-cube data set, $p = 0.6$, from local PH with $\delta = 0.2$. The horizontal axis represents the outlierness scores, the vertical axis shows the number of points.

in the local dimension 1 barcode. We show a histogram of the distribution of the PH outlierness scores in Fig. 5.7. We observe that the clearest difference between the

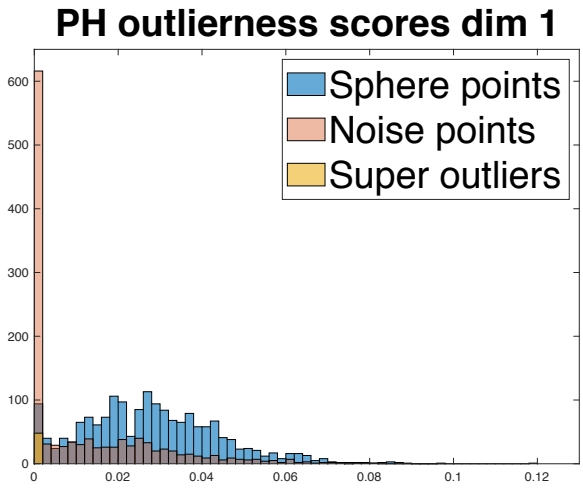


Figure 5.7: Histograms of the PH outlierness $out_{\mathrm{PH}}(y)$ values obtained on the sphere-cube data set, $p = 0.6$, from local PH with $\delta = 0.2$, considering only features in dimension 1. The horizontal axis represents the outlierness scores, the vertical axis shows the number of points.

sphere points and the noise points is the fact that a large proportion of noise points has $out_{\mathrm{PH}}(y) = 0$, while a clear majority of the sphere points has $out_{\mathrm{PH}}(y) > 0$. This can again be explained by the fact that sphere points have more neighbours within their δ-neighbourhood and therefore are more likely to form features in dimension 1. From these observations it seems more beneficial to use approach 2, described in

Subsubsection 5.2.2.1, for landmark selection based on dimension 1 (and analogously probably in higher dimensions in other data sets). For PH landmarks based on dimension 1 we thus choose points with large PH outlierness scores as landmarks and discard points with low PH outlierness scores from the data set as outliers whose removal does not alter the PH of the data set much.

5.2.5.2 Comparison of persistent homology landmarks and $k--$ landmarks to standard landmark selection methods

We compare our proposed methods for landmark selection, the $k--$ landmarks and the PH landmarks, to the current standard methods for landmark selection, i.e. random landmarks and maxmin landmarks. As mentioned before, we use the Euclidean distance as our distance function for all methods. Using the different techniques we choose m landmarks from N data points. This corresponds to a landmark sampling density of $\frac{m}{N}$. For the $k--$ landmarks, we define the number of clusters to be $k = pm$ and the number of outliers to be $j = (1-p)m$, where p is the probability with which a point in the respective data set was sampled from the signal data, i.e. the sphere, torus, or Klein bottle. We consider both the case where we choose the $k--$ cluster centres and the outliers found by the algorithm as our landmarks as well as the case where we only consider the $k--$ cluster centres to be landmarks. For the PH landmarks we choose $\delta = 0.2$ for the 3-dimensional data sets, $\delta = 0.5$ for the torus data and $\delta = 0.6$ for the Klein bottle. In all our data sets, we find that when looking for the maximal persistence of a feature across all dimensions[10], the value of $out_{\mathrm{PH}}(y)$ is exclusively determined by dimension 0. We therefore also include a variation of PH landmarks that only considers the local dimension 1 barcode for the calculation of $out_{\mathrm{PH}}(y)$. For the PH landmark version where we use all dimensions to determine the outlierness scores, we choose data points with $out_{\mathrm{PH}}(y) \approx$ small as our landmarks, for

[10]In practice, we compute dimensions 0, 1, and 2.

the version where we restrict ourselves to dimension 1, we choose points with large outlier scores as landmarks.

For all our data sets, our aim for the landmarks is to contain a high fraction of signal points, even when sampling only a small fraction of the data as landmarks. In Figures 5.8–5.13 we show plots of the fraction of signal points in the various landmark sets at different sampling densities. Since in the PH landmark selection we allow super outliers as landmarks once all other points are taken, we expect the fraction of signal landmarks for sampling density 1 to represent the probability of signal points in the data set, except in the variant of the $k--$ landmarks where we include only the cluster centres as landmarks (referred to as 'kMinusMinusOutlierFree' in the plots). Note that for this variant of $k--$ landmarks for data sets with $p \leq 0.5$, it is possible to obtain a signal fraction of 0 even for sampling density 1 if the algorithm selects all $k = pm$ cluster centres to be located among noise points. For the maxmin, random and $k--$ landmarks we show the average fraction of signal points and its standard deviation across 20 selections.

As expected, we observe that the maxmin algorithm performs very badly and tends to select noise points as landmarks for all data sets with only one exception (see Figures 5.8–5.13). For the sphere-Laplace data, the maxmin algorithm performs well (see Fig. 5.11), as the noise is located in a cluster far away from the signal and hence maximising the distance between landmarks results in many points being selected from the sphere. The fraction of signal landmarks for random selection also behaves as we expect for all data sets in Figures 5.8–5.13: the selected landmarks are representative for the whole data set with an almost constant fraction of signal points over all sampling densities that corresponds to the fraction of signal points in the data set. For the $k--$ landmarks, we can see a clear improvement in the signal fraction in most data sets in Figures 5.8–5.13 when considering only cluster centres as landmarks – the inclusion of outliers gives the $k--$ landmarks similarly

bad properties as the maxmin landmarks. The $k--$ landmarks that do not include outliers tend to perform well for high sampling densities, where the number outlier points corresponds roughly to the number of noise points in the data set. For low sampling densities however, the method only outperforms random selection for most of the sphere-cube and Klein bottle data sets (see Figures 5.8 and 5.13). For the sphere-plane, sphere-line and the sphere Laplace-line data sets the reason for this lies in the nature of the noise, which leads to the selection of cluster centres in the noise data. We also notice large standard deviations from the average fraction of signal points in the $k--$ landmark over 20 realisations.

With exception of the sphere-line and sphere-Laplace data sets (see Figures 5.10 and 5.11), the PH landmark selection techniques both outperform the standard methods as well as the $k--$ landmarks clearly for most cases, especially for low sampling densities. Interestingly, the $k--$ landmarks perform very well on the Klein bottle data set (see Fig. 5.13), beating PH landmarks for very small sampling densities. For the sphere-line and sphere-Laplace data sets (see Figures 5.10 and 5.11), the dimension 1 version of the PH landmarks outperforms all other methods for low sampling densities while the version considering all dimensions performs worse than all other methods in most cases. The noise in these data sets is located on lines in comparably dense regions of the data set where the local PH does not find any topological features in dimension 1, but many short lived features in dimension 0. In general, the two versions of PH landmarks start coinciding as soon as super outliers are added to the data set which we can observe in the plots as a rapid drop in the fraction of signal points. We add the super outliers to the landmarks in random order (once all other points are already selected as landmarks) and hence both PH landmark methods differ only slightly in the development of their signal fractions after the addition of super outliers. We note there are cases in which the PH landmarks thrive because the points that are not super outliers are predominantly signal points. This is not the case

for the sphere-line and the sphere-Laplace data sets with $p = 0.6$ (see Figures 5.10 and 5.11) for which the dimension 1 PH landmarks perform very well. Both of these data sets have less than 4 super outliers. Interestingly, there seems to be a trend for the dimension 1 landmarks to outperform the dimension 0 landmarks on data sets with high signal content, i.e. for $p > 0.6$ across all data sets. For lower signal content the PH landmarks considering all dimensions (although only dimension 0 in practice) performs strongly. For the Klein bottle data set, the dimension 1 version of the method outperforms the version considering all dimensions in almost all cases.

Overall, the results underline that PH landmarks represent the PH of the data set well and are robust to outliers, in particular for low sampling densities. They outperform standard methods in a large majority of cases and, moreover, give us a notion of how much a point can be considered an outlier for the respective variant of the method. $k--$ landmarks perform better than random selection in most cases and perform very well, in particular, for high sampling densities. Given the much higher computational cost and the large fluctuations in signal fraction between different realisations of the method, using random selection instead of $k--$ landmarks could however present a more practical approach.

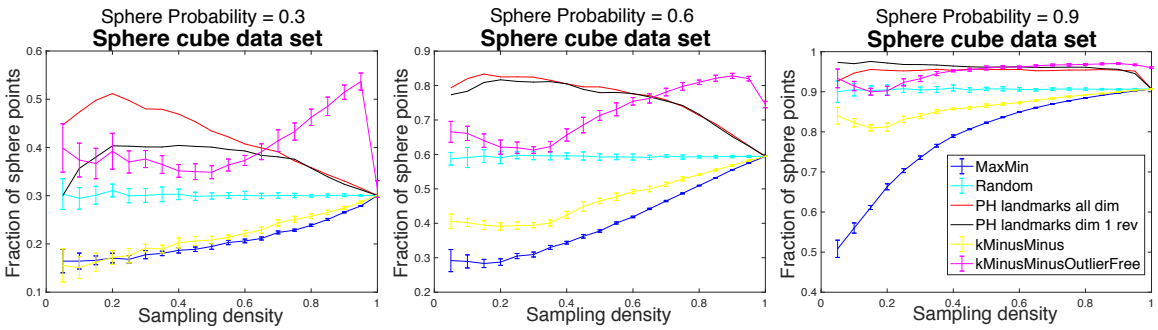


Figure 5.8: Comparison of the fraction of sphere points in selected landmark points for different landmark selection techniques on the sphere-cube data set for $\delta = 0.2$.

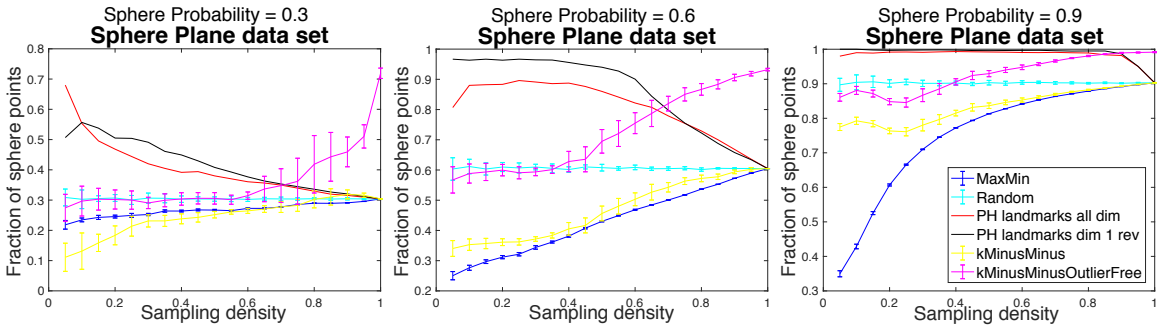


Figure 5.9: Comparison of the fraction of sphere points in selected landmark points for different landmark selection techniques on the sphere-plane data set for $\delta = 0.2$.

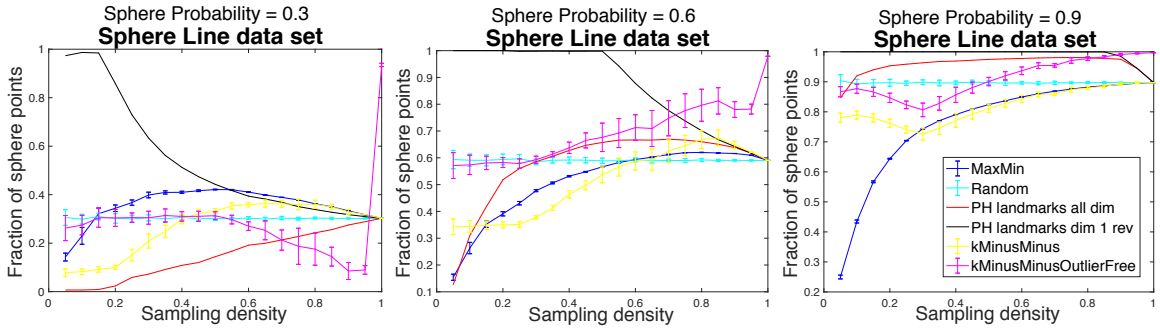


Figure 5.10: Comparison of the fraction of sphere points in selected landmark points for different landmark selection techniques on the sphere-line data set for $\delta = 0.2$.

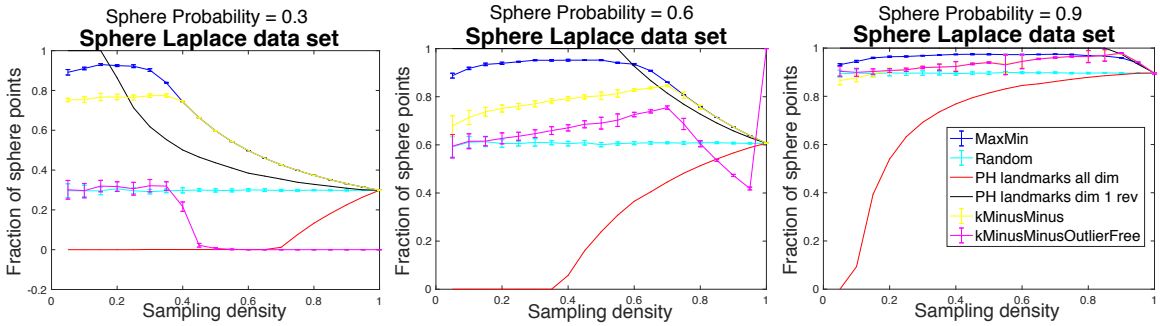


Figure 5.11: Comparison of the fraction of sphere points in selected landmark points for different landmark selection techniques on the sphere-Laplace data set for $\delta = 0.2$.

5.2.5.3 Comparisons between persistent homology landmark selection methods and dense core subsets

We now provide a more in detail study of the two PH landmark selection techniques, in particular concerning the influence of the δ parameter. We also compare the techniques to two dense core subsets, using for $K = 1$ and $K = 50$, which we consider as landmark sets. We present our results in Figures 5.14 – 5.19. We present only the

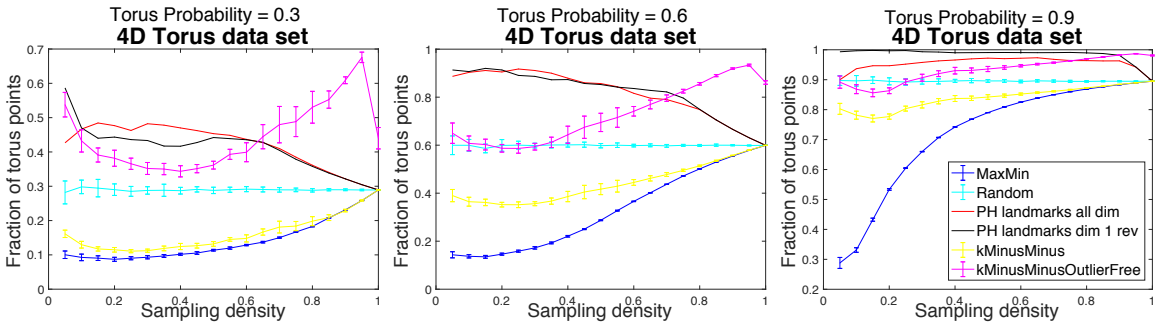


Figure 5.12: Comparison of the fraction of sphere points in selected landmark points for different landmark selection techniques on the Torus data set for $\delta = 0.5$.

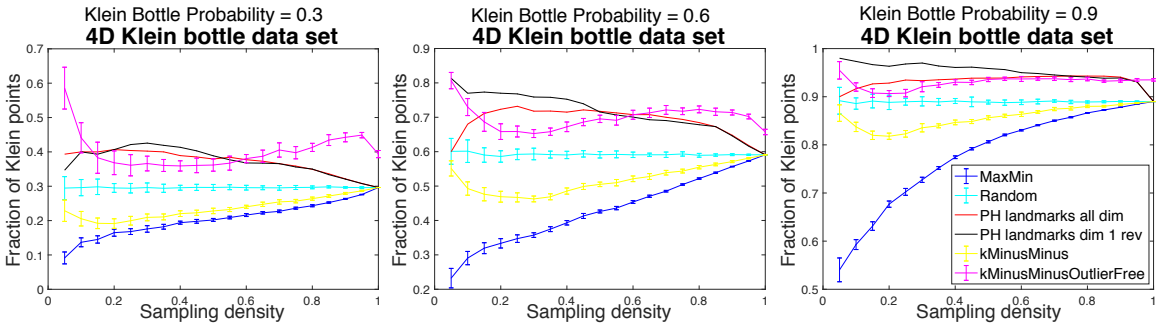


Figure 5.13: Comparison of the fraction of sphere points in selected landmark points for different landmark selection techniques on the Klein bottle data set for $\delta = 0.6$.

plots for data sets with a signal probability $p = 0.6$.

For the dense core subsets, we find that for all data sets except the sphere-plane, sphere-line, and sphere-Laplace data sets (see Fig. 5.15 – Fig. 5.17) the local density measure $K = 1$ outperforms the more global density measure $K = 50$. Indeed, in these cases, the dense core subset with $K = 1$ captures a larger fraction of signal points than most of our PH landmarks. For the sphere-cube data set (see Fig. 5.14), we seem to outperform the dense core subset with $K = 1$ for a small range of low sampling densities for $\delta = 0.05$, which is a δ value where most of the data points are classified as super outliers. Our definition of super outliers as points with less than two neighbours in their δ-neighbourhoods, seems to imply that, for this data set, the distance to the second closest neighbour is more relevant for low sampling densities than the distance to the closest neighbour. Interestingly, for the data sets where the local dense core subset with $K = 1$ performs well (see Figures 5.14, 5.18, and 5.19),

we also see that smaller values of δ give us better results, both for the all dimension PH landmark version and the restriction to dimension 1. For the sphere-line data set (see Fig. 5.16), where we have a better performance for the dense core subset with $K = 50$, larger δ-neighbourhoods are more advantageous for both PH landmark versions. In this case, we perform as well as the dense core subset with $K = 50$ for $\delta = 0.3, 0.35, 0.4$ in the dimension 1 PH landmark version. The sphere-Laplace data set (see Fig. 5.17) is the only data set where PH landmarks clearly outperform both dense core subsets, although only in dimension 1. Here again, we find the trend that larger δ values are advantageous for PH landmarks in dimension 1. That dense core subsets perform well on most of our data sets is determined by the fact that our signal points tend to be in denser regions than the noise points. It is only in the case of the sphere-Laplace data set (see Fig. 5.17), where the noise does not obey this characteristic and we observe that the local PH information in this case is richer than the distance to the K-th neighbour.

Overall, the PH landmarks outperform all standard techniques on most data sets as well as two dense core subsets on one data set for a broad range of δ-values. To obtain signal fractions that are high and stable over a long range of low sampling densities it appears to be advantageous to choose δ as small as possible without having too many super outliers in our data sets. We motivated our choices of δ on our data sets in Subsubsection 5.2.5.2 using this approach. For very small sampling densities, it can seem of advantage to choose a δ-resolution such that only a small number of points are non-super outliers. In these cases however, good results are only obtained based on density characteristics rather than local PH characteristics. We show how the number of super outliers depends on the choice of δ in Fig. 5.20. We observe that it is indeed the case that small δ values lead to a drastic increase in the proportion of super outliers in the data set. This underlines that, even though one can think of δ as a resolution parameter, depending on the data set, the proportion

of super outliers is an important factor to consider.

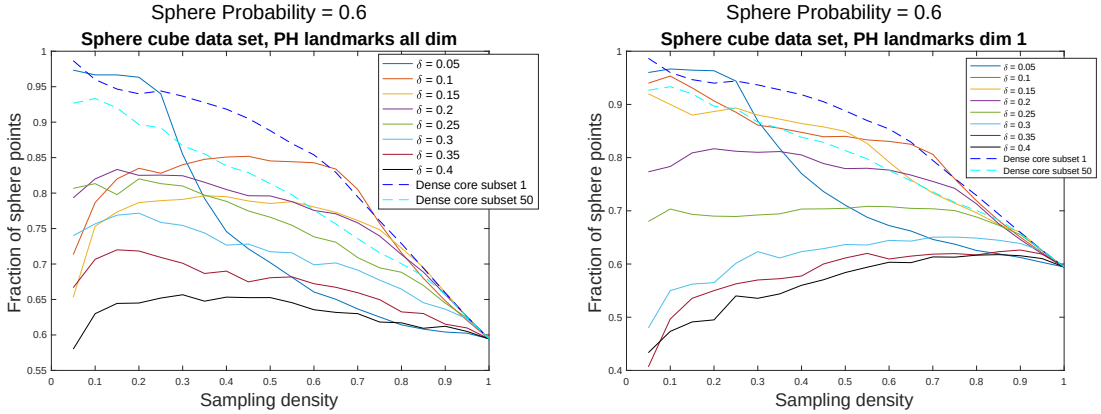


Figure 5.14: Comparison of the fraction of sphere points in selected PH landmark points for different values of δ and dense core subsets on the sphere-cube data set, $p = 0.6$. We show the PH landmarks version where we consider all dimensions in the PH outlierness score calculation (left) and the version restricted to dimension 1 (right).

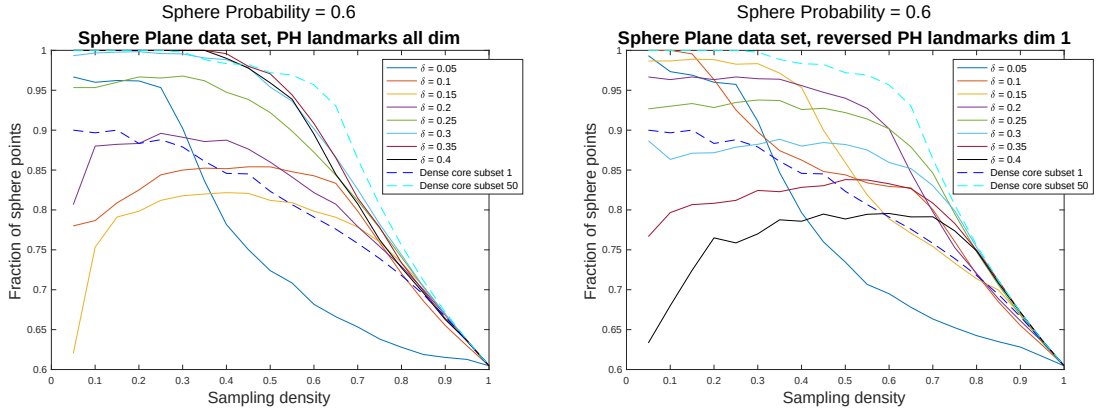


Figure 5.15: Comparison of the fraction of sphere points in selected PH landmark points for different values of δ and dense core subsets on the sphere-plane data set, $p = 0.6$. We show the PH landmarks version where we consider all dimensions in the PH outlierness score calculation (left) and the version restricted to dimension 1 (right).

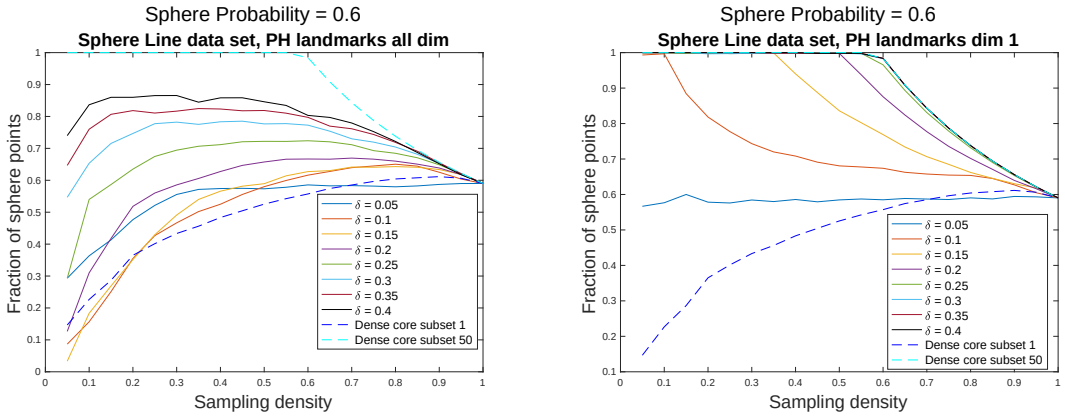


Figure 5.16: Comparison of the fraction of sphere points in selected PH landmark points for different values of δ and dense core subsets on the sphere-line data set, $p = 0.6$. We show the PH landmarks version where we consider all dimensions in the PH outlierness score calculation (left) and the version restricted to dimension 1 (right).

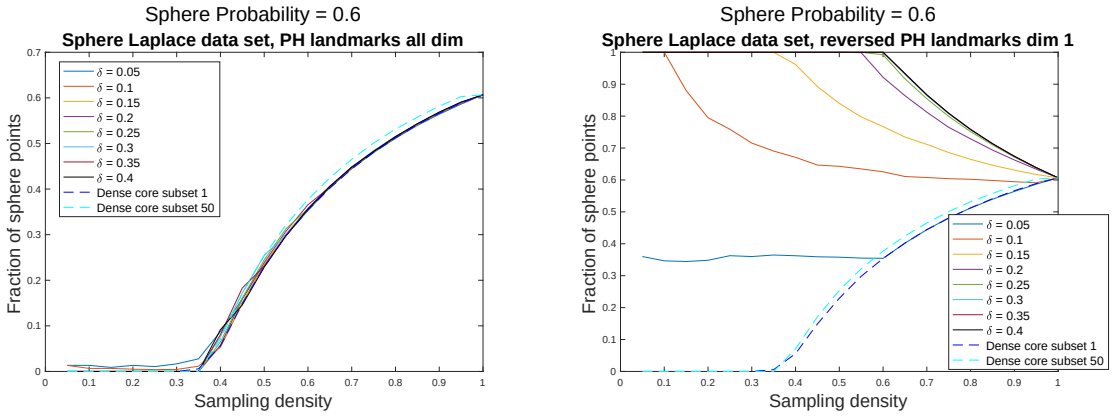


Figure 5.17: Comparison of the fraction of sphere points in selected PH landmark points for different values of δ and dense core subsets on the sphere-Laplace data set, $p = 0.6$. We show the PH landmarks version where we consider all dimensions in the PH outlierness score calculation (left) and the version restricted to dimension 1 (right).

