Topological Data Analysis of Organoids

Lewis Michael Marsh

St Cross College

University of Oxford

A thesis for the degree

Doctor of Philosophy in Mathematics

Trinity Term 2023

Acknowledgements

This thesis was made possible with the help of numerous people, to whom I am all grateful.

First and foremost, I would like to thank my supervisors Prof. Heather A Harrington, Prof. Helen M Byrne and Prof. Xin Lu for their guidance, feedback and continuous support throughout my DPhil project and for helping me with my goal to undertake an internship during my DPhil. I am grateful to my confirmation of status examiners Prof. Ruth Baker and Prof. Peter Grindrod for providing feedback which significantly improved this thesis. I want to thank all of my collaborators but in particular Dr Felix Zhou, David Beers, Prof. Stas Shvartsman, Dr Emilie Dufresne, Dr Renee S Hoekzema and Otto Sumray, as well as all of my co-members at the Centre for Topological Data Analysis and the Lu Lab, for their patience and valuable input throughout our joint research projects and for giving me confidence in applying mathematics to areas I was not previously familiar with. I thank Dr Xiao Qin, Dr Brittany-Amber Jacobs and Dr Xiaoyue Han for collecting the experimental data sets I analysed, Dr Huaming Yan for providing synthetic organoid data sets and Dr Thomas Carroll and Dr Joseph Kaplinski for their guidance on the pre-processing of scRNA-seq data. The discussions I had with Prof. Vidit Nanda on the ECT inference section greatly improved this part of my thesis.

Finally, I would like to express my gratitude to my family and in particular my parents for their immeasurable support and belief in me throughout my whole education. And a special thanks to Andreea – I could not have done this without you.

Contents

List of Figures **v**

List of Tables **vi**

1 Introduction **1**

1.1 Topological Data Analysis 3
1.2 scRNA-seq Data 4
1.3 Organoid Morphology 6
1.4 Theory of the Euler Characteristic Transform 8
1.5 Structure 9

2 Background and Literature Review **10**

2.1 Organoids 11
2.2 scRNA Sequencing Data 14
2.3 Previous Studies on Organoid scRNA-seq Data 16
2.4 Previous Studies on Organoid Morphology Data 17
2.5 Mathematical Preliminaries 18
2.5.1 Mapper 19
2.5.2 UMAP 22
2.5.3 Random Walks and Community Detection 26
2.5.4 Simplicial Complexes and Filtrations 33
2.5.5 The Euler Characteristic and Euler Characteristic Transform . 35
2.5.6 Kernels and Kernel Approximations 38
2.6 Summary 42

3 TDA of Single Cell RNA Sequencing Data **43**

3.1 scRNA-seq Data Structure and Analysis Methods 45
3.1.1 Unique Molecular Identifiers and Raw Data Structure 45
3.1.2 Variance Stabilizing Transform 45

3.1.3 Differential Expression and Its Generalisations 48
3.1.4 Trajectory Inference Methods 50
3.2 New Topological Analysis Methods for scRNA-seq Data 52
3.2.1 Multiscale Laplacian Score 52
3.2.2 UMAP Diffusion Cover 53
3.3 Data Sets 55
3.4 Results 56
3.5 Discussion 61

4 TDA of Experimental 2D Organoid Data 69
4.1 Statistical Analysis Methods 70
4.1.1 Random Forest Classification 70
4.1.2 Canonical Correlation Analysis 73
4.2 Temporal Shape Detection with DETECT 74
4.3 Data Set 75
4.4 Results 76
4.4.1 Regressing SECT to Classical Shape Statistics 77
4.4.2 Classification of Organoids with DETECT 79
4.5 Discussion 80

5 TDA of Synthetic 3D Organoid Data 83
5.1 The Model 84
5.2 Data Set 86
5.3 Results 88
5.4 Discussion 89

6 ECT Stability and Inference 91
6.1 Introduction 92
6.1.1 Problem Statement and Contributions 92
6.1.2 Outline 95
6.2 Background 96
6.2.1 Related Work 96
6.2.2 Topological Preliminaries 96
6.2.3 Gaussian Processes 98
6.3 ECT Stability of Non-Random Data 101
6.3.1 Stability for Smooth Curves 101
6.3.2 Stability of Piece-wise Linear Interpolation 103

6.4 ECT Stability of Random Data 104
6.5 Example 107
6.6 Discussion 108

7 Discussion and Outlook **111**

Bibliography **117**

Appendix **131**
A.1 Differential Expression Analyses 131
A.2 UMAP Theory 133
A.3 ECT Stability Proofs 141

List of Figures

I.1 Graphical Abstract viii

2.1 Overview of Transcription and Translation 15
2.2 Reeb Graph Example 20
2.3 Mapper Algorithm 23
2.4 Multiscale Clustering Through Markov Stability Example 34
2.5 Example of Simplicial Complexes 35
2.6 ECT Analysis Pipeline for Organoids 39

3.1 Example of the Multiscale Laplacian Score 63
3.2 Multiscale Laplacian Score on T Cell Data 64
3.3 Mapper and PAGA Graphs on T Cell Data 65
3.4 UMAP Plots of Mouse Colon Organoid scRNA Data 66
3.5 Multiscale Laplacian Score on Organoid Data 67
3.6 Mapper and PAGA Graphs on Organoid Data 68

4.1 Example of a Decision Tree. 71
4.2 Effects of VC Treatment on Organoids. 76
4.3 Example of Classical Shape Descriptors. 78
4.4 Canonical Correlation Analysis of DETECT 79

5.1 Example of Synthetic Organoid Data. 87
5.2 Synthetic Organoid Data DETECT Output 89

6.1 ECT Instability Example I 94
6.2 ECT Instability Example II 94
6.3 Example Curves and Samples 108
6.4 Posterior SECT Distributions 109

A.1 DE Clusters For T Cell Data 132
A.2 DE Clusters For Mouse Colon Data 132

List of Tables

3.1 Comparison of Trajectory Inference Methods. 51

4.1 R^2-Coefficients in ECT Regression. 78

A.1 Differentially Expressed Genes on T Cell Data Set 131
A.2 Differentially Expressed Genes on Mouse Colon Data Set 133

Abstract

Organoids are multi-cellular structures which are cultured *in vitro* from stem cells to resemble specific organs (e.g., colon, liver) in their three-dimensional composition. The gene expression and the tissue composition of organoids constantly affect each other. Dynamic changes in the shape, cellular composition and transcriptomic profile of these model systems can be used to understand the effect of mutations and treatments in health and disease. In this thesis, I propose new techniques in the field of topological data analysis (TDA) to analyse the gene expression and the morphology of organoids. I use TDA methods, which are inspired by topology, to analyse and quantify the continuous structure of single-cell RNA sequencing data, which is embedded in high dimensional space, and the shape of an organoid.

For single-cell RNA sequencing data, I developed the multiscale Laplacian score (MLS) and the UMAP diffusion cover, which both extend and improve existing topological analysis methods. I demonstrate the utility of these techniques by applying them to a published benchmark single-cell data set and a data set of mouse colon organoids. The methods validate previously identified genes and detect additional genes with known involvement cancers.

To study the morphology of organoids I propose DETECT, a rotationally invariant signature of dynamically changing shapes. I demonstrate the efficacy of this method on a data set of segmented videos of mouse small intestine organoid experiments and show that it outperforms classical shape descriptors. I verify the method on a synthetic organoid data set and illustrate how it generalises to 3D to conclude that DETECT offers rigorous quantification of organoids and opens up computationally scalable methods for distinguishing different growth regimes and assessing treatment effects. Finally, I make a theoretical contribution to the statistical inference of the method underlying DETECT.

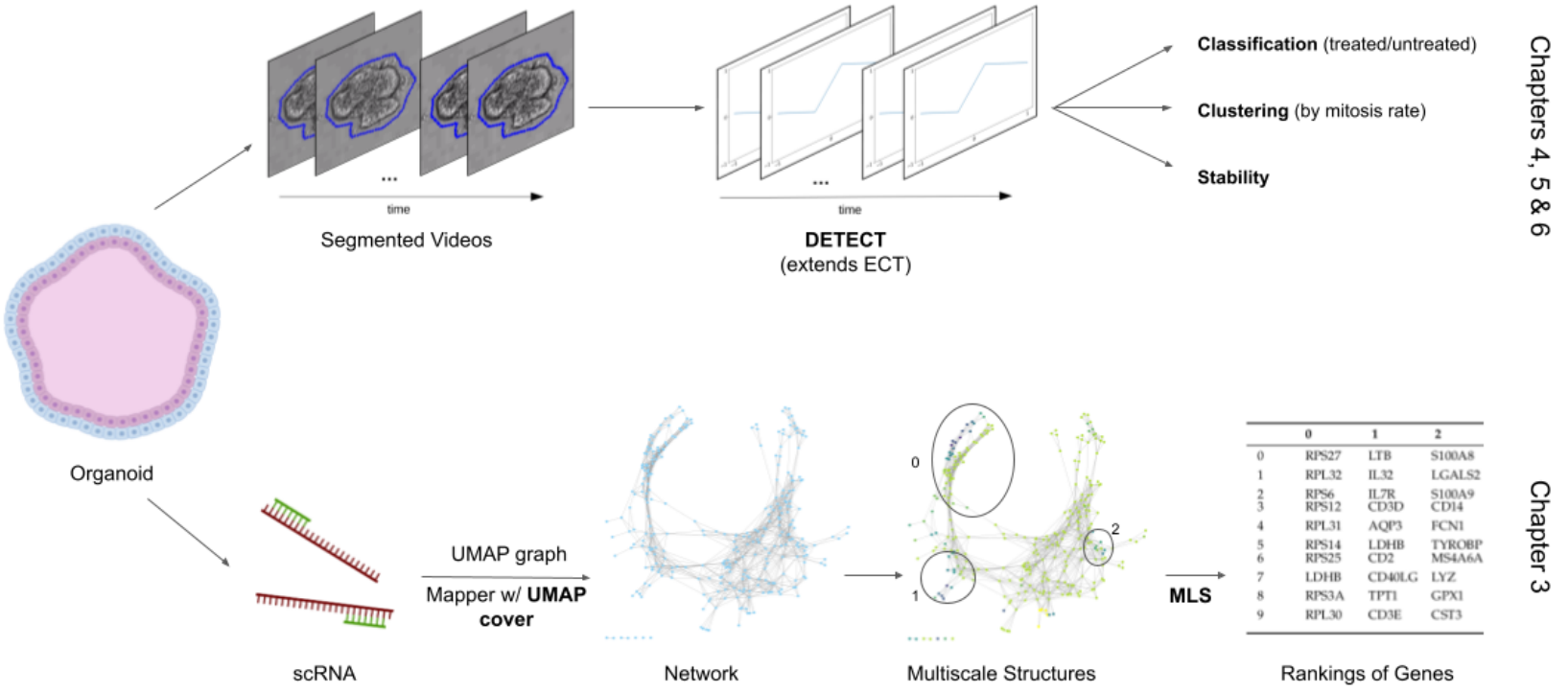


Figure I.1: Graphical abstract. I analyse two types of organoid data in this thesis: Organoid boundaries segmented from videos (top left) and single-cell RNA sequencing data (bottom left). The former is analysed using DETECT, a novel signature proposed in this thesis which extends the Euler characteristic transform. DETECT can classify experimental organoids into treated and untreated groups based on their morphology. These results are verified on synthetic organoid data, which DETECT clusters by mitosis rate. We can prove that the ECT and DETECT are stable (top right). The scRNA-seq data can be represented by a UMAP graph or a simplified Mapper graph. We can identify genes which are expressed consistently with this network structure at multiple scales using the novel multiscale Laplacian score (MLS; bottom right). **New contributions are in bold.**

List of Publications and Manuscripts

Peer-reviewed publications

1. Hoekzema RS, **Marsh L**, Sumray O, Carroll TM, Lu X, Byrne HM, Harrington HA. 'Multiscale Methods for Signal Selection in Single-Cell Data'. In: *Entropy* (2022) 24:1116. `https://doi.org/10.3390/e24081116`.

   Contributions: Developed and implemented the multiscale Laplacian score and analysed its output on two data sets.

2. **Marsh L**, Dufresne E, Byrne HM, Harrington HA. "Algebra, Geometry and Topology of ERK Kinetics". In: *Bull Math Bio* (2022) 84:137. `https://doi.org/10.1007/s11538-022-01088-2`.

   Contributions: Proposed and implemented the topological metric on posterior distributions. Developed and implemented the Bayesian data analysis. Implemented the simulations suggesting structural non-identifiability. Contributed to the proofs and implementations of the structural identifiability tests. Derived the algebraic model reductions and contributed to the QSSA model reductions.

3. Yeung E, McFann S, **Marsh L**, Dufresne E, Filippi S, Harrington HA, Wühr M, Shvartsman SY. "Inference of Multisite Phosphorylation Rate Constants and Their Modulation by Pathogenic Mutations". In: *Current Biology* (2019) 30:5. `https://doi.org/10.1016/j.cub.2019.12.052`.

   Contributions: Developed and implemented the Bayesian data analysis.

Submitted to peer-reviewed journal

1. **Marsh L**, Zhou FY, Qin X, Lu X, Byrne HM, Harrington HA. 'Detecting Temporal shape changes with the Euler Characteristic Transform'. Under review at: *Transactions of Mathematics and Its Applications*. Preprint: arXiv:2212.10883.

   Contributions: Developed and implemented DETECT and the pipeline it is used in in the above paper. Applied DETECT to both experimental and synthetic data.

2. Zhou FY, Jacobs BA, Han X, Ruiz Puig C, Carroll TM, Chadwick J, Qin X, Lisle R, **Marsh L**, Byrne HM, Harrington HA, Zhou L, Lu X. 'Modelling and detecting cellular plasticity with SAM (Shape, Appearance and Motion)'. Under review at: *Cell*. Preprint not available.

Contributions: Conducted the ECT analysis of organoids and contributed to the ECT being included and implemented in SAM.

3. **Marsh L**, Beers D. 'Stability and Inference of the Euler Characteristic Transform'. Submitted to: *Discrete and Computational Geometry*. Preprint: arXiv:2303.13200.

   Contributions: Derived and wrote all sections relating to the statistical inference of the ECT. Created all figures and designed and implemented all simulations.

Chapter 1

Introduction

Chapter Content

1.1 Topological Data Analysis 3
1.2 scRNA-seq Data 4
1.3 Organoid Morphology 6
1.4 Theory of the Euler Characteristic Transform 8
1.5 Structure 9

Organoids are *in vitro* cell cultures that mimic certain functions of mammalian tissues. They are also known as mini organs because their constituent cells can differentiate into various cell lineages with defined cellular functions. The increasing use of organoids for studying tissue development, disease progression and tissue responses to genetic and environmental perturbations can be attributed to their ability to recapitulate the three-dimensional (3D) cellular architecture and function of their tissue of origin [87]. This architecture distinguishes organoids from earlier two-dimensional (2D) tissue cultures. In particular, due to their 3D similarity to the tissue from which they were derived (the primary tissue), organoids model tissue-specific functions, such as signalling pathways of the primary tissue more accurately [148, 123]. As such, they can be used to study the impact of defined genetic and environmental manipulations *in vitro*. Analysing organoid responses to these manipulations makes them useful for understanding the development and disease progression as well as for drug testing

[28] while decreasing the cost and ethical implications of testing on animal or human subjects. Organoids also have the potential to play an important role in the future of precision medicine: For example, growing organoids derived from defective tissue of a patient and applying various drugs to such organoids could reveal which treatment promises the greatest therapeutic benefit. While organoids can now be generated to study multiple organs, including the brain, kidney and liver, the most widely studied organoids are derived from the intestinal epithelium, one of the fastest renewing mammalian tissues [123].

Organoids are grown from stem cells. Growth factors present in the culture medium (e.g. Matrigel) surrounding the organoids drive the stem cells to proliferate and the organoids to increase in size. During this early stage, many cells exhibit pluripotency (the ability to develop into any cell type found in a tissue). In response to developmental cues, these pluripotent cells migrate to their destined site in the organoid and differentiate to perform their function and maintain homeostasis (broadly: the functioning of the tissue/cell culture). Cell differentiation leads to various areas of an organoid being occupied by different cell types with distinct functional properties. At the organoid level, this process leads to changes in morphology. At the cell level, changes in the number of cell types and their spatial positions affect cross-talk and, by extension, molecular processes within the cell, such as gene expression and transcription.

Genetics, including cancer mutations, are known to affect tissue composition and thus cell fate, which in turn affects organoid morphology. Thus, tissue composition, which is closely linked to tissue morphology, and genetics are tightly coupled. Unravelling interactions between morphology, genetics and treatments is thus an essential aspect of understanding tissue fate and disease progression. For example, metaplasia,[1] an early stage of oncogenesis, requires an understanding of both tissue composition and genetics. Metaplasia is often linked to irregular cell morphology and cell plasticity

[1] 'A change of cells to a form that does not normally occur in the tissue in which it is found.' according to the National Cancer Institute. `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/metaplasia` (last checked: 21/02/23).

[163]. As it requires the de-differentiation of cells, it also affects gene transcription and expression.

Current methods for both analysing organoid morphology and transcriptomics (in the form of single-cell RNA sequencing data) are incomplete. For both types of data, I propose new methods using topological data analysis (TDA) to improve analysis methods for such data. In the case of morphology data, I show that the method I developed is scalable, interpretable and has high discriminative power. Most existing methods lack at least one of these properties. Similarly, I compare my methods for scRNA-seq data with existing methods on benchmark data and demonstrate how it yields new insights and identifies genes not found previously. All of the methods I propose are motivated by organoid data but generalise to other types of data.

1.1 Topological Data Analysis

Topological data analysis is a set of methods which use topology to study the shape of data. TDA has been successfully applied in biology [51, 29, 61, 101, 110]. In particular, it has been used to study morphology [3, 39, 155] and to analyse high-dimensional sequencing data of genetic material [60, 118, 122]. Related to TDA is network theory, as any network can be viewed as a special instance of a simplicial complex, a key notion in TDA. In particular, random walk theory on networks, the main method from network theory used in this thesis, generalises from networks to simplicial complexes [128]. Random walks on networks have been successfully applied to a number of research problems in biology [124, 36, 150, 8].

In this DPhil thesis, I demonstrate how TDA methods are effective in analysing organoid morphology and single-cell RNA-seq data at multiple resolutions. The topological analyses, presented across separate chapters, increase the understanding of the relationship between morphology and genetics. While integrating scRNA-seq and morphology data of organoids into a single study is beyond the scope of this DPhil project, I present a roadmap towards such an integrated study in the discussion chapter.

1.2 scRNA-seq Data

Single-cell sequencing methods provide transcriptomic data at an unprecedented resolution: unlike previous bulk sequencing methods, which aggregate the genetic material from multiple cells into one data point, single-cell sequencing methods generate one data point per cell. This increase in resolution has also affected how researchers view notions of cell type. While previously cell types were viewed as discrete - or binary - (e.g., a cell could be either completely 'cancerous' or completely 'healthy'), single-cell sequencing has revealed many intermediate cell states, which can be viewed as a continuous trajectory from one cell type to another [122].

A class of methods for analysing transcriptomic data are differential expression (DE) tests. In the over-simplified, but nonetheless illustrative example of a data set containing points derived from 'cancerous' and 'healthy' cells, DE tests use statistics to identify genes which are expressed at significantly higher levels on the 'cancerous' than the 'healthy' cells (or vice-versa). However, in many instances of scRNA-seq data sets, it is impossible to stably assign cells into such discrete categories. In practice, clustering algorithms are often used to assign cells to cell types before a DE test is conducted. If the structure underlying the data is continuous, such clustering is unstable (with respect to small perturbations of the data, subsampling, small changes in the parameters of the algorithm, etc.). By extension, any DE analysis that builds on such clustering is also unstable.

Govek and co-authors proposed a generalised DE test using the Laplacian score [71] on a cell-similarity network, which assumes that the sequencing data has a continuous structure [60]. They model the continuous structure using simplicial complexes, a key notion from TDA. In particular, their method does not require the data to be clustered and thus does not suffer from clustering-related instabilities. Their test can also be applied to spatial-transcriptomic data (RNA data annotated by spatial coordinates of a cell within a tissue or cell culture) to identify genes highly expressed in particular spatial regions.

A drawback of using the Laplacian score for generalised DE testing is that it

operates only at a single scale (both in terms of gene expression similarity and spatial distance). However, many data sets exhibit structure at multiple scales. In our simple example, the 'cancerous' and 'healthy' cells may comprise multiple sub-groups (e.g. epithelial, muscular, neuronal and connective cells). In such cases, clustering can be performed at multiple resolutions (many clustering algorithms, including k-means and single-linkage, have a parameter controlling how coarse the resulting clustering is) and the DE test is applied to each clustering to obtain a multiscale description of the data.

In this thesis, I propose a novel *multiscale Laplacian score* (MLS) which generalises the Laplacian score analysis introduced by Govek et al. [60]. The MLS uses insights from network theory and random walk theory to perform generalised DE tests at multiple resolutions. Further, I propose to use variation of information (VI), a heuristic commonly used in multiscale community detection on networks, to detect resolutions at which the data exhibits interesting structures. While this heuristic is motivated by community detection/clustering methods, the MLS itself, like the Laplacian score, does not use any assignment of cells into clusters.

I apply this method to two data sets: a benchmark data set of lung-tumour infiltrating human T cells [91] and a data set collected from mouse colon organoids which have various cancer-related mutations induced. First, I demonstrate that VI identifies resolutions of interest in both data sets. On the mouse colon data, I show that the identified resolutions relate to cell types identified by biomarkers and to genetic conditions, respectively. Second, I show that the MLS identifies differentially expressed genes (which are known to be involved in various types of cancer) at each resolution in both data sets and have not been identified by the Laplacian score.

Trajectory inference methods, unlike DE testing, assume that the structure of single-cell data is continuous. They attempt to infer continuous trajectories between different cell states. The performance of such methods varies significantly across data sets and depends on their topology [125]. Further, many methods are biased either through the limited topology they can model or the large number of hyperparameters which need to be set by the user [125]. A method from TDA that has been used

successfully for trajectory inference [122, 118] is Mapper [134]. Mapper builds a graph representation of high-dimensional data by adapting the concept of a Reeb graph, a notion from pure topology. However, the cover, a key hyperparameter in the Mapper algorithm, has been chosen manually rather than by a biological or computational principal in these studies. The results therefore may be biased.

I propose a novel heuristic that enables the cover in Mapper to be specified algorithmically. This heuristic, called the *UMAP diffusion cover*, is again inspired by network theory and random walks. I illustrate that, in conjunction with Mapper, it returns promising results on the two aforementioned data sets by comparing it to the state-of-art trajectory inference method PAGA [160].

1.3 Organoid Morphology

Organoids in their initial stages of growth comprise clusters of stem cells and are typically approximately circular in their shape (in the top-down, 2D view). As they grow, they undergo morphogenesis, which is the emergence of increasingly complex geometric and topological structures [78]. Morphogenesis typically proceeds as a series of size and shape changes and topological transitions [78, 107]. These transitions play a key role in tissue function and are perturbed in a number of pathological conditions [78, 11]. However, the mechanisms underlying these transitions are not well understood [78]. Quantifying changes in morphology in detail and understanding their connections with genetics and disease progression would yield an increased understanding of the underlying mechanisms and give a fast, inexpensive method to infer the genetics and the tissue health of an organoid.