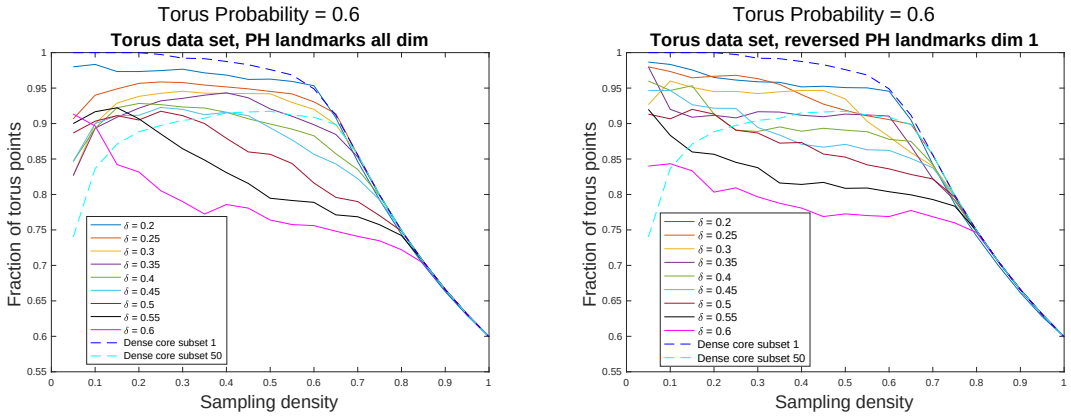


Figure 5.18: Comparison of the fraction of sphere points in selected PH landmark points for different values of δ and dense core subsets on the torus data set, $p = 0.6$. We show the PH landmarks version where we consider all dimensions in the PH outlierness score calculation (left) and the version restricted to dimension 1 (right).

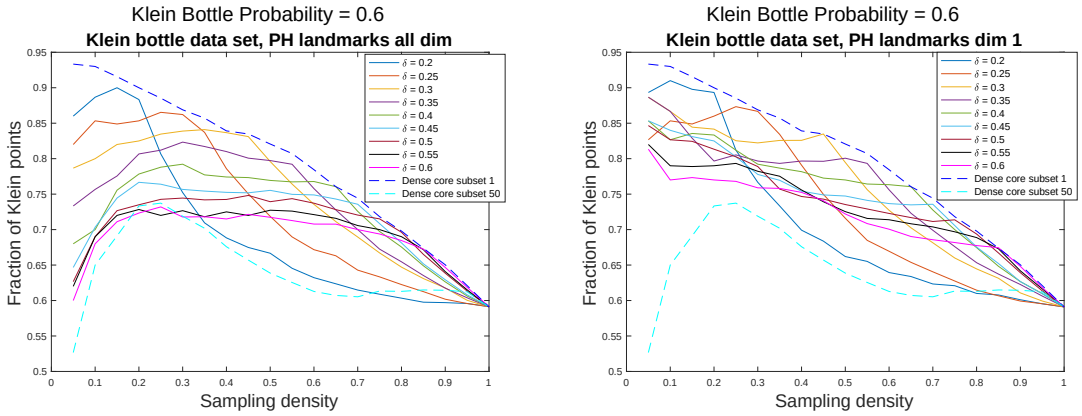


Figure 5.19: Comparison of the fraction of sphere points in selected PH landmark points for different values of δ and dense core subsets on the Klein bottle data set, $p = 0.6$. We show the PH landmarks version where we consider all dimensions in the PH outlierness score calculation (left) and the version restricted to dimension 1 (right).

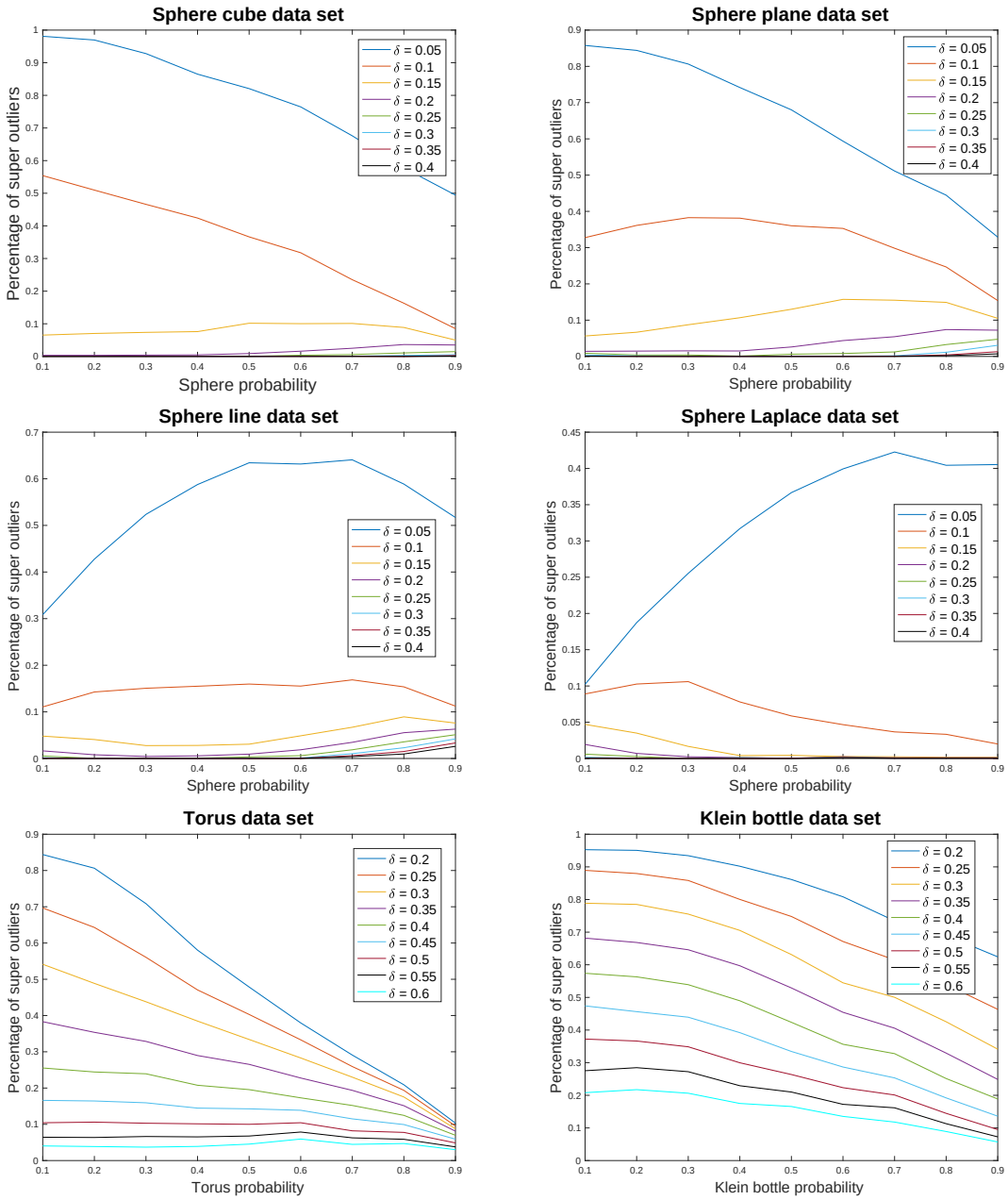


Figure 5.20: Influence of the choice of δ on the number of super outliers for different signal probabilities.

5.2.6 Summary and discussion

We proposed two outlier-robust landmark selection techniques, $k--$ landmarks and PH landmarks. PH landmark selection is the first landmark selection method developed specifically for the use of PH.

The $k--$ landmarks outperformed existing standard landmark selection methods – random selection and maxmin selection – in many cases. They however tended to do so for large sampling densities, which are not the most relevant for a good landmark selection technique. While $k--$ landmarks did meet our goals for an outlier-robust landmark selection technique, this algorithm has a high computational cost and is difficult to use on a data set with unknown properties as one has to predetermine the number of outliers for the algorithm to find.

We found that PH landmarks outperformed the existing standard landmark selection techniques on data sets containing noise, in particular for low landmark sampling densities. We observed this for a wide range of δ resolution values. In most of our data sets, the restriction to dimension 1 for calculating the outlierness values slightly outperformed the version considering all dimensions (which coincides with a restriction to dimension 0 in our cases) for data sets with a high signal content and a low noise content, while dimension 0 performed better for higher noise content. Note that for dimension 0 PH landmarks, we used points with small outlierness values as landmarks, while for dimension 1 PH landmarks we used points with large outlierness values as landmarks. Based on the Mayer-Vietoris sequence 5.24, the latter approach seems more intuitive and the difference between the two approaches could be a reflection of the fact that our observation 5.24 for the sequence only holds for dimensions $n > 0$.

PH landmark selection restricted to dimension 1 was able to outperform dense core subsets showing that the method can capture richer information than just considering the K-th nearest neighbour as shown on the sphere-Laplace data set. Unfortunately,

on many of our data sets, PH landmarks were outperformed by dense core subsets due to the fact that the signal points tend to be located in denser parts of the data set than noise points. In the future, it would be beneficial to study more data sets, for example, more sparse data sets, to observe whether PH landmarks can outperform dense core subsets in these cases. It is also important to note that, in some cases, the success of PH landmarks was caused by the exclusion of super outliers with fewer than two neighbours in their δ-neighbourhood, which can be interpreted as being equivalent to setting a threshold on the distance to the second closest neighbour of a point. Overall, we however consider PH landmarks to be a more practical approach than dense core subsets since they only require the choice of one parameter, δ. In contrast, when applying dense core subsets as suggested in [4], one needs to choose the parameter K to obtain a density estimate, followed by the number of densest points to be chosen from the data set, which are then used to obtain maxmin landmarks. In [86], the authors observed that when studying dense core subsets the choices of parameters can indeed result in strong topological differences in the selected point clouds.

PH landmark selection could further be improved, for example, by considering a different definition of the measure of outlierness. Another approach could be to remove the point considered the most extreme outlier in the data set (excluding super outliers) according to its PH outlier value, to recalculate the PH outlierness values for the remaining data points, and to repeat the computation until all data points are assigned an outlierness value. As we currently only calculate the outlierness values with respect to removing the points from the full data set, we do not take the changes into account that occur by removing points within the same δ-neighbourhood.

In summary, of our two proposed landmark selection techniques, $k--$ landmarks and PH landmarks, PH landmarks performed very well on our test data sets. PH landmark selection fully met our goals of finding landmarks that represent the topol-

ogy of a data set well, as well as being robust to outliers. In addition, the method provides a measure for how much we consider a specific point to be an outlier and a set of super outliers that are so far away from other data points, that they should be ignored as landmarks. The PH outlierness values as well as the number of super outliers with respect to the δ resolution could also be used to provide interesting insight into data sets for exploratory data analysis. Although further investigations will be necessary to test the applicability to real-world data sets, PH landmarks contribute a valuable alternative to the current standard landmark selection techniques for PH, in particular for noisy data sets.

5.3 Classification of data points on intersecting surfaces

Building on our notion of local PH described in Section 5.1, we use a similar idea to classify points on intersecting surfaces. Consider a data set that consists of points sampled from two intersecting bounded planes. Based on the location of the points on the intersecting planes, we can distinguish three different types of points: points close to a plane boundary (we refer to these as *boundary points*), points close to or in the intersection of the planes (we refer to these as *intersection points*), and points that are neither close to the intersection nor close to a boundary (we refer to these as *inner points*). Identifying points close to intersections is of great interest, for example in fields such as algebraic geometry, where intersections are relevant for sampling techniques from algebraic manifolds (see for example [51]), but also in a wider data science context for data which does not satisfy the manifold hypothesis [105]. We develop a method that can detect points close to geometric anomalies such as intersections and boundaries which is fully based on local PH. We take two approaches to classify the points: classification via distance measures on the local PH output and classification via features in the local dimension 1 barcodes. Since intersections

can also be detected using a local version of principal component analysis (PCA), see for example [169, 170], we further compare our classification via features in the local dimension 1 barcodes to this approach.

5.3.1 Classification via distance measures

For classification via distance measures, we consider a neighbourhood of size $\delta > 0$ around every point $y \in D$, which includes all data points that are within distance δ of y. As for the PH landmarks in Subsection 5.2.2.1, we define the local PH of the point to be the δ-Link of the point, which we obtain by calculating the Vietoris–Rips filtration on the points within the δ-neighbourhood excluding the point itself. We then map every point to its local barcodes[11] $\mathcal{B}_n(y) = \{[\eta_i, \zeta_i)\}_{i=1}^{I(n)}$ in dimensions $n = 0, 1$:

$$y \mapsto \mathcal{B}_n(y) = \{[\eta_i, \zeta_i)\}_{i=1}^{I(n)}, \tag{5.27}$$

We use the barcodes to compute distance matrices $\mathcal{D}_0$ for dimension 0 and $\mathcal{D}_1$ for dimension 1, where $\mathcal{D}_n = (d(\mathcal{B}_n(y_i), \mathcal{B}_n(y_j)))_{i,j=1}^{N}$ for a given distance measure d. We use the Bottleneck distance[12] and the Wasserstein distance (using L_∞ and $\mathtt{p} = 2$ in the definition provided in Chapter 2, Subsubsection 2.4.1). In addition, we convert the barcodes to persistence landscapes[13] and compute the L_2 distance. We apply k-medoids clustering [148] to the obtained distance matrices.

5.3.2 Classification via local dimension 1 persistent homology features

Our second classification method is based on the following observation: consider again data points sampled from two intersecting bounded planes. For a 'reasonable' choice of δ, we expect the local dimension 1 barcodes $\mathcal{B}_1(y)$ to exhibit a different number of persistent features depending on the local geometry around the point y:

[11] See Chapter 2, Subsection 2.2.2 for a definition.
[12] See Chapter 2, Subsubsection 2.4.1 for definition.
[13] See Chapter 2, Subsubsection 2.4.2 for definition.

- For a boundary point y_b, we expect no persistent bars in $\mathcal{B}_1(y_b)$.
- For an inner point y_i, we expect one persistent bar in $\mathcal{B}_1(y_i)$.
- For an intersection point y_{int}, we expect at least two[14] persistent bars in $\mathcal{B}_1(y_{int})$.

However, the 'reasonable' choice for δ is not trivial: it needs to be large enough to enclose enough data points to produce features in dimension 1, but also small enough to only produce the loop around the data point y as a persistent feature in the barcode. Both of these factors depend heavily on the properties of the data set and such a δ value might not always exist. In fact, for the intersecting planes data set, we were not able to find a suitable δ value. We therefore choose a strategy to enforce the desired persistent features in dimension 1: instead of considering the points within a δ-neighbourhood around every point, we consider the points within a δ-annulus around every point with an outer radius $r_{out} = \delta$ and an inner radius of $r_{in} = \alpha\delta$ for $\alpha < 1$ (see Fig. 5.21).

We map every data point $y \in D$ to the number of features $\mathcal{N}_1(y)$ in dimension 1 of its local annulus barcode $\mathcal{B}_{1,r_{out},r_{in}}(y) = \{[\eta_i, \zeta_i)\}_{i=1}^{I(n)}$:

$$y \mapsto \mathcal{N}_1(y). \tag{5.28}$$

In Fig. 5.22 we show examples of the three types of points and the corresponding dimension 1 barcodes from the Vietoris–Rips filtration on the points in their δ-annuli. We classify the data points $y \in D$ according to the value of $\mathcal{N}_1(y)$. We show our classification procedure in Algorithm 5. We will later also refer to this method as *local PH*.

[14] In a very small neighbourhood of an intersection point, the points around it form two circles intersecting in two points. The dimension 1 homology group of two circles intersecting in two points consists of three loops. However, since we are working with data, we often do not achieve such an ideal scenario and we therefore expect to see intersection points with only two bars.

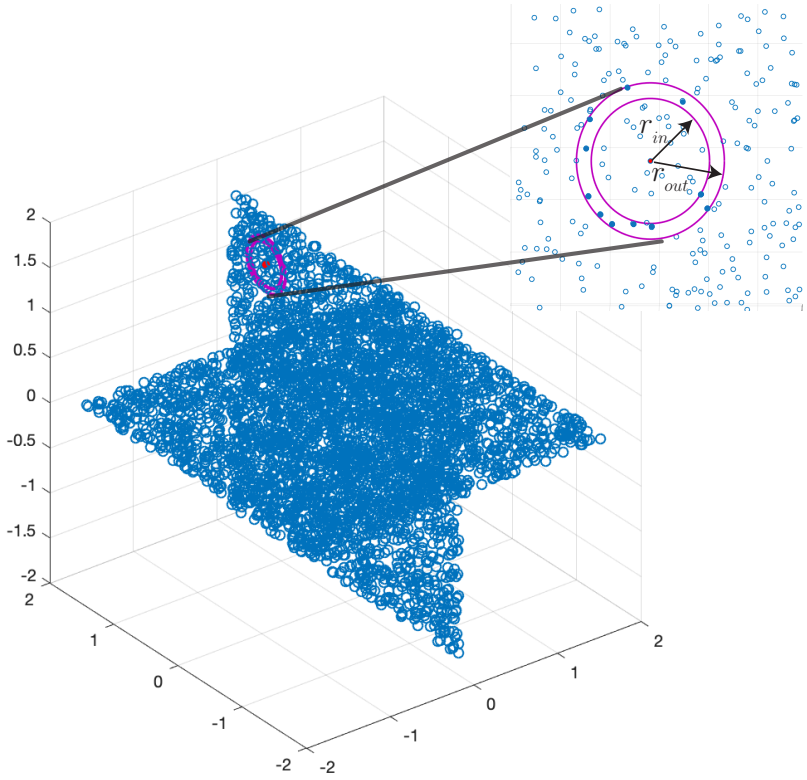


Figure 5.21: Visualisation of a δ-annulus around a point (shown in red) on the intersecting planes data set for $r_{out} = \delta = 0.3$ and $r_{in} = \frac{4}{5} \cdot 0.3$. We show the points within the δ-annulus in dark blue.

5.3.3 Implementation

We implement our method in MATLAB using RIPSER [32] for the computation of the local Vietoris–Rips complexes. For the calculation of the Bottleneck and Wasserstein distances we use DIONYSUS [173], for the calculation of the persistent landscape distance we use the PERSISTENT LANDSCAPE TOOLBOX [89]. We gratefully acknowledge the use of a code for local principal component analysis (PCA) written by Jared Tanner which we use for the comparison of our method to the detection of the singularities via local PCA [169, 170]. For the computation of the Hausdorff distance, we use the MATLAB function `HausdorffDist` [84].

5.3.4 Data sets

We use three different data sets of increasing geometric complexity. All data sets contain points close to intersections, in two data sets we also have points close to

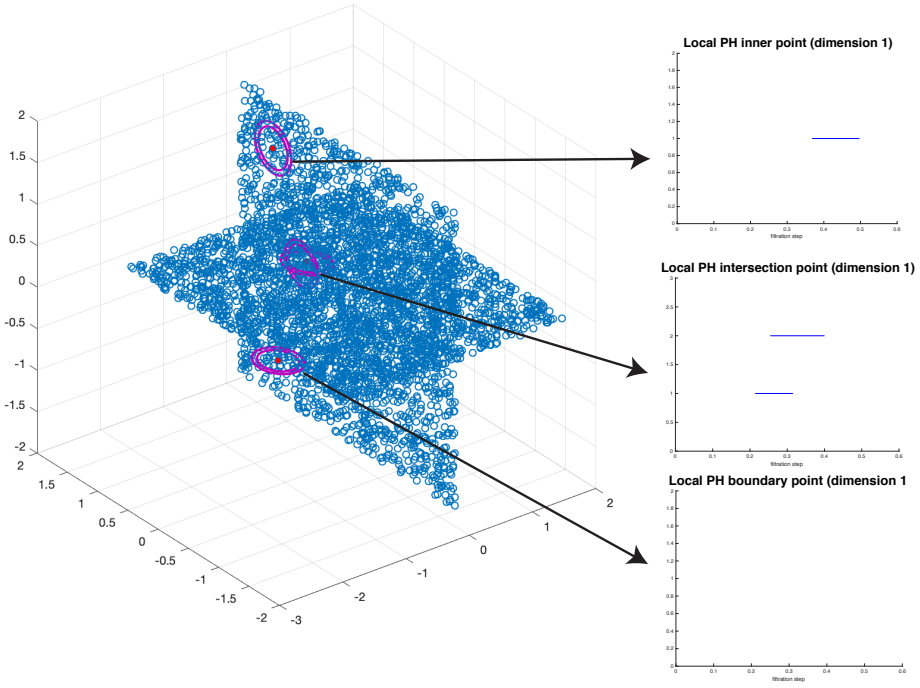


Figure 5.22: An example of a boundary point, an inner point and an intersection point and their corresponding local dimension 1 barcodes obtained from a Vietoris–Rips filtration on the points in the δ-annulus for $r_{out} = \delta = 0.3$ and $r_{in} = \frac{4}{5} \cdot 0.3$. As we expect, the barcodes of the three points show a different number of features. Note that the intersection point example only has two persistent bars rather than three as we would expect in an ideal scenario. This is often the case for data.

boundaries. Note that the points were not sampled to explicitly include points on intersections or boundaries.

5.3.4.1 Intersecting planes

We sample 4000 points from two intersecting planes: the xy-plane $[-2, 2]^2 \subset \mathbb{R}^2$ and the yz-plane $[-2, 2]^2 \subset \mathbb{R}^2$. The probability for every point to be sampled from either of the two planes is $p = 0.5$.

5.3.4.2 Cyclo-octane conformation space

The chemical molecule cyclo-octane C_8H_{16} consists of eight carbon atoms that form a ring. Each carbon atom is bound to two other carbon atoms and two hydrogen atoms. Taking into account chemical and physical forces, cyclo-octane can take different forms, called conformations, in 3-dimensional space depending on the location of

Algorithm 5 Classification of points on intersecting surfaces via dimension 1 persistent homology

Input: Data points $D = \{y_1, \ldots, y_N\}$, a distance function $d : D \times D \to \mathbb{R}$, local annulus outer radius r_{out}, local annulus inner radius r_{in}.
Output: A set of points with no dimension 1 persistent homology $D_b \subset D$,
a set of points with one persistent feature in dimension 1 $D_i \subset D$,
a set of points with two or more persistent features in dimension 1 $D_{int} \subset D$.

```
for all y ∈ D do
    Find annulus points D̃_y ⊂ D such that r_in ≤ d(y, ỹ) ≤ r_out for every ỹ ∈ D̃_y
    if | D̃_y |> 2 then
        Compute dimension 1 persistent homology of Vietoris–Rips filtration on
        D̃_y \ {y}
        Calculate number of bars N_1(y) with persistence > r_out − r_in
        if N_1(y) == 0 then
            D_b ← D_b ∪ y
        else if N_1(y) == 1 then
            D_i ← D_i ∪ y
        else if N_1(y) ≥ 2 then
            D_int ← D_int ∪ y
        end if
    end if
end for
```

the carbon atoms. As the positions of the hydrogen atoms are fully determined by the locations of the carbon atoms, every conformation can be represented as a point $(C_{1,1}, C_{1,2}, C_{1,3}, \ldots, C_{8,1}, C_{8,2}, C_{8,3}) \in \mathbb{R}^{24}$, where $C_{i,1}, C_{i,2}, C_{i,3}$ correspond to the coordinates of the i-th carbon atom in $\mathbb{R}^3$. Martin *et al.* [169, 170] found that the conformation space of cyclo-octane corresponds to the union of a Klein bottle and a sphere intersecting in two circles of singularities. We use the data set introduced by Martin *et al.* [169] to test whether we can recover the singularities using our classification approaches. The data set consists of 6040 points in $\mathbb{R}^{24}$ sampled from a larger data set consisting of 1 031 644 cyclo-octane conformations. This data set is publicly available as part of the JavaPlex package [247].

5.3.4.3 Henneberg surface

Henneberg's minimal surface is a self-intersecting surface with geometrically complicated boundaries. It is an immersion of $\mathbb{R}P^2$ in $\mathbb{R}^3$. We use a data set kindly provided by Martin *et al.* [170], which consists of 5456 points sampled from the Henneberg surface using the following parametrisation:

$$\begin{aligned} x &= \frac{2(\beta^2-1)\cos(\phi)}{\beta} - \frac{2(\beta^6-1)\cos(3\phi)}{3\beta^3}, \\ y &= -\frac{6\beta^2(\beta^2-1)\sin(\phi) + 2(\beta^6-1)\sin(3\phi)}{3\beta^3}, \\ z &= \frac{2(\beta^4+1)\cos(2\phi)}{\beta^2}, \end{aligned}$$

where $\beta \in [0.4, 0.6]$ and $\phi \in [0, 2\pi]$. Note that for this restriction of β the surface does not have a triple intersection. The points were sampled to maintain a pairwise minimal distance. Moreover, 32 points closest to the boundary intersections were moved to be within a set distance of their closest boundary. Further particulars for the sampling process that was applied can be found in [170].

5.3.5 Results

First we present the results for our two approaches to classifying data points on the intersecting plane data set. We then apply the classification via local dimension 1 PH features to the cyclo-octane conformation space and the Henneberg surface.

5.3.5.1 Intersecting planes

We first classify the points via distance measures as described in Subsection 5.3.1. We use $\delta = 0.3$ to calculate the local PH both for dimensions 0 and 1. We ignore points with no local dimension 1 PH (with the chosen δ radius there are no points with no local dimension 0 PH) and apply the Bottleneck distance, the Wasserstein distance, and the persistent landscape distance to calculate pairwise distances between the barcodes of the remaining data points in dimensions 0 and 1 separately. We apply

the k-medoids algorithm for $k = 2, 3, 4$ to the resulting distances matrices $\mathcal{D}_0$ and $\mathcal{D}_1$. We do not obtain the expected classes of points in any of these cases. We show an example of the obtained classes for dimension 1 barcodes, the Bottleneck distance, and $k = 2$ in Fig. 5.23. We observe that the only class of points that we can identify correctly are the boundary points as they are excluded from the cluster analysis due to a lack of signal in the dimension 1 barcode. Note that changing the value of δ does not influence the quality of resulting clusters. Applying other clustering algorithms such as average linkage clustering or community detection [143, 174] does not yield any improvements either.