Therefore, there is a need to study the morphology of organoids and how it changes over time. Ideally, any measure of the morphology should be interpretable and yield a meaningful distinction between different morphologies induced by different genetic conditions and environmental stimuli. Further, to accommodate the ever-increasing size of data sets and to allow for future use in precision medicine, any method should also allow for high throughput. I.e., the method should have sufficiently low com-

putational complexity, but also should not depend on the staining of different tissue regions or on any other manual annotation of the data.

Several studies have focused on quantifying and analysing organoid morphology. Simple measures, such as cell numbers, organoid volume and surface area, diameter, shape factor (ratio of surface area to volume), and growth rate have been used to relate morphology, genotype and drug responses [27, 69, 84, 161]. Furthermore, deep learning methods can segment organoid images and extract morphological features including organoid perimeter and eccentricity [62, 83]. At the same time, mechanistic models have been developed to investigate the relationship between stem cell proliferation, cell fate specification, organoid growth and morphology. These agent-based and continuum models have been compared with growth curves derived from experimental data [69, 140, 161]. However, these approaches lack either discriminative power (simple measures, which I illustrate in Chapter 4), interpretability (deep learning approaches) or scalability (model-based approaches). The complexity of organoid morphology lends itself to more sophisticated analysis. For example, genus and average curvature [78], measures from geometry and topology, have been used to distinguish shape changes in organoids. Here, I propose studying the geometry and topology of organoids with topological data analysis.

The Euler characteristic transform (ECT) [146] is a method from TDA that compares shapes embedded in Euclidean space. It is a sufficient statistic of shapes (i.e., any two distinct shapes give distinct ECT signatures), is fast to compute and amenable to further statistical analysis. Through its rigorous motivation by notions of pure topology, it is also easy to interpret.

In this thesis, I extend the Euler characteristic transform so that it can analyse shapes that change over time and it is rotationally invariant. The temporal evolution of shape is key to understanding morphogenesis. Rotational invariance is necessary as the rotation of an organoid only depends on non-biological factors, such as the initial placement of an organoid inside a well. Making the ECT rotationally invariant is inspired by a theorem of Curry et al. [41]. To the best of my knowledge, this work is the first instance of this theorem being turned into a computable signature. I call my

method DEtecting Temporal shape changes with the Euler Characteristic Transform (DETECT).

I then show on a data set of experimental mouse small intestine organoids (segmented videos of the experiments) that it is possible to regress a number of classical shape descriptors (e.g. diameter, major/minor axis lengths, area, convex area) from the ECT with high accuracy. Further, I show that DETECT can accurately classify organoids into treated and untreated groups. By contrast, the aforementioned collection of classical shape descriptors is unable to classify these organoids with an accuracy that exceeds guessing. We conclude that DETECT outperforms classical shape descriptors in analysing the morphogenesis of organoids.

Next, I validate DETECT on synthetic data. This data set has been generated by a mechanistic model describing organoid growth [161]. I show that on this data, in which we have complete control over model parameters, DETECT clusters organoids by parameter values. Further, I illustrate how DETECT generalises to 3D data. This is particularly relevant as it is becoming easier to perform 3D segmentations of experimental data [78].

1.4 Theory of the Euler Characteristic Transform

While the ECT is a sufficient statistic on a broad class of shapes [41, 59] and is fast to compute, it is not stable with respect to standard metrics. Even small perturbations to the input can lead to large distortions in the output. Hence, while the ECT encodes all information about a shape, it may also encode large amounts of noise.

In joint work with David Beers, a DPhil student in the TDA research group at Oxford, I have proposed a metric on the space of shapes with respect to which the ECT is, in fact, stable. This metric is non-standard in the sense that it also depends on arc lengths and curvature. I show that, in the presence of ambient Gaussian noise, a smoothing procedure on a simplicial complex, relying merely on taking weighted averages of vertices, leads to probabilistic convergence in the metric we proposed.

As the ECT is continuous in this metric, the smoothing procedure I propose gives a consistent estimator of the ECT.

I do not apply this estimator to the experimental organoid data as the organoid boundaries are already smoothed by the segmentation algorithm. However, the result on the consistent estimation of the ECT supports the analysis of organoid data as it indicates that smoothing by methods as simple as weighted averages efficiently removes noise from the ECT signature.

1.5 Structure

This thesis is structured as follows. In Chapter 2, I give the background on the biology of organoids and key concepts of single-cell RNA sequencing (scRNA-seq) data. I then review relevant methods of topological data analysis and network theory and existing work on scRNA-seq and morphology analysis of organoids. In Chapter 3, I introduce the multiscale Laplacian score and UMAP diffusion cover, apply them to two scRNA-seq data sets and discuss the findings. Further, in Chapter 4, I introduce DETECT and build a pipeline for organoid morphology analysis around it. I use this pipeline to show that the DETECT outperforms classical shape descriptors on an experimental data set in Chapter 4 before using synthetic data to validate DETECT and to illustrate how it generalises to 3D data in Chapter 5. The theoretical contribution to the statistical estimation of the ECT follows in Chapter 6 before the thesis concludes with a discussion of the key findings and possible future work in Chapter 7.

Chapter 2

Background and Literature Review

Chapter Content

2.1 Organoids 11
2.2 scRNA Sequencing Data 14
2.3 Previous Studies on Organoid scRNA-seq Data 16
2.4 Previous Studies on Organoid Morphology Data 17
2.5 Mathematical Preliminaries 18
2.5.1 Mapper 19
2.5.2 UMAP 22
2.5.3 Random Walks and Community Detection 26
2.5.4 Simplicial Complexes and Filtrations 33
2.5.5 The Euler Characteristic and Euler Characteristic Transform . 35
2.5.6 Kernels and Kernel Approximations 38
2.6 Summary 42

I start this chapter by introducing the notion of an organoid. First, I explain stem cells, which are the cells organoids are derived from, and how organoids are grown. I further explain why organoids are useful in cancer research and present examples of their use (Section 2.1). Next, I describe the relevance of RNA data and single-cell methods (Section 2.2). Morphology and scRNA-seq data are the two types of organoid data analysed in this thesis. Therefore, I review existing studies of scRNA-

seq organoid data (Section 2.3) and organoid morphology data (Section 2.4). Finally, I introduce mathematical notions that are of relevance to several methods introduced or several chapters of this thesis (Section 2.5).

2.1 Organoids

Organoids are cultured from stem cells *in vitro* and mimic the functions of mammalian tissues [28]. To understand organoids in general and the types of organoid data (shape and RNA sequencing) I analyse in this thesis in particular, we first need to understand the function of stem cells within a tissue.

Stem cells are defined by two properties: a stem cell can reproduce itself (it can undergo *mitosis*) indefinitely and a stem cell can generate daughter cells of a different functional cell type (this process is called *differentiation*) [136]. Differentiated cells, by contrast, can be characterised by the genes which are transcribed in that cell. After a stem cell differentiates, the cell type of a cell can differentiate further. Cells which can change their cell type further are called *progenitor cells*. While progenitor cells can differentiate into a cell type different to their own, they cannot self-replicate indefinitely. We call the sequence of cell types along which a cell differentiates a *lineage* and the cells at the end of a lineage a *matured cells*.

Tissues in which dead cells are replaced by the differentiation of stem cells are called *renewal tissues*. Examples of renewal tissues in the human body include the skin, the liver, kidneys, the colon, blood and the brain. However, not all tissues are renewal tissues. E.g., muscles, the heart or bones do not renew via stem cells and their differentiation. While differentiation is an effective way for a tissue comprising diverse cell types to self-renew, differentiation also increases the risk of mutation of cells. For this reason, human cancers are almost exclusively found in self-renewal tissues [136] and, therefore, the study of cancer includes the study of stem cells.

Stem cells are not uniformly distributed across self-renewal tissues. They are typically located in micro-environments called *niches*. The environment in niches is favourable to the persistence and functions of stem cells [136]. Niches make stem

cells available to external stimuli, which includes both cell-to-cell and cell-to-matrix signalling [54].

To grow an organoid, stem cells are first extracted from live tissues and then placed in a medium (such as matrigel) to which growth factors are added. At this stage, the organoid is round (circular in the 2D top-down view) as its cellular composition is typically homogeneous. The stem cells then proliferate and differentiate and self-organise during morphogenesis, i.e. the cells spatially rearrange to form distinct micro-environments, including stem cell niches [123]. This rearrangement is driven by developmental cues, which are system-autonomous mechanisms; i.e., a spatially non-uniform distribution of cell types forms, even if the cells are exposed to a spatially uniform signalling environment [123]. Self-organisation usually occurs via a sequence of self-patterning events, typically starting with a symmetry-breaking event. Several symmetry-breaking mechanisms have been proposed and all of them involve positive and negative feedback loops in sub-cell signalling pathways [123]. Cell rearrangements are further mediated by physical cell-to-cell interactions and are thus affected by adhesion, cortical contractility, cortical tension and cell motility [123].

Genetics affect cell signalling and cross-talk between signalling pathways and, by extension, tissue composition, self-patterning and symmetry-breaking. Conversely, the tissue composition and the distribution of cell types affect cell signalling and cell microenvironments, thereby altering the cells' ability to proliferate and their molecular functions. Hence, tissue composition, self-patterning and genetics constantly affect each other. Unravelling the interplay between morphology, genomics and treatments is thus a key aspect of understanding tissue fate and disease progression. For example, metaplasia, an early stage of oncogenesis, requires an understanding of both tissue composition and genetics. Metaplasia is often linked to irregular cell morphology and cell plasticity [163]. As it requires the de-differentiation of cells, it also affects gene transcription and expression.

What distinguishes organoids from traditional 2D mammalian tissue cultures is that they more closely resemble the tissue they were derived from in terms of 3D tissue composition rather than merely in terms of 2D composition. In particular,

organoids recapitulate tissue-critical features in terms of architecture, differentiated cells and tissue-specific function. Thereby, organoids can be compared to traditional genetically engineered mouse models, cell lines and patient-derived xenografts [148]. Unlike organoids, all of these tissue models are *in vivo* and come with higher costs and ethical hurdles. Organoids generally take less time to culture and are usually cheaper to culture compared to these tissue cultures, while maintaining good success rates [148].

These favourable features of organoids have been exploited to address a number of important research questions. Firstly, organoids have been used to study normal tissue development and the development of cancer (carcinogenesis) [148]. For example, a series of studies [99, 49, 56] used organoids to show that a specific sequence of mutations leads to the developmental independence of niche-specific signals in colorectal cancers. These studies also showed that this developmental independence directly facilitates tumour growth, migration and metastatic colonisation. Organoids were also xenotransplanted into mice for verification, where they established invasive colorectal cancers [56]. In similar studies, organoids were used to highlight the roles of specific genes in Barrett's oesophagus [97] and colorectal cancer [93].

Secondly, organoids have also been used to identify and study the properties of the tumour microenvironment for immune therapy in cancer. While tumour organoid models generally lack intact microenvironments, recent findings have shown that co-cultured organoids (i.e., multiple organoids grown in the same well) accurately replicate some aspects of the tumour microenvironment [148]. For example, organoid co-cultures can accurately mimic aspects of the tumour microenvironment of pancreatic carcinomas [109, 15]. Further, it has been demonstrated that patient-derived organoids from specific tumours can be cultured together with peripheral blood lymphocytes from the same patients, which then are able to generate T cells [46]. In principle, such co-cultures can help optimise the response of effector T cells against the patient's neoplastic cells [148].

Finally, organoids promise important advances in precision medicine. In some limited cohort clinical studies, the response of patient-derived organoids (PDOs) largely

replicated the initial response of these patients to the same treatments [96, 72, 141, 154, 113]. If these findings are verified in larger trials, organoids could help to identify the most effective treatment options for an individual cancer patient (by applying several treatments to several PDO cultures and identifying the best response). At present, such an approach could only help with optimising second-line or adjuvant therapies, as it typically takes 4-6 weeks to derive PDOs [148]. In either case, organoids can be used to identify biomarkers indicative of drug response in a given patient. Tiriac et al. [141] compared standard cytotoxic drug responses *in vivo* to the drug responses of *in vitro* PDOs that had the same drugs applied. By sequencing the cells of the organoids, they thereby managed to derive a transcriptional signature of common responders to different chemotherapies. While it is unclear whether this transcriptional signature reflects differences in drug response or drug pharmacology, the signature has been shown to correctly identify a patient sub-group with improved therapeutic response.

2.2 scRNA Sequencing Data

Since single-cell sequencing was named 'Method of the Year' in 2013 by Springer Nature, it has become a benchmark for investigating the genetic heterogeneity of tissues and other cell cultures [90]. In this section, I compare RNA sequencing with DNA sequencing and protein expression data and single-cell methods with bulk measurements. More detail on the structure and typical size of the data as well as the pre-processing methods I use is given in Chapter 3.

Why RNA?

Genetic research often compares genotype and phenotype. Genotype refers to all of the DNA of an organism. Phenotype refers to all observable traits of an organism. In the context of molecular biology, these traits are primarily determined by protein expression, although other traits, such as morphology, may be considered as well.

DNA, which forms the genotype of an organism, is the underlying blueprint for all cellular processes. If one of these processes is needed, RNA copies relevant sections of the DNA, which in return are translated into proteins (see Figure 2.1). Hence, unlike DNA, RNA provides a snapshot of the processes active in a cell at a given time. RNA, therefore, lies at the interface between geno- and phenotype: on the one hand, it is a transcript of a DNA segment, on the other hand, it determines gene expression and as such directly contributes to the phenotype.

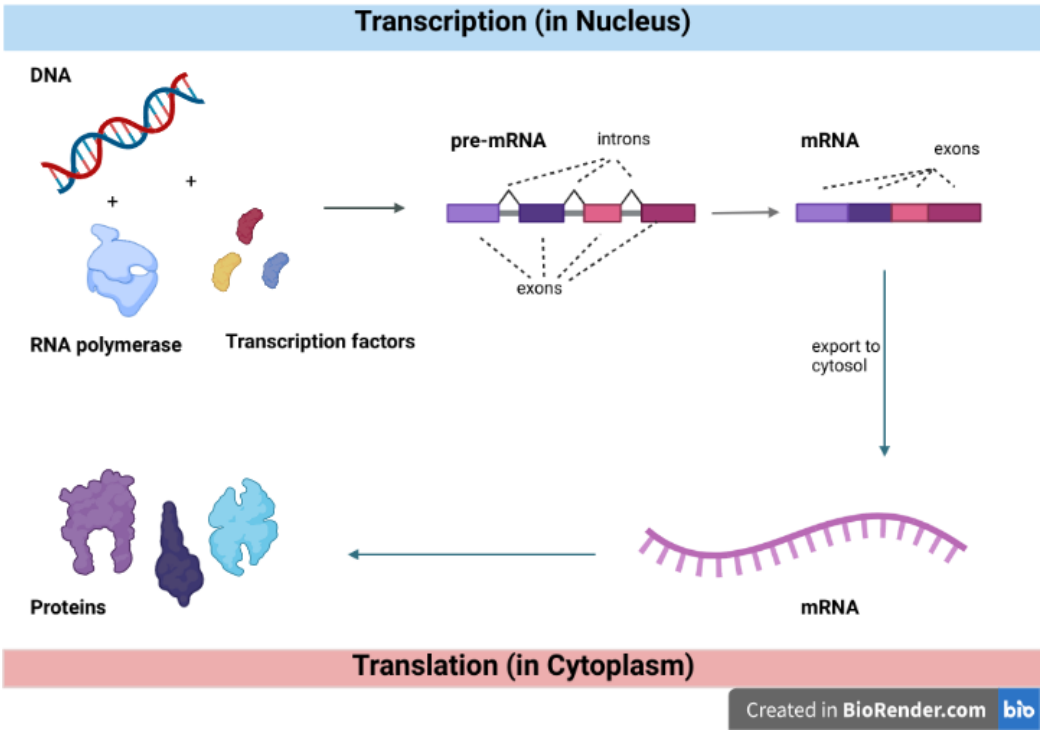


Figure 2.1: In the first step of transcription (of mRNA), RNA polymerase and transcription factors bind to the DNA strand in the nucleus of the cell (top left). The polymerase opens the DNA double-strand to allow the copying of a section of one side of the strand. The copied single strand is called pre-mRNA (top centre). The pre-mRNA contains sequences of nucleotides which get removed by the cell (called introns) and sequences which are not (called exons). The process of removing introns is called splicing and turns pre-mRNA into (mature) mRNA (top right). The mRNA is then exported to the cytoplasm (bottom right), where it is translated into proteins (bottom left).

For example, diseases with a genetic link can be caused by a non-functional enzyme, which can be the result of incorrect RNA splicing (see Figure 2.1). In this case, the mutated and original enzymes derive from the same DNA. In such a scen-

ario, RNA sequencing gives more detailed information about the underlying issue than protein expression levels while DNA sequencing would yield little insight.

Why Single Cells?

Until the development of single-cell sequencing (sc-seq) techniques, genetic material was sequenced in bulk, i.e. fragments of DNA or RNA strands of many cells were sequenced simultaneously and aggregated to produce a data point. While there are advantages to bulk sequencing data (e.g., it is typically less sparse as it is less susceptible to so-called dropout effects), sc-seq enables genetic analyses at a finer resolution, at the cell level. Sc-seq allows for cell-type identification, the arrangement of cell populations into hierarchies and the identification of cells transitioning between states [90] to be distinguished. Analysis of sc-seq data provides a detailed view of tissue development and its underlying dynamics [90].

2.3 Previous Studies on Organoid scRNA-seq Data

Since the first applications of scRNA-seq technology to organoids were published, such as Gruen et al.'s study on mouse small intestine organoids in 2015 [63], a vast array of similar studies have followed. I point the interested reader to [162] for a comprehensive review. A common problem with early studies was that the number of cells harvested from organoids was relatively small compared to the high-throughput scRNA-seq methods are designed for ($>$ 1000 cells), which may make them inaccurate [26]. Optimised procedures for lower throughput have been developed [26] and the number of cells sequenced has also increased in later organoid studies [22] (both data sets analysed in this thesis contain well over 1000 cells).

ScRNA-seq technology applied to organoids can yield important insights into how similar an organoid is to primary tissue in terms of cellular composition and subsequently help to improve protocols for growing organoids. Similarly, it allows for *in vitro* testing cell-specific responses to environmental variables, such as treatments or

changes in the microenvironment [22]. Examples of successful applications include a number of brain organoid studies [30, 117], where scRNA-seq methods showed that organoids derived from different areas of the brain accurately model the cellular composition of the primary tissue. Brain organoids have also been used to model a neurodevelopmental disorder [16]. Here, scRNA-seq verified that the iPSC lineages in organoids were compatible with the expected disease phenotype.

The applications of scRNA-seq to other types of organoids are similar: scRNA-seq was used to optimise the protocols for growing mouse small intestine [63] and liver organoids [31]. Notably, a cell atlas for mouse small intestine organoids has been created [65]. In kidney organoids, an scRNA-seq study has shown that the mapping of disease-related genes from primary tissue to organoids is consistent, but while the cellular composition of primary tissue and organoids is similar, cell types occur in different proportions and stages of maturation in the organoids [112]. Tiriac et al. [141] compared standard cytotoxic drug responses *in vivo* to the drug responses of *in vitro* PDOs that had the same drugs applied. They used scRNA-seq methods on organoids to identify biomarkers indicating a positive drug response (see Section 2.1).

2.4 Previous Studies on Organoid Morphology Data

Several studies have focused on quantifying and analysing organoid morphology. Simple measures, such as cell numbers, organoid volume and surface area, diameter, shape factor (ratio of surface area to volume), and growth rate have been used to relate morphology, genotype and drug responses [27, 42, 69, 84, 161]. OrganoSeq provides a software package that segments organoid images and extracts simple measures from these segmentations [19].

Furthermore, deep learning methods can segment organoid images and extract morphological features including organoid perimeter and eccentricity [62, 83]. OrganoID is a software package that uses deep learning to segment and track organoids

across a whole video [100]. Abdul et al. [1] used image-classification deep neural networks to classify images of human colon organoids into opaque/non-opaque and budding/non-budding classes with high accuracy.

Beck et al. have studied the morphological properties of organoids at the cell level [7]: They measure the number and size of cells and the size of cysts. In a data set of Madin-Darby canine kidney cysts organoids, they collect the above data for organoids exposed to different genetic perturbations and observe a number of constraints. These constraints vary with age, genetics and applied treatments and growth factors [7]. While their data is partially collected by software, it requires manual correction and thus does not allow for high throughput.

Furthermore, mechanistic models have been developed to investigate the relationship between stem cell proliferation, cell fate specification, organoid growth and morphology. These agent-based and continuum-based models have been compared with experimental data using simple growth curves [69, 74, 140, 161]. The complexity of organoid morphology lends itself to more sophisticated analysis. For example, genus and average curvature [78], measures from geometry and topology, have been used to distinguish shape changes in organoids. Bremond-Martin et al. used Vietoris-Rips filtrations, a common TDA method, together with vectorisations of persistence diagrams to cluster organoids by developmental stage [25].

2.5 Mathematical Preliminaries

Throughout this section, we will use the concept of a graph:

Definition 2.1. Let V be a finite set of vertices. A *graph* is a tuple $G = (V, E, w)$, where $E \subset V \times V$ are called edges and $w : E \to \mathbb{R}_{>0}$ is a weight function.

We call a graph *undirected* if $(a, b) \in E$ whenever $(b, a) \in E$ and call it *directed* otherwise.

We call a graph *unweighted* if $w(e) = 1$ for all $e \in E$ and will possibly omit w from the definition of G in this case. We call a graph *weighted* if it is not unweighted.

For a given graph, consider $\mathbb{R}^V$, the free vector space generated by V. The adjacency operator of G is the linear operator $A : \mathbb{R}^V \to \mathbb{R}^V$ defined by

$$A(v) = \sum_{(v,v') \in E} w((v, v')) \cdot v'.$$

We usually consider the matrix representation of A we get by considering the canonical basis on $\mathbb{R}^V$, which is called the *adjacency matrix*.

2.5.1 Mapper

A topological method used to analyse scRNA-seq data in Chapter 3 is Mapper [134], which was developed by Singh et al. in 2017. Mapper is commonly used to summarise topological properties, in particular the connectivity, of high-dimensional data sets. It has been successfully used for trajectory inference in scRNA-seq data [118, 122]. It is theoretically well-motivated and is able to model a wide range of different topologies in data [122].

Mapper is based on the so-called *Reeb graph*, a notion from pure topology which tracks the topological evolution of level sets of a topological space endowed with a real-valued function:

Definition 2.2. Let (X, f) be a pair of a topological space X and a continuous function $f : X \to \mathbb{R}$. Define the set of points $G_{(X,f)} = \{\pi_0 (f^{-1}(c)) \mid c \in \mathbb{R}\}$, where $\pi_0(Y)$ denotes the set of path-components of a topological space Y. Furthermore, define $F : X \to G_{(X,f)}$ as the function mapping each $x \in X$ to the unique path-component in which it is contained. We call $G_{(X,f)}$, endowed with the quotient topology induced by the surjection F, the Reeb graph of the pair (X, f).