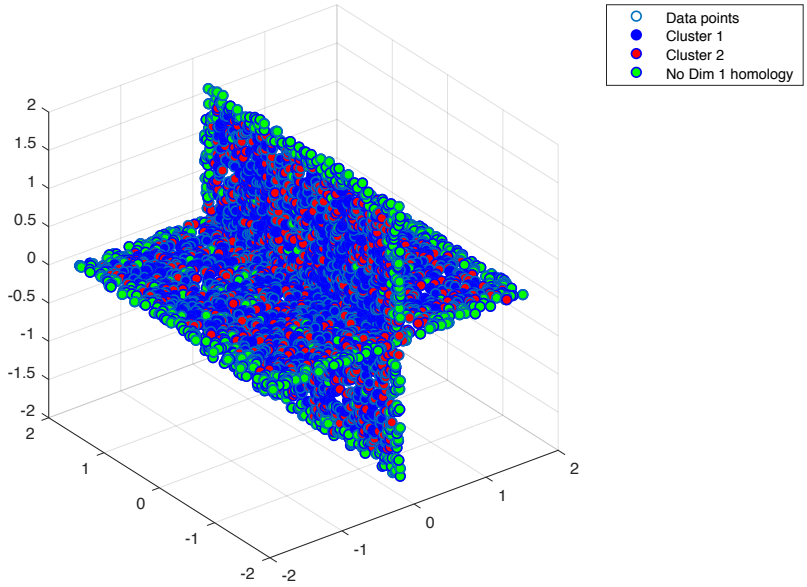
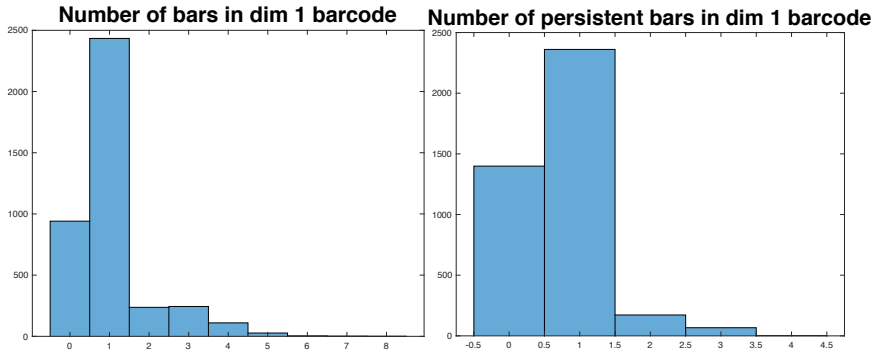


Figure 5.23: Clusters obtained from applying k-medoids for $k = 2$ to the pairwise Bottleneck distance matrix. We exclude points with no local PH in dimension 1 for $\delta = 0.3$.

We now proceed to classify the points based on their dimension 1 PH features (local PH) as described in Subsection 5.3.2. We first check whether our assumptions about the dimension 1 barcodes are correct by calculating the full number of bars. We show the histogram of the number of bars in Fig. 5.24 (a). We find that there are barcodes with many more bars than we expect. We now consider the number of persistent bars. We define a bar $[\eta_i, \zeta_i)$ to be persistent if $\zeta_i - \eta_i \geq r_{out} - r_{in}$.

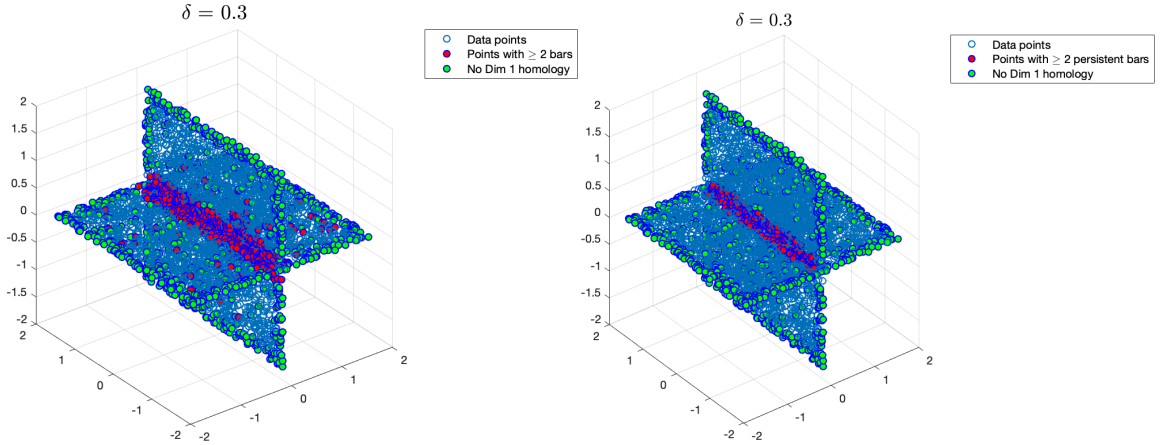
We show the histogram of the number of persistent bars in Fig. 5.24 (b). We find that this represents our expectations with at most three persistent bars in each local barcode. We proceed to the classification of the data points using both the number of bars and the number of persistent bars in the barcodes. Fig. 5.25 shows the data



(a) Total number of bars. **(b)** Number of persistent bars.

Figure 5.24: Histograms of the number of bars in dimension 1 barcodes of the intersecting plane data set. The horizontal axis shows the number of bars in the local dimension 1 barcodes, the vertical axis represents the number of points.

with the points coloured depending on whether they have no dimension 1 PH, or two or more (persistent) bars in the 1-dimensional barcode. We find that the number of



(a) Classification based on the total number of bars.

(b) Classification based on the number of persistent bars.

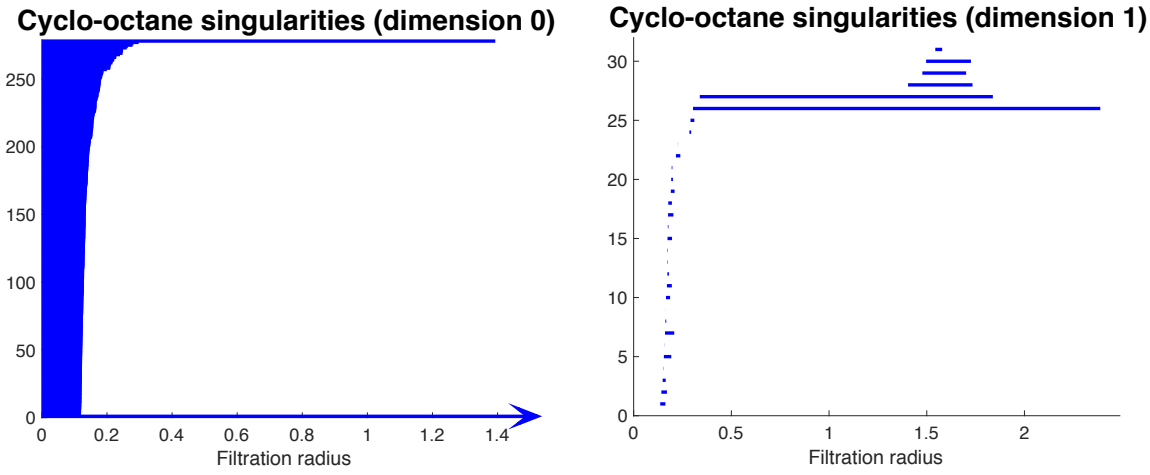
Figure 5.25: Points on the intersecting planes data set coloured by their local PH properties. We use $r_{out} = \delta = 0.3$ and $r_{in} = \frac{3}{4} \cdot 0.3$ for the δ-annuli.

persistent bars can identify the three expected classes very well. In particular, the intersection points are correctly classified. For the given δ-radius there are some inner

plane points that are classified as boundary points as they have no dimension 1 PH. Overall, the result is very promising.

5.3.5.2 Cyclo-octane conformation space

We apply our classification method via dimension 1 PH features to detect the points around the circular intersections of the Klein bottle and the sphere in the cyclo-octane data set. We use $r_{out} = 0.4$ and $r_{in} = \frac{5}{8} \cdot 0.4$ and identify points that have more than one persistent bar in their 1-dimensional barcode. We then perform PH on the identified points in $\mathbb{R}^{24}$ and show the corresponding barcodes in Fig. 5.26. We



(a) Dimension 0 barcode.

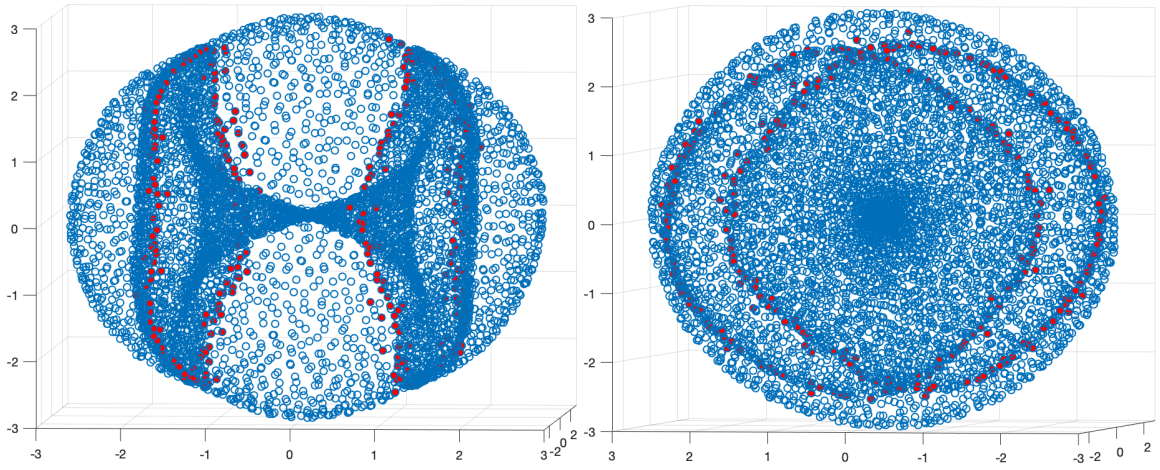
(b) Dimension 1 barcode.

Figure 5.26: Barcodes from a Vietoris–Rips filtration performed on the intersection points detected in the cyclo-octane data set.

find that the barcodes behave as we would expect for points sampled from two circles: in both dimension 0 and dimension 1, we have two persisting bars corresponding to the two circles. In dimension 1, we can also observe that, immediately after the two circles connect in dimension 0, loops appear which die as soon as there is only one loop left in the data set, i.e., when the two circles are connected to each other fully by 2-simplices. The points that we identify as being close to the intersection of the Klein bottle and the sphere therefore exhibit the topological properties of two circles.

Martin *et al.* [169, 170] used an isomap projection to 3D to visualise the data set

and its intersections. To further verify that we are finding the correct intersection points in the data set, we apply the isomap algorithm [249,250] with k-nearest neighbours for the construction of the graph that underlies isomap[15] with $k = 5$. We show the result including our identified intersection points in Fig. 5.27. We clearly see that



(a) **(b)**

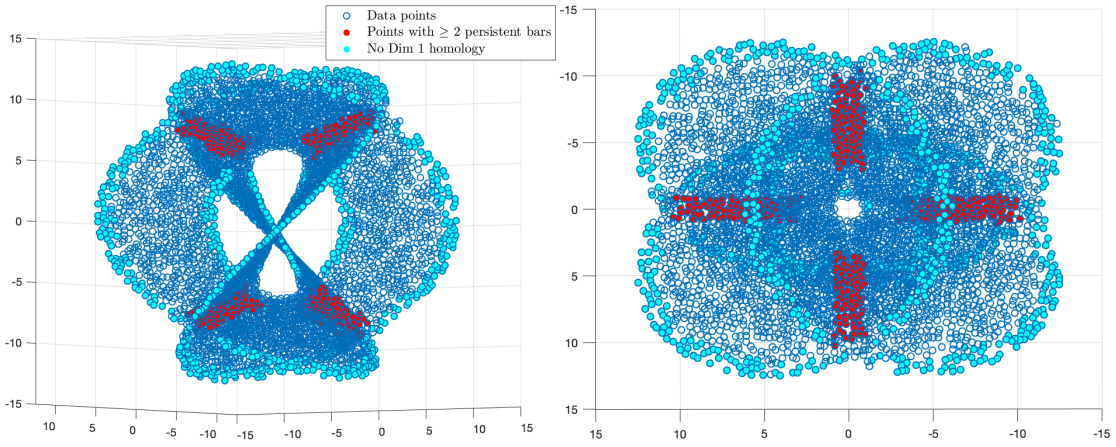
Figure 5.27: Points classified by their local PH properties in the cyclo-octane data set viewed from two different perspectives. We show points with two or more persistent features in dimension 1 for $r_{out} = \delta = 0.4$ and $r_{in} = \frac{5}{8} \cdot \delta$ coloured in red.

the points with more than one persistent bar in their local dimension 1 barcodes are indeed located close to the circular intersection of the Klein bottle and the sphere. Our method thus appears to be successful on the data set.

5.3.5.3 Henneberg surface

We apply our classification method via dimension 1 PH features to the Henneberg surface data set. Our aim is to detect both boundary points and intersection points. We use $r_{out} = 2$ and $r_{in} = \frac{3}{4} \cdot 2$ and show our results in Fig. 5.28. We find that the method successfully detects boundary and intersection points on the Henneberg surface. The method fails for some boundary points that are located close to intersections or other boundaries and are neither classified as boundaries nor as intersections.

[15]We thank Barbara Mahler for tuning the parameters in the isomap projection.



(a) (b)

Figure 5.28: Two perspectives on the points classified by their local PH properties in the Henneberg surface data set. We show points with two or more persistent features in dimension 1 for $r_{out} = \delta = 2$ and $r_{in} = \frac{3}{4} \cdot \delta$ coloured in red and points with no persistent features in dimension 1 in cyan.

Overall, our method however performs very well on the geometrically complex data set.

5.3.5.4 Comparison to local principal component analysis

Martin *et al.* [169, 170] used the cyclo-octane data set (and the Henneberg surface data) to demonstrate a novel method to triangulate non-manifold-like surfaces. In a first step, the authors apply a local version of principle component analysis (PCA) to the data. For every point, the authors use local PCA to determine whether the data points in a neighbourhood around this point can be approximated by a 3-dimensional or 2-dimensional affine space. They thereby identify points close to intersections as points whose neighbourhood is approximated by a 3-dimensional rather than a 2-dimensional space. We compare our local PH method to the author's local PCA approach.

Our first observation is the following: even though both approaches are built on local neighbourhoods of points in a data set, these neighbourhoods can be substantially different. See, for example, Fig. 5.29 where we depict a data set sampled from two cones that intersect in one point (note that we do not specifically include this

intersection point in out data set). While we see that both methods can successfully

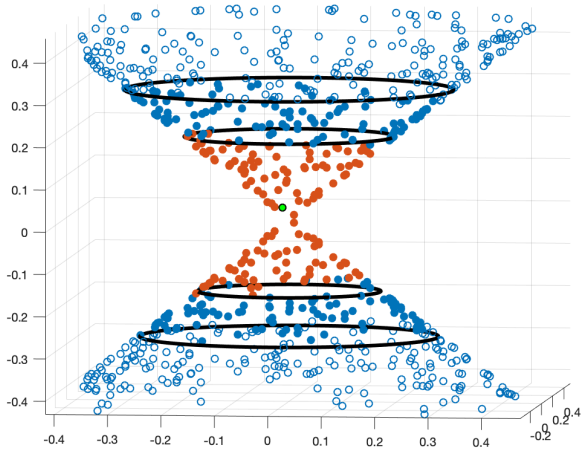


Figure 5.29: Local neighbourhoods considered for the detection of intersections using local PH versus local PCA on data sampled from intersecting cones. We show the point whose neighbourhood we study coloured in green. In local PH we perform computations on an annular neighbourhood around the green point which, for example, consists of the two sets of points coloured in blue. In contrast, local PCA [169, 170] for the green point studies the union of the orange and blue points together with the green point.

identify points close to the intersection of the cones, they each consider a substantially different set of points to arrive at their conclusions.

We now compare the robustness of both methods with respect to the choice of the size of neighbourhood δ in Fig. 5.30. For each neighbourhood radius δ, we compute local PH on the cyclo-octane data set with $r_{in} = \delta$ and $r_{out} = \frac{5}{8}\delta$ and apply local PCA for a local neighbourhood size δ (this corresponds to the parameter ϵ in [170]) keeping all other parameters as in [170]. We observe that, for each value of δ, the points detected as intersection points by local PH are subsets of the points identified by local PCA. We further use the intersection points detected at $\delta = 0.2$ with local PCA and intersection points detected at $\delta = 0.4$ with local PH as reference point sets for the respective method. As we increase δ, we compute the Hausdorff distance to the reference data set for each method for each set of identified intersection points. We find that, for $\delta = 0.45$, the Hausdorff distance increases steeply for local PCA. The reason for this is that local PCA begins to include points from the curved region of the Klein bottle in the set of intersection points (see upper righthand corner of Fig. 5.30).

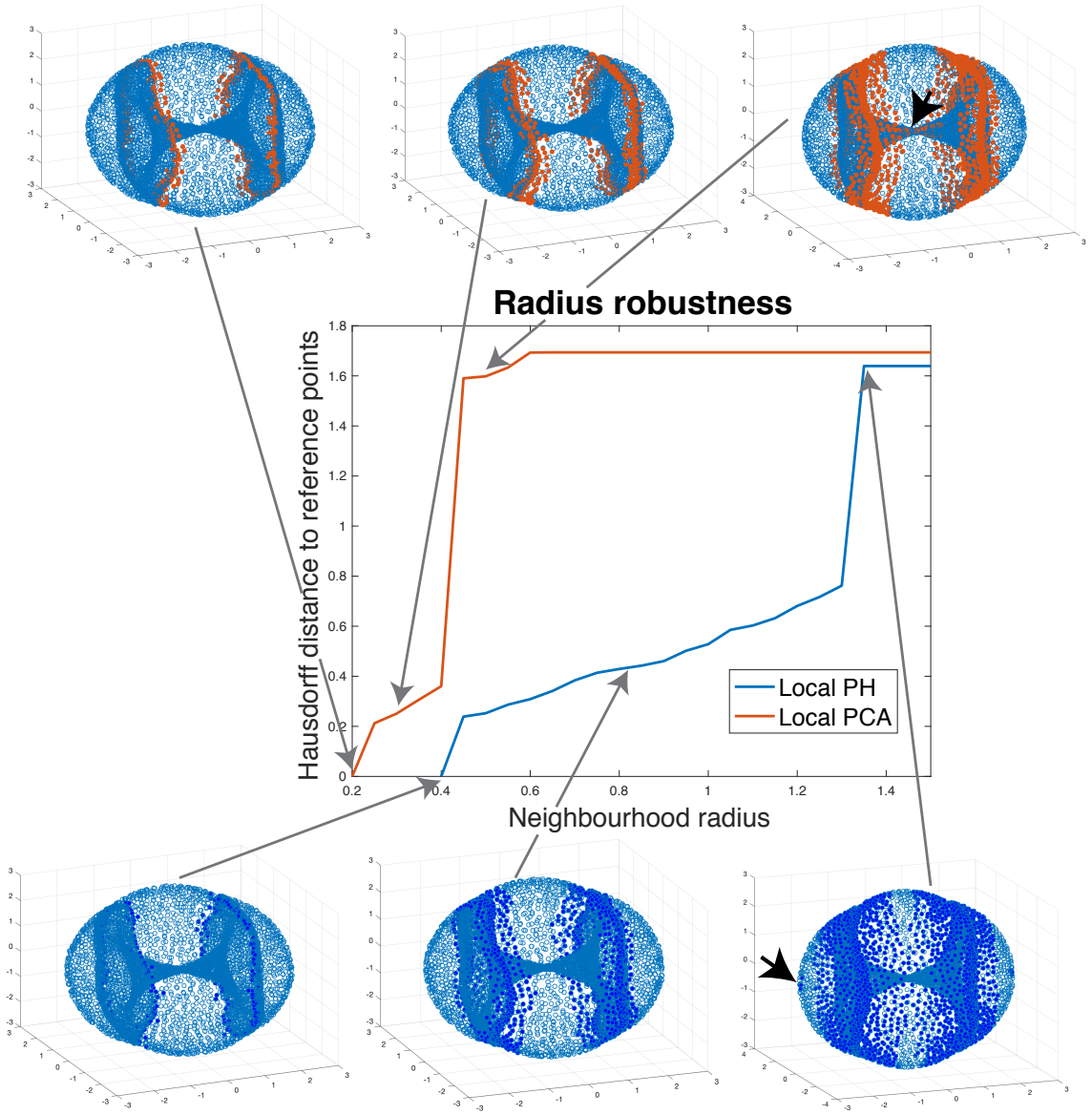


Figure 5.30: Robustness with respect to the choice of local neighbourhood size for the detection of intersections in the cyclo-octane data set using local PH versus local PCA. The horizontal axis of the plot represents the size of the local neighbourhood, the vertical axis corresponds to the Hausdorff distance between the selected intersection points and the set of reference points for the respective method. We show several sets of intersection points detected by local PCA (upper row) and local PH (lower row) and their locations in the isomap projections of the data set. We depict the set of reference points for local PCA in the upper lefthand corner and the reference points for local PH in the lower lefthand corner. We illustrate examples of points that are responsible for the steep increase in the Hausdorff distance for the respective method with black arrows.

In contrast, local PH seems to gradually increase selected intersection points in the region around the intersections up to $\delta = 1.3$. From $\delta = 1.35$ onwards, local PH also starts to include points that are far away from the reference point set, but these

originate from a different region of the Klein bottle than for local PCA (see lower righthand corner of Fig. 5.30). Based on these results, we note that for the cyclo-octane data set a) local PH is more robust to changes in the neighbourhood radius, and b) local PCA seems to be more prone to falsely identify points from regions with high curvature as intersection points. Observation b) implies that local PCA would misclassify points sampled from a surface shaped like an egg carton as belonging to a 3-dimensional region of the data and therefore being close to an intersection, while local PH would in such a case only detect one persistent loop around every point and correctly conclude that the points are sampled from a locally 2-dimensional object.

5.3.6 Summary and discussion

We were able to successfully classify points sampled from intersecting surfaces based on their geometric properties which we measured by determining the number of persistent features in their local dimension 1 barcodes. Specifically, we could identify points close to boundaries, points close to intersections and points that were neither close to an intersection nor close to a boundary. We demonstrated that the method can be applied to geometrically complex data sets. For example, in the 24-dimensional cyclo-octane conformation space data which is the union of a Klein bottle with a sphere intersecting in two circles, we were able to identify the points close to the intersection and verify that they have the expected PH and the correct location in an isomap projection to $\mathbb{R}^3$. For data sampled from Henneberg's minimal surface, we correctly detected boundary points and intersection points with few points that were not correctly classified.

Classifying data points based on their persistent topological features in dimension 1 is a very simple approach. For the method to be successful we did not require points to be sampled specifically from the intersection of the surfaces. It was sufficient that there were points around intersections and that the data points were sampled

reasonably uniformly. Choosing the optimal radii for the local annuli requires further investigation and could in fact pose an interesting application for multidimensional PH as there are two natural parameters.

We compared our method to an existing approach using local PCA [170] which can also detect intersections in data. For the cyclo-octane data set, we showed that local PH differs from the local PCA approach and has clear advantages, in particular with respect to robustness to the local neighbourhood and curvature of the underlying surface from which the data was sampled.

In the future, it will be interesting to see whether our method can provide new insight into real-life data sets where intersections or boundaries of surfaces carry relevant information. The method could further provide an approach to reduce the size of data sets for the computation of PH in cases where only the shape of the boundary of the data is relevant, for example, to detect whether certain cell types tend to be located around circular tissue structures. It would also be interesting to investigate whether the local PH barcodes carry more geometrically relevant information, for example, on intersection angles.

In summary, local PH can provide a simple yet powerful tool to detect points in geometrically interesting locations of data surfaces. We expect the method to be relevant for numerous applications in data science and applied algebraic geometry.

"ich meße ouch der sunnen rat,
wie wit das uf den wolken gat
der regenbogen unde blibt,
wo im ein zil min linge schribt."[a]

From the speech by the personified art of geometry (Geometria) in *Der meide kranz*, Heinrich von Mügeln, 14th century. Edition by [263].

[a]Translation: "I also measure the wheel of the sun; [I measure] how far above the clouds the rainbow goes and [where] it stops, when my ruler sets the limit for it. " [263].

6 Discussion and Outlook

In the present thesis, we applied persistent homology (PH) to study biological data, using both existing techniques and developing novel approaches. We thereby went through every step in the PH pipeline, which we outlined in Fig. 1.2 in the Introduction chapter. To characterise tumour blood vessel networks spatially, we developed the radial filtration (see Chapter 3). On two different experimental data sets, we showed that the number of loops in tumour blood vessel networks and their spatial distribution reflect effects of tumour treatment with vascular targeting agents and radiotherapy. We further investigated the number of connected components with short persistence for one of the data sets and found that it follows similar trends for the different treatment groups, as the number of loops. We concluded that the radial filtration is suitable to quantify tumour vasculature spatially, both via topological features in dimension 0 and dimension 1. For functional neuronal networks, we applied the weight rank clique filtration and observed that it is useful for the study of human motor learning and schizophrenia (see Chapter 4). For the human motor learning data, we identified concrete barcode features that represent underlying

communities in the networks. We compared two PH output analysis methods, persistence landscapes and persistence images, on functional networks from schizophrenia patients, healthy siblings, and healthy controls. We highlighted the different types of insights that the two methods can give. In Chapter 5, we developed two novel techniques that use a local version of PH. To circumvent computational complexity of PH on large and noisy data sets, we introduced a novel method that selects landmarks for PH from such data sets. To create a measure of how well suited a particular point is as a landmark, we extracted features from its local PH barcode. We ranked all data points using this measure to identify a landmark set. Our landmark selection process proved more robust to outliers than existing standard techniques, as well as a clustering-based technique, on all of our synthetic data sets, in particular for low sampling densities. Finally, we explored the use of local PH for the detection of geometric anomalies in data sampled from intersecting surfaces. We found that our approach is well suited to the task and that it has distinct advantages over an existing method (see Chapter 5).

PH is a promising technique for many different applications in biology where being able to quantify shape and how it changes is important. An example of such an application is the tumour blood vessel data studied here. PH also considers higher order interactions which is an advantage over existing techniques when studying brain networks. In both biological applications in this thesis, however, while PH provided interesting and relevant insight, we needed prior knowledge of the system to inform the right choice of filtration or the interpretation of PH output. For example, when we developed the radial filtration, we specifically wanted to capture vessel loops and tortuosity, which are known to be relevant in tumour vasculature. We built our filtration based on existing filtrations which have been used to capture tortuosity in other structural data sets. Similarly, for the functional neuronal networks, we knew that the motor-learning data included community structure, which we found to be reflected

in short-lived loops in the beginning of the filtration. For the schizophrenia data set, we required subject group knowledge to separate schizophrenia patients, healthy siblings of schizophrenia patients, and controls. As discussed in Subsection 1.2.1 of the Introduction chapter, the choice of filtration determines the information that we can obtain. Moreover, interpreting barcodes or even persistence landscape or persistence images summaries beyond persistent features is far from intuitive. Consequently, while PH does identify biologically relevant features successfully, it is not as versatile as other methods. PH is, in particular, difficult to use for exploratory data analysis. Although, for point clouds, PH is stable with respect to noise, it is currently also not clear how noise affects the application of PH to networks, which again makes it difficult to use in biological applications. For example, imaging noise or errors in the data extraction pipeline for our blood vessel data could introduce artefacts which systematically lead to loops in the extracted data which are not present in the biological networks. Particularly in small blood vessel networks, this could change the outcome of the analysis. Therefore, there are many limiting factors for the application of PH to biological data.

A major challenge for the use of PH in biology is the computational complexity of the calculation. In both our biological applications, we had to find strategies to reduce the data set to make PH computable. As we studied networks, we were able to make use of comparatively straightforward, yet effective, reduction techniques. For the tumour blood vessels, our data contained multiple points sampled from vessel segments between two branching points. We were therefore able to reduce the number of points on every segment while maintaining branching points and preserving loops in the network. One could imagine scenarios where this strategy fails and has a significant effect on loops in the vessel networks, but we did not encounter such cases in practice. We are yet to investigate the effect of this reduction on dimension 0 barcodes of the radial filtration which capture tortuosity in blood vessel networks.

Due to the spatial nature of the radial filtration, we were also able to approximate the full filtration by considering filtrations of two separate regions around the tumour centre. In the case of one of our data sets, these reduction approaches were not enough to ensure timely PH computation on the full data set. In future, we could, however, employ further combinations of these strategies to finish the computations. For the functional neuronal networks, we computed filtrations up to a threshold of edge weights and, in the case of the schizophrenia data, only considered positive edge weights. This was sufficient to enable computation on the data sets. Since one can argue that features arising from highly synchronised nodes in these networks are more relevant than those arising from weakly synchronised nodes, this strategy should still enable PH to capture relevant features despite an inevitable loss of information.

Biological data, however, is not always in the form of networks. Moreover, many data sets, such as gene expression data (see, for example, the data studied in [165]), are very large, high-dimensional, and noisy. This is a combination that currently poses major challenges for PH. We provide a novel method to select landmarks for the computation of PH on noisy point clouds based on the computation of local PH around data points (see Chapter 5). While our results on simple artificial data sets are very promising, it would be interesting to test the method on biological data sets and study the PH of the landmarks. It could further be insightful to study the set of points that local PH identifies as being least suitable as landmarks and observe whether these points are considered to be outliers by other methods. While we provide a novel preprocessing approach for large and noisy data sets, many others will need to be developed to make PH more applicable, in particular for clinical settings where fast analysis of data is crucial. Simultaneously, it is important to further simplify PH computation, as reduction techniques are not useful in all cases. For our blood vessel data, for example, it would have been informative to study the vascularisation of the tumours by considering the nodes of the vessel networks, or specific subsets thereof,

as input points for a Vietoris–Rips filtration. Even after we reduced the data points to the smallest subsets which would still give us biologically interesting information, i.e., branching points and end points of vessels, the computation of the Vietoris–Rips filtration proved infeasible. Applying landmark selection techniques would no longer ensure biological interpretability of results in this case.

In addition to informing landmark selection, we found that local PH can identify geometric anomalies such as intersections and boundaries in data sampled from intersecting surfaces (see Chapter 5). We observed that our method is more robust with respect to the choice of neighbourhood size and regions of high curvature in the data in comparison with detecting intersections via local PCA. We therefore found local PH to be useful for the study of geometrical problems that are not intrinsically topological in nature. PH has previously been observed to carry geometrical information [56] and we believe that there are many further connections yet to be explored. The local focus of our method means that PH in this case is easily computable and parallelisable over large data sets. As biological data can be very heterogeneous, being able to detect intersections could lead to interesting insights.

In the present thesis, we have found both global and local PH to be applicable to a variety of interesting data problems. In all applications, however, PH only provides one part of the puzzle. In future, we therefore believe that it will become more important to understand PH as one of many ways to generate interpretable features, which characterise aspects of a data set, rather than a data analysis method by itself. One can then feed these features into other existing methods from data science, see Fig. 6.1 for a modified version of the PH pipeline. While methods that vectorise PH output, such as persistence landscapes [53, 55], persistence images [3], kernel-based approaches (see, for example, [69,155,156,211]), or persistence codebooks [278], allow a combination of PH with machine learning, they do not, in most cases, enable us to trace specific topological features that are relevant to a biological problem. Even

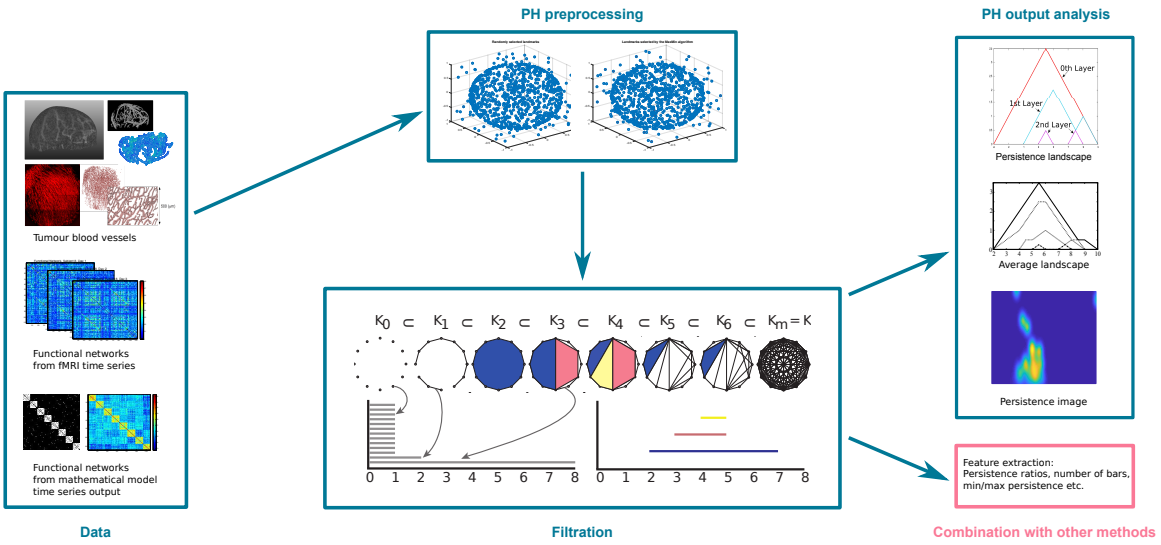


Figure 6.1: Combination of persistent homology with other methods from data science. The figure contains modified versions of our images in [237, 238, 240] as well as images of experimental data provided by Russel Bates and Bostjan Markelc/Jakob Kaeppler (with permission).

though these methods are very useful, they thus also reduce interpretability of results. As we saw in Chapter 4, persistence images [3] allow us to identify discriminative features in PH output, but these are also limited by choices, for example, the resolution of the pixel grid. Hofer *et al.* [135] use deep learning to detect which topological features in a persistence diagram are important to a specific task, but they require large training sets in their supervised tasks. This is not always possible when working with biological data, see, for example, the data in Chapter 3. We therefore believe that using PH to generate interpretable feature vectors, such as, for example, in [18, 43], is a promising approach, in particular since these features can then be combined with other biological parameters in the data in a larger analysis.

In summary, PH has proven its potential to provide valuable insight into biological problems. A lot of research effort is, however, still necessary until PH can become an easily applicable and interpretable technique.

" ich mak wol sin der künste dach,
sint das min zirkel in sich sloß
alles das uß naturen floß."[a]

From the speech by the personified art of geometry (Geometria) in *Der meide kranz*, Heinrich von Mügeln, 14th century. Edition by [263].

[a]Translation: "I may well be the cloak around all the arts, since my compasses have encircled everything which flowed out of nature." [263].

7 Conclusions

The emergence of data science as its own field has had a great impact on structural biology. Conversely, there is great potential for structural biology to directly influence developments in data science [179]. This thesis contributed to currents in both directions. We highlighted some of the impact that PH can have on two specific biological problems (tumour blood vessels and functional neuronal networks). Conversely, we developed a novel outlier-robust subsampling method in response to computational problems that arose when applying PH to biological data (PH landmark selection). We then found that we could extend the topological tools that we used for subsampling to detect geometric anomalies in non-manifold like data.

We conclude that:

1. PH is a powerful tool to study shape in biological data, but there are still many practical hurdles. In particular, PH only provides insights on one aspect of the data and will benefit when combined with other techniques.

2. Our developed radial filtration for tumour blood vessel networks shows great potential to enhance spatial understanding of structural characteristics in vessel

networks arising from tumour-induced angiogenesis as well as from the effects of tumour treatment. However, computation of PH on this data remains challenging.

3. The number of loops and their distribution in a tumour blood vessel network provide useful measures for structural abnormality of the vasculature.

4. The number of short bars in dimension 0 barcodes of the radial filtration also captures abnormal characteristics of tumour blood vessels.

5. PH can provide insight into functional neuronal networks during motor-learning and in disease.

6. Whether persistence is a good measure for the relevance of a topological feature in a barcode depends on the filtration. When interpreting PH output from a new data set or filtration one should consider all features, regardless of their persistence.

7. Local PH can be used for landmark selection before applying PH. Local PH based landmark selection outperforms existing standard methods with respect to robustness to outliers on large and noisy data sets. The measure for topological outlierness that we compute for our selection could be investigated for other applications.

8. Local PH can detect geometrical properties such as the presence of intersections or boundaries in data sampled from intersecting surfaces.

A

Useful Additional Mathematical Definitions

We list definitions that can be helpful for understanding some of the mathematical background of this thesis. Most of these definitions can be found in [79]. For basic topological definitions we used [178], but note that [79] and [150] also contain equivalent but more complicated formulations in most cases. We most of the definitions also appear in [237] with minor modifications.

A.1 Topology definitions

A.1.1 Topological spaces

Definition A.1.1 (topology)**.** A *topology* on a set $\mathbb{X}$ is a collection $\mathcal{T}$ of subsets of $\mathbb{X}$ with the following properties:

i. $\emptyset$ and $\mathbb{X}$ are in $\mathcal{T}$.

ii. The union of elements of any subcollection of $\mathcal{T}$ is in $\mathcal{T}$.

iii. The intersection of elements of any finite subcollection of $\mathcal{T}$ is in $\mathcal{T}$.

We call sets that belong to the collection $\mathcal{T}$ *open sets* of $\mathbb{X}$.

Definition A.1.2 (topological space)**.** A *topological space* is an ordered pair $(\mathbb{X}, \mathcal{T})$ that consists of a set $\mathbb{X}$ and a topology $\mathcal{T}$ on $\mathbb{X}$.

Remark 6. We often refer to $\mathbb{X}$ as the topological space.

Definition A.1.3 (separation)**.** A *separation* of a topological space $\mathbb{X}$ is a disjoint partition $\mathbb{X} = U \dot{\cup} W$ into two non-empty, open subsets. We say that a topological space is *connected* if there exists no separation of $\mathbb{X}$.

Connectedness is an important topological property that stays invariant under continuous functions.

A.2 General mathematical definitions

Definition A.2.1 (Hausdorff space)**.** A topological space $\mathbb{X}$ is called a *Hausdorff space* if for every pair $x_1, x_2 \in \mathbb{X}$ of distinct points there exist neighbourhoods U_1 and U_2 of x_1 and x_2 respectively that are disjoint.

Definition A.2.2 (compact)**.** A topological space $\mathbb{X}$ is said to be *compact* if for every covering of $\mathbb{X}$ by open sets we can find a finite subcollection of open sets that also covers $\mathbb{X}$.

Definition A.2.3 (m-manifold)**.** An *m-manifold* is a compact, connected Hausdorff space $\mathbb{X}$ such that each point $x \in \mathbb{X}$ has a neighbourhood that is homeomorphic to an open subset of $\mathbb{R}^m$.

B

Additional Information and Results

B.1 Persistent homology applied to tumour blood vessel networks

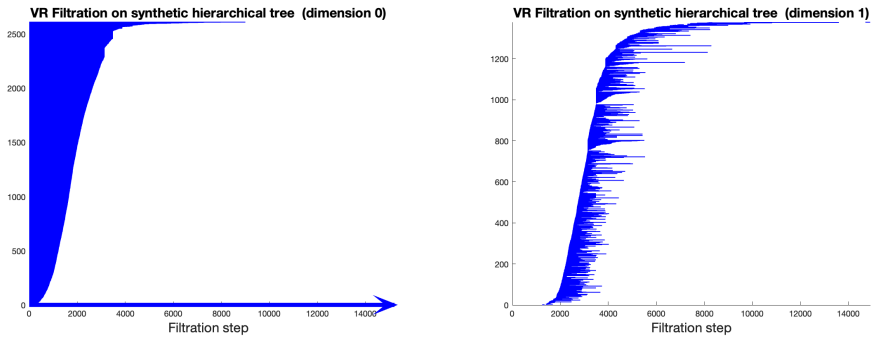
B.1.1 Vietoris-Rips filtration on branching points

We perform the Vietoris-Rips filtration on branching points (both inner branching points and end points) of our vessel point clouds to assess the degree of vascularisation of the tumours. Due to the size of our blood vessel netoworks we only obtain partial results from the software Ripser, which we show below, but no barcodes for the full filtration. This was also the case when we restricted the Vietoris-Rips filtration to inner branching points and when we used a reduced version of the algorithm. We use a modified version of the barcode plot function from [191] to visualise our barcodes.

B.1.1.1 Example barcodes for synthetic hierarchical tree

In Fig. B.1 we show the barcodes for the Vietoris-Rips filtration on the branching points of the synthetic hierarchical tree. We find a small number of persistent bars

in the dimension 0 barcode as well as two filtration steps (around 3500) where many connected components seem to merge at once. In dimension 1 we observe many short bars, several medium persistent bars as well as one persistent bar, which we interpret as signs of good vascularisation. For dimension 2 we did not obtain any results. Note that the filtration was aborted by Ripser and we are therefore only observing partial results.



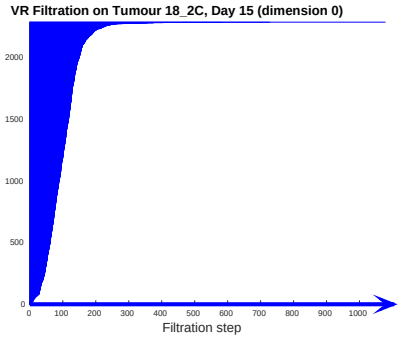
(a) **(b)**

Figure B.1: Barcodes from the Vietoris-Rips filtration performed on branching points of the synthetic hierarchical tree.

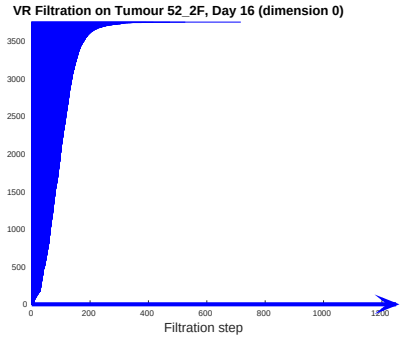
B.1.1.2 Example barcodes for the multiphoton intravital 3D imaging.

We show example barcodes for four of our data categories in Fig. B.2 and Fig. B.3. Again, we are only seeing partial results as Ripser aborted the computations. In dimension 0, taking into account the different scaling, we do not see a large difference between the different treatment regimes. There is one notable persistent bar in the control branching points, the other tumours do not exhibit bars that are as persistent. For dimension 1, we find a very striking difference in the barcode of the anti-Dll4 treated tumour that exhibits one very persistent feature that occurs halfway through the filtration. While we did observe a persistent feature for the hierarchical tree, in the case of the anti-Dll4 treated tumour it occurs a small gap of no features in the filtration. For the other tumour categories we find several medium scale persistent

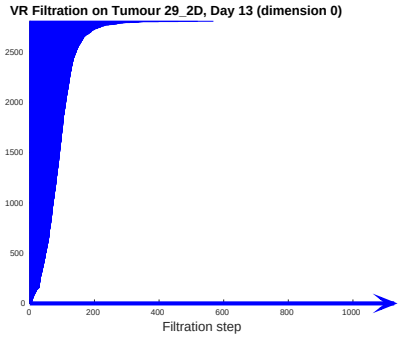
features in the barcodes. For the irradiated tumour we observe a total smaller number of features in dimension 0 and 1 than for the other tumour categories.



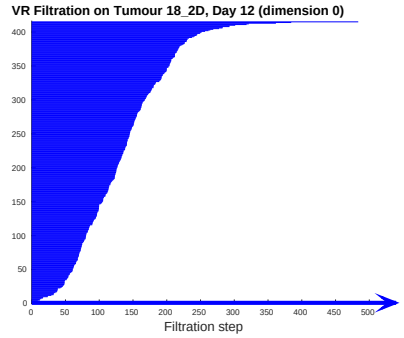
(a) Control, day 3 of observation



(b) DC101 treated, day 3 after treatment

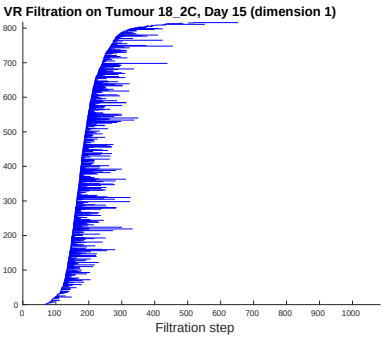


(c) Anti-Dll4 treated, day 3 after treatment

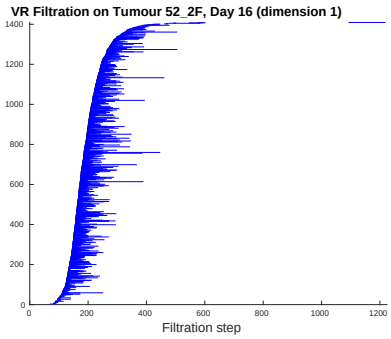


(d) Single-dose irradiated, day 3 after treatment

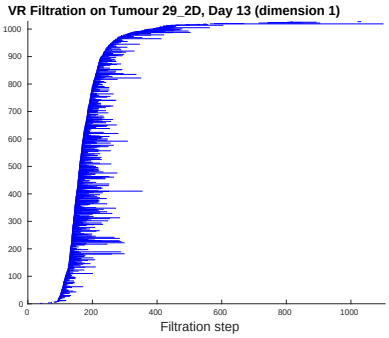
Figure B.2: Vietoris-Rips barcodes dimension 0



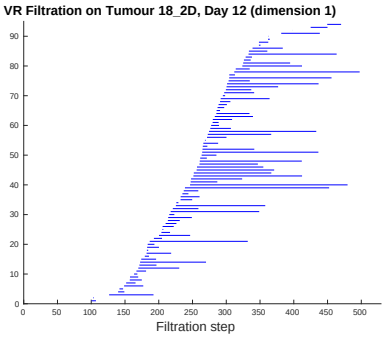
(a) Control, day 3 of observation



(b) DC101 treated, day 3 after treatment

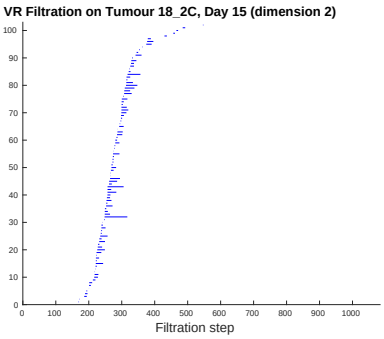


(c) Anti-Dll4 treated, day 3 after treatment

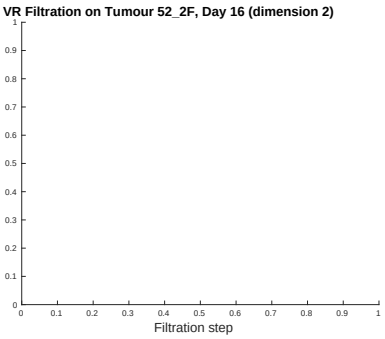


(d) Single-dose irradiated, day 3 after treatment

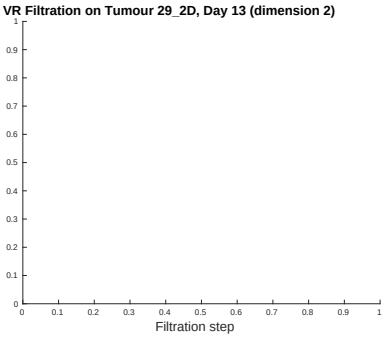
Figure B.3: Vietoris-Rips barcodes dimension 1



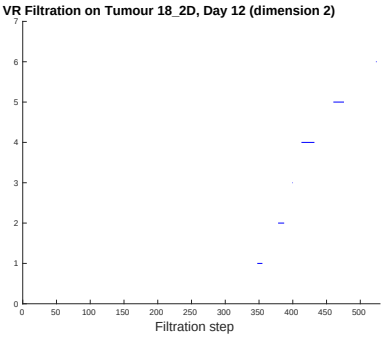
(a) Control, day 3 of observation



(b) DC101 treated, day 3 after treatment

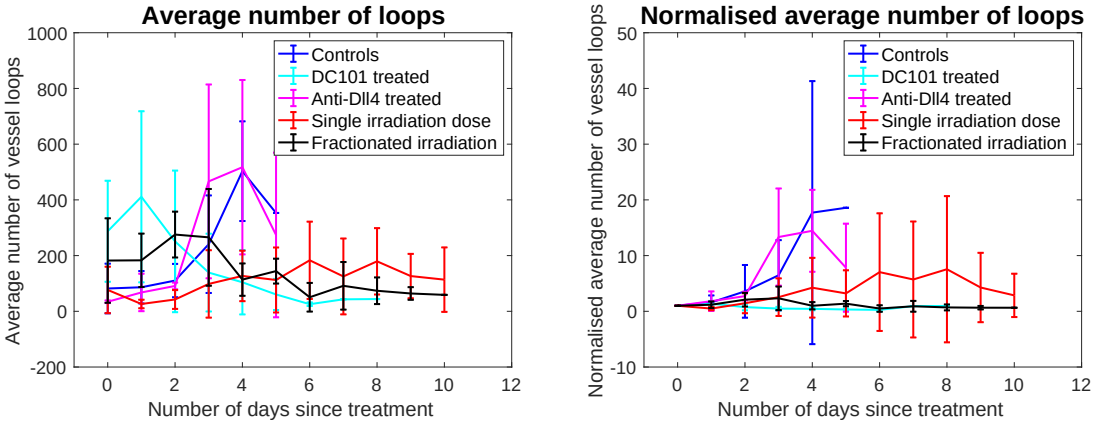


(c) Anti-Dll4 treated, day 3 after treatment



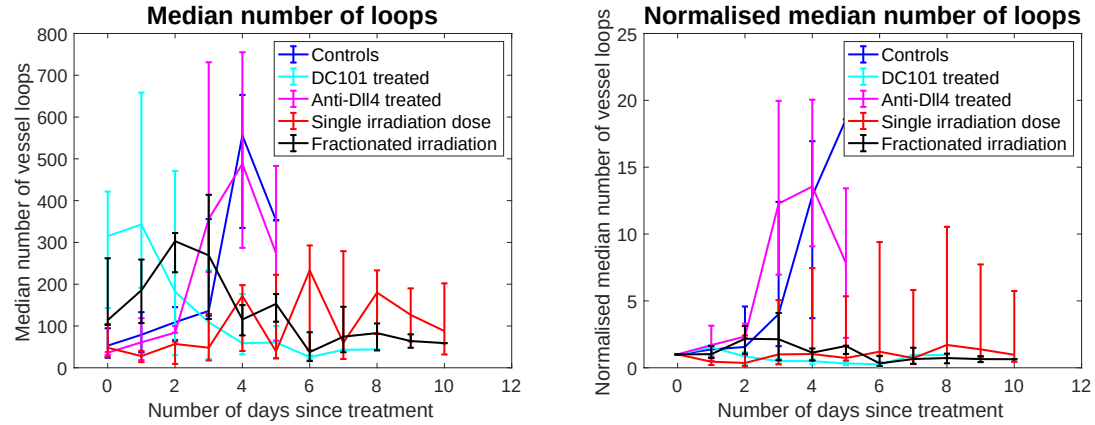
(d) Single-dose irradiated, day 3 after treatment

Figure B.4: Vietoris-Rips barcodes dimension 2



(a) **(b)**

Figure B.5: Average total number of loops captured by the radial filtration in dimension 1.



(a) **(b)**

Figure B.6: Median total number of loops captured by the radial filtration in dimension 1.

B.1.2 Number of loops and their distribution

B.1.2.1 Multiphoton intravital 3D imaging

We show the figure versions from Chapter 3, Subsubsection 3.5.4.1 when including the full data set and all available days in Figures B.5, B.6, B.7, and B.8.

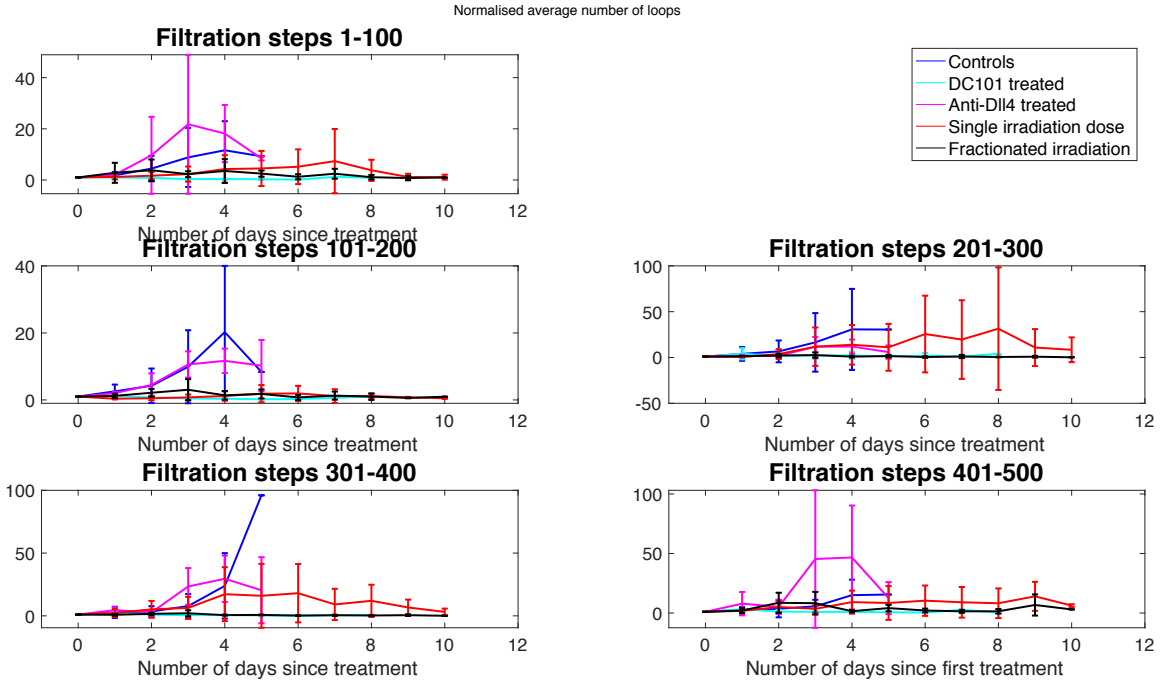


Figure B.7: Normalised average number of loops for different filtration intervals. We perform normalisation with respect to the number of loops in the specified filtration interval on day 0.