As shown in Figure 2.2, a Reeb graph is often topologically equivalent to the geometric realisation of a graph. Indeed, the Reeb graph in Figure 2.2 captures that the original topological space is path-connected and contains at least one non-contractible loop. Furthermore, each point in the topological space can be projected onto the Reeb graph (see the height-lines on the torus in Figure 2.2, each height-line mapping to a unique point on the Reeb graph).

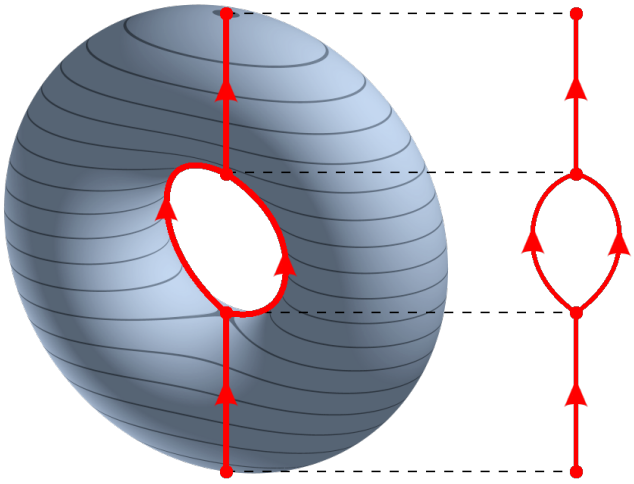


Figure 2.2: A Reeb graph (right) extracted from a torus (left) with the z-axis used as a lens function (illustrated by horizontal height lines). Taken from `https://en.wikipedia.org/wiki/Reeb_graph#/media/File:3D-Leveltorus-Reebgraph.png` (last checked 10/05/2023).

When considering real-world data, it is common to assume that data points are sampled from a manifold or a similarly well-behaved topological space embedded within some ambient space (usually $\mathbb{R}^N$). Mapper approximates the Reeb graph of the underlying topological space from a finite, discrete sample. Working only with a discrete sample $\{x_i\}_{i=1,\dots,n} \subseteq \mathbb{R}^N$, rather than the continuous space, gives rise to two challenges:

1. Consider a real-valued function f, called the *lens*, defined on the ambient space and, by restriction, on the intrinsic topological space and the sample. By the finite nature of the sample, it is unlikely that $f(x_i) = f(x_j)$ for two distinct sample points.

2. Even if point 1. were resolved and samples were mapped to the same value by f, none of these samples would lie in the same path-component, due to the discrete nature of the sample.

Both issues imply that, generically, the Reeb graph of a discrete sample is (topologically equivalent to) the sample itself. Thereby, the Reeb graph of the sample

does not yield a simplification of the data. I now describe the Mapper algorithm by explaining how it resolves the above two issues. Other challenges, including how to choose f and other parameters, will be addressed in a later chapter of this thesis (Section 3.2.2), as their resolution requires the context of data.

To address point 1., Mapper does not ask whether $f(x_i) = f(x_j)$ exactly for two distinct sample points, but whether $f(x_i)$ and $f(x_j)$ both lie in a common open interval out of a pre-defined collection of intervals. Instead of looking at the pre-image of a continuously varying c, Mapper considers the pre-images of a sequence of overlapping intervals, U_i say. More formally, Mapper takes a finite open cover $\mathcal{U} := \{U_i\}_{i \in A}$ of the image of the data under f as one of its inputs, where A is some indexing set. The pre-images $f^{-1}(U_i)$ are constructed to contain several points of a generic sample X. In practice, ensuring that the pre-images $f^{-1}(U_i)$ contain a usable number of samples is an important consideration when choosing a cover $\mathcal{U}$.

We now consider point 2. Each pre-image $f^{-1}(U_i)$ is a discrete set of points. As for the full sample, we assume that this set of points approximates a continuous geometric object. To determine which points should be considered as lying in the same path-component of the underlying topological space, Mapper applies a clustering algorithm to each pre-image $f^{-1}(U_i)$. For each generated cluster, Mapper creates a node in the Reeb-graph approximation. Mapper is best suited to use density-based or hierarchical clustering methods (e.g. linkage clustering, DBSCAN, HDBSCAN, ToMaTo) whose notion of a cluster is more similar to a path-connected component (compared to centroid-based methods, such as k-means).

Finally, to model the notion of quotient topology used in the definition of a Reeb-graph on a discrete data set, Mapper connects any two nodes via an edge if their associated clusters share at least one data point. Note that clusters can share data points, as the cover elements U_i overlap and, thus, each point can lie in several pre-images and can be clustered several times.

To formalise the intuition described above, we begin by describing the input to the algorithm. Let $X := \{x_1, ..., x_n\} \subset \mathbb{R}^N$ be a finite sample and let $f : X \to \mathbb{R}^d$ be a function (note that the domain can be any Euclidean space $\mathbb{R}^d$ for the Mapper

algorithm). Let $\mathcal{U} := \{U_i\}_{i \in A}$ be a finite open cover of the image of the data, i.e. a finite collection of open sets U_i such that $f(X) \subset \cup_{i \in A} U_i$. Assume that CA is a clustering algorithm that is compatible with X.

Remark. Both f (often determined by a dimension-reduction algorithm) and CA have their own hyper-parameters. These should be considered as further inputs.

The Mapper algorithm consists of the following steps [134]:

1. For each cover element $U_i \in \mathcal{U}$, construct the pre-image $f^{-1}(U_i)$.
2. Apply CA to each pre-image $f^{-1}(U_i)$. Denote the resulting set of clusters by $\{CA(U_i)_j\}$ and add them as nodes to an output-graph G.
3. Iterate through all possible pairs of distinct nodes in G. If the intersection of a pair of clusters is non-empty, i.e. if $CA(U_i)_j \cap CA(U_{i'})_{j'} \neq \varnothing$, then add the edge $(CA(U_i)_j, CA(U_{i'})_{j'})$ to G.
4. Return G.

The algorithm is illustrated by the example presented in Figure 2.3.

2.5.2 UMAP

An important input for the Mapper algorithm is the filter function. For some data sets the choice of filter function is obvious, e.g. one component of the data which is especially important or some relevant metadata. Where data comes without an obvious filter, it is common to use a dimension-reduction procedure as a filter function. In this thesis, I typically use UMAP to filter the data.

Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) is a non-linear dimension reduction method developed by McInnes et al. [102]. UMAP draws heavily on insights from topological data analysis and, as such, lends itself well to being used as part of a Mapper pipeline. In this section, I give a brief overview of the motivation of UMAP and after introducing all relevant notions state the UMAP algorithm.

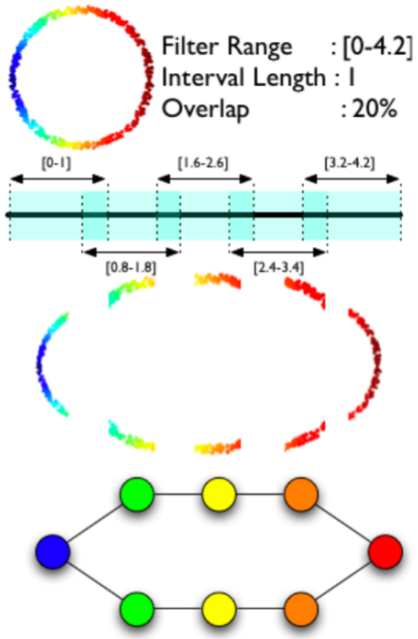


Figure 2.3: The Mapper algorithm applied to a random sample from S^1. The lens function f is indicated by the colouring of the sample. The algorithm uses a cover of the interval $[0, 4.2]$ with 5 intervals with 20% overlap each. Taken from [134].

In Appendix A.2, I present a more detailed motivation of the UMAP algorithm. In particular, I present a new version of the motivation presented by McInnes et al., which uses Vietoris-Rips filtrations, a central notion in TDA. This novel way of presenting the motivation for the UMAP algorithm further highlights the connection between UMAP and topological data analysis. I also present a new generalisation of the UMAP algorithm, which embeds a simplicial complex in low dimensional space, rather than a graph.

Manifold Hypothesis

Similarly to Mapper, UMAP assumes that the input data set has been sampled from an m-manifold $\mathcal{M}$ embedded into Euclidean space $\mathbb{R}^N$ and that, typically, $m \ll N$. Specifically, UMAP postulates that the data has been sampled from a Riemannian manifold $(\mathcal{M}, g)$ following the uniform distribution with respect to g. Notably, the embedding of $\mathcal{M}$ into $\mathbb{R}^N$ need not be an isometry with respect to g and the Euclidean metric; however, if g is constant on an open ball in $\mathcal{M}$, then the metric induced by g is

identical to a scalar multiple of the Euclidean metric in that neighbourhood (Lemma 1 in [102]). Therefore, the UMAP method locally approximates the unknown intrinsic metric g by scalar multiples of the extrinsic Euclidean metric.

Nearest Neighbour Graphs

UMAP approximates the manifold $\mathcal{M}$ underlying a sample X using nearest neighbour graphs:

Definition 2.3. Let (X, δ) be a finite metric space and k a positive integer. Define a directed graph $G_{knn} = (V, E)$ by $V = X$,

$$E = \{(x, y) \in X \times X \,|\, |\{z \in X \,|\, 0 < \delta(x, z) \leq \delta(x, y)\}| \leq k\}$$

That is, an ordered pair of points (x, y) is an edge if y is among the k closest points to x.[1] We call G_{knn} the *k-nearest-neighbour graph* of (X, δ).

Assuming X is a sample from a Riemannian manifold $(\mathcal{M}, g)$ embedded in $\mathbb{R}^N$ and that k is small relative to the size of X, we can view the k closest points to $x \in X$ as samples in a small neighbourhood of x. As the Euclidean distance approximates the Euclidean distance well in small neighbourhoods (up to scaling), the g-distances from x to its k-nearest-neighbours should be similar for all $x \in X$. Hence, we re-scale the weights of the edges in the following way to obtain a k-nn graph weighted by g:

Let $x \in X$ and let ρ_x denote the distance (in the ambient metric δ) to its closest neighbour. For each $x \in X$ define $\sigma_x > 0$ such that the equality

$$\sum_{(x,x') \in E} \exp\left(-\frac{\max\{0, \|x - x'\| - \rho_x\}}{\sigma_x}\right) = \log_2(k) \tag{2.1}$$

holds. The $\log_2$-term on the right-hand-side of Equation (2.1) has been chosen based on empirical experiments [102]. We can then define a weight-function

$$\begin{aligned} \tilde{w}(x, x') &= w(x, x') + w(x', x) - w(x, x')w(x', x), \\ w(x, x') &= \begin{cases} \exp\left(-\frac{\max\{0, \|x - x'\| - \rho_x\}}{\sigma_x}\right) & (x, x') \in E \\ 0 & \text{otherwise} \end{cases}. \end{aligned} \tag{2.2}$$

[1] Not considering x as a point close to itself.

The weighting w gives higher strength to pairs of points that lie close together in a re-scaled version of the ambient metric which approximates g. The subtraction of ρ_x in Equation (2.1) is for performance reasons and improves the behaviour of outliers [102]. We can interpret $w(x, x')$ as the probability that a directed edge exists from x to x' and $\tilde{w}(x, x')$ as the probability that a directed edge exists from x to x' or that a directed edge exists from x' to x. Henceforth, we call a k-nn graph with weighing $\tilde{w}$ a *UMAP graph*. By the symmetry of $\tilde{w}$, we consider the UMAP graph to be a symmetrised, undirected version of a k-nn graph.

The UMAP Algorithm

UMAP embeds a k-nn graph G_{knn} with weights $\tilde{w}$ into low-dimensional Euclidean space such that the embedding is approximately isometric by using stochastic gradient descent on the following loss function:

Definition 2.4. Let $G = (V, E)$ be a graph and let v and w be two weight functions on E (which take values in $(0, 1]$ only). Their *Kullback-Leibler divergence* is defined as

$$D((G, v), (G, w)) = \sum_{e \in E} v(e) \log \left(\frac{v(e)}{w(e)} \right) + (1 - v(e)) \log \left(\frac{1 - v(e)}{1 - w(e)} \right). \tag{2.3}$$

We note that while D is differentiable for changes in v, the weight-function v is not differentiable for small changes in its inputs in the low-dimensional space $\mathbb{R}^d$ if it follows a definition such as Equation (2.2). In the target embedding, UMAP aims to approximate the metric structure of the underlying Riemannian manifold. As the sample is assumed to be uniformly distributed with respect to this Riemannian metric, we may assume that σ_y and ρ_y are (approximately) the same for all $y \in \mathbb{R}^d$, the embedding space. Without loss of generality, we may set $\sigma_x = y$ for all y and ρ_y to a user-defined value min-dist. Then, for an edge connecting points $y, y' \in \mathbb{R}^d$ in the low-dimensional embedding, we have that its weight is

$$v(y, y') = \exp\left(-\max\left\{0, \|y - y'\| - \text{min-dist}\right\}\right).$$

UMAP approximates v by the smooth function

$$\Phi(y, y') := \left(1 + a\|y - y'\|^{2b}\right)^{-1},$$

where a and b are determined by least-squares fitting against v for a given value of min-dist.

Recall that our initial sample is $\{x_i\} \subset \mathbb{R}^N$. Let $\{y_i\} \subset \mathbb{R}^d$ be the dimension-reduced sample. A first guess for the coordinates of $\{y_i\}$ is obtained by a spectral embedding [102] and its positions are subsequently optimised by gradient descent. Using the above differentiable function Φ, the stochastic gradient descent then repeatedly samples edges $(x, x') \in E$ with probability $\tilde{w}(x, x')$ and updates the position of one of its contained vertices according to the gradient on $\log(\Phi)$. For each sampled vertex, it uniformly samples n-neg-samples (a user defined-parameter) other vertices and updates the position of x according to the gradient of $\log(1 - \Phi)$. The former is motivated by minimising the right-hand term in Equation (2.3), while the latter minimises the left-hand term in Equation (2.3). This procedure is repeated over several epochs. Formally, the optimisation algorithm is given by Algorithm 1, where we iterate over ordered edges (i.e. if $(a, b) \in E$ then $(b, a) \in E$ too) and the function Random() returns numbers in $[0, 1]$ uniformly at random.

Note that the negative sampling in Algorithm 1 employs the value one minus the weight of any hypothetical edge between y_a and y_b. However, negative sampling aims to minimise the right-hand term of Equation (2.3), which only considers edges in E. In practice, this discrepancy does not seem to negatively affect the UMAP output. We can then summarise the UMAP algorithm as Algorithm 2.

2.5.3 Random Walks and Community Detection

In this section I introduce the concept of community detection on graphs, using the concepts of modularity maximisation and Markov stability. Random walks allow us to interpret the Laplacian score in a way which naturally motivates its generalisation to a multiscale Laplacian score in Chapter 3. Further, the interpretation of the (multiscale) Laplacian score in terms of random walks links these scores back to

Algorithm 1 Stochastic gradient descent

procedure OPTIMISEEMBEDDING($G_{knn} = (V, E, \tilde{w})$, $\{y_i\}$, min-dist, n-epochs, n-neg-samples)
    $\alpha \leftarrow 1.0$
    Φ is fitted from min-dist
    for $i \leftarrow 1, ...,$ n-epochs **do**
        for $(a, b) \in E$ **do**
            if Random() $\leq \tilde{w}(a, b)$ **then**
                $y_a \leftarrow y_a + \alpha \cdot \nabla(\log(\Phi))(y_a, y_b)$
                for $j \leftarrow 1, ...,$ n-neg-samples **do**
                    $c \leftarrow$ random vertex in V
                    $y_a \leftarrow y_a + \alpha \cdot \nabla(\log(1 - \Phi))(y_a, y_c)$
                end for
            end if
        end for
        $\alpha \leftarrow 1.0 - i/$n-epochs
    end for
    return $\{y_i\}$
end procedure

Algorithm 2 UMAP

procedure UMAP($\{x_i\}$, k, d, min-dist, n-epochs, n-neg-samples)
    Construct $G_{knn} = (V, E, \tilde{w})$ from $\{x_i\}$ using k
    $\{y_i\} \leftarrow$ spectral embedding of $G_{knn} = (V, E, \tilde{w})$ in $\mathbb{R}^d$
    $\{y_i\} \leftarrow$ OPTIMISEEMBEDDING(G_{knn}, $\{y_i\}$, min-dist, n-epochs, n-neg-samples)
    return $\{y_i\}$
end procedure

clusterings of cells, which are key to classical DE tests. In this section, we assume that any graph is undirected unless it is stated otherwise.

Given a graph G, with nodes V, and (possibly weighted) adjacency matrix A, a *community structure* on G is a partition of the graph's nodes

$$V = C_1 \sqcup C_2 \sqcup \cdots \sqcup C_k.$$

We call each C_i a community and assume that it is non-empty. Moreover, we denote by c_v the community a node v belongs to and assume that the sub-graphs induced by the C_i are all connected (i.e., between any two nodes in C_i there is a path only along nodes in C_i connecting the two nodes). We consider two community structures to be equivalent if one can be obtained from the other by permuting the labels $1, ..., k$.

Modularity

One can view a community structure on a graph as a clustering of its nodes. A commonly used function to assess the quality of a community structure and, hence, a clustering of nodes, is modularity:

Definition 2.5. For a given graph G, endowed with a community structure $\{c_v\}_{v\in V}$, let k_v be the degree of node v and let m denote the sum of the degrees of all nodes. Then the *modularity* of $\{c_v\}_{v\in V}$ is defined as

$$Q\left(\{c_v\}_{v\in V}\right) = \frac{1}{2m} \sum_{v,w\in V} \left[A_{vw} - \frac{k_v k_w}{2m}\right] \delta(c_v, c_w),$$

where δ denotes the Kronecker delta.

In the expression $A_{vw} - k_v k_w/(2m)$, the adjacency matrix entry gives the number of edges between nodes v and w (0 or 1). By contrast, to understand the fraction consider a random graph model in which we are given nodes V. Each node v has k_v 'stubs' of an edge attached to it (stubs being halves of edges which can be connected to form a whole edge). In total, there are $2m$ stubs. We now reconnect these stubs uniformly at random to form edges in a new, random graph. Given a node v, the probability of its i-th stub connecting to one of the stubs of node w is then $k_w/(2m-1)$

(the stub cannot connect to itself). By extension, the probability that any of the k_v stubs of v connects to a stub of w is given by $k_v k_w/(2m-1)$. Hence, we can view the second term as an approximation to the expected number of edges between v and w under a random re-wiring of G (if m is large). It follows that $A_{vw} - k_v k_w/(2m)$ approximates the difference between the actual and the expected number of edges between v and w. Hence, modularity is relatively large if the difference between the actual and expected number of edges within communities is high and is low for edges between communities.

In practice, community structures on a graph in which connections within communities are significantly more frequent than connections between communities, are found by maximising modularity. Modularity extends to directed graphs.

Markov Stability

Markov stability employs Markov chain methods to generalise the above interpretation of modularity [45].

In addition to the notation of the previous subsection, we define $d := (d_{v_1}, ..., d_{v_n})^T$, $D := \mathrm{diag}(d)$ and $M := D^{-1}A$. Note that M is the transition matrix of a Markov chain (i.e. all entries are non-negative and all rows sum to 1). If G is connected and not bipartite, the Markov chain corresponding to M is ergodic and, thus, has a unique stationary distribution π [24].

Furthermore, $(M^s)_{vw}$ is the probability of a random walker on G, starting at node v, landing on node w after s discrete time-steps. Thus, $(M^s)_{vw} - \pi_w$ gives the difference in probability that a random walker starting at v walks to w in s time steps minus the probability that the walker is at w in steady-state.

In a similar vein to modularity, we can define the *discrete stability of a partition of* G as

$$r_{\mathrm{disc}}(t, \{c_v\}_{v\in V}) = \min_{0<s\leq t} \sum_{v,w\in V} \pi_v \left[(M^s)_{vw} - \pi_w\right] \delta(c_v, c_w).$$

That is, we sum over the differences $(M^s)_{vw} - \pi_w$ and weight them by π_v. Note that in practice, we can often omit the min-expression, from the above definition without

altering the stability score significantly [45]. Discrete stability generalises modularity, as $Q(\{c_v\}_{v\in V}) = r_{\text{disc}}(1, \{c_v\}_{v\in V})$ [129].

The above notion can be refined for continuous-time Markov chains. Continuous-time Markov chains do not change their state at each time $t \in \mathbb{N}$, but at random events $\{t_i\}_{i\in\mathbb{N}} \subset \mathbb{R}$. Their distribution is given by setting $t_0 = 0$ and assuming that $\Delta_i = t_i - t_{i-1}$ are i.i.d. exponential random variables with rate 1 for $i \geq 1$. At each t_i, the probability of the walker to transition from their current node v to another node w is, again given by M_{vw}. The probability of a random continuous-time walker to being on node v at time 0 and being on node w at time t is given by the (v, w)-entry of the matrix $P(t) := \exp(-tL^{\text{rw}})$, where exp is the matrix exponential function and $L^{\text{rw}} := I - M$ [45]. The matrix L^{rw} is called the *random walk-normalised Laplacian* of G. We can then define the *continuous stability of a partition of* G as

$$r_{\text{cont}}(t, \{c_v\}_{v\in V}) = \sum_{v,w\in V} \pi_v \left[P(t)_{vw} - \pi_w\right] \delta(c_v, c_w).$$

The discrete and continuous stability of a partition allow us to consider 'stable' partitions across different resolutions. We view Markov stability as follows: We pick a node v at random from the stationary distribution π and place a random walker at this node. The stability of a partition at time t is then the expected value of the random variable indicating whether the walker has remained in c_v after walking for time t (it is permissible for the walker to leave and re-enter c_v) minus the probability of the walker being in c_v at steady state. For small t, the walker is likely to still be close to v, while for large t its location becomes increasingly independent of its starting point. Hence, the stability of a partition decreases with increasing t. It is therefore important to only compare stability scores calculated at the same time t.

Using the above intuition, we can illustrate why maximising Markov stability at different times t is useful for analysing community structures at different resolutions. At small t, a community structure obtained by maximising stability will contain many relatively small communities, while for large t a few relatively large communities will be found. When increasing t from small to large, the small communities will merge to form larger ones. In particular, Markov stability can find sub-communities (at small

t) of larger communities (found at larger t) on an array of benchmark data sets [129]. A further advantage of being able to scale t is that it enables Markov stability, unlike other state-of-art methods, to detect non-clique-like community structures [129]. Going forward, we only consider continuous Markov stability in this chapter.

Louvain Method

Finding interesting community structures on a graph via modularity or Markov stability involves maximising a function over all partitions of a given graph. It is known that finding global optima of functions such as modularity or Markov stability is NP-hard [45]. A scalable method that does not guarantee finding a global optimum but yields good results in practice is the Louvain method [18].

Originally developed for modularity optimisation, the Louvain method exploits the insight that while computing the modularity score of a large graph is costly, computing the difference in modularity arising from moving a node v in an existing community structure from one community to another can be efficiently computed. Concretely, if we assume that a node v is moved from its current community C_{old} to a new community C_{new}, the difference in modularity given by this move is

$$\Delta Q = \frac{1}{m}\left[\left(\sum_{w\in C_{new}} A_{vw} - \frac{k_v k_w}{2m}\right) - \left(\sum_{w\in C_{old}} A_{vw} - \frac{k_v k_w}{2m}\right) + \frac{k_v^2}{2m}\right].$$

When computing ΔQ, we can record the sum of degrees within a community to speed up computations. I.e., if Σ_{new} and Σ_{old} are the sum of degrees within C_{new} and C_{old}, respectively, then ΔQ simplifies to

$$\Delta Q = \frac{1}{m}\left[\sum_{w\in C_{new}} A_{vw} - \sum_{w\in C_{old}} A_{vw} + \frac{k_v(\Sigma_{old} - \Sigma_{new} + k_v)}{2m}\right].$$

Analogously, for continuous Markov stability at a fixed t, let $\Pi = \mathrm{diag}(\pi)$ and $B = \Pi P(t) - \pi^T\pi$. Then the difference in r_{cont} when a single node v is moved from its current community C_{old} to a new community C_{new}, is given by

$$\Delta r_{\mathrm{cont}} = \sum_{w\in C_{new}} (B_{vw} + B_{wv}) - \sum_{w\in C_{old}} (B_{vw} + B_{wv}) + 2B_{vv}.$$

Note that for ΔQ we assume that the underlying graph is undirected. However, the expression can be generalised to the setting of directed graphs.

The Louvain method uses easily computable difference functions together with an initial community structure and iterates through the nodes of a graph in order. It then performs the following steps at each node v:

1. For each neighbour v' of v, compute the difference in objective function, $\Delta_{v'}$ say, if v were hypothetically moved from its current community to that of v'.

2. Of all alternative communities considered in step 1., let C^*_{new} be the community that leads to the largest difference Δ^* in the objective function. If $\Delta^* > 0$, move i to C^*_{new}. If not, leave v in its current community.

These two steps are repeated until no nodes are moved in a full iteration over the graph's nodes. In practice, the procedure is typically randomised by permuting the order of the nodes at random. Once no further improvement is observed, a new graph is created: each node in this graph is a community arising from the previous step. A weighted edge between two nodes is created with a weight equal to the sum of all weights of edges connecting the two communities in the previous graph. This newly constructed graph can also contain self-loops. Steps 1. and 2. are then again repeated until no further improvement is attained. New graphs are created and steps 1. and 2. are re-applied until no improvement in modularity is obtained.

Note that some current implementations of the Louvain method terminate if the increase in the objective function after an iteration over the nodes falls below a user-defined threshold.[2]

Variation of Information

When using Markov stability to detect communities at different resolutions, (i.e. different values of t), it may not be immediately clear what constitutes a 'good' resolution. While for a given graph more than one resolution may yield insight into

[2]E.g. Generalized Louvain optimization `https://github.com/michaelschaub/generalizedLouvain` (visited 28/4/23) and find communities `https://sites.google.com/site/findcommunities/` (visited 28/4/23).

the graph structure, this does not imply that all resolutions are equally insightful. In particular, identifying the scales at which we can detect interesting structures is informative in its own right.

Studies using Markov stability have addressed this problem by computing partitions at a large number of time points [9, 45, 129]. At each time-point t, they compute a large number of partitions. As they use the Louvain method, which is randomised, these partitions need not be the same. For resolutions at which there is an obvious community structure, we expect each iteration of the Louvain method to yield a partition that approximates the optimal structure. By contrast, at resolutions where there is no evident partitioning, we would expect different iterations of the Louvain method to generate partitionings that are rather dissimilar. Thus, if the average dissimilarity of the partitionings obtained at a fixed t is small, t should be viewed as a resolution at which an interesting structure exists. A dissimilarity measure widely used [9, 45, 129] is *variation of information* (VI):

Definition 2.6. Let $\{C_i\}_{1\leq i\leq k}$ and $\{C'_j\}_{1\leq j\leq k'}$ be two community structures on a graph with N nodes. Then their variation of information is defined to be

$$\mathrm{VI}\left(\{C_i\},\{C'_j\}\right) = 2H\left(\{C_i\},\{C'_j\}\right) - H(\{C_i\}) - H(\{C'_j\}),$$

where

$$H\left(\{C_i\},\{C'_j\}\right) = -\sum_{i,j}\frac{\left|C_i\cap C'_j\right|}{N}\log_2\left(\frac{\left|C_i\cap C'_j\right|}{N}\right),$$

$$H\left(\{C_i\}\right) = -\sum_{i=1}^{k}\frac{|C_i|}{N}\log_2\left(\frac{|C_i|}{N}\right).$$

An example of how VI can be used together with Markov stability optimisation to find community structures at different scales is given in Figure 2.4.

2.5.4 Simplicial Complexes and Filtrations

To undertake our work, we require a mathematical definition of a shape which we can utilise for fast computations. In topological data analysis (TDA) and computational geometry, simplicial complexes are widely used for this purpose:

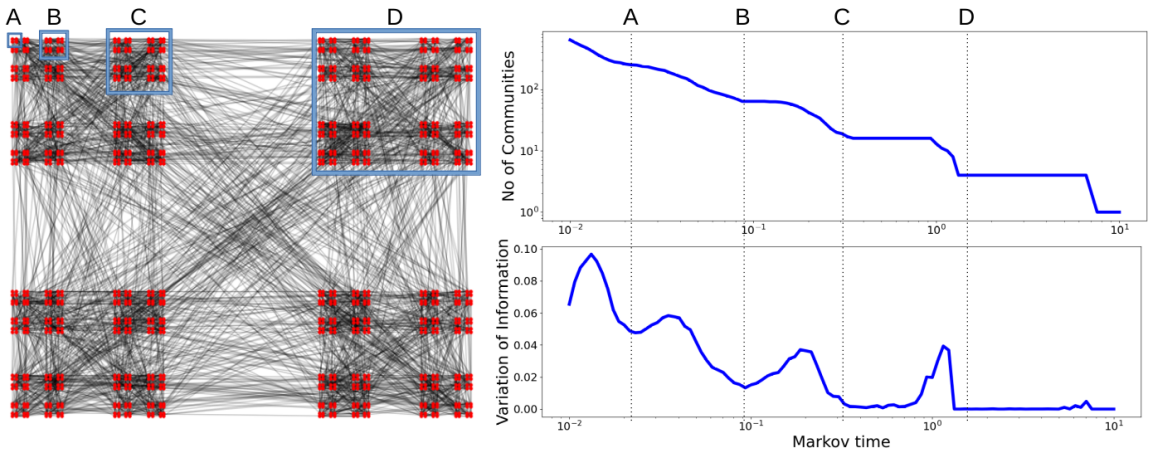


Figure 2.4: The graph on the left displays community structures at four different scales, exemplified by the groups A, B, C and D. When computing the mean pairwise variation of information (bottom right) as a function of scale (Markov time), we find local minima corresponding to resolutions A (256 communities), B (64 communities), C (16 communities) and D (4 communities). Figure inspired by [5].

Definition 2.7. Given a finite set of vertices X, an *abstract simplicial complex* is a set of subsets of X, denoted $\mathcal{K}$, such that for any $\tau \in \mathcal{K}$ and $\sigma \subseteq \tau$, we have $\sigma \in \mathcal{K}$. We call $\sigma \in \mathcal{K}$ a *simplex* of $\mathcal{K}$. Moreover, for any simplex σ, we define $\dim(\sigma) = |\sigma| - 1$ and $\mathcal{K}_i = \{\sigma \in \mathcal{K} \mid \dim(\sigma) = i\}$. We call the elements of $\mathcal{K}_i$ the i-simplices of $\mathcal{K}$.

A *geometric (or embedded) simplicial complex* $\mathcal{K}$ is an abstract simplicial complex that is endowed with an embedding in $\mathbb{R}^d$. That is, $X \subset \mathbb{R}^d$ and for all $\sigma, \tau \in \mathcal{K}$ with $\sigma \neq \tau$ we have $\text{relint}(\text{cvx}(\sigma)) \cap \text{relint}(\text{cvx}(\tau)) = \varnothing$. Here, relint is the relative interior[3] and cvx the convex hull.[4] We can then think of $\mathcal{K}$ equivalently as the union of the convex hulls of all of its simplices, which is a topological subspace of $\mathbb{R}^d$.

Given two simplicial complexes $\mathcal{K}$ and $\mathcal{K}'$, a *simplicial map* f is a function $f : X \to X'$, extending to a map $f : \mathcal{K} \to \mathcal{K}'$ by $f(\sigma) = \{f(x) \,|\, x \in \sigma\}$.

Note that any abstract simplicial complex can be viewed as a geometric simplicial complex by considering V to be a subset of the free vector space generated by V. This geometric simplicial complex is called the *geometric realisation* of $\mathcal{K}$. Conversely, each

[3]For a convex set $C \subset \mathbb{R}^d$, that is $\{x \in C \,|\, \forall y \in C : \exists \lambda > 1 : \lambda x + (1-\lambda)y \in C\}$.

[4]I.e., the intersection of all convex subsets of $\mathbb{R}^d$ that contain the given set of points.

Figure 2.5: Example of two simplicial complexes, $\mathcal{K}_1$ and $\mathcal{K}_2$, embedded in $\mathbb{R}^2$. Vertices as blue dots, 1-simplices as red lines.

geometric simplicial complex has an underlying abstract simplicial complex which is obtained by forgetting any geometric information.

In most settings, abstract simplicial complexes are more amenable to computations while geometric simplicial complexes, perhaps unsurprisingly, contain geometric information. The organoid boundaries extensively studied in the following section are all equivalent when modelled as abstract simplicial complexes. We use filtrations to study and compare geometric simplicial complexes representing organoid boundaries.

Definition 2.8. A *filtration* of a simplicial complex $\mathcal{K}$ is a function $f : \mathcal{K} \to \mathbb{R}$ such that $f(\sigma) \leq f(\tau)$ for $\sigma \subseteq \tau$. For $j \in \mathbb{R}$, we then define

$$\mathcal{K}^j = \{\sigma \in \mathcal{K} \mid f(\sigma) \leq j\}.$$

All $\mathcal{K}^j$ are simplicial complexes in their own right and $\mathcal{K}^j \subseteq \mathcal{K}^i$ for $j \leq i$.

2.5.5 The Euler Characteristic and Euler Characteristic Transform

The *Euler characteristic* is an invariant of topological spaces. Any two topological spaces that are (homotopy) equivalent have the same Euler characteristic. Conversely, if two topological spaces have different Euler characteristics, we can conclude that they are topologically different (i.e. not homotopy equivalent).

Definition 2.9. The Euler characteristic of a topological space X with finitely generated homology is defined as the following alternating sum

$$\chi(X) := \sum_{i=0}^{\infty} (-1)^i \operatorname{rank}\ H_i(X; \mathbb{Z}).$$

Here, the rank of a finitely generated abelian group is the number of $\mathbb{Z}$ summands in its canonical decomposition.

If a topological space X is homeomorphic to the geometric realisation of a simplicial complex $\mathcal{K}$, we can compute the Euler characteristic entirely from the combinatorial information of an abstract simplicial complex. We define the Euler characteristic of $\mathcal{K}$ to be the Euler characteristic of its geometric realisation.

Lemma 2.10 (E.g. Theorem 2.44 in [70])**.** *Let $\mathcal{K}$ be a geometric simplicial complex. Then its Euler characteristic is*

$$\chi(\mathcal{K}) = \sum_{i=0}^{\infty} (-1)^i \cdot |\, \mathcal{K}_i \,|.$$

Given a subset $X \subset \mathbb{R}^d$, such as a simplicial complex $\mathcal{K}$ embedded in $\mathbb{R}^d$, using a sequence of Euler characteristics induced by a sub-level sets yields additional discriminative information:

Definition 2.11. For a subset $X \subseteq \mathbb{R}^d$, we define the *Euler characteristic transform* (ECT) of X to be the following map:

$$\begin{aligned} \mathrm{ECT}_X : S^{d-1} \times \mathbb{R} &\longrightarrow \mathbb{Z} \\ (v, t) &\longmapsto \chi(\{x \in X : \langle x, v \rangle \leq t\}). \end{aligned}$$

In words, the Euler characteristic transform of a shape X encodes the Euler characteristic of the intersection of X with every closed half-space with affine boundary. When $\mathrm{ECT}_X(v,t)$ is not well defined, we set $\mathrm{ECT}_X(v,t) = \infty$. In this context, we set $\infty + \infty = \infty$, $\infty - \infty = \infty$, and $\infty + n = \infty$ for any integer n.

If X is an embedded simplicial complex $\mathcal{K}$, we can compute the ECT of X as follows: Let $v \in S^{d-1}$ be a fixed direction in $\mathbb{R}^d$. We then call the filtration on $\mathcal{K}$ induced by

$$f_v : \mathcal{K} \to \mathbb{R}, \quad \sigma \mapsto \max_{x \in \sigma} \{\langle x, v \rangle\}$$

the sub-level set filtration of $\mathcal{K}$ in direction v, where $\langle \cdot, \cdot \rangle$ is the standard inner product in $\mathbb{R}^d$. We denote the above filtration $\mathcal{K}^{\langle \cdot, v \rangle}$ and each sub-level-set at $t \in \mathbb{R}$ as $\mathcal{K}^{\langle \cdot, v \rangle \leq t}$. Then, using Lemma 2.10, we have that

$$\begin{aligned} \mathrm{ECT}_{\mathcal{K}} : S^{d-1} \times \mathbb{R} &\longrightarrow \mathbb{Z} \\ (v, t) &\longmapsto \chi\left(\mathcal{K}^{\langle \cdot, v \rangle \leq t}\right). \end{aligned}$$

Assume that $a \in \mathbb{R}$ is larger than the diameter of X in $\mathbb{R}^d$. Then $\{x \in X : \langle x, v \rangle \leq -a\} = \varnothing$ and $\{x \in X : \langle x, v \rangle \leq t\} = X$ for all $v \in S^{d-1}$ (similar statements hold for f_v if X is a simplicial complex). We can now define the smooth Euler characteristic transform and related constructions [146]:

Definition 2.12. First, let the *Euler characteristic curve* in a fixed direction v be

$$\mathrm{ECC}_X^v : [-a, a] \to \mathbb{Z}, \qquad t \mapsto \chi\left(\{x \in X : \langle x, v \rangle \leq a\}\right).$$

Secondly, this curve is smoothed by defining the *smooth Euler characteristic curve* (SEC) as follows:

$$\mathrm{SEC}_X^v : [-a, a] \to \mathbb{R}, \qquad t \mapsto \int_{-a}^{t} \left(\mathrm{ECC}_X^v(x) - \overline{\mathrm{ECC}_X^v}\right) \mathrm{d}x,$$

where $\overline{\mathrm{ECC}_X^v}$ is the mean of the function $\mathrm{ECC}_{\mathcal{K}}^v$ over the interval $[-a, a]$.

Finally, we can define the *smooth Euler characteristic transform* (SECT):

$$\mathrm{SECT}_X : S^{d-1} \times [-a, a] \to \mathbb{R}, \qquad (v, t) \mapsto \mathrm{SEC}_{\mathcal{K}}^v(t).$$

Turner and colleagues have shown that both the ECT and SECT are injective on a broad class of shapes (including embedded simplicial complexes) embedded in $\mathbb{R}^2$ and $\mathbb{R}^3$ [146]. Ghrist et al. [59] and Curry et al. [41] independently extended this injectivity result to general $\mathbb{R}^d$. The ECT or SECT therefore also discriminate between embedded simplicial complexes that are equivalent up to translation, rotation, reflection and combinations thereof. The issue of discriminating between shapes equivalent up to translation can be overcome by re-centring simplicial complexes by subtracting the mean of all vertices from each vertex in the simplicial complex. However, resolving rotation and reflection requires more care. Fortunately, Curry et al. [41] present a result on the ECT which yields a sufficient statistic on the space of embedded simplicial complexes modulo actions of the orthogonal group $O(d)$.

Before introducing their result, we recall the notion of a pushforward measure:

Definition 2.13. Let (X_1, Σ_1) and (X_2, Σ_2) be measurable spaces, $f : X_1 \to X_2$ be a measurable function, and $\mu : \Sigma_1 \to [0, \infty]$ be a measure on X_1. Then $f_*\mu$, the pushforward of μ along f, is the measure on X_2 defined by $(f_*\mu)(U) = \mu(f^{-1}(U))$ for each $U \in \Sigma_2$.

Then Theorem 6.6 in [41] states:

Theorem 2.14. *Let $\mathcal{K}$ and $\mathcal{K}'$ be generic simplicial complexes embedded in $\mathbb{R}^d$. Let μ be the Lebesgue measure on S^{d-1}. If $(\mathrm{ECT}_{\mathcal{K}})_* (\mu) = (\mathrm{ECT}_{\mathcal{K}'})_* (\mu)$, then there exists a $\phi \in O(d)$ such that $\mathcal{K} = \phi (\mathcal{K}')$.*

Note that the converse implication of the above theorem is trivial, as the Lebesgue measure on S^{d-1} is invariant under the action of $O(d)$. Note that the function mapping ECC to SEC is injective. Therefore, Theorem 2.14 generalises to the SECT.

For any $x \in [-a, a]$, define $\delta_x : L^2([-a, a]) \to \mathbb{R}$ by $\delta_x(f) = f(x)$ for all $f \in L^2([-a, a])$. Then if two embedded simplicial complexes, $\mathcal{K}$ and $\mathcal{K}'$ say, satisfy $K = \phi(\mathcal{K}')$ for some $\phi \in O(d)$, we get

$$\int_{S^{d-1}} \delta_x \circ \mathrm{SECT}_{\mathcal{K}} \, \mathrm{d}\mu = \int_{L^2([-a,a])} \delta_x \, \mathrm{d}((\mathrm{SECT}_{\mathcal{K}})_*(\mu))$$

$$= \int_{L^2([-a,a])} \delta_x \, \mathrm{d}((\mathrm{SECT}_{\mathcal{K}'})_*(\mu)) = \int_{S^{d-1}} \delta_x \circ \mathrm{SECT}_{\mathcal{K}'} \, \mathrm{d}\mu$$

for all $x \in [-a, a]$ by the change of variable formula for integrals with measure pushforwards and Theorem 2.14. Hence, the mean of such SECT evaluations, or collections thereof, form a statistic that can be used for distinguishing shapes modulo $O(d)$ actions.

2.5.6 Kernels and Kernel Approximations

Finally, we briefly discuss kernels and their associated Hilbert spaces - a theory we will use in later chapters when comparing topological signatures. Kernels are generalisations of inner products. The motivation for generalising inner products is twofold. First, we may want to use data science methods requiring an inner product (e.g. support vector classification, principal component analysis or k-means) on data in spaces

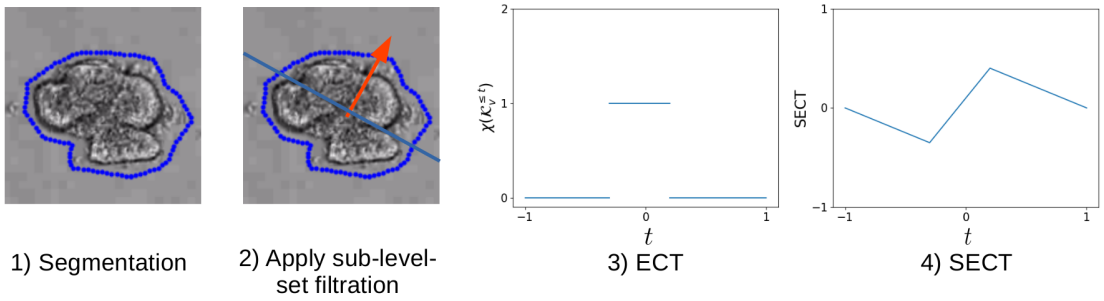


Figure 2.6: Standard ECT pipeline visualised on segmented organoid boundaries. 1) Segmented input data (in blue) over a video frame. 2) Illustration of the sub-level-set filtration in the direction given by the arrow. 3) The ECT in the direction given in 2). 4) The SECT in the direction given in 2).

not endowed with an inner product. Second, even if it is possible to define an inner product in data space, features in the data may not be linear. For example, not every labelled data set can be separated by a plane, resulting in inaccurate classification. In such an instance, support vector classification (SVC) in combination with kernels could, by contrast, allow a separation.

Definition 2.15. Let X be a non-empty set. A kernel on X is a symmetric function $k : X \times X \to \mathbb{R}$ such that for all $x_1, ..., x_n \in X$ and all $a_1, ..., a_n \in \mathbb{R}$ we get

$$\sum_{i,j=1}^{n} a_i a_j k\left(x_i, x_j\right) \geq 0. \tag{2.4}$$

Given finite subsets $X' = \{x'_1, ..., x'_m\}$ and $X^* = \{x^*_1, ..., x^*_n\}$ of X, we denote by $K(X', X^*)$ the $m \times n$ matrix with i, j-entry $K(X', X^*)_{ij} = k\left(x'_i, x^*_j\right)$, which is called the *Gram matrix* of k at X' and X^*.

Note that Equation (2.4) is equivalent to each Gram matrix of the form $K(X', X')$ being positive-definite.

For general $\mathcal{X}$, the Kronecker delta gives a kernel. If $\mathcal{X}$ is a subset of an inner product space, the inner product is a kernel. If $\mathcal{X} \subseteq \mathbb{R}^d$, then the function

$$k(x, y) := \exp\left(-\frac{\|x - y\|^2}{\lambda}\right),$$

where $\lambda > 0$ is a hyperparameter, is a kernel called the *Gaussian kernel*, which we employ later in this thesis.

Definition 2.16. Let k be a kernel on some set $\mathcal{X}$. If we define $\mathcal{H}_0 = \text{span}_{\mathbb{R}}\{k(x, \cdot) \,|\, x \in \mathcal{X}\}$, then $\mathcal{H}_0$ is a vector space of functions from $\mathcal{X}$ to $\mathbb{R}$. Note that

$$\langle k(x, \cdot), k(y, \cdot)\rangle_{\mathcal{H}_0} := k(x, y) \tag{2.5}$$

defines an inner product on $\mathcal{H}_0$ by bi-linear extension. We define $\mathcal{H}_k = \overline{\mathcal{H}_0}$, the completion of $\mathcal{H}_0$. Then $\mathcal{H}_k$ is a Hilbert space, called the *Reproducing Kernel Hilbert Space* (RKHS) of k.

In the above construction, $\mathcal{H}_k$ is still a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$ (rather than a Hilbert space of equivalence classes of functions) by the Moore-Aronszajn theorem [120]. The name RKHS derives from the fact that for any $f \in \mathcal{H}_k$ and $x \in \mathcal{X}$, we have $\langle k(x, \cdot), f\rangle_{\mathcal{H}_k} = f(x)$, which is called the *reproducing property* of k. In particular, the reproducing property of an RKHS together with the Cauchy-Schwarz inequality gives that the linear functionals $\delta_x : \mathcal{H}_k \to \mathbb{R}$ defined by $\delta_x(f) = f(x)$ at $f \in \mathcal{H}_k$ are continuous for all $x \in \mathcal{X}$.