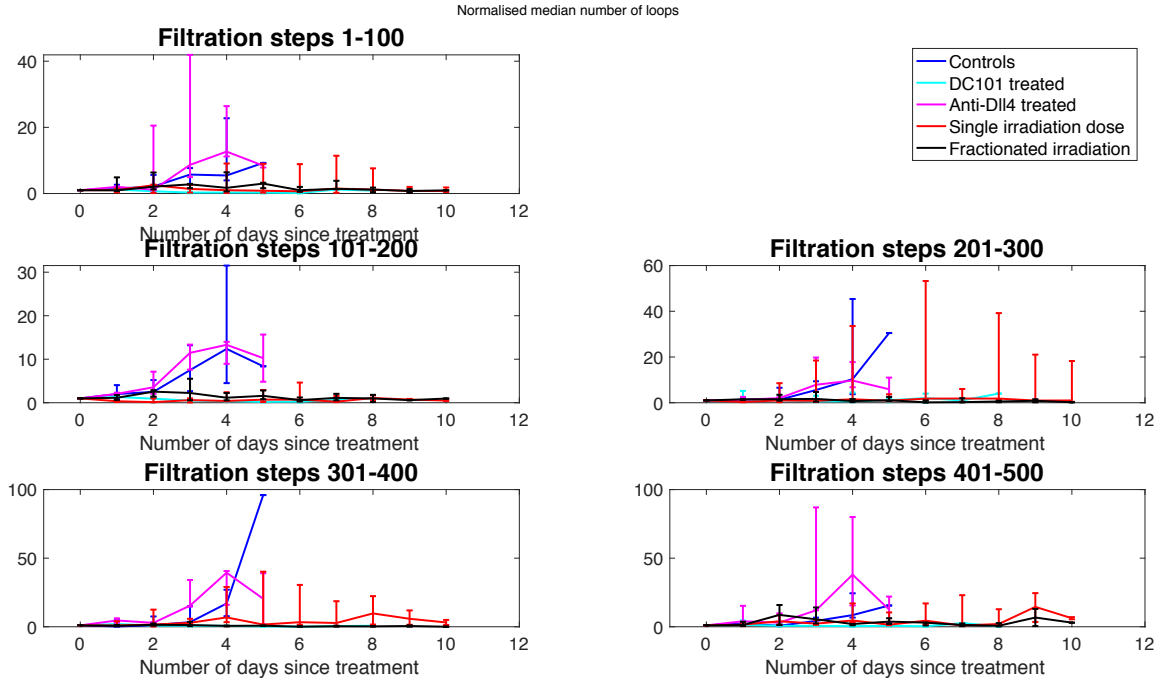
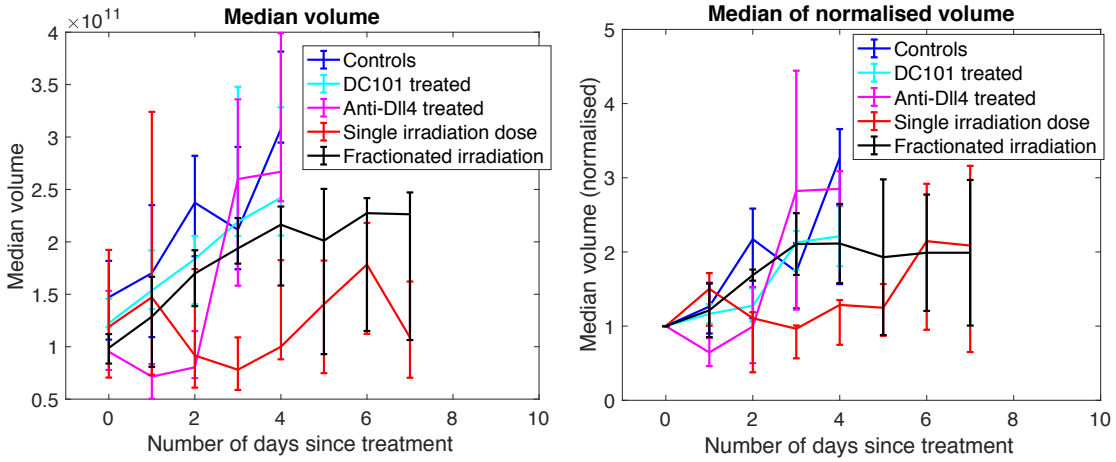


Figure B.8: Normalised median number of loops for different filtration intervals. We perform normalisation with respect to the number of loops in the specified filtration interval on day 0.

B.1.3 Tumour volume approximations

We compute approximations of tumour volume by calculating the volume of the sphere whose radius is the maximal radius used in the radial filtration. We present results in Fig. B.9 for the intravital data and Fig. B.10 for the ultramicroscopy data.



(a) (b)

Figure B.9: Medians of approximated tumour volumes for the intravital data in μm^3. We show median values and interquartile distances.

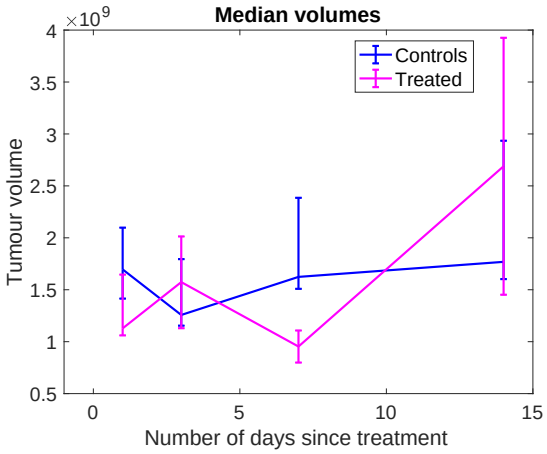


Figure B.10: Medians of approximated tumour volumes for the ultramicroscopy data. Since we do not know the true length scales in the data, the radii used to compute the approximate volumes are only proportional to the 'true' radii. We show median values and interquartile distances.

B.2 Persistent homology applied to task-based fMRI data

B.2.1 Application to motor-learning data

B.2.1.1 Table with often-occurring brain regions in 1-dimensional loops

In Tables B.1 and B.1, we indicate the brain regions that often occur in 1-dimensional loops.

Table B.1: Loops that consist of edges that occur in loops of functional networks at least 50 times over all subjects (part I). We list the loops that we find in the left column. (We start with one of the nodes, which we choose arbitrarily, and end with the node that is adjacent to the starting node in the loop.) We denote an occurrence of a loop on a specific day with the symbol x in the table and present variations of the loop that we interpret as representing the same loop. We use the following abbreviations for the brain regions: l: left; r: right; ant: anterior; post: posterior; AnGy: Angular gyrus; CinGy: Cingulate gyrus; COC: Central opercular cortex; FOC: Frontal operculum cortex; FMedC: Frontal medial cortex; FP: Frontal pole; HG: Heschl's gyrus; IC: Insular cortex; InfFGyPT: Inferior frontal gyrus pars triangularis; IntCalC: Intracalcrine cortex; LinGy: Lingual gyrus; OFG Occipial fusiform gyrus; OFC: Orbital frontal cortex; OP: Occipial pole; PaCinGy: Paracingulate gyrus; ParOpC: Parietal operculum cortex; PHGy: Parahippocampal gyrus; PostGy: Postcentral gyrus; PP: Planum polare; PreGy: Precentral gyrus; PT: Planum temporale; Put: Putamen; SupCalC: Supercalcrine Cortex; SuppMA: Supplemental motor area; SupMargGy: Supramarginal gyrus; SupPL: Superior parietal lobule; SupTempGy: Superior temporal gyrus; InfFGyPO: Inferior frontal gyrus pars opercularis; MTGy: Middle temporal gyrus.

Loop	Day 1	Day 2	Day 3
–lSuppMA–rSuppMA–rPreGy–lPreGy–	x	x	x
–lOFG–lOP–rOP–rOFG–	x	x	x
–lSupTempGy ant–lPP–lHG–lPT–lSupTemGy post–	x	x	x
–lIC–rIC–rPP–lPP–	x	variant: –rIC–rPP–lPP–lHG–lCOC–lIC–	variant: –rIC–rPP–rHG–lPP–lIC–
–rIC–lIC–lPut–rPut–	x	x	x
–lIntCalC–lLinGy–lOFG–rOFG–rLinGy–rIntCalC–rSupCalC–lSupCalC–	x		variant: –lIntCalC–lLinGy–rLinGy–rIntCalC–
–lFP–lPaCinGy–rPaCinGy–rFP–		x	x
–rPP–lPP–lHG–lCOC–lIC–lFOC–lInfFGyPO–lInfFGyPT–lFP–LSuppMA–lPreGy–RPreGy–rPostGy–rSupMargGyAnt–rParOpC–rPT–rSupTempGy post–rSupTempGy ant–		x	variant: –rPP–rHG–rPT–rSupTempGy post–rSupTempGy ant–
–rOFG–lOFG–lLinGy–rLinGy–	x		
–lPaCinGy–rPaCinGy–rCinGy ant–lCinGy ant–		x	
–lIC–lCOC–lHG–lPP–		x	

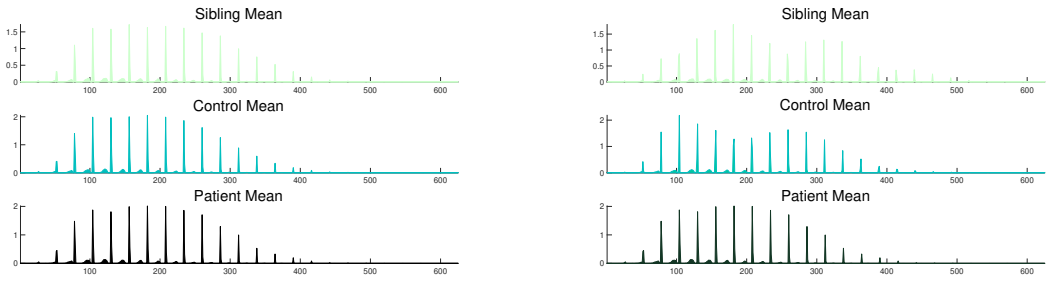
Table B.2: Loops that consist of edges that occur in loops of functional networks at least 50 times over all subjects (part II). We list the loops that we find in the left column. (We start with one of the nodes, which we choose arbitrarily, and end with the node that is adjacent to the starting node in the loop.) We denote an occurrence of a loop on a specific day with the symbol x in the table and present variations of the loop that we interpret as representing the same loop. We use the following abbreviations for the brain regions: l: left; r: right; ant: anterior; post: posterior; AnGy: Angular gyrus; CinGy: Cingulate gyrus; COC: Central opercular cortex; FOC: Frontal operculum cortex; FMedC: Frontal medial cortex; FP: Frontal pole; HG: Heschl's gyrus; IC: Insular cortex; InfFGyPT: Inferior frontal gyrus pars triangularis; IntCalC: Intracalcrine cortex; LinGy: Lingual gyrus; OFG Occipial fusiform gyrus; OFC: Orbital frontal cortex; OP: Occipial pole; PaCinGy: Paracingulate gyrus; ParOpC: Parietal operculum cortex; PHGy: Parahippocampal gyrus; PostGy: Postcentral gyrus; PP: Planum polare; PreGy: Precentral gyrus; PT: Planum temporale; Put: Putamen; SupCalC: Supercalcrine Cortex; SuppMA: Supplemental motor area; SupMargGy: Supramarginal gyrus; SupPL: Superior parietal lobule; SupTempGy: Superior temporal gyrus; InfFGyPO: Inferior frontal gyrus pars opercularis; MTGy: Middle temporal gyrus.

Loop	Day 1	Day 2	Day 3
–lFP–lFMedC–rFMedC–rFP–		x	
–lPHGy ant–lPHGy–rPHGy–rPHGy ant–		x	
–lInfFGyPT–lInfFGyPO–lFOC–lIC–lPP–lSupTempGy ant–lSupTemGy post–lMTGy post–lMTGy ant–lFP–lOFC–		x	
–rSupPL–rSupMargGy post–rSupMargGy ant–rPostGy–RPreGy–			x
–rPostGy–rSupPL–lSupPL–lSupMargGy ant–rSupMargGyAnt–			x
–lIntCalC–lLinGy–rLinGy–rIntCalC–			x
–lSupMargGy post–lAnGy–rAnGy–rSupMargGy post–rSupMargGy ant–lSupMargGy ant–			x
–lPT–lHG–lPP–lSupTempGy ant–lSupTemGy post–			x
Total number of loops	7	12	13

B.2.2 Application to schizophrenia data

B.2.2.1 Results from analysis of persistence image data set specific parameters by Tegan Emerson

In the left image of Fig. B.11, we see the mean vectorised persistence image for each subject group when we generate the persistence images using the maximum birth time and persistence across all subjects. (We create the mean vectorised persistence image for each subject group by taking the mean of each vector entry.) Observe that, other than a slight amplitude variation, the means look very similar. These mean persistence images are the mean of the vectorised persistence images for all samples from each group. The top row of Fig. B.12 contains the mean persistence images in image form when one selects the maximum birth and persistence across all subjects. On the right of Fig. B.11, we show the mean vectorised persistence image for each



(a) **(b)**

Figure B.11: (a) Mean vectorised persistence image — the horizontal axis corresponds to individual pixels in the persistence images, and the vertical axis indicates their intensity values — for each subject group generated using the maximum values of birth and persistence across all subjects to create all persistence images. We then take the means over the persistence images of each group. (b) Mean vectorised persistence image for each subject group generated using maximum values of birth and persistence determined by calculating the maximum birth and persistence for each of the three groups separately and using this group-specific information to create the persistence images for each subject within its group. We then take the means over the persistence images of each group. Image source: [238].

subject group, where we set the maximum birth and persistence values separately for each subject group (instead of setting the maximum birth and persistence values to be the same for all subjects). Observe that the sibling and control means both have two humps, whereas the patients have one that is clearly discernible. Similarly, in

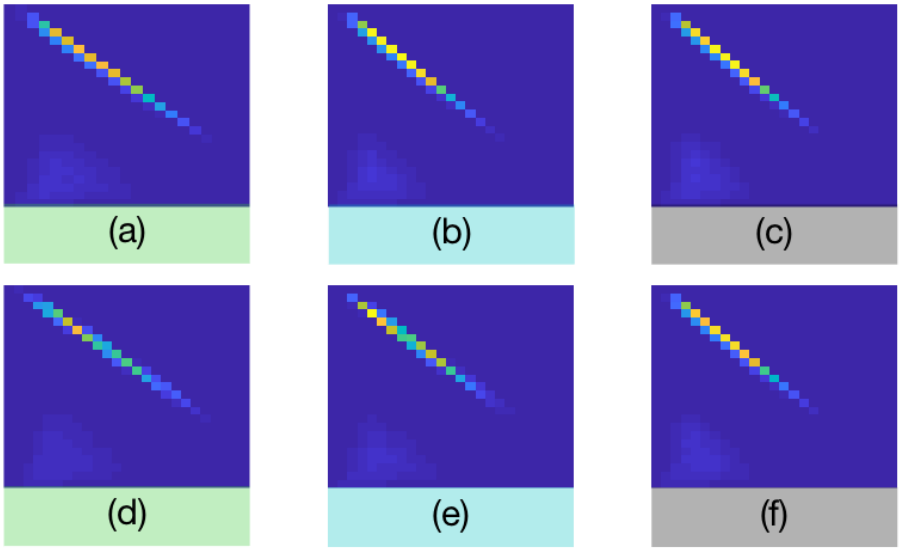
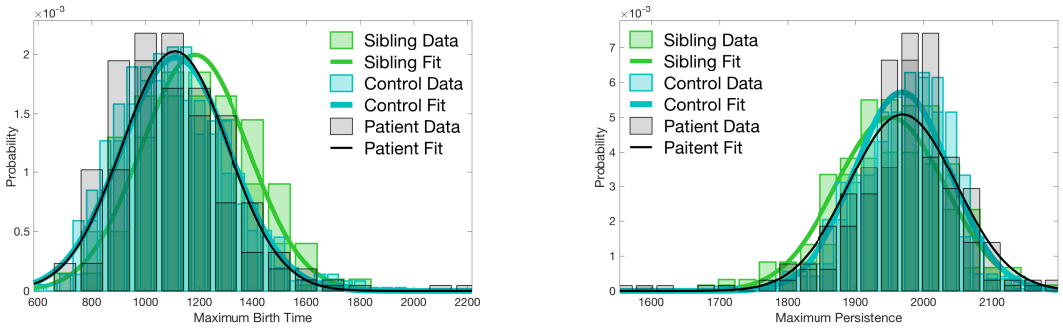


Figure B.12: Mean persistence image for each subject group. We generate panels (a)–(c) using the maximum values of birth and persistence determined across all subjects to compute all persistence images before creating the depersistence imagected means over the persistence images in each subject group. We generate panels (d)–(f) using maximum values of birth and persistence determined using maximum birth and persistence based on subject-group membership to compute all persistence images within each group before creating the depersistence imagected means over the persistence images in each group. The color axis is the same across rows. From left to right across each row are sibling, control, and patient averages. Image source: [238].

Fig. B.12, we observe two patches along the prominent diagonal with high intensity for the sibling and control means; however, in the bottom row, we only observe one clear (and elongated) hot spot for the patient mean. Therefore, there are multiple, smaller regions where loops often occur in the filtrations of the functional networks of siblings and controls, whereas there is seemingly a single, larger region of loops in the filtrations of the networks of the patients.

It is also worth noting the locations of the local maxima for each subject type. Relative to the maximum values across each class, groupings of loops occur at different locations. From the values of the vectors, we see that the controls and patients have more similar maximum magnitudes than do the patients and their siblings. Based on these similarities and differences, we conclude that we are able to accurately separate the populations using persistence images. Surprisingly, despite the pronounced difference in persistence image performance when we use different maximum values for

each class, the distributions of the maximum birth times and persistences for each subject type are not statistically-significantly different from each other. In Fig. B.13, we show Gaussian fits to the set of maximum birth times and maximum persistences for each subject type. Observe the strong similarity across all classes and the especially close similarity between the control and patient distributions. Because the maximum values are linked closely to the preprocessing of the data, it is important to conduct further research into how to account for these observations.



(a) **(b)**

Figure B.13: The distribution of (a) the maximum birth times across all samples for each subject type and (b) the maximum persistences across all samples for each subject type. Image source: [238].

B.2.2.2 Top brain regions in the distinguishing pixel birth–persistence bounds found by Tegan Emerson

In Fig. B.14, we show the distinguishing pixels from each of the three binary classifiers. We obtain the complete accuracy using the set of 41 unique pixels from the total of 625 pixels in the persistence images.

In Fig. B.15, we present the relative importances of different brain regions for each pixel. In the left panel, we show the top nodes for each subject type based on frequency (proportion of the subject type for which that top node is involved in the generation of a loop in the distinguishing-pixel region). In the right panel, we show the proportion of the subjects for which the top node(s) is (are) present. The vertical gaps in each plot signify that there are no nodes that are consistently involved in loops for that distinguishing pixel. We can make several observations from the left

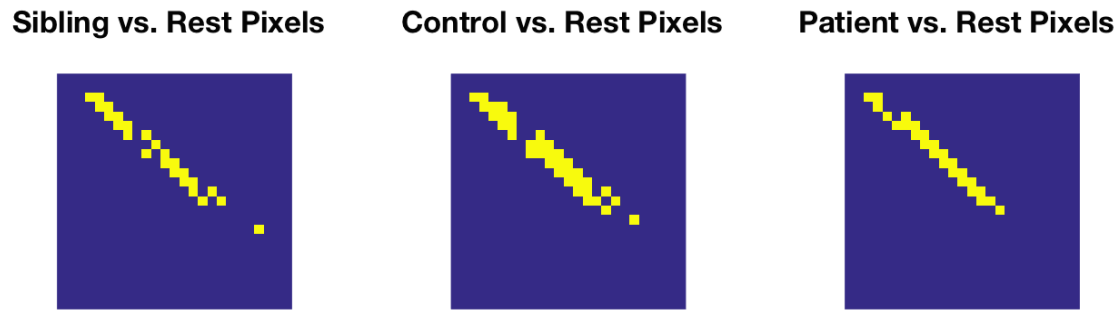
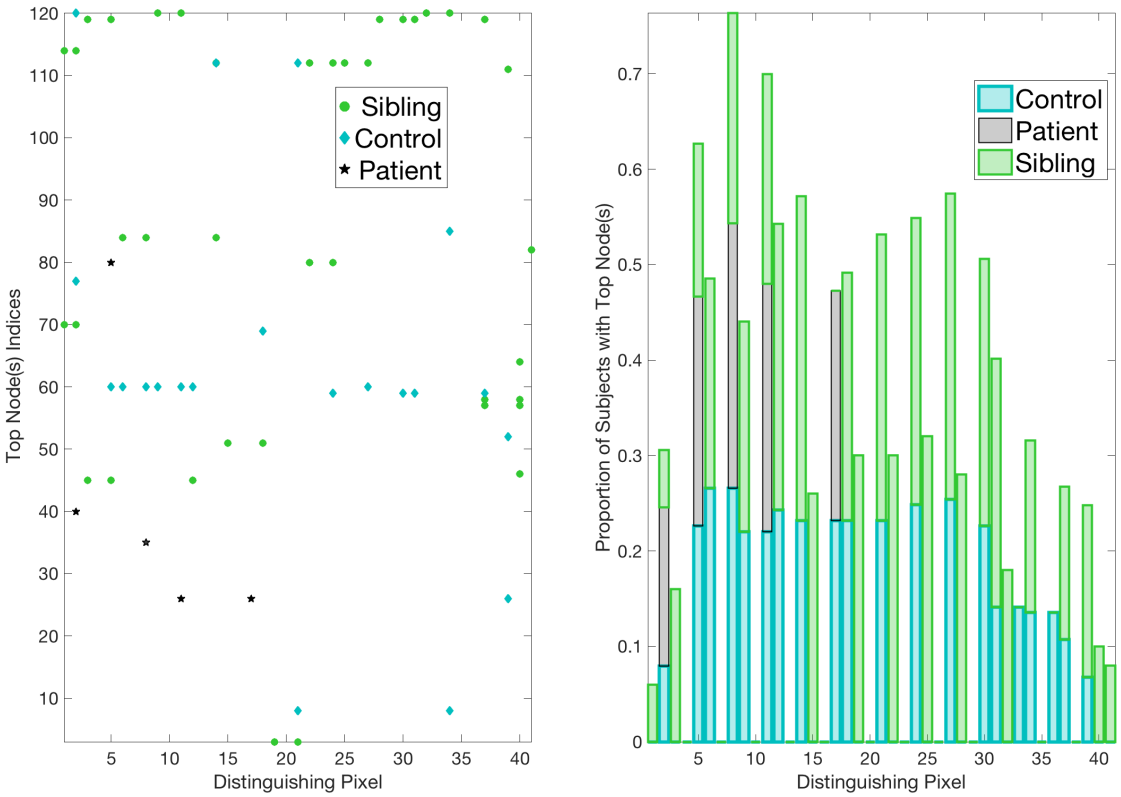


Figure B.14: The set of distinguishing pixels determined via SSVM as critical for obtaining 100% classification accuracy on the testing set. Image source: [238].

panel of Fig. B.15. First, there are only five distinguishing pixels for which we find



(a) **(b)**

Figure B.15: (a) The index (indices) of the top node(s) associated with each distinguishing pixel that we determine via SSVM. (b) A stacked bar graph of the proportion of each subject type with the corresponding node(s). Pale green indicates siblings, greenish blue indicates controls, and black indicates patients. Image source: [238].

top nodes for the patients. We are thus unable to predict which brain regions are involved in loops in the functional networks during the given task for schizophrenia patients. By contrast, there are many distinguishing pixels for which we find top nodes for the siblings. The control group lies between the other two in terms of its number of distinguishing pixels with top nodes, but there are still few top nodes, relative to the number of distinguishing pixels that have top nodes. In Tables B.3–B.6 and Figs. B.16–B.18, we indicate which brain regions (as well as their locations) we identify as top nodes. We include only the distinguishing pixels at which top nodes exist within a cohort.

An equivalent way to identify a top node is to calculate the percentage of a given subject class that has a topological feature in the corresponding pixel region (see the bar graph in Fig. B.15) and determine if a specific node is in the group of representatives for all of the subjects that have a topological feature in the pixel region. If a node occurs in the list of representatives for a topological feature for every subject of the class with a topological feature in the pixel region, then it is identified as a top node. Thus, when considering Tables B.3–B.6, it is possible to see the same brain regions listed for more than one distinguishing pixel index. This is also reflected in Fig. B.15 by the occurrence of multiple markers along the same horizontal line.

B.2.2.3 List of brain regions that represent nodes in the functional networks

In Tables B.7–B.11, we give the numbering of the brain regions and their corresponding IDs.

Table B.3: Top nodes involved in loop representatives over distinguishing pixel birth–persistence bounds (part I). We include only distinguishing pixels for which there is (are) top node(s) within a cohort. 'Left' and 'Right' refer to the hemispheres of the brain. We use the following abbreviations: superior frontal gyrus medial segment (MSFG), superior temporal gyrus (STG), opercular part of the inferior frontal gyrus (OpIFG), transverse temporal gyrus (TTG), frontal operculum (FO), gyrus rectus (GRe), middle frontal gyrus (MFG), orbital part of the inferior frontal gyrus (OrIFG), precuneus (PCu), cuneus (CC), anterior insula (AIns), superior parietal lobule (SPL), lingual gyrus (LiG), cerebellum exterior (CE), parahippocampal gyrus (PHG), medial frontal cortex (MFC), medial orbital gyrus (MOrG), and posterior cingulate gyrus (PCgG).

Pixel Indices	Siblings		Controls		Patients	
	Node	Location	Node	Location	Node	Location
1	70	Left MSFG	–	–	–	–
	114	Left STG	–	–	–	–
2	70	Left MSFG	77	Left OpIFG	40	Left FO
	114	Left STG	120	Left TTG	–	–
3	45	Right GRe	–	–	–	–
	119	Right TTG	–	–	–	–
5	45	Right GRe	60	Left MFG	80	Left OrIFG
	119	Right TTG	–	–	–	–
6	84	Left PCu	60	Left MFG	–	–
8	84	Left PCu	60	Left MFG	35	Right CC
9	120	Left TTG	60	Left MFG	–	–
11	120	Left TTG	60	Left MFG	26	Left AIns
12	45	Right GRe	60	Left MFG	–	–

Table B.4: Top nodes involved in loop representatives over distinguishing pixel birth–persistence bounds (part II). We include only distinguishing pixels for which there is (are) top node(s) within a cohort. 'Left' and 'Right' refer to the hemispheres of the brain. We use the following abbreviations: superior frontal gyrus medial segment (MSFG), superior temporal gyrus (STG), opercular part of the inferior frontal gyrus (OpIFG), transverse temporal gyrus (TTG), frontal operculum (FO), gyrus rectus (GRe), middle frontal gyrus (MFG), orbital part of the inferior frontal gyrus (OrIFG), precuneus (PCu), cuneus (CC), anterior insula (AIns), superior parietal lobule (SPL), lingual gyrus (LiG), cerebellum exterior (CE), parahippocampal gyrus (PHG), medial frontal cortex (MFC), medial orbital gyrus (MOrG), and posterior cingulate gyrus (PCgG).