The main point - in the context of this work - of defining an RKHS is to illustrate that applying a kernel to elements of $\mathcal{X}$ can be viewed as first embedding $\mathcal{X}$ into some Hilbert space of functions $\mathcal{H}_k$ (by $x \mapsto k(x, \cdot)$) and then taking an inner product of such embedded elements. While $\mathcal{H}_k$ may be infinite-dimensional and thus the embedding of $\mathcal{X}$ into $\mathcal{H}_k$ is intractable in general, we never need to compute the (exact) embedding itself - computing $k(x, y)$ for all $x, y \in \mathcal{X}$ is sufficient for any downstream method relying on the inner product of $\mathcal{H}_k$ only. This insight is called the *kernel trick*.

If $\mathcal{X}$ is of finite size n, then computing $k(x, y)$ for all pairs $k(x, y)$ requires only finitely many computations and scales as $\mathcal{O}(n^2)$. To enable computations for large n, a number of approximation methods of lower computational complexity have been developed. One such method is the *Nystroem approximation*:

Definition 2.17. Let $\mathcal{X} = \{x_1, ..., x_n\}$ be of finite size n and let $m < n$, where m is a natural number. Denote by K the $n \times n$-matrix with i, j-entry $k(x_i, x_j)$. Then K can be written in block-form as

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}.$$

For $C := [K_{11}, K_{12}]^T$, the m-th Nystroem approximation of K is the matrix

$$\tilde{K} = C K_{11}^{\dagger} C^T,$$

where † denotes the Moore-Penrose pseudo-inverse.

If K is of approximate rank m (or less), then $\tilde{K} \approx K$ [89]. In practice, the approximate rank of a Gram matrix is in many cases much less than n if n is large [89]. Computing $\tilde{K}$ only requires $\mathcal{O}(mn)$ evaluations of k, where m is typically fixed. In practice, we only compute $C(K_{11}^{\dagger})^{1/2}$ (to save computer memory), as the standard inner product of the i-th and j-th rows of $C(K_{11}^{\dagger})^{1/2}$ approximately gives $k(x_i, x_j)$. This matrix $C(K_{11}^{\dagger})^{1/2}$ can be used as a non-linear transformation and can be further analysed. In particular, we take the row vectors and feed them to a method using inner products. Computing $C(K_{11}^{\dagger})^{1/2}$ has a runtime complexity of $\mathcal{O}(nm^2 + m^3)$. There exist sampling heuristics for picking an optimal set of m points from $\mathcal{X}$ [89].

Previous Studies Using ECTs and Kernel Methods

Kernel methods have been used successfully in conjunction with the SECT for both shape regression and classification problems [39, 155, 139]. Both of these studies use Gaussian process models. Gaussian processes include the inversion of a Gram matrix and thus have a runtime of $\mathcal{O}(n^3)$. Hence, the Nystroem method we employ scales better to large data sets. These models are also conceptually more complex than linear regression and random forest classification, which we use for regression and classification. To the best of our knowledge, neither the ECT nor SECT have previously been used in combination with kernel approximation methods.

2.6 Summary

In this chapter, I introduced organoids, how they are grown, how they have been used in cancer research and what potential they might have for the future. I highlighted the interplay of the tissue composition, which to a large extent determines organoid morphology, and scRNA-seq data. Further, I gave an overview of previous studies using morphology or scRNA-seq data of organoids. I then introduced some mathematical preliminaries which are used throughout this thesis: Firstly, Mapper and UMAP are two TDA-related methods that create graph representations of data. Further, I introduced random walks on graphs and how these can be used for (multiscale) community detection on graphs. These methods are used and extended in Chapter 3 to structure scRNA-seq data and genes that drive the differentiation of cells along lineages. Finally, I explained the Euler characteristic transform, a sufficient statistic of shapes embedded in $\mathbb{R}^d$, and the fundamentals of kernel methods. The Euler characteristic transform is extended in Chapter 4 to a new method called DETECT and used in conjunction with kernel methods to classify organoids with cancer mutations, segmented from videos of experiments in a 2D view, into treated and untreated groups. I show that DETECT generalises to 3D data by using synthetic data generated by a mechanistic model of organoid growth in Chapter 5. I prove a new result on the stability of the ECT in Chapter 6.

Chapter 3

TDA of Single Cell RNA Sequencing Data

Chapter Content

3.1 scRNA-seq Data Structure and Analysis Methods 45
3.1.1 Unique Molecular Identifiers and Raw Data Structure 45
3.1.2 Variance Stabilizing Transform 45
3.1.3 Differential Expression and Its Generalisations 48
3.1.4 Trajectory Inference Methods 50
3.2 New Topological Analysis Methods for scRNA-seq Data 52
3.2.1 Multiscale Laplacian Score 52
3.2.2 UMAP Diffusion Cover 53
3.3 Data Sets 55
3.4 Results 56
3.5 Discussion 61

Single-cell RNA sequencing (scRNA-seq) provides an unprecedented level of detail about the dynamic process of transcription in cells. The high-resolution data challenges the traditional notion of cell types as discrete entities and suggests that instead they should be viewed as a continuum [60, 122]. Differential expression (DE) tests are commonly used to analyse transcriptomic data. However, DE tests rely on clustering algorithms to assign cells to discrete categories on such data sets, which are unstable

on high-resolution scRNA-seq data. Govek et al. propose a clustering-independent Laplacian score to generalise DE tests for scRNA-seq data with a continuous underlying structure [60]. I extend their single-scale method to a multiscale method. To this end, I propose a novel multiscale Laplacian score (MLS) in this chapter. The MLS combines ideas from network and random walk theory to perform generalised DE tests at multiple resolutions. I propose to use variation of information (VI), a common method in community detection on networks, to identify at which scales a data set exhibits an interesting structure.

Trajectory inference (TI) methods assume a continuous structure of single-cell data and aim to infer continuous trajectories between different cell states [125]. The performance of such methods varies significantly across data sets and depends on both the topology underlying a data set and the topology a trajectory inference can model [125]. One method for trajectory inference is a TDA algorithm called Mapper (see Section 2.5.1), which can model a range of different data topologies and has been successfully used in practice [118, 122]. In these studies, the cover, a key hyperparameter in the Mapper algorithm, is chosen manually rather than by a biological or computational principal. The results, therefore, may be biased. I propose a novel heuristic called the UMAP diffusion cover, inspired by network theory and random walks, that can be used to algorithmically specify the cover in Mapper for TI.

In this chapter, I first introduce the raw data structure of scRNA-seq data, its pre-processing methods and review common DE tests and TI methods (Section 3.1). After introducing the MLS and the UMAP diffusion cover (Section 3.2), the new methods are applied to two scRNA-seq data sets (introduced in Section 3.3). The results demonstrate that the MLS identifies differentially expressed genes at multiple resolutions, including genes that have not been identified by the Laplacian score or a classical DE test. Furthermore, I show that the VI identifies resolutions of interest and relates them to cell types and genetic conditions. The UMAP diffusion cover in conjunction with Mapper gives promising results on two data sets compared to the state-of-the-art method PAGA (all in Section 3.4). The chapter finishes with a

discussion of the results and potential directions for future research (Section 3.5).

3.1 scRNA-seq Data Structure and Analysis Methods

3.1.1 Unique Molecular Identifiers and Raw Data Structure

In RNA sequencing, it is common to break up the molecular strands of RNA into short fragments before counting the occurrences of each gene (i.e. fragments with an identical sequence of nucleotides). For technical reasons, it is not feasible to read the genetic material of a single, isolated cell directly. Hence, the genetic material is amplified first using polymerase chain reaction (PCR) protocols.

PCR protocols amplify some genes at a higher frequency than others which biases the data. The technique of unique molecular identifiers (UMIs) addresses this problem: short RNA sequences, called UMIs, are attached to each fragment at random prior to PCR amplification. When sequencing is performed, multiple occurrences of the same gene with the same UMI attached are counted as a single occurrence. The resulting counts are called *UMI counts* and the collection of all UMI counts for all genes of interest forms the raw version of our data.

The total number of UMI counts measured for a single cell (summed over all genes) is called the *sequencing depth*. It represents the amount of genetic material that has been sequenced from a cell. Random effects and technical details mean that the sequencing depth can vary significantly between cells. Thus, UMI counts should always be interpreted in the context of the sequencing depth of each cell. For this and other reasons, such as batch effects, it is essential to pre-process the raw UMI count data prior to its analysis. We describe a state-of-art pre-processing procedure that accounts for variable sequencing depth and batch effects in the next subsection.

3.1.2 Variance Stabilizing Transform

The *variance stabilising transform* (VST), introduced by Hafemeister and Satija in [67], attempts to learn the distribution of each UMI count. It uses information such

as sequencing depth and experimental batch to transform each count so that the distribution of the resulting data only depends on biological information present in the data. When describing the VST, we write $k_{ij} \in \mathbb{N}$ for the raw UMI count of gene i for cell j. The VST outputs a transformed z_{ij} corresponding to each k_{ij}.

The UMI counts are commonly assumed to follow a negative binomial (NB) distribution [67, 4]. This assumption allows for a negative binomial regression in a generalised linear model (GLM), a regression model commonly applied to count data [108, Documentation Chapter 326]. The negative binomial regression is an extension of the Poisson regression. The Poisson regression model assumes that the variance equals the mean of the explained data at each value of the independent variable. Similar to the Poisson distribution, NB random variables can be interpreted as the number of events occurring in a fixed time period, each event occurring at the same rate, independently of other events. For a Poisson random variable, this rate is deterministic, whereas, for an NB random variable, the rate is sampled at random from a Gamma distribution. As a result, variance and mean do not need to be equal in the NB distribution and are independent parameters. The work of [67] interprets sequencing depth as time and occurrences as detections of UMIs in any NB distributions they employ.

Let m_j denote the sequencing depth of cell j. Formally, the negative binomial regression under a generalised linear model then assumes for each gene i,

$$k_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma_{ij})$$

with

$$\mu_{ij} = \exp\left(\alpha_i + \beta_i \log_{10} m_j\right), \qquad \sigma_{ij} = \sqrt{\mu_{ij} + \frac{\mu_{ij}^2}{\theta_i}}, \tag{3.1}$$

where α_i, β_i and θ_i, the regression model parameters, are to be inferred [67]. We call θ_i the dispersion parameter associated with gene i. The negative binomial model then predicts that for cells of sequencing depth m on average $\mu_i = \exp\left(\alpha_i + \beta_i \log_{10} m\right)$ UMIs of gene i are detected. In practice, the parameters of the negative binomial

regression can be inferred for each gene i using a generalised linear model (GLM) and maximum likelihood methods [67].

Hafemeister and Satija [67] find that performing a negative binomial regression for each gene typically results in overfitting. They demonstrate, via bootstrapping, that such a procedure leads to unsatisfactory results on benchmark data sets, especially for genes with low average expression levels [67]. They pool information across genes with similar average expressions. Their procedure can be summarised in three steps:

1. Infer α_i, β_i, and θ_i for each gene i.

2. For all α_i obtained in Step 1., perform a kernel regression against the average expression value $\overline{k_i} := \sum_j^N k_{ij}/N$ using a Gaussian kernel. The bandwidth of the kernel is fixed using the Sheather and Jones method [133]. Repeat this process for β_i and θ_i.

   For each gene i, denote by α_i^*, β_i^* and θ_i^* the parameters predicted for gene i under the above regression models. Let μ_{ij}^* and σ_{ij}^* be defined as in Equation (3.1), but using the starred expressions for the negative binomial model parameters instead.

3. Transform the UMI counts as

$$z_{ij} = \frac{k_{ij} - \mu_{ij}^*}{\sigma_{ij}^*}.$$

The above procedure is called the variance stabilising transform (VST). Since its inception, it has been included in the standard genetics R-package `Seurat` [68] and in many peer-reviewed studies [81, 66, 64, 20, 82].

Performing the VST on distinct batches of the same experiment separately is sufficient for removing batch effects [67]. Hafemeister and Satija [67] demonstrate on a benchmark data set that the VST effectively removes the correlation of (transformed) UMI counts with sequencing depth and gene abundance. The VST also removes the effects of the sequencing depth on downstream analyses [67].

3.1.3 Differential Expression and Its Generalisations

A common problem in biology, and science more generally, is establishing whether a measurable quantity varies significantly across two or more experimental groups. In the case of gene expression, such a test is called *differential gene expression*. The experimental groups may be determined by the experimental set-up (e.g. different treatments or different mutations), meta-data or through clustering algorithms.

Common methods for differential gene expression, such as DESeq [4], assume that the UMI counts are distributed according to a parameterised probability distribution (often following an NB distribution, similar to Sub-Section 3.1.2). Testing for differential expression is then equivalent to testing the statistical hypothesis that the parameters are the same across experimental groups. Alternatively, a non-parametric test can be performed, which typically requires the counts to already be corrected for sequencing depth, but is considered to be more stable [166].

Laplacian Score

Differential gene expression requires a partition of the cells into discrete groups or clusters. However, scRNA-seq data is typically continuous in nature, leading to unstable clusterings and, by extension, to unstable downstream differential gene expression analyses [60, 122]. I.e., even small perturbations to the input data or clustering parameters could lead to different clusterings and, by extension, to different DE analyses. To overcome this issue, Govek et al. [60] propose to first construct a graph on the scRNA-seq data (e.g. through a neighbourhood graph or by trajectory inference) which models scRNA-seq data as a continuous, connected structure. In particular, the graph endows cells with a continuous notion of similarity via the graph distance, as opposed to a discrete, binary notion of similarity (cells being either in the same cluster or not). Their method, a variation of the Laplacian score first defined in [71], tests the 'consistency' of a gene's expression with the graph structure. A gene is considered to be consistent with the graph structure if cells that are nearby on the graph have similar expression levels of this gene. While the score favours genes which have

locally similar expression levels in every neighbourhood of the graph, it also penalises genes with low variance (and thus express at similar levels across the whole data set).

Mathematically, the method of Govek et al. works as follows: Given a graph $G = (V, E)$ with N nodes, let $f \in \ell^2(V)$ be a signal on G. We define the *graph mean* of f as

$$\mu_G(f) = \frac{1}{\sum_{v \in V} d_v} \sum_{v \in V} d_v f(v)$$

and the *graph variance* of f as

$$\mathrm{Var}_G(f) = \sum_{v \in V} d_v \left(f(v) - \mu_G(f)\right)^2.$$

The graph mean of a signal f can be interpreted as the expected value of $f(v)$, where v is the (random) location of a random walker at steady-state. Similarly, the graph variance is $\sum_v d_v$ times the variance of this random variable. For any signal f on G, we then introduce its re-centred form

$$\tilde{f}(v) := f(v) - \mu_G(f).$$

Definition 3.1. Let $G = (V, E)$ be a graph with adjacency matrix A and $f : V \to \mathbb{R}$ a signal on G. Then the *Laplacian score* of f (in the sense of [60]) is defined as

$$LS(f) = \frac{\left\langle D^{1/2}\tilde{f}, D^{1/2} L^{\mathrm{rw}} \tilde{f} \right\rangle}{\left\langle D^{1/2}\tilde{f}, D^{1/2}\tilde{f} \right\rangle} = \frac{\sum_{v,w \in V} A_{vw} \left(f(v) - f(w)\right)^2}{\mathrm{Var}_G(f)}.$$

Remark. When the Laplacian score was first introduced in [71] its definition used an adjacency matrix of a k-nn graph with a specific weighting. Govek et al. [60] used a generalised version of the Laplacian score (without the weighting of [71]) on scRNA-seq data. We use the weighting of the UMAP graph, which is key to all of our scRNA-seq data analyses.

We interpret the Laplacian score as the expected squared difference in the signal f when a random walker at steady-state takes one (discrete time) step, divided by the variance of the signal at the node at which the walker is based at steady state.

The Laplacian score is low if the signal f takes similar values on nodes connected by an edge and exhibits high variance. Equivalently, the score is low whenever the

signal on the node at which a discrete-time random walker is located is expected to change by a small amount whenever they walk for time $t = 1$ (and the graph variance of the signal is high). In such a case, we say that the signal f is consistent with the graph structure of G.

In community detection, one-step discrete-time Markov stability is known to favour clique-like structures [129]. However, real-world networks often contain communities with large diameters, which are thus not cliques. Similarly, favours signals which are consistent on cliques. Most importantly, the Laplacian score cannot capture consistent expressions at different scales.

3.1.4 Trajectory Inference Methods

Single-cell RNA sequencing enables researchers to study dynamic molecular processes at a cellular level. Such processes include the cell cycle, cell differentiation and cell activation [125]. To model and infer such processes from the data, a range of trajectory inference methods have been developed. These methods typically order cells along a tree- or graph-like structure according to their levels of gene expression. Arranging cells in such a way can reveal temporal evolution (e.g. in differentiation or activation) or periodicity (e.g. in a cell cycle[1]).

Recent methods for trajectory inference vary significantly with respect to the topology they can model (ranging from being restricted to a linear or tree-like topology to any topology that can be represented by a graph or simplicial complex), the format of their output and their scalability.

In this chapter, I focus on PAGA [160], a scalable method that returns a weighted graph. In this graph, each node represents a cluster of cells and weighted edges between nodes capture the similarity between clusters. There are no restrictions on the structure of the returned graph. PAGA was identified as one of the most flexible and well-performing methods [125] and has been widely used [2, 6, 116, 137, 165]. A summary of alternative methods can be found in Table 3.1.

[1] A series of periodic events as cells grow and undergo mitosis.

Method	Platform	Topology	Accuracy	Scalability	Stability
PAGA	Python	Graph	Very Good	Excellent	Very Good
RaceID	R	Graph	Sufficient	Poor	Good
Slingshot	R	Tree	Excellent	Good	Very Good
pCreode	Python	Tree	Very Good	Poor	Good
Monocle ICA	R	Tree	Very Good	Poor	Good
STEMNET	R	Multifurcation	Very Good	Very Good	Good
SCORPIUS	R	Linear	Excellent	Good	Excellent
Wanderlust	Python	Linear	Good	Good	Excellent

Table 3.1: Comparison of some common trajectory inference methods. Based on Figure 2 in [125]. Here, 'Multifurcation' means a tree with one main branch in which all other branches connect to the main branch and do not have branches of their own. Each method was tested on synthetic and real data. Each method has only been tested on synthetic data with underlying ground truth it can infer. E.g. while SCORPIUS has a better accuracy result than PAGA, it has only been tested on data with linear topology. By contrast, PAGA has also been tested on data with a more complex topology. Stability was tested by sub-sampling real and synthetic data sets and assessing the resulting changes in output. See [125] for details.

Initially, PAGA constructs a neighbourhood graph (i.e. a k-nearest-neighbour graph; c.f. Section 2.5.2) from a pre-processed scRNA-seq data set and then clusters nodes in this graph based on their connectivity. The original paper [160] suggests using the Louvain method [18] (c.f. Section 2.5.3), but the documentation of the main implementation [159] now uses the Leiden method [143] as default.

With both community detection methods, PAGA proceeds to construct a new graph in which the nodes are given by the clusters obtained in the previous step. It then uses a random model on the original graph to determine the probability that for any two clusters A and B, there are fewer than n edges connecting nodes in A with nodes in B. Using the true number of edges connecting A and B in the original graph, the random model is then used to assign a probability to the pair of nodes A and B in the new graph. If this probability is larger than some user-defined threshold, (i.e. if the two clusters have stronger connectivity than what one would expect at random), PAGA inserts an edge between A and B in the new graph. The resulting graph, also called a PAGA graph, is then returned.

The hyperparameters PAGA requires are the k-nn parameter k, the probability

threshold, the clustering algorithm and all hyperparameters of that algorithm. Typically, k is set to some value between 15 and 40 and the original publication of PAGA suggests that 0.025 is generally a good value for the probability threshold [160]. The packages `scanpy` [159] and `Seurat` [68] use the Louvain [18] or Leiden [143] algorithm as default clustering algorithms.

3.2 New Topological Analysis Methods for scRNA-seq Data

3.2.1 Multiscale Laplacian Score

As with the application of random walk-normalised Laplacians to finding community structures on G, there is reason to believe that the Laplacian score favours consistent expression on clique-like structures. While these clique-like structures exist in scRNA-seq data, they only represent a single scale and it has been suggested that in scRNA-seq data clique-like structures are often distorted by common pre-processing methods [33]. Moreover, it cannot capture consistent expression at different scales. However, many data sets exhibit structure at multiple scales. Classical DE tests, in which cell types are typically determined by clustering algorithms, easily generalise to a multiscale analysis: Clustering can be performed at multiple resolutions (many clustering algorithms, including k-means and single-linkage, have a parameter controlling how coarse the resulting clustering is) and the DE test is applied to each clustering to obtain a multiscale description of the data. We, therefore, introduce the novel multiscale Laplacian score (MLS), which extends the definition of [60]:

Definition 3.2. Let $G = (V, E)$ be a graph with adjacency matrix A, $f : V \to \mathbb{R}$ be a signal (e.g. the expression of a gene) on G and $t \in \mathbb{R}_{\geq 0}$. Then the *multiscale Laplacian score* of f at resolution t is defined as

$$MLS(f,t) = \frac{\left\langle D^{1/2}\tilde{f}, D^{1/2}(I - P(t))\tilde{f} \right\rangle}{\left\langle D^{1/2}\tilde{f}, D^{1/2}\tilde{f} \right\rangle} = \frac{\sum_{v,w \in V} d_v P(t)_{vw} \left(f(v) - f(w)\right)^2}{\mathrm{Var}_G(f)},$$

where $P(t) := \exp\left(-tL^{\mathrm{rw}}\right)$.

Remark. The multiscale Laplacian graph kernel (MLGK) by Kondor and Pan [86] has a similar name to the MLS, but uses different constructions and serves a different purpose. The MLGK is multiscale by considering a sequence of nested sub-graphs, while the MLS is multiscale by considering random walks of different lengths. The MLGK compares *different graphs* while the MLS compares different signals on the *same graph*.

The MLS of a signal f at time t can be interpreted as the expected squared difference in signal a continuous-time random walker at steady state walking for time t is exposed to, divided by the variance of the signal such a random walker is exposed to. As for the original Laplacian score, the multiscale score of f at time t is low if the signal to which a continuous-time random walker is exposed is expected to change only by a small amount whenever they walk for time t (and the graph variance of the signal is high). In such a case, we say that the signal f is consistent with the graph structure of G at scale t. An example of different signals that are consistent with a graph at different scales, and thus can be identified by the MLS, is given in Figure 3.1.

In our MLS analysis pipeline, we take a pre-processed transcriptomic data set and construct a k-nn graph, with each node representing a cell. We then calculate partitions of the graph into communities at a large number of different resolutions (Markov times) using the Louvain algorithm [18]. We re-calculate the partitioning at each resolution several times, to obtain a mean pairwise variation of information (VI) at each Markov time. Next, we select a small set of resolutions at which the VI attains local minima. Finally, we calculate the MLS at each of these resolutions for each gene in the data set. By plotting the MLSs at subsequent resolution against each other it is possible to identify genes which are particularly consistent with the topological structure at a given scale by observing deviation from mean behaviour.