Pixel Indices	Siblings		Controls		Patients	
	Node	Location	Node	Location	Node	Location
14	84	Left PCu	112	Left SPL	–	–
	112	Left SPL	–	–	–	–
15	51	Right LiG	–	–	–	–
17	–	–	112	Left SPL	26	Left AIns
18	51	Right LiG	69	Right MSFG	–	–
19	3	Right Amyg.	–	–	–	–
21	3	Right Amyg.	8	Left CE	–	–
	–	–	112	Left SPL	–	–
22	80	Left OrIFG	–	–	–	–
	112	Left SPL	–	–	–	–
24	80	Left OrIFG	59	Right MFG	–	–
	112	Left SPL	–	–	–	–
25	112	Left SPL	–	–	–	–

Table B.5: Top nodes involved in loop representatives over distinguishing pixel birth–persistence bounds (part III). We include only distinguishing pixels for which there is (are) top node(s) within a cohort. 'Left' and 'Right' refer to the hemispheres of the brain. We use the following abbreviations: superior frontal gyrus medial segment (MSFG), superior temporal gyrus (STG), opercular part of the inferior frontal gyrus (OpIFG), transverse temporal gyrus (TTG), frontal operculum (FO), gyrus rectus (GRe), middle frontal gyrus (MFG), orbital part of the inferior frontal gyrus (OrIFG), precuneus (PCu), cuneus (CC), anterior insula (AIns), superior parietal lobule (SPL), lingual gyrus (LiG), cerebellum exterior (CE), parahippocampal gyrus (PHG), medial frontal cortex (MFC), medial orbital gyrus (MOrG), and posterior cingulate gyrus (PCgG).

Pixel Indices	Siblings		Controls		Patients	
	Node	Location	Node	Location	Node	Location
27	112	Left SPL	60	Left MFG	–	–
28	119	Right TTG	–	–	–	–
30	119	Right TTG	59	Right MFG	–	–
31	119	Right TTG	59	Right MFG	–	–
32	120	Left TTG	–	–	–	–
33	–	–	59	Right MFG	–	–
34	120	Left TTG	8	Left CE	–	–
	–	–	85	Right PHG	–	–
36	–	–	8	Left CE	–	–
	–	–	85	Right PHG	–	–

Table B.6: Top nodes involved in loop representatives over distinguishing pixel birth–persistence bounds (part IV). We include only distinguishing pixels for which there is (are) top node(s) within a cohort. 'Left' and 'Right' refer to the hemispheres of the brain. We use the following abbreviations: superior frontal gyrus medial segment (MSFG), superior temporal gyrus (STG), opercular part of the inferior frontal gyrus (OpIFG), transverse temporal gyrus (TTG), frontal operculum (FO), gyrus rectus (GRe), middle frontal gyrus (MFG), orbital part of the inferior frontal gyrus (OrIFG), precuneus (PCu), cuneus (CC), anterior insula (AIns), superior parietal lobule (SPL), lingual gyrus (LiG), cerebellum exterior (CE), parahippocampal gyrus (PHG), medial frontal cortex (MFC), medial orbital gyrus (MOrG), and posterior cingulate gyrus (PCgG).

Pixel Indices	Siblings		Controls		Patients	
	Node	Location	Node	Location	Node	Location
37	57	Right MFC	59	Right MFG	–	–
	58	Left MFC	–	–	–	–
	119	Right TTG	–	–	–	–
39	111	Right SPL	26	Left AIns	–	–
	–	–	52	Left LiG	–	–
40	46	Left GRe	–	–	–	–
	57	Right MFC	–	–	–	–
	58	Left MFC	–	–	–	–
	64	Left MOrG	–	–	–	–
41	82	Left PCgG	–	–	–	–

Table B.7: Node numbers (NNs) of brain regions (BRs) and their corresponding IDs (part I).

NN	1	2	3	4	5	6
ID	23	30	31	32	36	37
BR	Right accumbens area	Left accumbens area	Right amygdala	Left amygdala	Right caudate	Left caudate
NN	7	8	9	10	11	12
ID	38	39	47	48	55	56
BR	Right cerebellum exterior	Left cerebellum exterior	Right hippocampus	Left hippocampus	Right pallidum	Left pallidum
NN	13	14	15	16	17	18
ID	57	58	59	60	61	62
BR	Right putamen	Left putamen	Right thalamus proper	Left thalamus proper	Right ventral diencephalon	Left ventral diencephalon
NN	19	20	21	22	23	24
ID	71	72	75	76	100	101
BR	Cerebellar vermal lobules I–V	Cerebellar vermal lobules VI–VII	Left basal forebrain	Right basal forebrain	Right anterior cingulate gyrus	Left anterior cingulate gyrus
NN	25	26	27	28	29	30
ID	102	103	104	105	106	107
BR	Right anterior insula	Left anterior insula	Right anterior orbital gyrus	Left anterior orbital gyrus	Right angular gyrus	Left angular gyrus

Table B.8: Node numbers (NNs) of brain regions (BRs) and their corresponding IDs (part II).

NN	31	32	33	34	35	36
ID	108	109	112	113	114	115
BR	Right calcarine cortex	Left calcarine cortex	Right central operculum	Left central operculum	Right cuneus	Left cuneus
NN	37	38	39	40	41	42
ID	116	117	118	119	120	121
BR	Right entorhinal area	Left entorhinal area	Right frontal operculum	Left frontal operculum	Right frontal pole	Left frontal pole
NN	43	44	45	46	47	48
ID	122	123	124	125	128	129
BR	Right fusiform gyrus	Left fusiform gyrus	Right gyrus rectus	Left gyrus rectus	Right inferior occipital gyrus	Left inferior occipital gyrus
NN	49	50	51	52	53	54
ID	132	133	134	135	136	137
BR	Right inferior temporal gyrus	Left inferior temporal gyrus	Right lingual gyrus	Left lingual gyrus	Right lateral orbital gyrus	Left lateral orbital gyrus

Table B.9: Node numbers (NNs) of brain regions (BRs) and their corresponding IDs (part III).

NN	55	56	57	58	59	60
ID	138	139	140	141	142	143
BR	Right middle cingulate gyrus	Left middle cingulate gyrus	Right medial frontal cortex	Left medial frontal cortex	Right middle frontal gyrus	Left middle frontal gyrus
NN	61	62	63	64	65	66
ID	144	145	146	147	148	149
BR	Right middle occipital gyrus	Left middle occipital gyrus	Right medial orbital gyrus	Left medial orbital gyrus	Right postcentral gyrus medial segment	Left postcentral gyrus medial segment
NN	67	68	69	70	71	72
ID	150	151	152	153	154	155
BR	Right precentral gyrus medial segment	Left precentral gyrus medial segment	Right superior frontal gyrus medial segment	Left superior frontal gyrus medial segment	Right middle temporal gyrus	Left middle temporal gyrus
NN	73	74	75	76	77	78
ID	156	157	160	161	162	163
BR	Right occipital pole	Left occipital pole	Right occipital fusiform gyrus	Left occipital fusiform gyrus	Right opercular part of the inferior frontal gyrus	Left opercular part of the inferior frontal gyrus

Table B.10: Node numbers (NNs) of brain regions (BRs) and their corresponding IDs (part IV).

NN	79	80	81	82	83	84
ID	164	165	166	167	168	169
BR	Right orbital part of the inferior frontal gyrus	Left orbital part of the inferior frontal gyrus	Right posterior cingulate gyrus	Left posterior cingulate gyrus	Right precuneus	Left precuneus
NN	85	86	87	88	89	90
ID	170	171	172	173	174	175
BR	Right parahippocampal gyrus	Left parahippocampal gyrus	Right posterior insula	Left posterior insula	Right parietal operculum	Left parietal operculum
NN	91	92	93	94	95	96
ID	176	177	178	179	180	181
BR	Right postcentral gyrus	Left postcentral gyrus	Right posterior orbital gyrus	Left posterior orbital gyrus	Right planum polare	Left planum polare
NN	97	98	99	100	101	102
ID	182	183	184	185	186	187
BR	Right precentral gyrus	Left precentral gyrus	Right planum temporale	Left planum temporale	Right subcallosal area	Left subcallosal area

Table B.11: Node numbers (NNs) of brain regions (BRs) and their corresponding IDs (part V).

NN	103	104	105	106	107	108
ID	190	191	192	193	194	195
BR	Right superior frontal gyrus	Left superior frontal gyrus	Right supplementary motor cortex	Left supplementary motor cortex	Right supramarginal gyrus	Left supramarginal gyrus
NN	109	110	111	112	113	114
ID	196	197	198	199	200	201
BR	Right superior occipital gyrus	Left superior occipital gyrus	Right superior parietal lobule	Left superior parietal lobule	Right superior temporal gyrus	Left superior temporal gyrus
NN	115	116	117	118	119	120
ID	202	203	204	205	206	207
BR	Right temporal pole	Left temporal pole	Right triangular part of the inferior frontal gyrus	Left triangular part of the inferior frontal gyrus	Right transverse temporal gyrus	Left transverse temporal gyrus

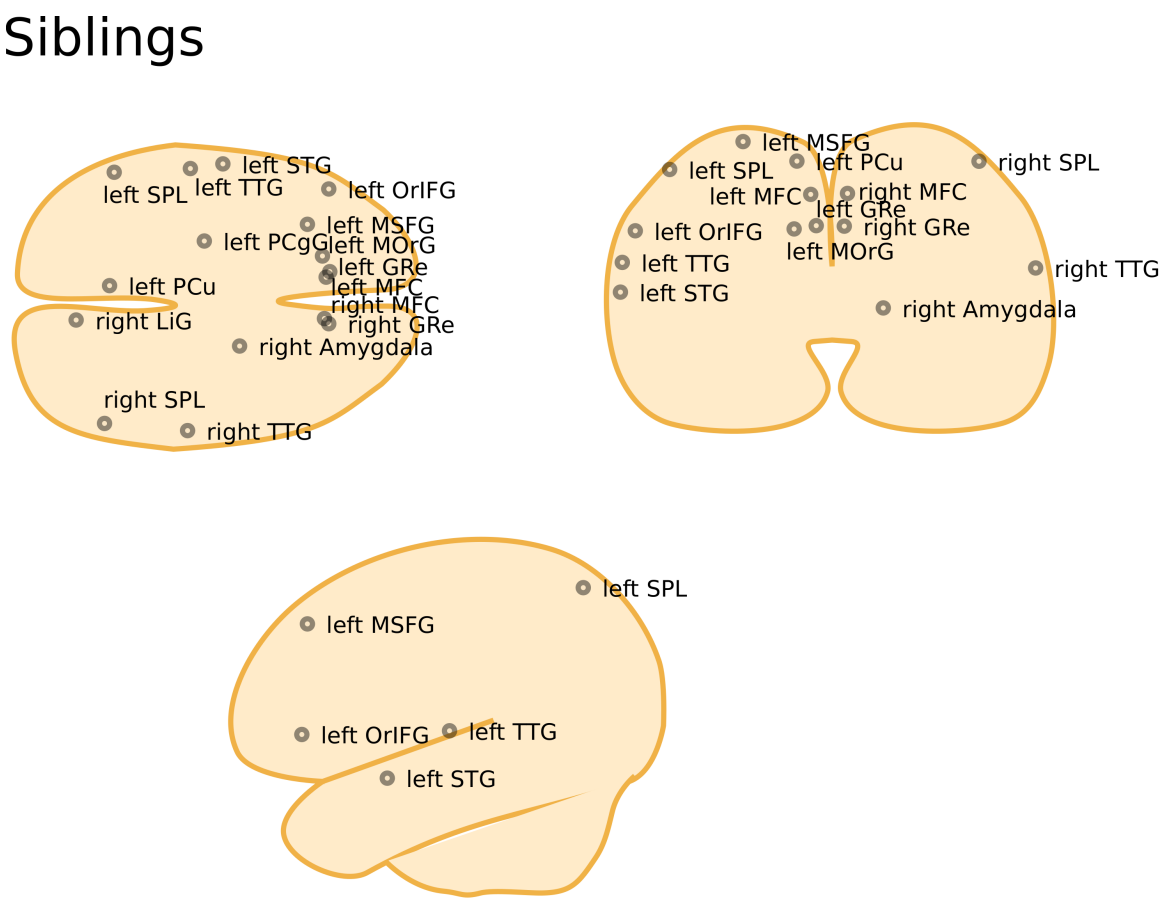


Figure B.16: Top nodes in representatives of loops in the distinguishing pixel birth–persistence bounds for siblings. Image source: [238].

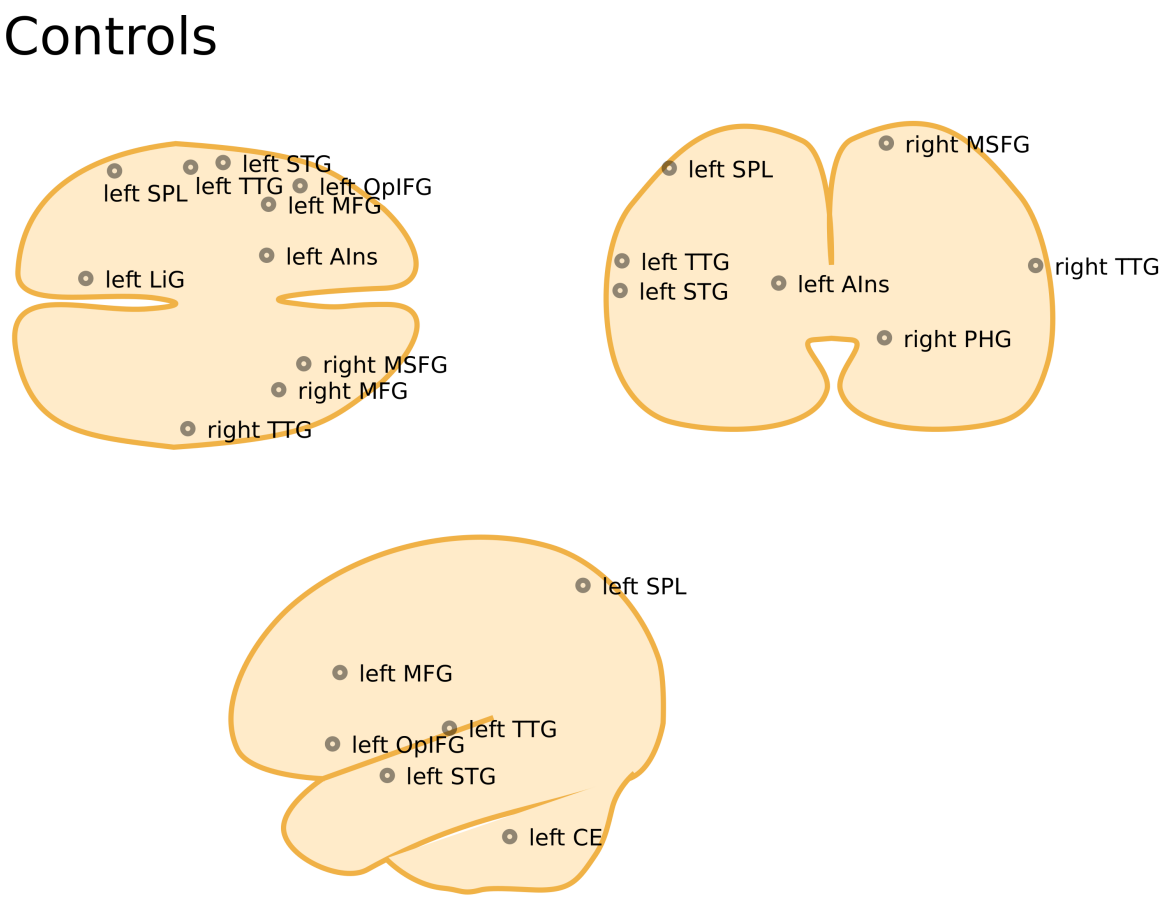


Figure B.17: Top nodes in representatives of loops in the distinguishing pixel birth–persistence bounds for controls. Image source: [238].

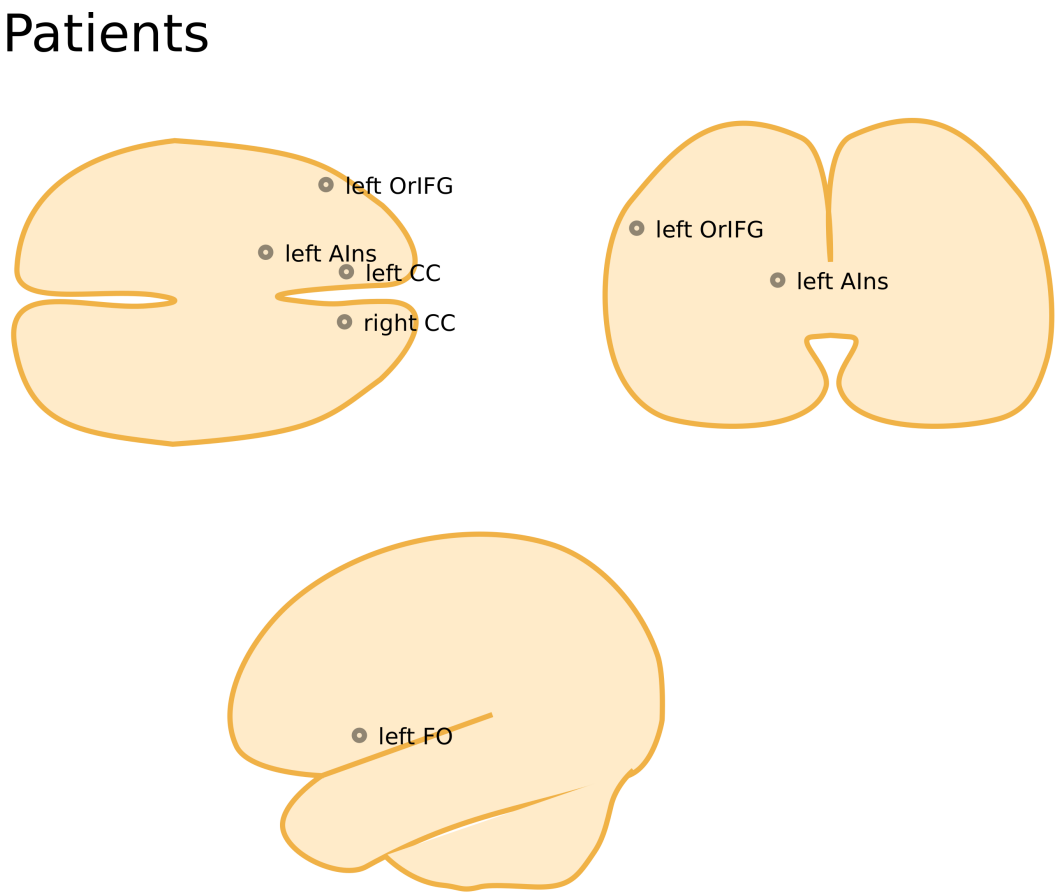


Figure B.18: Top nodes in representatives of loops in the distinguishing pixel birth–persistence bounds for patients. Image source: [238].

Bibliography

[1] Henry Adams and Gunnar Carlsson. On the nonlinear statistics of range image patches. *SIAM Journal on Imaging Sciences*, 2(1):110–117, 2009.

[2] Henry Adams, Sofya Chepushtanova, Tegan Emerson, Eric Hanson, Michael Kirby, Francis Motta, Rachel Neville, Chris Peterson, Patrick Shipman, and Lori Ziegelmeier. Persistence images. `https://github.com/CSU-TDA/PersistenceImages`, 2016.

[3] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *The Journal of Machine Learning Research*, 18(1):218–252, 2017.

[4] Henry Adams and Andrew Tausz. JavaPlex tutorial. Available at `http://javaplex.googlecode.com/svn/trunk/reports/javaplex_tutorial/javaplex_tutorial.pdf`, 2015.

[5] Mahmuda Ahmed, Brittany Terese Fasy, and Carola Wenk. Local persistent homology based distance between maps. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 43–52. ACM, 2014.

[6] Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Essential Cell Biology*. Garland Science, New York and London, 2014.

[7] Aaron F. Alexander-Bloch, Nitin Gogtay, David Meunier, Rasmus Birn, Liv Clasen, Francois Lalonde, Rhoshel K. Lenroot, Jay N. Giedd, and Edward T. Bullmore. Disrupted modularity and local connectivity of brain functional networks in childhood-onset schizophrenia. *Frontiers in Systems Neuroscience*, 4(147):1–16, 2010.

[8] Aaron F. Alexander-Bloch, Renaud Lambiotte, Ben Roberts, Jay Giedd, Nitin Gogtay, and Edward T. Bullmore. The discovery of population differences in network community structure: New methods and applications to brain functional networks in schizophrenia. *NeuroImage*, 15:3889–3900, 2012.

[9] Alexander R. A. Anderson and Mark A. J. Chaplain. Continuous and discrete mathematical models of tumor-induced angiogenesis. *Bulletin of Mathematical Biology*, 60(5):857–899, 1998.

[10] Ariana Anderson and Mark S. Cohen. Decreased small-world functional network connectivity and clustering across resting state networks in schizophrenia: An fMRI classification tutorial. *Frontiers in Human Neuroscience*, 7:1–18, 2013.

[11] Alex Arenas, Albert Díaz-Guilera, Jurgen Kurths, Yamir Moreno, and Changsong Zhou. Synchronization in complex networks. *Physics Reports*, 469(3):93–153, 2008.

[12] Alex Arenas, Albert Díaz-Guilera, and Conrad Pérez-Vicente. Synchronization reveals topological scales in complex networks. *Physical Review Letters*, 96(11):114102, 2006.

[13] Peter Ashwin, Stephen Coombes, and Rachel Nicks. Mathematical framework for oscillatory network dynamics in neuroscience. *The Journal of Mathematical Neuroscience*, 6(2):1–92, 2016.

[14] Andrey Babichev and Yuri Dabaghian. Persistent memories in transient networks. In *Emergent Complexity from Nonlinearity, in Physics, Engineering and the Life Sciences*, pages 179–188. Springer, Cham, 2017.

[15] Andrey Babichev and Yuri Dabaghian. Persistent memories in transient networks. In Giorgio Mantica, Ruedi Stoop, and Sebastiano Stramaglia, editors, *Emergent Complexity from Nonlinearity, in Physics, Engineering and the Life Sciences. Proceedings of the XXIII International Conference on Nonlinear Dynamics of Electronic Systems, Como, Italy, 7 – 11 September 2015*, volume 191 of *Springer Proceedings in Physics*, pages 179–188. Springer Nature, Cham, 2017.

[16] Andrey Babichev, Dmitriy Morozov, and Yuri Dabaghian. Replays of spatial memories suppress topological fluctuations in cognitive map. *Network Neuroscience*, 3(3):707 – 724, 2018.

[17] James W. Baish, Yuval Gazit, David A. Berk, Mutsumi Nozue, Laurence T. Baxter, and Rakesh K. Jain. Role of tumor vascular architecture in nutrient and drug delivery: an invasion percolation-based network model. *Microvascular research*, 51(3):327–346, 1996.

[18] Jean-Baptiste Bardin, Gard Spreemann, and Kathryn Hess. Topological exploration of artificial neuronal network dynamics. *Network Neuroscience*, 3(3):725 — 743, 2018.

[19] Danielle S. Bassett, Edward T. Bullmore, Beth A. Verchinski, Venkata S. Mattay, Daniel R. Weinberger, and Andreas Meyer-Lindenberg. Hierarchical orga-

nization of human cortical networks in health and schizophrenia. *The Journal of Neuroscience*, 28(37):9239–9248, 2008.

[20] Danielle S. Bassett, Brent G. Nelson, Bryon A. Mueller, Jazmin Camchong, and Kelvin O. Lim. Altered resting state complexity in schizophrenia. *NeuroImage*, 59(3):2196–2207, 2012.

[21] Danielle S. Bassett, Mason A. Porter, Nicholas F. Wymbs, Scott T. Grafton, Jean M. Carlson, and Peter J. Mucha. Robust detection of dynamic community structure in networks. *Chaos*, 23:013142, 2013.

[22] Danielle S. Bassett and Olaf Sporns. Network neuroscience. *Nature Neuroscience*, 20(3):353–364, 2017.

[23] Danielle S. Bassett, Nicholas F. Wymbs, Mason A. Porter, Peter J. Mucha, Jean M. Carlson, and Scott T. Grafton. Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences of the United States of America*, 108(18):7641–7646, 2011.

[24] Danielle S. Bassett, Nicholas F. Wymbs, Mason A. Porter, Peter J. Mucha, and Scott T. Grafton. Cross-linked structure of network evolution. *Chaos*, 24(1):013112, 2014.

[25] Danielle S. Bassett, Nicholas F. Wymbs, M. Puck Rombach, Mason A. Porter, Peter J. Mucha, and Scott T. Grafton. Task-based core–periphery organization of human brain dynamics. *PLoS Computational Biology*, 10(4):e1003171, 2013.

[26] Danielle S. Bassett, Muzhi Yang, Nicholas F. Wymbs, and Scott T. Grafton. Learning-induced autonomy of sensorimotor systems. *Nature Neuroscience*, 18:744–751, 2015.

[27] Danielle S. Bassett, Perry Zurn, and Joshua I. Gold. On the nature and use of models in network neuroscience. *Nature Reviews Neuroscience*, 2018. Available at `https://doi.org/10.1038/s41583-018-0038-8`.

[28] Russell Bates. *Learning to Extract Tumour Vasculature: Techniques in Machine Learning for Medical Image Analysis*. PhD thesis, University of Oxford, 2017.

[29] Russell Bates. Russ-learn: set of tools for application and training of deep learning methods for image segmentation and vessel analysis. Software available at `https://ibme-gitcvs.eng.ox.ac.uk/RussellB/unet-test`, software retrieved in 2018.

[30] Russell Bates, Benjamin Irving, Bostjan Markelc, Jakob Kaeppler, Graham Brown, Ruth J. Muschel, Michael Brady, Vicente Grau, and Julia A. Schnabel. Segmentation of vasculature from fluorescently labeled endothelial cells in multi-photon microscopy images. *IEEE transactions on medical imaging*, 38(1):1 –10, 2019.

[31] Russell Bates, Benjamin Irving, Bostjan Markelc, Jakob Kaeppler, Ruth Muschel, Vicente Grau, and Julia A. Schnabel. Extracting 3D vascular structures from microscopy images using convolutional recurrent networks. arXiv:1705.09597, 2017.

[32] Ulrich Bauer. Ripser: a lean C++ code for the computation of Vietoris–Rips persistence barcodes. Software available at `https://github.com/Ripser/ripser`, software retrieved in 2017.