3.2.2 UMAP Diffusion Cover

When analysing data with Mapper, a key hyperparameter is the cover. When the lens f maps into a one-dimensional space, then the choice of cover is straightforward:

a sequence of intervals between the minimum and maximum values. The only choices which remain to be set are the number and size of the intervals and the amount of overlap between consecutive intervals (typically chosen manually).

In practice, however, Mapper graphs obtained through one-dimensional filter functions are not as expressive as those obtained through two-dimensional filter functions, and, indeed, a large number of publications use Mapper with two-dimensional filter functions (e.g. [94, 122, 126, 142]). However, using intervals as covers does not generalise well to the 2D setting (using cubes instead of intervals), as the resulting Mapper graphs vary significantly with the rotation of the image of the filter function (e.g. in UMAP this rotation is arbitrary) and Mapper graphs are more likely to be disconnected. Therefore, a cover is often selected manually [122]. Such judiciously chosen covers are undesirable as they increase the risk of reverse-engineering the desired result. I propose a novel heuristic for choosing covers in high dimensions, which uses ideas from random walks and is invariant under rotations of the filter function image.

The heuristic starts by considering the UMAP weighted k-nn graph. The UMAP-weighted k-nn graph is a good model for the manifold structure of scRNA-seq data, as it can model the continuous structure of the data. Furthermore, re-scaling distances to account for local changes in density reduces the likelihood of rare cell states being treated as outliers. The UMAP graph is also fast to compute [102].

After constructing the UMAP graph, we pick a node at random and calculate the continuous-time or discrete-time transition probabilities to all nodes in the graph for a random walker that walks for some pre-defined amount of time t. We then take the m nodes with the highest transition probabilities and place them into a new cover element. We repeat this procedure by sampling from the remaining nodes which are not in a cover element until all nodes are covered.

Repeated simulations suggest that this method is stable with respect to the introduced randomness. It also performs better than sampling nodes at random and creating cover elements with their m nearest neighbours, as it is more stable with respect to outliers. The heuristic is summarised by Algorithm 3. While in theory,

we do not have to compute a filter function when using this cover, there is a direct analogy to using UMAP as a filter. The UMAP algorithm (c.f. Section 2.5.2) consists of two parts: the construction of a k-nn graph and its embedding. The first part can be viewed as the actual dimension reduction, while the second part is merely a transformation of the dimension-reduced sample to Euclidean coordinates. The UMAP diffusion cover places samples which are 'close' in the dimension-reduced sample into the same cover element. While the embedding of this graph gives a filter function in a classical sense, the cover described here is invariant with respect to this embedding.

Algorithm 3 Random Walk Cover

procedure RandomWalkCover(G (UMAP graph), m, t)
  $N \leftarrow$ list containing $\{y_i\}$, the nodes of G
  $T \leftarrow \exp(-tL_G^{\text{rw}})$ or $(I - L_G^{\text{rw}})^t$ # cont. or disc. random walk
  $C \leftarrow \varnothing$
  while N not empty **do**
    $y \leftarrow$ random sample from N
    $P \leftarrow e_y T$ # Here e_y is a vector with 1 in y-entry and 0's elsewhere
    $C' \leftarrow \{y'_1, ..., y'_{m-1}, y\}$, # the $m-1$ largest indices in P and y.
    for $y' \in C'$ **do**
      Remove y' from N
    end for
    $C \leftarrow C \cup \{C'\}$
  end while
  return C
end procedure

3.3 Data Sets

I illustrate the UMAP diffusion cover and the multiscale Laplacian Score outlined above on two experimental scRNA-seq data sets. First, I analyse a data set of 24,911 human T cells infiltrating lung tumours and adjacent normal tissue, published in [91]. This T cell data set was originally used to demonstrate the utility of the standard Laplacian score in [60] and is, therefore, useful for benchmarking the multiscale Laplacian Score. Second, I study scRNA-seq data derived from mouse colon organoids. The organoids, grown and sequenced by the Lu Lab at the Ludwig Institute at

the University of Oxford, have different genetic backgrounds (Wild Type, APCmin, KRAS, p53 Null, p53 Mutant) and contain different cell types. This data set contains 3,958 cells.

Pre-processing of Data Sets

We normalised the T cell and mouse colon organoid scRNA data sets using the Variance Stabilizing Transform (VST) [67]. Compared to log-normalisation,[2] the VST has been shown to be more effective at removing noise induced by factors such as differences in sequencing depth across cells while retaining biological heterogeneity [67]. For each data set, the VST returns the 3,000 genes with the highest dispersion.

The variance stabilised data set is then reduced to the 30 principal components with the highest variance using PCA. We use 30 components following the recommendation given in the manual of the `R`-library Seurat [68]. We then construct UMAP-weighted k-nn graphs on both data sets, using $k = 15$ and the `umap-learn` Python package ($k = 15$ is the recommended default in both `Seurat` and `umap-learn`). Following [60], we use the Pearson correlation distance for both data sets. We sample 3,000 cells at random between the PCA and k-nn graph steps in the T cell data set to improve runtime. Notably, our UMAP plots (see Figure 3.2) of the T cell data look somewhat different from the t-SNE[3] plots presented in [91, 60]. This difference is mainly a result of using the VST instead of log-normalisation, rather than UMAP instead of t-SNE. As we use the UMAP graphs, we present the UMAP plots instead of the t-SNE plots for consistency.

3.4 Results

In this section, I apply Mapper [32] with a UMAP diffusion cover and the multiscale Laplacian score (MLS), as described in the previous sections, to the T cell and organoid data sets for trajectory inference and feature selection. I compare the trajectory inference obtained using Mapper with the novel diffusion cover to trajectories ob-

[2]Equivalent to $z_{ij} := \log(1 + k_{ij} \cdot 10^4/m_j)$ in the notation of Section 3.1.2.
[3]A non-linear dimension-reduction method, see [152].

tained using PAGA [160]. I also apply the MLS to both data sets and compare the results to the standard Laplacian score presented in [60]. The multiscale Laplacian score allows for feature selection at different scales and naturally separates consistently expressed genes into different resolutions, which is not possible (without further analysis) when using the standard Laplacian score alone. For the organoid data, the resolutions identified by the MLS correspond to partitionings into cell types and genetic conditions. We conclude that the MLS provides a natural way to identify genes that are consistently expressed by different cell types and across different genetic backgrounds. I use `scanpy` [159] to compute the PAGA graphs and `KeplerMapper` [153] to compute the Mapper graphs.

T Cell Data

First, we apply the MLS to the human T cell data set [91]. This data does not partition into stable clusters, as remarked by [60] and highlighted by the UMAP plots in Figure 3.2. We identify three resolutions of interest based on the variation of information. Genes with a relatively low MLS at the finest resolution, t_1, include IGKC and IFI27. Both are highly expressed by a small group of cells (in the centre of the left-hand side and top right of the UMAP plot respectively, compare Figure 3.2 (B)). The gene IGKC is an immunoglobulin gene, an antibody component found in B cell subsets, particularly plasma cells [95]. Cells expressing IGKC are also JCHAIN+ and positive for antibody subtypes suggestive of class switching (e.g., IGHG1 and IGHA1). Since this data set comprises T cells, this switching behaviour suggests that the cells in question are doublets (two cells in the same experimental droplet), specifically T cells binding B cells. While not representing single-cell states, it is important that these readings are picked up in the analysis. The gene IFI27 is part of an antiviral/interferon-induced (IFI) response signature. It is particularly interesting that MLS can detect a specific transcriptional programme shared across multiple cell types (CD4+ and CD8+ T cells). This programme could be a shared T cell programme directed against viruses or induced during stress responses (e.g., for scRNA-seq processing) [149].

Similarly, at resolution t_2 AREG and GZMB show high consistency with the topological structure, both in relation to other genes and to other time scales. In particular, AREG is expressed highly on a group of cells which connects a cluster of natural killer T cells (bottom centre in UMAP plots) with most of the remaining cells. Within the immune system, AREG is expressed by subsets of NK cells and other types of innate lymphoid cells (ILCs) [40, 105], where it plays an important role in mediating type 2 immunity [105, 164]. Despite bridging different clusters in our global clustering and that of the original authors [91], this population likely represents cells in various states of transition between two previously described AREG+ NK cell phenotypes: one with high levels of secreted molecules associated with effector functions (CCL3, CCL4) and the other expressing homing receptors associated with a more circulatory phenotype (CD44, SELL) [40] (see Figure 3.2). Therefore, while the original authors identified these two subsets as discrete NK and type 1 ILC-like cell types, respectively (Figure S13 in [91]), the MLS-informed approach highlights AREG as a shared feature, supporting the notion that these two populations may be consistent with a more continuous transition between CCL3+ and SELL+ states within the NK cell population. This interpretation is further supported by the preserved expression of NK cell markers (e.g., CD94/KLRD1, NKG2A/KLRC1) in the SELL+AREG+ cells, which is often considered to be a feature of NK cells that is not shared by otherwise closely related type 1 ILCs [10, 13].

Similarly, GZMB is highly expressed on the intersection of exhausted and proliferating T cells, two clusters which are visible in the community structure at t_2 but merge at t_3.

Finally, at Markov time t_3 FGFBP2 and NKG6 are examples of genes with relatively low expression that are highly and consistently expressed on the cluster of natural killer T cells.

A comparison to a standard differential expression test is given in Table A.1 and Figure A.1 in the appendix.

Next, we compare the trajectory inference obtained by Mapper (with a UMAP Diffusion cover with $t = 1$, $m = 1000$ and an agglomerative clustering algorithm

using cosine dissimilarity, average linkage, and a distance threshold of 0.85) and the trajectory obtained by PAGA (with $k = 15$, the Leiden algorithm [143] with resolution parameter 0.8 and edge threshold 0.3). These graphs can be seen in Figure 3.3. The PAGA graph places a large number of cells in a single node, while Mapper places the same cells into several nodes connected by more edges. While this arguably makes the visualisation more crowded, as a result, the nodes placed further away from the main body of cells correspond to genuine outliers in the UMAP plot. Most notably, the group of five sparse nodes on the top of the Mapper plot corresponds to the group of cells with high IGKC expression (c.f. Figure 3.2). In contrast, the majority of leaf nodes in the PAGA graph do not seem to exhibit markedly different gene expression patterns and seem to instead correspond to outer areas of the main body of cells in the UMAP plots.

An advantage of the PAGA graph is that it contains fewer nodes and edges. While it is theoretically possible to reduce the number of nodes in the Mapper graph by increasing the number of cells in each cover element, the resulting graph does not capture much structure on this data set.

Mouse Colon Organoid Data

We perform the same analysis as for the T cell data on the mouse colon organoid data set. While this data set is smaller than the T cell data set, we have access to meta-data for the organoid data. In particular, we know which mutations are present in each organoid and can distinguish different cell types (see Figure 3.4).

Note that the APCmin cells form a cluster which is distinct from the rest of the data. The main cluster (of non-APCmin cells) contains cells of multiple different genotypes and cell types. These conditions do not form distinct clusters in the UMAP plot and there is a continuous transition between these genetic backgrounds and cell types.

When partitioning the UMAP graph into communities at different scales, we again observe three Markov times at which the variation of information attains a local minimum (see Figure 3.5). At the finest resolution, t_1, there is a correspondence

between clusters and (groups of) cell types. At the intermediate resolution, t_2, there is a correspondence between (groups of) genetic conditions and communities. The coarsest resolution, t_3, gives a partition into APCmin and non-APCmin cells. We conclude that the resolutions found by minimising the variation of information are biologically meaningful.

The multiscale Laplacian score also identifies a number of genes that are expressed at different resolutions: for example, at resolution t_1 (see Figure 3.5) Tff3 and Agr2 are expressed consistently at high levels on the Goblet-like cells. Agr2 and its associated protein are involved in metastasis and promote invasive behaviour of gastric cancer cells [145], while Tff3 is known to correlate with poor survival rates of gastric cancer patients [77, 103], but its function in cancer cells seems to be less well understood.

At the intermediate resolution, t_2, Krt18 and Krt7 are consistently expressed in a (connected) subset of the APCmin cluster. Krt7 supports the progression of cancerous cells in gastric cancers [75]. Moreover, Krt18, a gene involved in the activation of the MAPK pathway in gastric cancers [58], is expressed highly at the intersection between the Wild Type and KRAS cells, which merge into one community at the coarsest resolution t_3.

Finally, at resolution t_3, H2afj and Ccng1 have a relatively low MLS (compared to the standard Laplacian score). Both are expressed consistently on the APCmin cluster. H2afj is known to be involved in resistance to chemoradiation in human colorectal cancers and high H2afj expression correlates with significantly worse relapse-free survival in patients [156]. Similarly, Ccng1 is known to be a key gene involved in drug resistance and cell proliferation in gastric cancers [132, 76].

A comparison to a standard differential expression test is provided in Table A.2 and Figure A.2 in the appendix.

For the trajectory inference, both PAGA and Mapper identify the APCmin cells as a distinct cluster. However, the Mapper graph (parameters are identical to those used on the previous data set) seems to outperform the PAGA graph at capturing the linear arrangement of the main cluster of cells, as is clearly visible in the UMAP plots. Mapper also infers the circular arrangement of cells, seen in the UMAP plots

(Figure 3.4) between the Wild Type and KRAS cells.

3.5 Discussion

Single-cell RNA sequencing data allows us to study cell transitions between genotypes and disease states at an unprecedented resolution. With increasing resolution, we are increasingly able to detect the continuity of such transitions. Continuous transitions between cell states question the standard model of the cell type as a discrete notion.

In this chapter, I have presented two novel solutions to analyse the continuous structure of scRNA data. First, I present a novel method for selecting a cover in Mapper that is well-motivated for scRNA data: it is based on the UMAP graph, which can model the continuous structure of the data, and adapts for locally varying sampling density, thereby decreasing the likelihood of rare intermediate cell states getting lost in the analysis. Crucially, the cover is not picked by hand, as in previous applications to scRNA data [122], and uses only two hyperparameters. Visual inspection of the resulting trajectories on the lung-infiltrating T cell data [91] and the mouse colon organoid data shows that Mapper in conjunction with the proposed heuristic for picking a cover can infer trajectories that are similarly informative than analyses with PAGA, an established state-of-art trajectory inference algorithm.

Second, I proposed a novel method for feature selection, which takes into account the continuous data structure and extends work by Govek et al. [60]. The method, called the multiscale Laplacian Score, can select features consistent with the continuous structure of scRNA data sets at multiple resolutions. It also comes with a heuristic of identifying resolutions in a data set at which the data exhibits a stable structure. By applying this new method to two data sets, I demonstrated that it identifies a number of biologically relevant genes at each resolution and thus yields additional information to the method of Govek et al. [60].

While I compared the output of the Mapper graph using a UMAP diffusion cover to PAGA, a state-of-art method for trajectory inference, further benchmarking on additional data sets and additional methods is important for future work. The review

by Saelens et al. [125] provides a general framework and a number of metrics and data sets for such benchmarking. Unfortunately, comparison to existing Mapper methods, such as [122, 118], will be difficult as the covers are chosen by hand (and are not disclosed to the best of our knowledge) and use proprietary software.

When applying the multiscale Laplacian score on scRNA-seq data, I focused on modelling the geometry of cell space with a specific k-nn graph constructed using UMAP. The MLS is flexible for use on other graphs, such as Mapper graphs [122, 118], but the resulting analysis would change if the underlying cell graph changes. Similarly, a graph could be fitted to be most consistent with the graph signals given by the genes using a Laplacian, with such a graph fitting procedure having been proposed by Daitch, Kelner and Spielman [43]. The effects of using different graphs to model data geometry on the MLS remain future work. Some of these suggested graphs, including Mapper graphs, aggregate several cells into a node. Different methods for aggregating expression values and their effects on downstream analyses could be explored.

Recently, a method for automated scale detection in multiscale community detection has been proposed [131]. This method could also be applied to the MLS. The choice of resolution(s) for the MLS is not limited to Markov stability times (e.g. graph wavelets [144] could be an alternative). Future directions include extending these signal selection approaches to other signals more generally (e.g., epigenetic factors), other complex single-cell network structures [79] or other higher-order networks [130, 14], with a view towards data integration [88].

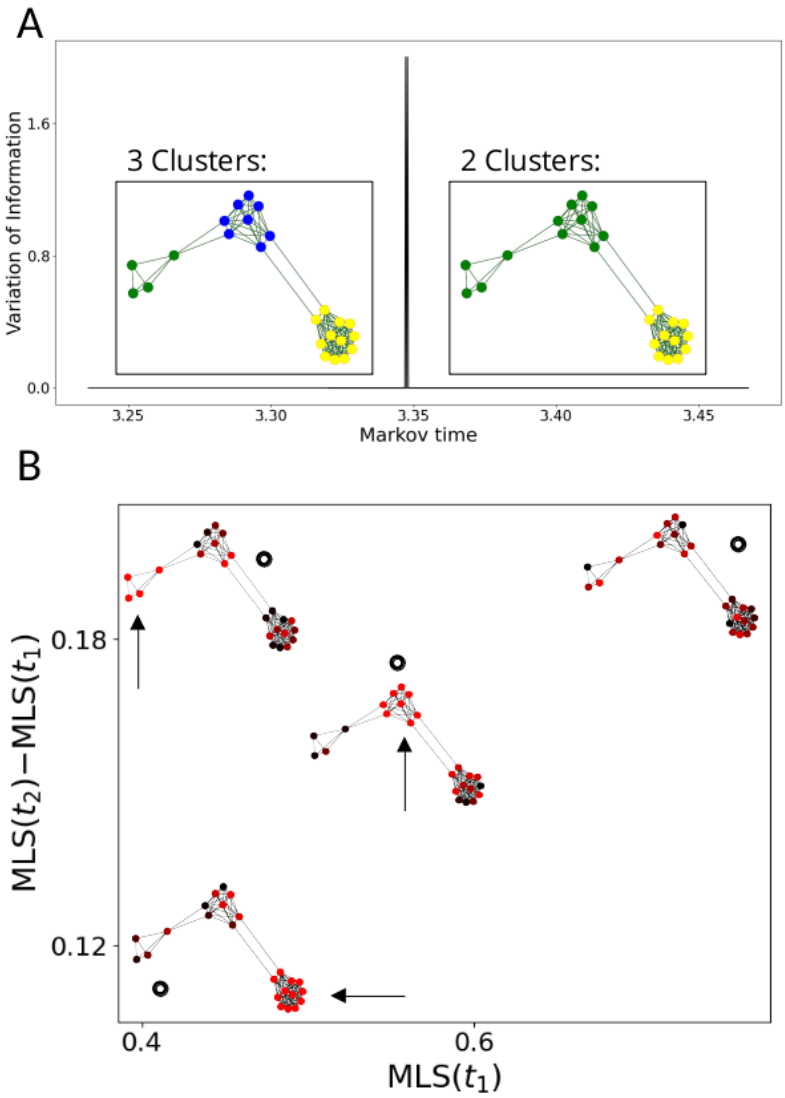


Figure 3.1: We construct a graph with three communities, all of different sizes. (A) the VI (on y-axis, VI is 0 except for a brief spike around $t = 3.35$) identifies resolutions t_1, at which all three communities are identified, and t_2, at which two communities are identified (note that due to the simplicity of the graph, there are intervals of local minima instead of points; we pick t_1 before the spike and t_2 after). In (B), we calculate the MLS at t_1 and t_2 (given by black circles) of three signals that are equal to 1 on one of the t_1-communities (constant part of the signal is highlighted by arrows) and uniformly random elsewhere, and one completely random signal. The signal that is constant on the largest cluster (bottom left) is identified as highly consistent at both times. The random signal (top right) is identified as inconsistent at both times. Conversely, the signal constant on the smallest community (top left) has a high MLS at t_2 relative to the MLS at t_1, separating it from the signal constant on the community of intermediate size (centre).