[33] Francisco Belchi, Mariam Pirashvili, Joy Conway, Michael Bennett, Ratko Djukanovic, and Jacek Brodzki. Lung topology characteristics in patients with chronic obstructive pulmonary disease. *Scientific Reports*, 8(1):5341, 2018.

[34] Paul Bendich, James S. Marron, Ezra Miller, Alex Pieloch, and Sean Skwerer. Persistent homology analysis of brain artery trees. *Annals of Applied Statistics*, 10(1):198–218, 2016.

[35] Paul Bendich, Bei Wang, and Sayan Mukherjee. Local homology transfer and stratification learning. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1355–1370. SIAM, 2012.

[36] Austin R. Benson, David F. Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353:163–166, 2016.

[37] Gabriele Bergers and Douglas Hanahan. Modes of resistance to anti-angiogenic therapy. *Nature Reviews Cancer*, 8:592–603, 2008.

[38] Miguel O. Bernabeu, Jakub Köry, James A. Grogan, Bostjan Markelc, Albert Beardo Ricol, Mayeul d'Avezac, Jakob Kaeppler, Nicholas Daly, James Hetherington, Timm Krüger, Philip K. Maini, Joe M. Pitt-Francis, Ruth J. Muschel, Tomás Alarcón, and Helen M. Byrne. Abnormal morphology biases haematocrit distribution in tumour vasculature and contributes to heterogeneity in tissue oxygenation. BioRxiv: 640060v1, 2019.

[39] Alessandro Bertolino and Giuseppe Blasi. The genetics of schizophrenia. *Neuroscience*, 164:288–299, 2009.

[40] Alessandro Bertolino, Leonardo Fazio, Annabella Di Giorgio, Giuseppe Blasi, Raffaella Romano, Paolo Taurisano, Grazia Caforio, Lorenzo Sinibaldi, Gianluca Ursini, Teresa Popoloizo, Emanuele Tirotta, Audrey Papp, Bruno Dallapiccola, Emiliana Borrelli, and Wolfgang Sadee. Genetically determined interaction between the dopamine transporter and the D2 receptor on prefronto-striatal activity and volume in humans. *The Journal of Neuroscience*, 29(4):1224–1234, 2009.

[41] Alessandro Bertolino, Paolo Taurisano, Nicola Marco Pisciotta, Giuseppe Blasi, Leonardo Fazio, Raffaella Romano, Barbara Gelao, Luciana Lo Biancho, Madia Lozupone, Annabella Di Giorgio, Grazia, Fabio Sambataro, Artor Niccoli-Asabella, Audrey Papp, Gianluca Ursini, Lorenzo Sinibaldi, Teresa Popoloizo, Wolfgang Sadee, and Giuseppe Rubini. Genetically determined measures of striatal D2 signalling predict prefrontal activity during working memory performance. *PloS ONE*, 5(2):e9348, 2010.

[42] Richard F. Betzel and Danielle S. Bassett. Multi-scale brain networks. *NeuroImage*, 160:73–83, 2017.

[43] Dhananjay Bhaskar, Angelika Manhart, Jesse Milzman, John T. Nardini, Kathleen Storey, Chad M. Topaz, and Lori Ziegelmeier. Analyzing collective motion with machine learning and topology. arXiv:1908.09081, 2019.

[44] Subhrait Bhattacharya, Robert Ghrist, and Vijay Kumar. Persistent homology for path planning in uncertain environments. *IEEE Transactions on Robotics*, 31(3):578–590, 2015.

[45] Rasmus M. Birn, Jason B. Diamond, Monica A. Smith, and Peter A. Bandettini. Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. *NeuroImage*, 31(4):1536–1548, 2006.

[46] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[47] Béla Bollobás. *Modern Graph Theory*. Springer, New York, 1998.

[48] Urs Braun, Axel Schäfer, Danielle S. Bassett, Franziska Rausch, Janina Schweiger, Edda Bilek, Susanne Erk, Nina Romanczuk-Seiferth, Oliver Grimm,

Leila Haddad, Kristina Otto, Sebastian Mohnke, Andreas Heinz, Mathias Zink, Henrik Walter, Andreas Meyer-Lindenberg, and Heike Tost. Dynamic reconfiguration of brain networks: A potential schizophrenia genetic risk mechanism modulated by nmda receptor function. *Proceedings of the National Academy of Sciences of the United States of America*, 113(44):12568–12573, 2016.

[49] Michael Breakspear, Stewart Heitmann, and Andreas Daffertshofer. Generative models of cortical oscillations: Neurobiological implications of the Kuramoto model. *Frontiers in Human Neuroscience*, 4(190):1 – 14, 2010.

[50] Paul Breiding, Sara Kališnik, Bernd Sturmfels, and Madeleine Weinstein. Learning algebraic varieties from samples. *Revista Matemática Complutense*, 31(3):545–593, 2018.

[51] Paul Breiding and Orlando Marigliano. Sampling from the uniform distribution on an algebraic manifold. arXiv:1810.06271, 2018.

[52] Coen Bron and Joep Kerbosch. Algorithm 457: Finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577, 1973.

[53] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16:77–102, 2015.

[54] Peter Bubenik. *Personal Communication*, 2016.

[55] Peter Bubenik and Paweł Dłotko. A persistence landscapes toolbox for topological statistics. *Journal of Symbolic Computation*, 78:91–114, 2017.

[56] Peter Bubenik, Michael Hull, Dhruv Patel, and Benjamin Whittle. Persistent homology detects curvature. arXiv:1905.13196, 2019.

[57] Elizabeth Bullitt, Guido Gerig, Stephen M. Pizer, Weili Lin, and Stephen R. Aylward. Measuring tortuosity of the intracerebral vasculature from MRA images. *IEEE Transactions on Medical Imaging*, 22(9):1163–1171, 2003.

[58] Elizabeth Bullitt, Donglin Zeng, Guido Gerig, Stephen Aylward, Sarang Joshi, J. Keith Smith, Weili Lin, and Matthew G. Ewend. Vessel tortuosity and brain tumor malignancy: a blinded study. *Academic Radiology*, 12(10):1232–1240, 2005.

[59] Edward T. Bullmore and Danielle Bassett. Brain graphs: Graphical models of the human brain connectome. *Annual Review of Clinical Psychology*, 7:113–140, 2011.

[60] Edward T. Bullmore, S. Frangou, and Richard M. Murray. The dysplastic net hypothesis: An integration of developmental and dysconnectivity theories of schizophrenia. *Schizophrenia Research*, 28(2–3):143–156, 1997.

[61] Edward T. Bullmore and Olaf Sporns. Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews*, 10:186–198, 2009.

[62] Edward T. Bullmore and Olaf Sporns. The economy of brain network organization. *Nature Reviews Neuroscience*, 13:336–349, 2012.

[63] Helen M. Byrne and Mark A. J. Chaplain. Mathematical models for tumour angiogenesis: numerical simulations and nonlinear wave solutions. *Bulletin of Mathematical Biology*, 57(3):461–486, 1995.

[64] Helen M. Byrne, Heather A. Harrington, Ruth Muschel, Gesine Reinert, Bernadette J. Stolz, and Ulrike Tillmann. Topology characterises tumour vasculature. *Mathematics Today*, 55(5):206 – 210, 2019 (in press).

[65] Joseph H. Callicott, Michael F. Egan, Venkata S. Mattay, Alessandro Bertolino, Ashley D. Bone, Beth Verchinksi, and Daniel R. Weinberger. Abnormal fMRI response of the dorsolateral prefrontal cortex in cognitively intact siblings of patients with schizophrenia. *American Journal of Psychiatry*, 160(4):709–719, 2003.

[66] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46:255–308, 2009.

[67] Gunnar Carlsson, Tigran Ishkhanov, Vin De Silva, and Afra Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76(1):1–12, 2008.

[68] Gunnar Carlsson, Afra Zomorodian, Anne Collins, and Leonidas J. Guibas. Persistence barcodes for shapes. *International Journal of Shape Modeling*, 11(2):149–187, 2005.

[69] Mathieu Carriere, Marco Cuturi, and Steve Oudot. Sliced wasserstein kernel for persistence diagrams. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 664–673, 2017.

[70] Sainsbury Wellcome Centre. MATLAB toolbox for analysis of output from the software aMAP (optimized automated mouse atlas propagation). Toolbox available at `www.gatsby.ucl.ac.uk/~test/matlabTools.zip`. See `https://github.com/SainsburyWellcomeCentre/aMAP/wiki` for description, software retrieved in 2019.

[71] Sanjay Chawla and Aristides Gionis. k-means--: A unified approach to clustering and outlier detection. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 189–197. SIAM, 2013.

[72] Elvis Chen. Laplacian random number generator. Available at: https://www.mathworks.com/matlabcentral/fileexchange/13705-laplacian-random-number-generator, March 2019.

[73] Fan R. K. Chung. *Spectral Graph Theory*. Number 92 in Regional Conference Series in Mathematics. AMS and CBMS, 1997.

[74] Moo K. Chung, Peter Bubenik, and Peter T. Kim. Persistence diagrams of cortical surface data. In J.L. Prince, D.L. Pham, and K.J. Myers, editors, *Information Processing in Medical Imaging. IMPI 2009*, volume 5636 of *Lecture Notes in Computer Science*, pages 386–397. Springer, Berlin, Heidelberg, 2009.

[75] Moo K. Chung, Hyekyoung Lee, Hernando Ombao, and Victor Solo. Exact topological inference of the resting-state brain networks in twins. *Network Neuroscience*, 3(3):674 – 694, 2019.

[76] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete and Computational Geometry*, 37(1):103–120, 2007.

[77] Alex Cole and Gary Shiu. Persistent homology and non-gaussianity. *Journal of Cosmology and Astroparticle Physics*, 2018(03):025, 2018.

[78] Guusje Collin, René S. Kahn, Marcel A. de Reus, Wiepke Cahn, and Martijn P. van den Heuvel. Impaired rich club connectivity in unaffected siblings of schizophrenia patients. *Schizophrenia Bulletin*, 40(2):438–448, 2014.

[79] Fred H. Croom. *Basic Concepts of Algebraic Topology*. Springer, New York, Heidelberg, Berlin, 1978.

[80] Carina Curto. What can topology tell us about the neural code? *Bulletin of the American Mathematical Society*, 54(1):63–78, 2017.

[81] Carina Curto and Vladimir Itskov. Cell groups reveal structure of stimulus space. *PLoS Computational Biology*, 4(10):e000205, 2008.

[82] Yuri Dabaghian, Facundo Mémoli, L. Frank, and Gunnar E. Carlsson. A topological paradigm for hippocampal spatial map formation using persistent homology. *PLoS ONE*, 8(8):e1002581, 2012.

[83] Mandeep S. Dagli, John E. Ingeholm, and James V. Haxby. Localization of cardiac-induced signal change in fMRI. *NeuroImage*, 9(4):407–415, 1999.

[84] Zachary Danziger. Hausdorff distance. Code available at `https://www.mathworks.com/matlabcentral/fileexchange/26738-hausdorff-distance`, code retrieved in August 2019.

[85] Neil Dawson, Xiaolin Xiao, Martin McDonald, Desmond J. Higham, Brian J. Morris, and Judith A. Pratt. Sustained NMDA receptor hypofunction induces compromised neural systems integration and schizophrenia-like alterations in functional brain networks. *Cerebral Cortex*, 24:452–464, 2014.

[86] Vin de Silva and Gunnar Carlsson. Topological estimation using witness complexes. In Markus Gross, Hanspeter Pfister, Marc Alexa, and Szymon Rusinkiewicz, editors, *SPBG'04 Symposium on Point - Based Graphics 2004*, pages 157–166. The Eurographics Association, 2004.

[87] A. Stan Deakin. Model for initial vascular patterns in melanoma transplants. *Growth*, 40(2):191–201, 1976.

[88] Ruud P. M. Dings, Melissa Loren, Hanke Heun, Elizabeth McNiel, Arjan W. Griffioen, Kevin H. Mayo, and Robert J. Griffin. Scheduling of radiation with angiogenesis inhibitors anginex and avastin improves therapeutic outcome via vessel normalization. *Clinical Cancer Research*, 13(11):3395–3402, 2007.

[89] Paweł Dłotko. The persistence landscape toolbox. Software available at `https://www.math.upenn.edu/~dlotko/persistenceLandscape.html`, software retrieved in 2015.

[90] Paweł Dłotko, Kathryn Hess, Ran Lavi, Max Nolte, Michael Reimann, Martina Scholamiero, Katharine Turner, Eilif Muller, and Henry Markram. Topological analysis of the connectome of digital reconstructions of neural microcircuits. arXiv:1601.01580, 2016.

[91] Paweł Dłotko and Thomas Wanner. Topological microstructure analysis using persistence landscapes. *Physica D: Nonlinear Phenomena*, 334:60 – 81, 2016.

[92] Michael Dobosz, Vasilis Ntziachristos, Werner Scheuer, and Steffen Strobel. Multispectral fluorescence ultramicroscopy: Three-dimensional visualization and automatic quantification of tumour morphology, drug penetration, and antiangiogenic treatment response. *Neoplasia*, 16(1):1 – 13, 2014.

[93] Hans-Ulrich Dodt, Ulrich Leischner, Anja Schierloh, Nina Jährling, Christoph Peter Mauch, Katrin Deininger, Jan Michael Deussing, Matthias Eder, Walter Zieglgänsberger, and Klaus Becker. Ultramicroscopy: three-dimensional visualization of neuronal networks in the whole mouse brain. *Nature methods*, 4(4):331–336, 2007.

[94] Jonathan F. Donges, Yong Zou, Norbert Marwan, and Jürgen Kurths. The backbone of the climate network. *Europhysics Letters (EPL)*, 87:48007, 2009.

[95] Andrew C. Dudley. Tumor endothelial cells. *Cold Spring Harbor Perspectives in Medicine*, 2(3):a006536, 2012.

[96] Emilie Dufresne, Parker B. Edwards, Heather A. Harrington, and Jonathan D. Hauenstein. Sampling real algebraic varieties for topological data analysis. arXiv:1802.07716, 2018.

[97] Herbert Edelsbrunner and John L. Harer. Persistent homology — A survey. *Contemporary Mathematics*, 453:257–282, 2008.

[98] Herbert Edelsbrunner and John L. Harer. *Computational Topology*. American Mathematical Society, Providence R. I., 2010.

[99] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete and Computational Geometry*, 28:511–533, 2002.

[100] Josef Ehling, Benjamin Theek, Felix Gremse, Sarah Baetke, Diana Möckel, Juliana Maynard, Sally-Ann Ricketts, Holger Grüll, Michal Neeman, Ruth Knuechel, Wiltrud Lederle, Fabian Kiessling, and Twan Lammers. Micro-CT imaging of tumour angiogenesis: quantitative measures describing micromorphology and vascularisation. *American Journal of Pathology*, 184(2):431 – 441, 2014.

[101] Anders Eklund, Thomas E. Nichols, and Hans Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences of the United States of America*, 113(28):7900–7905, 2016.

[102] Jeff Erickson. Combinatorial optimization of cycles and bases. In Afra Zomorodian, editor, *Advances in Applied and Computational Topology*, volume 70 of *Proceedings of Symposia in Applied Mathematics*, pages 195–228, Providence, Rhode Island, 2012. American Mathematical Society.

[103] Fabrizio De Vico Fallani, Jonas Richiardi, Mario Chavez, and Sophie Archard. Graph analysis of functional brain networks: Practical issues in translational neuroscience. *Philosophical Transactions of the Royal Society B Biological Sciences*, 369(1653):0130521, 2014.

[104] Brittany Terese Fasy and Bei Wang. Exploring persistent local homology in topological data analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6430–6434. IEEE, 2016.

[105] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[106] Daniel J. Fenn, Mason A. Porter, Mark McDonald, Stacy Williams, Neil F. Johnson, and Nick S. Jones. Dynamic communities in multichannel data: An application to the foreign exchange market during the 2007–2008 credit crisis. *Chaos*, 19(3):033119, 2009.

[107] Aasa Feragen, Francois Lauze, Pechin Lo, Marleen de Bruijne, and Mads Nielsen. Geometries on spaces of treelike shapes. In *Computer Vision. ACCV 2010*, volume 6493 of *Lecture Notes in Computer Science*, pages 160 – 173. Springer, Berlin, Heidelberg, 2011.

[108] Jacques Ferlay, Morten Ervik, F. Lam, Murielle Colombet, Les Mery, Marion Piñeros, Ariana Znaor, Isabelle Soerjomataram, and Freddie Bray. Global cancer observatory: Cancer today. Lyon, France: International agency for research on cancer. Available from: `https://gco.iarc.fr/today`, accessed on 25.11.2018.

[109] Ryan Flanagan, Lucas Lacasa, Emma K. Towlson, Sang Hoon Lee, and Mason A. Porter. Effect of antipsychotics on community structure in functional brain networks. *Journal of Complex Networks*, 2019. advanced access, `doi:10.1093/comnet/cnz013`.

[110] Judah Folkman. Tumour angiogenesis: Therapeutic implications. *The New England Journal of Medicine*, 285:1182 – 1186, 1971.

[111] Alex Fornito and Edward T. Bullmore. Reconciling abnormalities of brain network structure and function in schizophrenia. *Current Opinion in Neurobiology*, 30:44–50, 2015.

[112] Alex Fornito, Andrew Zalesky, Christos Pantelis, and Edward T. Bullmore. Schizophrenia, neuroimaging and connectomics. *NeuroImage*, 62:2296–2314, 2012.

[113] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, 2016.

[114] Michael D. Fox and Marcus E. Raichle. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience*, 8(9):700, 2007.

[115] Michael D. Fox, Dongyang Zhang, Abraham Z. Snyder, and Marcus E. Raichle. The global signal and observed anticorrelated resting state brain networks. *Journal of Neurophysiology*, 101(6):3270–3283, 2009.

[116] Celso Freitas, Elbert Macau, and Arkady Pikovsky. Partial synchronization phenomena in networks of identical oscillators with non-linear coupling. *Chaos*, 24:024402, 2014.

[117] Karl J. Friston, Steven Williams, Robert Howard, Richard S. J. Frackowiak, and Robert Turner. Movement-related effects in fMRI time-series. *Magnetic Resonance in Medicine*, 35(3):346–355, 1996.

[118] Marcio Gameiro, Yasuaki Hiraoka, Shunsuke Izumi, Miroslav Kramar, Konstantin Mischaikow, and Vidit Nanda. A topological measurement of protein compressibility. *Japan Journal of Industrial and Applied Mathematics*, 32:1–17, 2015.

[119] Guojun Gan, Chaoqun Ma, and Jianhong Wu. *Data Clustering: Theory, Algorithms, and Applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2007.

[120] Amanmeet Garg, Donghuan Lu, Karteek Popuri, and Mirza Faisal Beg. Brain geometry persistent homology marker for parkinson's disease. In *IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 525–528. IEEE, 2017.

[121] Yuval Gazit, James W. Baish, Nina Safabakhsh, Michael Leunig, Laurence T. Baxter, and Rakesh K. Jain. Fractal characteristics of tumor vascular architecture during tumor growth and regression. *Microcirculation*, 4(4):395–402, 1997.

[122] Caleb Geniesse, Olaf Sporns, Giovanni Petri, and Manish Saggar. Generating dynamical neuroimaging spatiotemporal representations (dyNeuSR) using topological data analysis. *Network Neuroscience*, 3(3):763 – 778, 2019.

[123] Robert Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45:61–75, 2008.

[124] Chad Giusti, Robert Ghrist, and Danielle S. Bassett. Two's company and three (or more) is a simplex. *Journal of Computational Neuroscience*, 41:1–14, 2016.

[125] Chad Giusti, Eva Pastalkova, Carina Curto, and Vladimir Itskov. Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences of the United States of America*, 112(44):13455–13460, 2015.

[126] Shom Goel, Dan G. Duda, Lei Xu, Lance L. Munn, Yves Boucher, Dai Fukumura, and Rakesh K. Jain. Normalization of the vasculature for treatment of cancer and other diseases. *Physiological Reviews*, 91(3):1071–1121, 2011.

[127] Daniel J. Gould, Tegy J. Vadakkan, Ross A. Poché, and Mary E. Dickinson. Multifractal and lacunarity analysis of microvascular morphology and remodeling. *Microcirculation*, 18(2):136–151, 2011.

[128] David Robert Grimes, Pavitra Kannan, Daniel R. Warren, Bostjan Markelc, Russell Bates, Ruth J. Muschel, and Mike Partridge. Estimating oxygen distribution from vasculature in three-dimensional tumour tissue. *Journal of The Royal Society Interface*, 13(116):20160070, 2016.

[129] James A. Grogan, Bostjan Markelc, Anthony J. Connor, Ruth J. Muschel, Joe M. Pitt-Francis, Philip K. Maini, and Helen M. Byrne. Predicting the influence of microvascular structure on tumour response to radiotherapy. *IEEE Transactions on Biomedical Engineering*, 64(3), 2016.

[130] Shuixia Guo, Lena Palaniyappan, Peter F. Liddle, and Jianfeng Feng. Dynamic cerebral reorganization in the pathophysiology of schizophrenia: A MRI-derived cortical thickness study. *Psychological Medicine*, 46(10):2201–2214, 2016.

[131] Shamik Gupta, Alessandro Campa, and Stefano Ruffo. Kuramoto model of synchronization: Equilibrium and nonequilibrium aspects. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(8):R08001, 2014.

[132] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144:646 – 674, 2011.

[133] Allen Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi, Dubai, Tokyo, 2001.

[134] Tony Hey, Stewart Tansley, and Kristin Tolle. The fourth paradigm: Data-intensive scientific discovery. Microsoft Research, available at:

https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/, 2009.

[135] Christoph Hofer, Roland Kwitt, Marc Niethammer, and Andreas Uhl. Deep learning with topological signatures. In *Advances in Neural Information Processing Systems*, pages 1634–1644, 2017.

[136] Thomas W. Hungerford. *Algebra*, volume 73 of *Graduate Texts in Mathematics*. Springer, New York, Heidelberg, Berlin, 1974.

[137] Herbert Hurwitz, Louis Fehrenbacher, William Novotny, Thomas Cartwright, John Hainsworth, William Heim, Jordan Berlin, Ari Baron, Susan Griffing, Eric Holmgren, Napoleone Ferrara, Gwen Fyfe, Beth Rogers, Robert Ross, and Fairooz Kabbinavar. Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. *New England Journal of Medicine*, 350(23):2335–2342, 2004.

[138] Esther Ibáñez-Marcelo, Lisa Campioni, Angkoon Phinyomark, Giovanni Petri, and Enrica L. Santarcangelo. Topology highlights mesoscopic functional equivalence between imagery and perception: The case of hypnotizability. *NeuroImage*, 200:437 – 449, 2019.

[139] Rakesh K. Jain. Normalizing tumor vasculature with anti-angiogenic therapy: a new paradigm for combination therapy. *Nature Medicine*, 7(9):987–989, 2001.

[140] Rakesh K. Jain. Normalization of tumor vasculature: an emerging concept in antiangiogenic therapy. *Science*, 307(5706):58–62, 2005.

[141] Rakesh K. Jain. Antiangiogenesis strategies revisited: from starving tumors to alleviating hypoxia. *Cancer Cell*, 26(5):605–622, 2014.

[142] Rakesh K. Jain, Dan G. Duda, Jeffrey W. Clark, and Jay S. Loeffler. Lessons from phase iii clinical trials on anti-vegf therapy for cancer. *Nature Clinical Practice Oncology*, 3(1):24–40, 2006.

[143] Lucas G. S. Jeub, Marya Bazzi, Inderjit S. Jutla, and Peter J. Mucha. A generalized Louvain method for community detection implemented in MATLAB, version 2.0. `https://github.com/GenLouvain/GenLouvain`, 2011–2016.

[144] Jakob Kaeppler. *Personal Communication*, 2019.

[145] Lida Kanari, Paweł Dłotko, Martina Scolamiero, Ran Levi, Julian Shillcock, Kathryn Hess, and Henry Markram. A topological representation of branching neuronal morphologies. *Neuroinformatics*, 16(1):3–13, 2018.

[146] Lida Kanari, Srikanth Ramaswamy, Ying Shi, Sebastien Morand, Julie Meystre, Rodrigo Perin, Marwan Abdellah, Yun Wang, Kathryn Hess, and Henry Markram. Objective morphological classification of neocortical pyramidal cells. *Cerebral Cortex*, 29(4):1719–1735, 2019.

[147] Raffi Karshafian, Peter N. Burns, and Mark R. Henkelman. Transit time kinetics in ordered and disordered vascular trees. *Physics in Medicine and Biology*, 48(19):3225 – 3237, 2003.

[148] Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344 of *Wiley Series in Probability and Satistics*. John Wiley & Sons, Hoboken, New Jersey, 2009.

[149] Moritz A. Konerding, Wolf Malkusch, B. Klapthor, Claudia van Ackern, E. Fait, Sally A. Hill, Charles Parkins, David J. Chaplin, Marco Presta, and Juliana Denekamp. Evidence for characteristic vascular patterns in solid tumours: quantitative studies using corrosion casts. *British Journal of Cancer*, 80(5-6):724–732, 1999.

[150] Czes Kosniowski. *A First Course in Algebraic Topology*. Cambridge University Press, Cambridge, London, New York, New Rochelle, Melbourne, Sydney, 1980.

[151] Violeta Kovacev-Nikolic. Persistent homology in analysis of point-cloud data. Master's thesis, University of Alberta, `https://era.library.ualberta.ca/files/cv43nx33b/Kovacev-Nikolic_Violeta_Fall2012.pdf`, 2012.

[152] Violeta Kovacev-Nikolic, Peter Bubenik, Dragan Nikolic, and Giseon Heo. Using persistent homology and dynamical distances to analyze protein binding. *Statistical Applications in Genetics and Molecular Biology*, 15(1):1–27, 2016.

[153] Miroslav Kramár, Arnaud Goullet, Lou Kondic, and Konstantin Mischaikow. Persistence of force networks in compressed granular media. *Physical Review E*, 87:042207, 2013.

[154] Yoshiki Kuramoto. *Chemical Oscillations and Waves and Turbulence*. Springer, Berlin, 1984.

[155] Genki Kusano, Yasuaki Hiraoka, and Kenji Fukumizu. Persistence weighted gaussian kernel for topological data analysis. In *International Conference on Machine Learning*, volume 48, pages 2004–2013, 2016.

[156] Roland Kwitt, Stefan Huber, Marc Niethammer, Weili Lin, and Ulrich Bauer. Statistical topological data analysis-a kernel perspective. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3070–3078. Curran Associates, Inc., 2015.

[157] Hyekyoung Lee, Moo K. Chung, Hongyoon Choi, Hyejin Kang, Seunggyun Ha, Yu Kyeong Kim, and Dong Soo Lee. Harmonic holes as the submodules of brain network and network dissimilarity. In R. Marfil, M. Calderón, F. Díaz del Río,

and A. Bandera P. Real, editors, *Computational Topology in Image Context. CTIC 2019*, volume 11382 of *Lecture Notes in Computer Science*, pages 110–122. Springer, Cham, 2019.