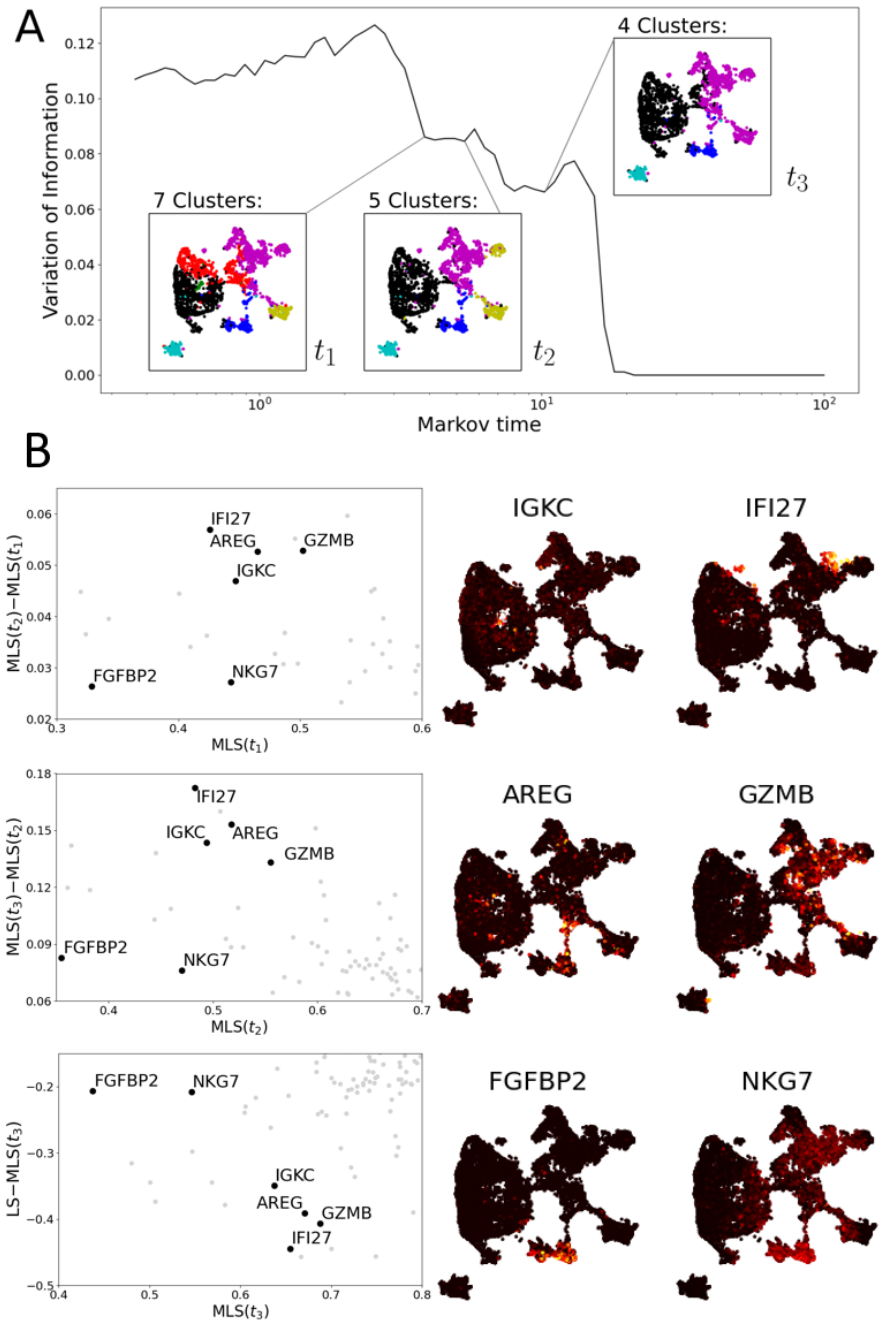
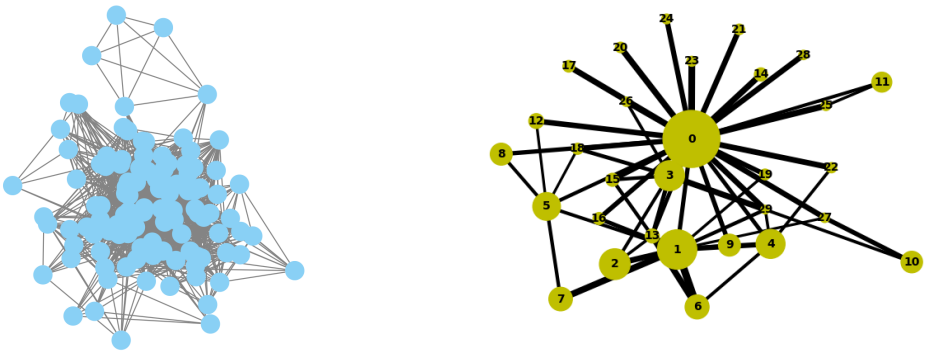


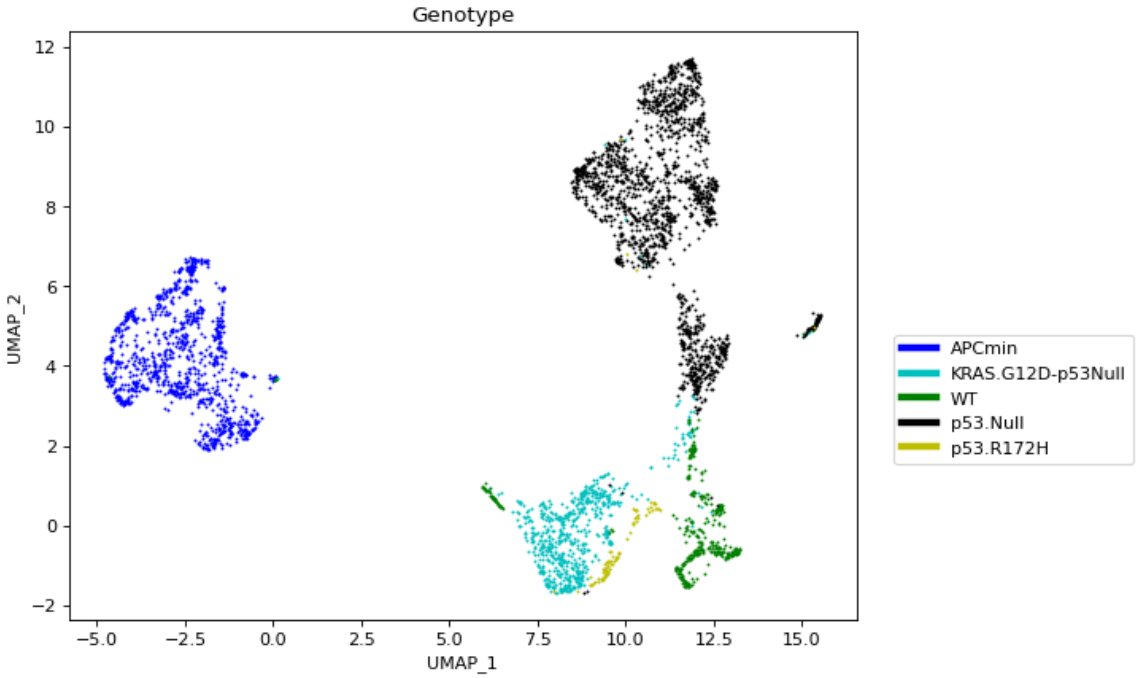
Figure 3.2: (A) The graph of the variation of information of the community structures returned by 100 iterations of the Louvain algorithm at each Markov time. The algorithm is run on the UMAP-weighted k-nn graph associated with the T cell data set. Local minima indicate stable community structures and thus scales of interest. The community structures at three such minima are by colourings of UMAP plots. (B) Left: three scatter plots comparing the multiscale Laplacian scores of genes (grey dots) at successive times to one another and of the final time to the combinatorial Laplacian score. We highlight 6 genes of interest (black dots; annotated). Middle and Right: UMAP plots visualising the expression of the genes of interest.



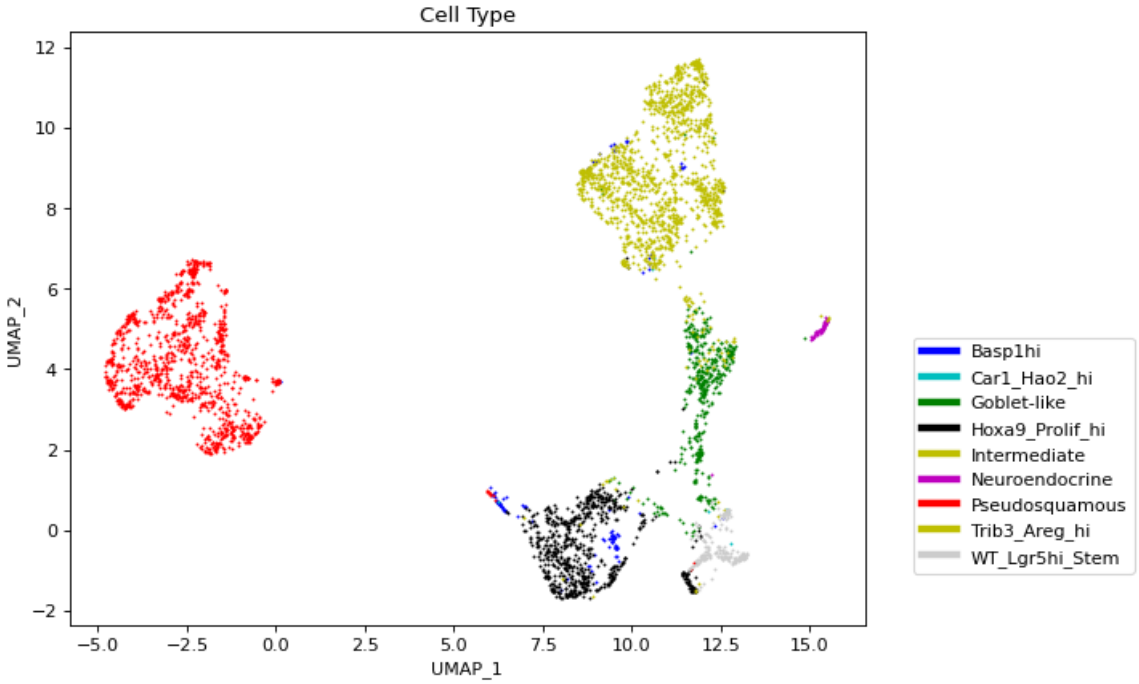
(a) Mapper graph

(b) PAGA graph

Figure 3.3: The Mapper graph (a) and the PAGA graph (b) as examples of inferred trajectories on the T cell data from [91]. In the Mapper graph, each node contains a similar amount of cells, while the number of cells in a node of the PAGA graph varies significantly (size of node indicates number of cells contained). The five sparse nodes visible at the top of the Mapper graph correspond to the group of cells with high IGKC expression (see Figure 3.2), while most leaves in the PAGA graph do not correspond to outliers visible in the UMAP plots.



(a) UMAP plot of mouse colon organoid scRNA data coloured by genotype.



(b) UMAP plot of mouse colon organoid scRNA data coloured by cell-type.

Figure 3.4: UMAP plots of mouse colon organoid scRNA data.

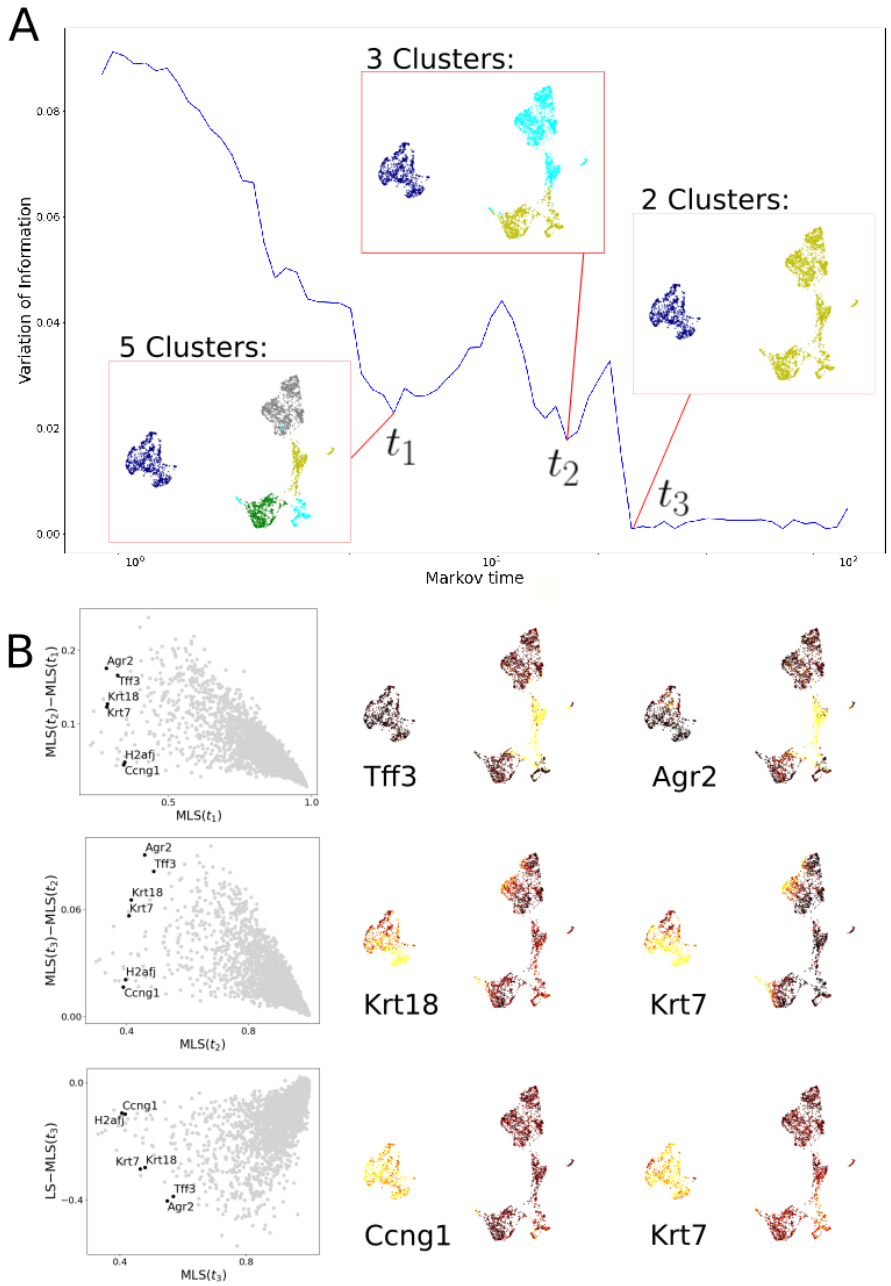
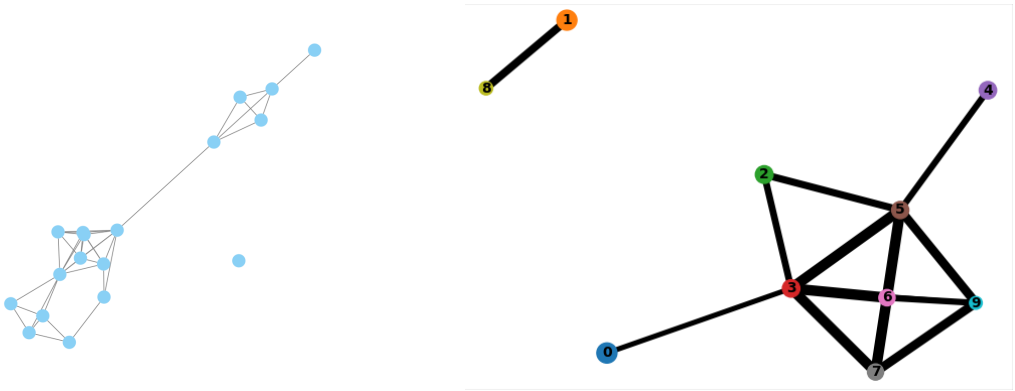


Figure 3.5: (A) The graph of the variation of information of the community structures returned by 100 iterations of the Louvain algorithm at each Markov time. The algorithm is run on the UMAP-weighted k-nn graph associated with the mouse colon organoid data set. Local minima indicate stable community structures and thus scales of interest. The community structures at three such minima are by colourings of UMAP plots. (B) Left: three scatter plots comparing the multiscale Laplacian scores of genes (grey dots) at successive times to one another and of the final time to the combinatorial Laplacian score. We highlight 6 genes of interest (black dots; annotated). Middle and Right: UMAP plots visualising the gene expression of the genes of interest.



(a) Mapper graph (b) PAGA graph

Figure 3.6: The Mapper graph (a) and the PAGA graph (b) as examples of inferred trajectories on the mouse colon organoid data.

Chapter 4

TDA of Experimental 2D Organoid Data

Chapter Content

4.1 Statistical Analysis Methods 70
4.1.1 Random Forest Classification 70
4.1.2 Canonical Correlation Analysis 73
4.2 Temporal Shape Detection with DETECT 74
4.3 Data Set 75
4.4 Results 76
4.4.1 Regressing SECT to Classical Shape Statistics 77
4.4.2 Classification of Organoids with DETECT 79
4.5 Discussion 80

An organoid is typically approximately circular in shape at the start of an experiment (in the 2D, top-down view). As organoids grow and stem cells differentiate, they can adopt different shapes that are altered by various pathological conditions [78]. Understanding how the morphology of an organoid changes over time can provide important insights into disease progression and treatment as well as an inexpensive and non-invasive quantification of tissue health. Previously, simple measures such as cell numbers, organoid volume, and shape factor were used to analyse organoid morphology [27, 69, 84, 161], but they often lack discriminative power. Deep learn-

ing methods can extract morphological features [62, 83], but they are not easily interpretable. Mechanistic models have been developed [69, 140, 161] and fitted to experimental data, but this approach lacks computational scalability.

In this chapter, I first introduce statistical methods used in this chapter (Section 4.1). I then propose a new signature called DETECT (DEtecting Temporal shape changes with the Euler Characteristic Transform; Section 4.2) that can accurately predict classical shape descriptors and classify organoids based on their morphology. DETECT is a novel extension of the SECT (see Section 2.5.5) and is scalable and interpretable through theoretical underpinnings from topology. I demonstrate the efficacy of DETECT using a data set of mouse small intestine organoids (introduced in Section 4.3) and show that DETECT outperforms simple shape descriptors at classifying organoids based on their morphology (Section 4.4). The chapter concludes with a discussion of the results and the potential implications of DETECT for future research in organoid morphology (Section 4.5).

4.1 Statistical Analysis Methods

4.1.1 Random Forest Classification

Random forests are an ensemble classification technique using a collection of tree-structured classifiers [23].

Definition 4.1. Let $\mathcal{X} = \mathbb{R}^d$ be a Euclidean space and $C = \{1, ..., c\}$ be a discrete set of categories. A *decision node* is a tuple (i, t, p, n) where $i = 1, ..., d$ is a feature index, $t \in \mathbb{R}$ is a decision threshold and p and n both either decision nodes or a category.

A collection of decision nodes $G = \{(i, t, p, n), ...\}$, we can endow G with a graph structure by adding a directed edge from $v = (i, t, p, n) \in G$ to $v' = (i', t', p', n') \in G$ whenever $p = v'$ or $n = v'$.

A collection of decision nodes $T = \{(i, t, p, n), ...\}$ is called a *decision tree* if its corresponding graph has the structure of a tree and has exactly one node with in-degree 0 (called the root).

A decision tree T classifies a data point $x \in \mathcal{X}$ into categories C as follows:

1. Find the root node $v = (i, t, p, n) \in T$.

2. If $x(i) \leq t$, then $v = p$. Else, set $v = n$.

3. If $v \in C$, terminate the classification and return the category v. Else, if v is a decision node, return to step 2.

This classification process is illustrated in an example in Figure 4.1.

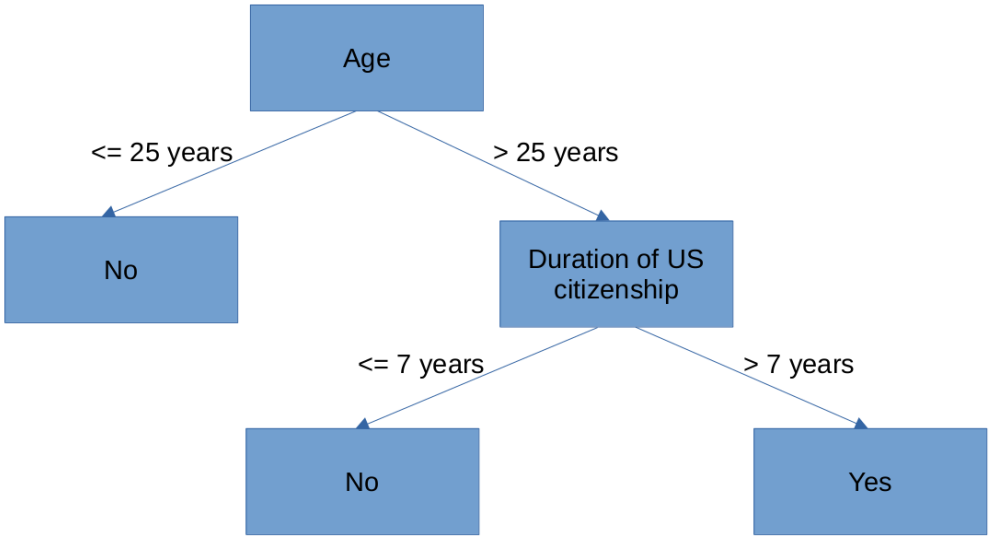


Figure 4.1: An example of a decision tree determining if a US national (represented as a tuple of their age and the duration for which they have held citizenship) is eligible to serve as a Representative. First, it is determined whether they are older than 25 years. If not, they cannot serve. If yes, the tree checks if they have held citizenship for more than 7 years, in which case they are eligible to serve.

Given a training data set $X \subset \mathcal{X}$, a decision tree can be trained by identifying the feature i and decision threshold t that minimises the overall classification error. The pair (i, t) splits the data into two: one subset of the data in which feature i is less or equal to t and another on which it is larger. As long as these subsets are of size two or larger, we can train further nodes on the two respective subsets by finding splits along a feature minimising the classification error. These two nodes are added to the pair (i, t) as p and n respectively. If all points in one of the two subsets are of the same category, we insert that category instead of inserting a new node. We iteratively apply this principle to all new nodes until the training procedure terminates.

In practice, early termination is often enforced in the above procedure. E.g. the procedure is halted at a node if the depth of a node (i.e. distance to the initial node) has reached a certain threshold or the size of the (iteratively split) data set at a node has fallen below a threshold. In that case, the category most frequent in the sub-data set at that node is inserted instead of a new node. Similarly, the optimisation problem of finding (i, t) with a minimal classification error may be further constrained to having to split the data into two subsets of at least some minimum size. All of the above rules regularise the training of decision trees.

However, even with regularising training procedures in place, single decision trees tend to overfit the training data [73]. To overcome this issue, Ho [73] introduced the method of random forests:

Definition 4.2. A *random forest* is a classifier consisting of multiple decision trees $RF = \{T_i\}_{i=1,\dots,k}$. At each input data $x \in \mathcal{X}$, RF returns the category most frequently returned by the decision trees T_i when they are applied to x.

In other words, RF decides the class of x by a majority vote of its trees. The name *random* forest derives from the common method for training random forests: For a given training data set of size n with m features, fix $n' < n$ and $m' \ll m$. Then each T_k is trained, using the procedure described above, on a random subsample of the data of size n' (sampled uniformly with replacement) and on m' random features of the data (sampled uniformly without replacement). Additional rules may still apply in the training of the individual trees T_i.

Training individual trees T_i on random subsets of the data and the features and then taking a majority vote makes random forests much more robust to outliers while retaining the capability of learning complex decision boundaries [73]. In particular, it can be proven that the generalisation error of random forests converges [23]; i.e., random forests do not overfit to training data.

4.1.2 Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a method for both identifying and quantifying the association of two sets of random variables [80]. It does so by maximising the correlation between linear combinations of both sets of variables. We can place both sets of random variables into vectors $\mathbf{X} = [X_1, ..., X_p]^T \in \mathbb{R}^p$ and $\mathbf{Y} = [Y_1, ..., Y_q] \in \mathbb{R}^q$. Then, assuming that all X_i and Y_j have finite second moment, we define

$$\mu_X = \mathbb{E}[\mathbf{X}] \in \mathbb{R}^p, \qquad \mu_Y = \mathbb{E}[\mathbf{Y}] \in \mathbb{R}^q,$$

$$\Sigma_{XX} = \mathbb{E}\left[(\mathbf{X} - \mu_X)(\mathbf{X} - \mu_X)^T\right] \in \mathbb{R}^{p\times p}, \quad \Sigma_{YY} = \mathbb{E}\left[(\mathbf{Y} - \mu_Y)(\mathbf{Y} - \mu_Y)^T\right] \in \mathbb{R}^{q\times q},$$

$$\Sigma_{XY} = \mathbb{E}\left[(\mathbf{X} - \mu_X)(\mathbf{Y} - \mu_Y)^T\right] \in \mathbb{R}^{p\times q}, \qquad \Sigma_{YX} = \Sigma_{XY}^T \in \mathbb{R}^{q\times p}.$$

In the first instance, CCA seeks (deterministic) coefficient vectors $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{b} \in \mathbb{R}^q$ such that

$$\text{Corr}\left(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y}\right) = \frac{\mathbf{a}^T\Sigma_{XY}\mathbf{b}}{\sqrt{\mathbf{a}^T\Sigma_{XX}\mathbf{a}}\sqrt{\mathbf{b}^T\Sigma_{YY}\mathbf{b}}} \tag{4.1}$$

is maximised. As the above formula clearly is scale-invariant in $\mathbf{a}$ and $\mathbf{b}$, we can introduce the further requirement that $\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2 = 1$. Call the vectors maximising Equation (4.1) $\mathbf{a}_1$ and $\mathbf{b}_1$ and define $U_1 = \mathbf{a}_1^T\mathbf{X}$ and $V_1 = \mathbf{b}_1^T\mathbf{Y}$. We call the maximal value of (4.1), denoted by ρ_1, the first *canonical correlation* of $\mathbf{X}$ and $\mathbf{Y}$. The variables U_1 and V_1 are called the first canonical variables.