[158] Hyekyoung Lee, Moo K. Chung, Hyejin Kang, Bung-Nyun Kim, and Dong Soo Lee. Discriminative persistent homology of brain networks. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 841–844, 2011.

[159] Hyekyoung Lee, Hyejin Kang, Moo K. Chung, Bung-Nyun Kim, and Dong Soo Lee. Weighted functional brain network modeling via network filtration. In *NIPS Workshop on Algebraic Topology and Machine Learning*, volume 3, 2012.

[160] Won Hee Lee, Edward T. Bullmore, and Sophia Frangou. Quantitative evaluation of simulated functional brain networks in graph theoretical analysis. *NeuroImage*, page Available at `http://dx.doi.org/10.1016/j.neuroimage.2016.08.050`, 2016.

[161] Yongjin Lee, Senja D. Barthel, Paweł Dłotko, Seyed Mohamad Moosavi, Kathryn Hess, and Berend Smit. High-throughput screening approach for nanoporous materials genome using topological data analysis: application to zeolites. *Journal of Chemical Theory and Computation*, 14(8):4427–4437, 2018.

[162] Lek-Heng Lim. Hodge laplacians on graphs. `https://www.stat.uchicago.edu/~lekheng/work/hodge-graph.pdf`, preprint downloaded July 2019.

[163] Jen-Yu Liu, Shyh-Kang Jeng, and Yi-Hsuan Yang. Applying topological persistence in convolutional neural network for music audio signals. arXiv:1608.07373v1, 2016.

[164] Yong Liu, Meng Linag, Yuan Zhou, Yong He, Yihui Hao, Ming Song, Chunshui Yu, Haihong Liu, Zhening Liu, and Tianzi Jiang. Disrupted small-world networks in schizophrenia. *Brain*, 131:945–961, 2008.

[165] Svetlana Lockwood and Bala Krishnamoorthy. Topological features in cancer gene expression data. In *Pacific Symposium on Biocomputing*, pages 108–119, 2015.

[166] Mary-Ellen Lynall, Danielle S. Bassett, Robert Kerwin, Peter J. McKenna, and Manfred Kitzbichler. Functional connectivity and brain networks in schizophrenia. *The Journal of Neuroscience*, 30(28):9477–9487, 2010.

[167] Robert MacPherson and Benjamin Schweinhart. Measuring shape with topology. *Journal of Mathematical Physics*, 53(7):073516, 2012.

[168] Barbara I. Mahler, Ulrike Tillmann, and Mason A. Porter. Analysis of contagion maps on a class of networks that are spatially embedded in a torus. arXiv:1812.09806, 2018.

[169] Shawn Martin, Aidan Thompson, Evangelos A. Coutsias, and Jean-Paul Watson. Topology of cyclo-octane energy landscape. *The Journal of Chemical Physics*, 132(23):234115, 2010.

[170] Shawn Martin and Jean-Paul Watson. Non-manifold surface reconstruction from high-dimensional point cloud data. *Computational Geometry*, 44(8):427–441, 2011.

[171] Vivien Marx. The big challenges of big data. *Nature Biology*, 498:255 – 260, 2013.

[172] Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, 2011.

[173] Dmitriy Morozov. Dionysus. Software available at `https://www.mrzv.org/software/dionysus/`, software retrieved in 2019.

[174] Peter J. Mucha, Thomas Richardson, Kevin Macon, Mason A. Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328:876–878, 2010.

[175] Mark R. Muldoon, Robert S. MacKay, Jeremy P. Huke, and David S. Broomhead. Topology from time series. *Physica D*, 65(1–2):1–16, 1993.

[176] Elizabeth Munch, Katharine Turner, Paul Bendich, Sayan Mukherjee, Jonathan Mattingly, and John Harer. Probabilistic Fréchet means and statistics on vineyards. *Electronic Journal of Statistics*, 9:1173 – 1204, 2015.

[177] James R. Munkres. *Elements of algebraic topology*. The Benjamin/Cummings Publishing Company, inc., Redwood City (California), Menlo Park (California), Reading (Massachusetts), Amsterdam, Don Mills (Ontario), Mexico City, Sydney, Bonn, Madrid, Singapore, Tokyo, Bogota, Santiago, San Juan, Wokingham (United Kingdom), 1984.

[178] James R. Munkres. *Topology*. Pearson Prentice Hall, New Jersey, 2000.

[179] Cameron Mura, Eli J. Draizen, and Philip E. Bourne. Structural biology meets data science: Does anything change? *Current opinion in structural biology*, 52:95–102, 2018.

[180] Kevin Murphy, Rasmus M. Birn, Daniel A. Handwerker, Tyler B. Jones, and Peter A. Bandettini. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? *NeuroImage*, 44(3):893–905, 2009.

[181] Janice A. Nagy, Sung-Hee Chang, and Harold F. Dvorak. Why are tumour vessels abnormal and why is it important to know? *British Journal of Cancer*, 100:865 – 869, 2009.

[182] Tomomichi Nakamura, Toshihiro Tanizawa, and Michael Small. Constructing networks from a dynamical system perspective for multivariate nonlinear time series. *Physical Review E*, 93:032323, 2016.

[183] Mark E. J. Newman. *Networks*. Oxford University Press, Oxford, UK, second edition, 2018.

[184] Monica Nicolau, Arnold J. Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270, 2011.

[185] John Nicponski and Jae-Hun Jung. Topological data analysis of vascular disease: A theoretical framework. BioRxiv: 637090, 2019.

[186] Jessica L. Nielson, Jesse Paquette, Aiwen W. Liu, Cristian F. Guandique, C. Amy Tovar, Tomoo Inoue, Karen-Amanda Irvine, John C. Gensel, Jennifer Kloke, Tanya C. Petrossian, Pek Y. Lum, Gunnar E. Carlsson, Geoffrey T. Manley, Wise Young, Jacqueline C. Beattie, Michael S.and Bresnahan, and Adam R. Ferguson. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nature Communications*, 6:8581, 2015.

[187] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry*, 39(1-3):419–441, 2008.

[188] World Health Organisation. Cancer fact sheet. `https://www.who.int/news-room/fact-sheets/detail/cancer`. Information retrieved 24.4.2019.

[189] Brent A. Orr and Charles G. Eberhart. Molecular pathways: Not a simple tube – the many functions of blood vessels in tumours. *Cinical Cancer Research*, 21(1):18 –23, 2015.

[190] Nina Otter, Mason A. Porter, Ulrike Tillmann, Peter Grindrod, and Heather A. Harrington. A roadmap for the computation of persistent homology. *European Physical Journal – Data Science*, 6(17):1–38, 2017.

[191] Nina Otter, Mason A. Porter, Ulrike Tillmann, Peter Grindrod, and Heather A. Harrington. A roadmap for the computation of persistent homology – tutorial. `https://arxiv.org/src/1506.08903v7/anc/tutorial10.pdf`, software available at `https://github.com/n-otter/PH-roadmap/tree/master/matlab`, software retrieved 2017, 2017.

[192] Markus R. Owen, Tomás Alarcón, Philip K. Maini, and Helen M. Byrne. Angiogenesis and vascular remodelling in normal and cancerous tissues. *Journal of Mathematical Biology*, 58(4-5):689, 2009.

[193] David Papo, Javier M. Buldú, Stefano Boccaletti, and Edward T. Bullmore. Complex network theory and the brain. *Philosophical Transactions of the Royal Society B*, 369(1688):20130520, 2014.

[194] David Papo, Massimiliano Zanin, José A. Pineda-Pardo, Stefano Boccaletti, and Javier M. Buldú. Functional brain networks: Great expectations and hard times and the big leap forward. *Philosophical Transactions of the Royal Society B*, 369(1653):20130525, 2014.

[195] Heon Joo Park, Robert J. Griffin, Susanta Hui, Seymour H. Levitt, and Chang W. Song. Radiation-induced vascular damage in tumors: implications of vascular damage in ablative hypofractionated radiotherapy (SBRT and SRS). *Radiation Research*, 177(3):311–327, 2012.

[196] Alice Patania, Francesco Vaccarino, and Giovanni Petri. Topological analysis of data. *European Physical Journal – Data Science*, 6(1):7, 2017.

[197] Vic Patrangenaru, Peter Bubenik, Robert L. Paige, and Daniel Osborne. Topological data analysis for object data. arXiv:1804.10255, 2018.

[198] Avi Peled, Amir B. Geva, William S. Kremen, Howard M. Blankfeld, Roberta Esfandiarfard, and Thomas E. Nordahi. Functional connectivity and working memory in schizophrenia: An EEG study. *International Journal of Neuroscience*, 106(1–2):47–61, 2001.

[199] Raimondo Penta and Davide Ambrosi. The role of the microvascular tortuosity in tumor transport phenomena. *Journal of Theoretical Biology*, 364:80–97, 2015.

[200] Steven E. Petersen and Olaf Sporns. Brain networks and cognitive architectures. *Neuron*, 88(1):207–219, 2015.

[201] Giovanni Petri, Paul Expert, Federico Turkheimer, Robin Carhart-Harris, David Nutt, Peter J. Hellyer, and Francesco Vaccarino. Homological scaffolds of brain functional networks. *Journal of the Royal Society Interface*, 11:20140873, 2014.

[202] Giovanni Petri, Martina Scolamiero, Irene Donato, and Francesco Vaccarino. Topological strata of weighted complex networks. *PLoS ONE*, 8(6):e66505, 2013.

[203] Mikael J. Pittet and Ralph Weissleder. Intravital imaging. *Cell*, 147(5):983–991, 2011.

[204] Mason A. Porter and James P. Gleeson. Dynamical systems on networks: A tutorial. *Frontiers in Applied Dynamical Systems: Reviews and Tutorials*, 4, 2016.

[205] Mason A. Porter, Jukka-Pekka Onnela, and Peter J. Mucha. Communities in networks. *Notices of the American Mathematical Society*, 56(9):1082–1097, 1164–1166, 2009.

[206] Marie Prewett, James Huber, Yiwen Li, Angel Santiago, William O'Connor, Karen King, Jay Overholser, Andrea Hooper, Bronislaw Pytowski, Larry Witte, Peter Bohlen, and Daniel J. Hicklin. Antivascular endothelial growth factor receptor (fetal liver kinase 1) monoclonal antibody inhibits tumor angiogenesis and growth of several mouse and human tumors. *Cancer Research*, 59(20):5209–5218, 1999.

[207] Dean Prichard and James Theiler. Generating surrogate data for time series with several simultaneously measured variables. *Physical Review Letters*, 73(7):951–954, 1994.

[208] Antonio Rampino, Rosie May Walker, Helen Scott Torrance, Susan Maguire Anderson, Leonardo Fazio, Annabella Di Giorgio, Paolo Taurisano, Barbara Gelao, Raffaella Romano, Rita Masellis, Ginaluca Ursini, Grazia Caforio, Giuseppe Blasi, J. Kirsty Millar, David John Porteous, Pippa Ann Thomson, Alessandro Bertolino, and Kathryn Louise Evans. Expression of DISC1-interactome members correlates with cognitive phenotypes related to schizophrenia. *PloS ONE*, 9(6):e99892, 2014.

[209] Wayne Rasband. Imagej. Image processing and analysis in JAVA. Software available at `https://imagej.nih.gov/ij/download.html`, software retrieved in 2019.

[210] Michael W. Reimann, Max Nolte, Martina Scolamiero, Katharine Turner, Rodrigo Perin, Giuseppe Chindemi, Paweł Dłotko, Ran Levi, Kathryn Hess, and Henry Markram. Cliques of neurons bound into cavities provide a missing

link between structure and function. *Frontiers in Computational Neuroscience*, 11(48):1–16, 2017.

[211] Jan Reininghaus, Stefan Huber, Ulrich Bauer, and Roland Kwitt. A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4741–4748, 2015.

[212] John Ridgway, Gu Zhang, Yan Wu, Scott Stawicki, Wei-Ching Liang, Yvan Chanthery, Joe Kowalski, Ryan J. Watts, Christopher Callahan, Ian Kasman, Mallika Singh, May Chien, Christine Tan, Jo-Anne S. Hongo, Fred de Sauvage, Greg Plowman, and Minhong Yan. Inhibition of Dll4 signalling inhibits tumour growth by deregulating angiogenesis. *Nature*, 444(7122):1083 – 1087, 2006.

[213] Laila Ritsma, Ernst J.A. Steller, Saskia I.J. Ellenbroek, Onno Kranenburg, Inne H. M. Borel Rinkes, and Jacco Van Rheenen. Surgical implantation of an abdominal imaging window for intravital microscopy. *Nature Protocols*, 8(3):583 – 594, 2013.

[214] Francisco Rodrigues, Thomas K. D. M. Peron, Peng Ji, and Jürgen Kurths. The Kuramoto model in complex networks. *Physics Reports*, 610:1–98, 2016.

[215] Avihai Ron, Xosé Luís Deán-Ben, Sven Gottschalk, and Daniel Razansky. Volumetric optoacoustic imaging unveils high-resolution patterns of acute and cyclic hypoxia in a murine model of breast cancer. *Cancer Research*, 3769, 2019 (in press).

[216] Mikail Rubinov and Edward T. Bullmore. Schizophrenia and abnormal brain network hubs. *Dialogues in Clinical Research*, 15(3):339–349, 2013.

[217] Erkki Ruoslahti. Specialization of tumour vasculature. *Nature Reviews Cancer*, 2:83 – 90, 2002.

[218] Fabio Sambataro, Giuseppe Blasi, Leonardo Fazio, Grazia Caforio, Paolo Taurisano, Raffaella Romano, Annabella Di Giorgio, Barbara Gelao, Luciana Lo Biancho, Apostolos Papazacharias, Teresa Popolizio, Marcello Nardini, and Allessandro Bertolino. Treatment with Olanzapine is associated with modulation of the default mode network in patients with schizophrenia. *Neuropsychopharmacology*, 35(4):904–912, 2010.

[219] Nicole Sanderson, Elliott Shugerman, Samantha Molnar, James D. Meiss, and Elizabeth Bradley. Computational topology techniques for characterizing time-series data. In Niall Adams, Allan Tucker, and David Weston, editors, *Advances in Intelligent Data Analysis XVI. IDA 2017*, volume 10584 of *Lecture Notes in Computer Science*, pages 284–296. Springer, Cham, 2017.

[220] Denisse Sciamarella and Gabo B. Mindlin. Topological structure of chaotic flows from human speech data. *Physical Review Letters*, 82(7):1450 – 1453, 1999.

[221] Denisse Sciamarella and Gabo B. Mindlin. Unveiling the topological structure of chaotic flows from data. *Physical Review E*, 64(3):036209, 2001.

[222] Martina Scolamiero. *Personal Communication*, 2016.

[223] Gianna Sepede, Antonio Ferretti, Mauro Gianni Perrucci, Francesco Gambi, Fiore Di Donato, Francesco Nuccetelli, Cosimo Del Gratta, Armando Tartaro, Rosa Maria Salerno, Filippo Maria Ferro, and Gian Luca Romani. Altered brain response without behavioral attention deficits in healthy siblings of schizophrenic patients: An event-related fMRI study. *NeuroImage*, 49(1):1080–1090, 2010.

[224] Ma Ángeles Serrano, Marián Boguná, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the Na-*

tional Academy of Sciences of the United States of America, 106(16):6483–6488, 2009.

[225] Felix Siebenhühner, Shennan A. Weiss, Richard Coppola, Daniel R. Weinberger, and Danielle S. Bassett. Intra-and inter-frequency brain network structure in health and schizophrenia. *PloS ONE*, 8(8):e72351, 2013.

[226] Dietmar W. Siemann. The unique characteristics of tumour vasculature and preclinical evidence for its selective disruption by tumor-vascular disrupting agents. *Cancer Treatment Reviews*, 37(1):63 – 74, 2011.

[227] Gurjeet Singh, Facundo Mémoli, and Gunnar E. Carlsson. Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *Eurographics Symposium in Point-Based Graphics*, pages 91–100, 2007.

[228] Gurjeet Singh, Facundo Mémoli, Tigran Ishkhanov, Guillermo Sapiro, Gunnar Carlsson, and Dario L. Ringach. Topological analysis of population activity in visual cortex. *Journal of Vision*, 8(11):1–18, 2008.

[229] Megha Singh and Ganesh Bagler. Network biomarkers of schizophrenia by graph theoretical investigations of brain functional networks. arXiv:1602.01191, 2016.

[230] Ann Sizemore, Chad Giusti, and Danielle S. Bassett. Classification of weighted networks through mesoscale homological features. *Journal of Complex Networks*, 2016.

[231] Ann E. Sizemore, Chad Giusti, Ari Kahn, Jean M. Vettel, Richard F. Betzel, and Danielle S. Bassett. Cliques and cavities in the human connectome. *Journal of Computational Neuroscience*, 44(1):115–145, 2018.

[232] Ann E. Sizemore, Jennifer E. Phillips-Cremins, Robert Ghrist, and Danielle S. Bassett. The importance of the whole: topological data analysis for the network neuroscientist. *Network Neuroscience*, 3(3):656 — 673, 2018.

[233] Stephen M. Smith, Karla L. Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F. Beckmann, Thomas E. Nichols, Joseph D. Ramsay, and Mark W. Woolrich. Network modelling methods for fMRI. *NeuroImage*, 54(2):875–891, 2011.

[234] Olaf Sporns. Contributions and challenges for network models in cognitive neuroscience. *Nature Reviews Neuroscience*, 17(5):652–660, 2014.

[235] Olaf Sporns. Graph-theoretical analysis of brain networks. In Arthur W. Toga, editor, *Brain Mapping: An Encyclopedic Reference*, volume 1, pages 629–633. Academic Press: Elsevier, Cambridge, Massachusetts, 2015.

[236] Gard Spreemann, Benjamin Dunn, Magnus Bakke Botnan, and Nils A. Baas. Using persistent homology to reveal hidden covariates in systems governed by the kinetic ising model. *Physical Review E*, 97(3), 2018.

[237] Bernadette J. Stolz. Computational topology in neuroscience. Master's thesis, University of Oxford, `http://www.math.ucla.edu/~mason/research/Dissertation-stolz2014-Corr.pdf`, 2014.

[238] Bernadette J. Stolz, Tegan Emerson, Satu Nahkuri, Mason A. Porter, and Heather A. Harrington. Topological data analysis of task-based fMRI data from experiments on schizophrenia. arXiv:1809.08504, 2018.

[239] Bernadette J. Stolz, Heather A. Harrington, and Mason A. Porter. The topological "shape" of Brexit. arXiv:1610.00752, 2016.

[240] Bernadette J. Stolz, Heather A. Harrington, and Mason A. Porter. Persistent homology of time-dependent functional networks constructed from coupled time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(4):047410, 2017.

[241] Bernadette J. Stolz and Barbara I. Mahler. H is for homology. `https://www.maths.ox.ac.uk/about-us/life-oxford-mathematics/oxford-mathematics-alphabet/h-homology`, 2016.

[242] Bernadette J. Stolz, Jared Tanner, Heather A. Harrington, and Vidit Nanda. Geometric anomaly detection in data. arXiv: 1908.09397.

[243] John Stout, Matthew R. Whiteway, Edward Ott, Michelle Girvan, and Thomas M. Antonsen. Local synchronization in complex networks of coupled oscillators. *Chaos*, 21(2):025109, 2011.

[244] Steven H. Strogatz. From Kuramoto to Crawford: Exploring the onset of synchronization in populations of coupled oscillators. *Physica D*, 143(1–4):1–20, 2000.

[245] Xiaoran Sun, Michael Small, Yi Zhao, and Xiaoping Xue. Characterizing system dynamics with a weighted and directed network constructed from time series data. *Chaos*, 24(2):024402, 2014.

[246] Jean Talairach and Pierre Tournoux. *Co-planar stereotaxic atlas of the human brain. 3-D proportional system: An approach to cerebral imaging.* Thieme, 1st edition, 1988.

[247] Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams. JavaPlex: A research software package for persistent (co)homology. In Han Hong and Chee Yap, editors, *Mathematical Software. ICMS 2014*, volume 8592 of *Lecture Notes*

in Computer Science, pages 129–136. Springer, Berlin, Heidelberg, 2014. Software available at `http://appliedtopology.github.io/javaplex/`.

[248] Dane Taylor, Florian Klimm, Heather A. Harrington, Miroslav Kramár, Konstantin Mishchaikow, Mason A. Porter, and Peter J. Mucha. Topological data analysis of contagion maps for examining spreading processes on networks. *Nature Communications*, 6:7723, 2015.

[249] Joshua B. Tenenbaum, Vin De Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[250] Joshua B. Tenenbaum, Vin De Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. Software available at `https://web.archive.org/web/20040411051530/http://isomap.stanford.edu/`, software retrieved in 2019.

[251] Giulio Tirabassi, Ricardo Sevilla-Escoboza, Javier M. Buldú, and Cristina Masoller. Inferring the connectivity of coupled oscillators from time-series statistical similarity analysis. *Scientific Reports*, 5(10829):1–14, 2015.

[252] Chad M. Topaz. Self-help homology tutorial for the simple(x)-minded. `https://drive.google.com/file/d/0B3Www1z6Tm8xV3ozTmN5RE94bDg/view`, accessed on 8.8.2019, 2015.

[253] Chad M. Topaz, Lori Ziegelmeier, and Tom Halverson. Topological data analysis of biological aggregation models. *PloS ONE*, 10(5):e0126383, 2015.

[254] Emma K. Towlson, Petra E. Vértes, Ulrich Müller, and Sebastian E. Ahnert. Brain networks reveal the effects of antipsychotic drugs on schizophrenia patients and controls. arXiv:1806.00128, 2018.

[255] Gillian M. Tozer, Simon M. Ameer-Beg, Jennifer Baker, Paul R. Barber, Sally A. Hill, Richard J. Hodgkiss, Rosalind Locke, Vivien E. Prise, Ian Wilson, and Borivoj Vojnovic. Intravital imaging of tumour vascular networks using multi-photon fluorescence microscopy. *Advanced drug delivery reviews*, 57(1):135–152, 2005.

[256] Vincent A Traag and Jeroen Bruggeman. Community detection in networks with positive and negative links. *Physical Review E*, 80(3):036115, 2009.

[257] Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distributions of persistence diagrams. *Discrete and Computational Geometry*, 52(1):44–70, 2014.

[258] Benedetta Ubezio, Raquel Agudo Blanco, Ilse Geudens, Fabio Stanchi, Thomas Mathivet, Martin L. Jones, Anan Ragab, Katie Bentley, and Holger Gerhardt. Synchronization of endothelial Dll4-notch dynamics switch blood vessels from branching to expansion. *eLife*, 5:e12167, 2016.

[259] M. Ulmer, Lori Ziegelmeier, and Chad M. Topaz. A topological approach to selecting models of biological experiments. *PloS ONE*, 14(3):e0213679, 2019.

[260] Jose L. P. Velazquez. Brain research: A perspective from the coupled oscillators field. *NeuroQuantology*, 4(2):155–165, 2006.

[261] Prakash Vempati, Aleksander S. Popel, and Feilim Mac Gabhann. Extracellular regulation of vegf: isoforms, proteolysis, and vascular patterning. *Cytokine and Growth Factor Reviews*, 25(1):1–19, 2014.

[262] Guillermo Vilanova, Ignasi Colominas, and Hector Gomez. Computational modeling of tumor-induced angiogenesis. *Archives of Computational Methods in Engineering*, 24(4):1071–1102, 2017.

[263] Annette Volfing. *Heinrich von Mügeln, >Der meide kranz<. A Commentary*, volume 111 of *Münchener Texte und Untersuchungen zur deutschen Literatur des Mittelalters*. Max Niemeyer Verlag, Tübingen, 1997.

[264] Liang Wang, Paul D. Metzak, William G. Honer, and Todd S. Woodward. Impaired efficiency of functional networks underlying episodic memory-for-context in schizophrenia. *The Journal of Neuroscience*, 30(39):13171–13179, 2010.

[265] Yuan Wang, Hernando Ombao, and Moo K. Chung. Topological seizure origin detection in electroencephalographic signals. In *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 351–354, April 2015.

[266] Larry Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5:501–532, 2018.

[267] Andrew Scott Waugh, Liuyi Pei, James H. Fowler, Peter J. Mucha, and Mason A. Porter. Party polarization in congress: A network science approach. arXiv: 0907.3509, 2009.

[268] Shmuel Weinberger. What is... persistent homology? *Notices of the American Mathematical Society*, 58(1):36–39, 2011.

[269] Andreas Weissenbacher, Christian Kasess, Florian Gerstl, Rupert Lanzenberger, Ewald Moser, and Christian Windischberger. Correlations and anticorrelations in resting-state functional connectivity MRI: a quantitative comparison of preprocessing strategies. *NeuroImage*, 47(4):1408–1416, 2009.

[270] Jonathan Welti, Sonja Loges, Stefanie Dimmeler, and Peter Carmeliet. Recent molecular discoveries in angiogenesis and antiangiogenic therapies in cancer. *The Journal of Clinical Investigation*, 123(8):3190–3200, 2013.

[271] Jefferey Wildmann. Bron–Kerbosch maximal clique finding algorithm. Code available at: `http://www.mathworks.co.uk/matlabcentral/fileexchange/30413-bron-kerbosch-maximal-clique-finding-algorithm`, code retrieved in July 2014.

[272] World Health Organization. Schizophrenia. `http://www.who.int/mental\_health/management/schizophrenia/en/`, 19 September 2015.

[273] Kelin Xia and Guo-Wei Wei. Persistent homology analysis of protein structure, flexibility, and folding. *International Journal for Numerical Methods in Biomedical Engineering*, 30(8):814–844, 2014.

[274] Lalita Yadav, Naveen Puri, Varun Rastogi, and Vandana Sharma. Tumour angiogenesis and angiogenic inhibitors: A review. *Journal of Clinical and Diagnostic Research*, 9(6):1 – 5, 2015.

[275] Andrew Zalesky, Alex Fornito, Gary F. Egan, Christos Pantelis, and Edward T. Bullmore. The relationship between regional and inter-regional functional connectivity deficits in schizophrenia. *Human Brain Mapping*, 33:2535–2549, 2012.

[276] Erik C. Zeeman. The topology of the brain and visual perception. In M. K. Fort, editor, *The Topology of 3-Manifolds*, pages 240–256. Prentice Hall, Englewood Cliffs, NJ, 1962.

[277] Dongli Zhou, Wesley K. Thompson, and Greg Siegle. MATLAB toolbox for functional connectivity. *NeuroImage*, 47:1590–1607, 2009.

[278] Bartosz Zielinski, Michał Lipiński, Mateusz Juda, Matthias Zeppelzauer, and Paweł Dłotko. Persistence codebooks for topological data analysis. arXiv:1802.04852, 2018.

[279] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete and Computational Geometry*, 33(2):249–274, 2005.