We then iteratively define U_k and V_k by maximising Equation (4.1) with the additional constraints that $\text{Cov}\left(\mathbf{a}_k^T\mathbf{X}, U_i\right) = \text{Cov}\left(\mathbf{b}_k^T\mathbf{Y}, V_i\right) = 0$ for all $i = 1, ..., k-1$. Similarly, the k-th canonical correlation ρ_k is defined to be the correlation of the k-th canonical variables U_k and V_k. We can compute $\min\{p, q\}$ canonical correlations in this fashion (assuming that Σ_{XX} and Σ_{YY} are both invertible).

The coefficients $\mathbf{a}_i$ and $\mathbf{b}_i$ can be computed as follows:

Lemma 4.3 (Result 10.1 in [80])**.** *Let $\mathbf{e}_i$ and $\mathbf{f}_i$ be eigenvectors corresponding to the i-th largest eigenvalues of $\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1/2}$ and $\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}$, respectively. Then,*

$$\mathbf{a}_i \propto \Sigma_{XX}^{1/2}\mathbf{e}_i, \qquad \mathbf{b}_i \propto \Sigma_{YY}^{1/2}\mathbf{f}_i.$$

Further,

$$\mathbf{f}_i \propto \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}\mathbf{e}_i.$$

Interpreting canonical variables

Let A be the matrix with columns $\mathbf{a}_i$ and E the orthogonal matrix with rows $\mathbf{e}_i$. Then using the above result and the spectral decomposition of $\Sigma_{XX} = P^T\Lambda P$, we get

$$A = E\Sigma^{1/2} = EP^T\Lambda^{1/2}P$$

up to a constant scalar. We can thus interpret the action of the matrix A on $\mathbf{X}$ as follows: it first transforms $\mathbf{X}$ to its principal components, then standardises the along the principal axes, before applying a rotation [80].

Further, CCA can be viewed as a generalisation of other correlation analyses, such as Pearson correlation, multiple correlations and principal component analysis (PCA) [80, p. 547].

Via the interpretation of CCA as a generalised multiple correlation analysis, it can be shown that ρ_k^2 gives the proportion of the variance of U_k that can be explained by (linear combinations of) $\mathbf{Y}$ (and vice-versa between V_k and $\mathbf{X}$). However, U_k need not represent a lot of the variance in $\mathbf{X}$, which needs to be checked separately. The proportion of the variance in $\mathbf{X}$ that can be explained by U_i, $i = 1, ..., k$ can be computed by

$$R_X = \frac{\operatorname{tr}\left(\sum_{i=1}^{k}\mathbf{a}_i\mathbf{a}_i^T\right)}{\operatorname{tr}\left(\Sigma_{XX}\right)}.$$

The formula for R_Y is defined analogously. These quantities should be checked by investigators before drawing conclusions from a CCA.

4.2 Temporal Shape Detection with DETECT

The SECT transforms a fixed, static shape into a functional signature. We now extend the definition of the SECT to get a rotationally invariant temporal signature of a sequence of shapes:

Definition 4.4. Let $T = [0, c]$ or $T = \{0, 1, ..., n-1\}$, where $c \in \mathbb{R}$ and $n \in \mathbb{N}$, be a set of time points. Let $\{\mathcal{K}(t)\}_{t \in T}$ be a sequence of simplicial complexes embedded in $\mathbb{R}^d$. Then the *DETECT* (DEtecting Temporal shape changes with the Euler Characteristic Transform) of this sequence is the transform

$$\text{DETECT}\left(\{\mathcal{K}(t)\}\right) : T \to L^2([-a, a]), \qquad t \mapsto \left(x \mapsto \int_{S^{d-1}} \text{SECT}_{\mathcal{K}(t)}(v, x)\, \mathrm{d}v\right).$$

Here, we use DETECT with $T = \{0, ..., n-1\}$. In practice, after integrating out the dependence on the direction, DETECT is a function from T to $L^2([-a, a])$. Such functions form an infinite dimensional vector space and thus a finite presentation is not possible in general. We, therefore, evaluate DETECT at any fixed $t \in T$ on a finite number of evenly spaced points P in $[-a, a]$. DETECT is then represented approximately by a $|T| \times |P|$-matrix. In this chapter, we apply DETECT to two time-course data sets of organoid boundaries. We consider the space of such matrices to be endowed with the $\|\cdot\|_2$-norm.

4.3 Data Set

We first acquired a set of imaging data derived from time-lapse imaging of mouse small intestine organoids. In total, we have 176 organoids and 320 video frames for each organoid. The data set comprises of 74 wild-type (WT) and 102 p53 knock-out (KO/mutant) genetics organoids. Both groups of organoids further split into untreated organoids (CNT) and organoids treated with valproic acid and GSK3 inhibitor CHIR99021 (VC). These organoids have been filmed throughout their growth and the resulting videos have been segmented. After segmentation, we have 100 points summarising the boundary of each organoid at each video frame (frames starting at the beginning of the experiment and are taken every 15 minutes henceforth). We discard some videos which for technical reasons have fewer than 320 video frames, as most videos below this threshold still seem to change their morphology at the end of the video.

The organoid boundaries in this data set are, in the 2D video view, close to being perfectly circular in the early stages of the videos across all experimental conditions.

This simple geometry is a result of cellular homogeneity in the early phases. As time progresses, cells proliferate and differentiate. Through proliferation, organoids grow in size, and through stem-cell differentiation, the cellular composition of organoids changes. As different cell types have different mechanistic properties and differentiation is not spatially uniform, organoids cease to be spherical. Most notably, they elongate and their boundary buckles, possibly leading to the growth of finger-like protrusions. Such growth behaviour is illustrated by the examples of final video frames presented in Figure 4.2.

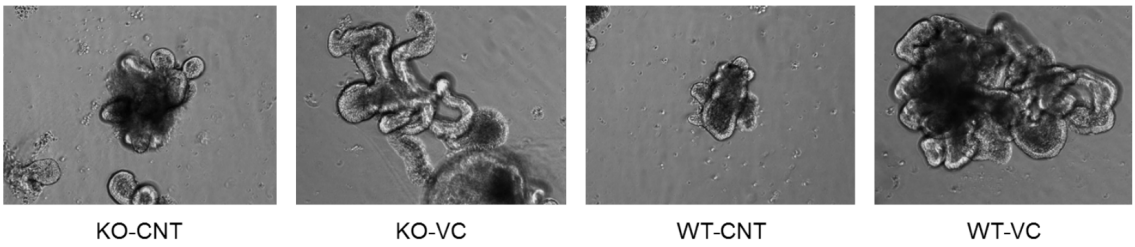


Figure 4.2: Phenotype effect of VC treatment to intestinal organoids. Static video snapshots of the final frame. One example is shown for each condition.

Each collection of boundary points is transformed into a simplicial complex representing the organoid boundaries, yielding a sequence of simplicial complexes indexed by $t = 0, ..., 319$ for each organoid. We re-centre each simplicial complex such that the mean of all vertices is the origin. We then compute the radius of the simplicial complex at $t = 0$ (i.e. the largest norm of all vertices after re-centring) and divide all vertices in the sequence of simplicial complexes by that value. We translate and scale the data in this way to simplify it, given its limited size. As a result, the initial size of organoids or any movement throughout time is not considered by any downstream analysis, including DETECT.

4.4 Results

We analyse the shape of organoid boundaries of experimental (2D) data by first building a simplicial complex representation. We then compute the SECT of each organoid at each time point and integrate out its S^1 component to obtain DETECT.

When computing DETECT, we use $a = 6$ for the 2D organoids and P to be 100 evenly spaces points in $[-a, a]$. We then compute a 100-dimensional feature embedding. We apply a Gaussian kernel with $\lambda = |N| \times |P|$).

4.4.1 Regressing SECT to Classical Shape Statistics

We first compute the SECT of static images of experimental organoids and demonstrate that the SECT (after its S^1 component has been integrated out) includes information conveyed by classical shape statistics. Henceforth, we call this signature the *marginalised SECT*. The classic shape statistics, diameter, mean and max centroid distances, the equivalent diameter, the major and minor axis lengths and the area of the convex hull, quantify the geometric properties of a 2D shape. These statistics are widely used and invariant under translation and $O(2)$-actions. As each of these statistics is calculated for a static shape, i.e. for an organoid boundary at a fixed time frame, we compare these statistics to the marginalised SECT at fixed time frames.

To compare the aforementioned shape statistics with the marginalised SECT, we apply a standard linear regression model. The marginalised SECT of each organoid at each time is represented by a Nystroem feature (c.f. Section 2.5.6; we set $m = 100$). We project the feature embedding to 50 dimensions using PCA [114]. The PCA vectors give the independent variables, while the classic shape statistics listed above are viewed as the dependent variables. We pass the square-root values of the convex hull area to the regression model, as it has a squared relationship with the remaining metrics in the (default) symmetric cases. An illustration of the main notions is given in Figure 4.3. We perform a 50-fold cross-validation for each metric and report mean coefficients of determination and standard deviations of the coefficient of determination in Table 4.1.

We find that the marginalised SECT regresses multiple classical shape statistics with high accuracy. The marginalised SECT has high predictive accuracy of equivalent diameters and perimeter and, as a result, can also detect symmetry breaking. The lower accuracy of the minor axis length and convex area suggests that the marginal-

	R^2-scores	std
equivalent diameter	0.880	0.073
max. centroid distance	0.916	0.041
mean centroid distance	0.894	0.067
major axis length	0.880	0.052
minor axis length	0.630	0.246
perimeter	0.972	0.024
$\sqrt{\text{convex area}}$	0.868	0.105

Table 4.1: Mean coefficients of determinations (R^2-scores) and their standard deviations (std) in a 50-fold cross-validation of a Linear regression in which the Nystroem embedding of the marginalised SECT gives the independent variables and the 8 variables above give the dependent variables.

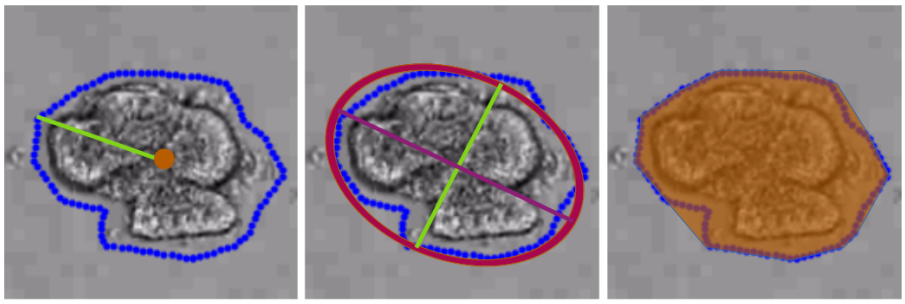


Figure 4.3: Left: The red point gives the centroid (point which minimises mean squared distance to boundary) of the organoid and the green line gives the distance to a boundary point. The maximal and mean lengths of such green lines give the max. and mean centroid distance. Centre: The red ellipse gives the ellipse with the best fit to the boundary. The purple line gives the major axis and the green line the minor axis. Right: the area of the convex hull (convex area) is visualised in opaque red.

ised SECT is more limited in its ability to capture the (mean) size of indentations, compared to detecting size and elongation. We remark that this limitation may be related to segmentation accuracy, and therefore, we consider organoid shapes with known segmentation (i.e. synthetic data) in Chapter 5.

In addition to the above regression analysis, we can decompose the covariance matrix of the aforementioned classical shape statistics by its singular values. We remark that we standardise the data in each feature before computing the covariances. We observe that the first four principal components of the classical shape statistics explain over 90% of the variance in this data set (see Figure 4.4). We then

perform a canonical correlation analysis (see Section 4.1.2) [157] between these four principal components and the PCA-transformed marginalised SECT data. We find that the first four pairs of canonical variables have a perfect correlation score of 1.0 and conclude that the marginalised SECT can explain over 90% of the variance in the classical shape statistics on the given data set.

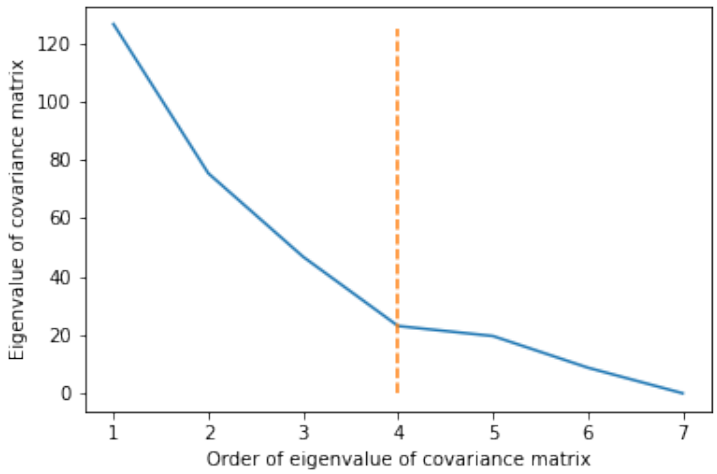


Figure 4.4: The eigenvalues (y-axis) of the covariance matrix of the classical shape descriptors in decreasing order (indices of eigenvalues on x-axis). The data was standardised in each component before the covariance matrix was computed. The first four eigenvalues (to the left of the dashed line) account for approximately 90.5% of the variance in the data.

4.4.2 Classification of Organoids with DETECT

We focus on classifying p53-knock-out organoids into treated and untreated groups. Before training a classifier, we seek to exclude some organoids where the segmentation does not accurately trace the organoid boundary in the video. In these cases, the segmented boundary is significantly larger than the true organoid boundary. Therefore, we exclude organoids whose radius grows by a factor of more than two. For p53-knock-out organoids, this procedure excludes two out of 98 organoids. Of the remaining 96 organoids, 55 are untreated and 41 are treated.

We use random forest classification (see Section 4.1.1) [23] to classify p53-knock-out organoids into untreated and VC-treated experimental groups. Random forest

classification then trains an ensemble of decision trees trained on random subsets of the Nystroem-transformed DETECT data. The trees classify data points by majority vote. We use the `scikit-learn` [115] implementation of random forest classification and optimise the hyper-parameters of the maximum tree depth, the minimum number of samples allowed to define a split and the minimum number of samples per tree leaf by cross-validation grid search (`GridSearchCV` in `scikit-learn`). Based on this optimisation, the maximum tree depth is five, the minimum number of samples to define a split is five and the minimum number of samples per tree leaf is three. The 5-fold cross-validation for these parameters gives a mean classification accuracy of 68.8% with a standard deviation of 2.7%. We remark that setting the number of Nystroem features to $m = 500$ increases the mean accuracy further to 70.0% but also increases the standard deviation to 7.5%. The higher standard deviation suggests that setting $m = 500$ could result in overfitting.

The accuracy of classification results based on DETECT exceeds those based on all classical statistics (e.g. area and perimeter) which give a mean classification accuracy of 60.5% and a standard deviation of 4.8% when we use the pipeline and cross-validation method described above. Classification based on DETECT also exceeds the baseline accuracy associated with guessing, which is 57.4% as there are slightly more untreated than treated organoids in our data set. Further, we have shown that combining DETECT with machine learning can distinguish organoids treated with valproic acid and GSK3 inhibitor based on quantification of their shape dynamics as well as regress out classical shape statistics.

4.5 Discussion

In this chapter, I have introduced a new technique from the field of topological data analysis for detecting temporal shape changes with the Euler characteristic transform (DETECT). I have highlighted its utility by studying organoid morphology. I first showed that several classical shape descriptors, including the diameter, the mean and maximum centroid distances, the equivalent diameter, the major and minor axis

lengths and the area of the convex hull, can be regressed from the (marginalised) smooth Euler characteristic transform (SECT) with high accuracy. Then I applied DETECT, with kernel approximation methods and random forests, to a data set of experimental p53 knock-out mouse small intestine organoids and showed that this approach can distinguish VC-treated organoids from untreated organoids. This integration with kernel approximations enables larger data sets to be analysed as when kernel methods are used the runtime complexity of ECT can be reduced from being cubic to approximately linear in the number of data points. Such improvements in runtime also allow for the application of a wider range of statistical methods.

It remains future work to extend our findings to data sets of different types of organoids (derived from different organs, with different genetic backgrounds and/or cultured under different conditions). Further, it is important to study information loss between 3D data and their 2D projections. One possible way of approaching this problem is by considering synthetic data generated from 3D mechanistic models (such as [161]) and comparing it to 2D projections of the same data by using DETECT. However, these models neglect certain biophysical processes (e.g. the effects of gravity, the production of extracellular matrix, mechanical stress) [161]. Such a comparison is further complicated by the fact that even if there were no information loss between 3D data and its 2D projection, the DETECT signatures of 3D and 2D would be very different (due to S^2 and S^1 having different Euler characteristics).

Further worthwhile research includes extending this analysis to other types of morphological data that do not have regularised and smooth boundaries. In practice, data sets analysed by the ECT and its extensions may be noisier than the data set analysed in this chapter.

Finally, a feature selection in (kernelised) DETECT space followed by a reconstruction of a dynamic shape along that feature would be worthwhile future research: One would first identify features which vary most across different organoid categories. Second, one would then reconstruct how the temporal evolution of an organoid shape changes along that feature. Wang et al. [155] give a blueprint for such an analysis in their SINATRA pipeline. However, further theoretical work is needed to extend

their method to DETECT to account for both a shape changing over time and the signature being rotationally invariant. Once such a theoretical extension is accomplished, such an inversion of DETECT would help to gain further information on how genetics and treatments are associated with organoid morphology.

Chapter 5

TDA of Synthetic 3D Organoid Data

Chapter Content

5.1 The Model 84
5.2 Data Set 86
5.3 Results 88
5.4 Discussion 89

Organoids are inherently three-dimensional entities. Thus, the 2D view given by the videos in Chapter 4 may lose information about organoid morphology and growth. Experimental 3D segmentations of organoids have already been collected (e.g. [78]). Although we did not have access to such data for our studies, it is likely that 3D segmentations will become the predominant type of experimental organoid morphology data.

In this chapter, I analyse 3D data generated by a mechanistic model developed by Yan et al. [161] to illustrate how DETECT generalises to 3D shapes. Further, I use this data set, in which we have tight control over the factors driving perturbations to the morphology, to show that DETECT captures biologically meaningful information. In particular, I show that we can accurately cluster organoids by the proliferation rates of their constituent cells based on the DETECT signature of their 3D shape evolution.

This chapter is structured as follows: I first give a summary of the model used to generate the organoid data (Section 5.1). I then describe the data set and how it has been pre-processed (Section 5.2) before presenting the results of applying DETECT to this data set (Section 5.3). The chapter concludes with a brief discussion of the results and potential future work (Section 5.4).

5.1 The Model

The data analysed in this chapter has been generated from a continuum mechanistic model published by Yan, Konstorum and Lowengrub in [161] which describes the growth of intestinal organoids in 3D. In contrast to agent-based models (e.g. [27, 92, 140]), which resolve individual cells of an organoid and their interactions, continuum models describe organoids in terms of volume fractions of different cell types. These volume fractions are continuous functions of space and time, which can be thought of as the fraction of cells of a given type occupying an infinitesimally small neighbourhood of a point in space. Their evolution is described by a set of partial differential equations (PDEs).

The model described in [161] distinguishes three cell types:

Stem Cells (SCs): They proliferate slowly and secrete short-range self-renewal promoters (e.g. Wnt) and corresponding long-range inhibitors (e.g. Dkk). When SCs differentiate, they produce committed progenitor cells.

Committed progenitor cells (CPs): They proliferate more rapidly than SCs. When they differentiate, they produce terminally differentiated cells.

Terminally differentiated cells (TDs): They secrete differentiation promoters, forming a negative feedback loop on SC and CP renewal.

In addition to the cell types listed above, volume can also be occupied by dead material (D) and the host region in which the organoid is being cultured (H). The host region may be gel. Volume can also be occupied by non-solid material, such as water (W). The volume fraction of each material is summarised in functions

$\phi_i(x, y, z, t)$, where the subscript i indicates one of the aforementioned abbreviations ($i = SC, CP, TD, D, H, W$).

For $i = SC, CP, TD, D$, the model assumes

$$\frac{\partial \phi_i}{\partial t} = -\nabla \cdot (\mathbf{u}_s \phi_i) - \nabla \cdot \mathbf{J}_i + \mathrm{Src}_i,$$

where $\mathbf{J}_i$ is a mass flux taken to be the generalised Fick's law (see [161, Eq. (3)]) and $\mathbf{u}_s$ is the mass-averaged velocity of all solid components (see [161, Eq. (5)]). Thus, the first term on the right-hand-side of the above equation, $\nabla \cdot (\mathbf{u}_s \phi_i)$, models passive cell movement (advection), i.e. motion induced by movement of other nearby material. The second term, $\nabla \cdot \mathbf{J}_i$, models active movement of the cells, i.e. movement induced by the cells themselves. The terms Src_i represent net the net rate of production of new material. All parameters that we vary in our data set are included in these source terms. In particular,

$$\begin{aligned}
\mathrm{Src}_{SC} =& \lambda_m^{SC} n \phi_{SC} (2p_0 - 1) - \lambda_n^{SC} \mathbf{H}(\tilde{n}_{SC} - n)\, \phi_{SC} \\
\mathrm{Src}_{CP} =& \lambda_m^{SC} n \phi_{SC} 2(1 - p_0) + \lambda_m^{CP} n \phi_{CP} (2p_1 - 1) - \lambda_n^{CP} \mathbf{H}(\tilde{n}_{CP} - n)\, \phi_{CP} \\
\mathrm{Src}_{TD} =& \lambda_m^{CP} n \phi_{CP} 2(1 - p_1) - \lambda_n^{TD} \mathbf{H}(\tilde{n}_{TD} - n)\, \phi_{TD} - \lambda_a^{TD} \phi_{TD}, \\
\mathrm{Src}_{D} =& \lambda_n^{SC} \mathbf{H}(\tilde{n}_{SC} - n)\, \phi_{SC} + \lambda_n^{CP} \mathbf{H}(\tilde{n}_{CP} - n)\, \phi_{CP} + \lambda_n^{TD} \mathbf{H}(\tilde{n}_{TD} - n)\, \phi_{TD} \\
& + \lambda_a^{TD} \phi_{TD} - \lambda_L \phi_D.
\end{aligned}$$

In the above, $\mathbf{H}$ denotes the Heaviside function[1] and λ_j^i denotes the mitosis (m), necrosis (n), apoptosis (a) and lysis (L) ($i = m, n, a, L$) rates of the various cell types j, respectively. These rates are assumed to be constant in space and time. Mitosis refers to the proliferation (or self-renewal) of cells, necrosis to cell death induced by external factors (e.g. lack of nutrient availability) and apoptosis is self-induced, 'natural' cell death. Lysis refers to the disintegration of dead cells into non-solid material (here: water).

In the definition of Src_i, p_0 and p_1 denote the self-renewal probabilities of SCs and CPs, respectively, and $\tilde{n}_i$ denote the minimal nutrient levels needed to support

[1] $\mathbf{H}(x) = 1$ if $x > 0$, $\mathbf{H}(x) = 0$ otherwise.

cell viability. The term n represents the overall nutrient concentration and satisfies a reaction-diffusion equation [161, Eq. (14)]. Unlike the rate parameters λ, these probabilities and concentrations are not assumed to be constant and are governed by PDEs depending on additional model parameters. As we keep these parameters constant, I will not give further details about these PDEs, which can be found in [161, pp. 6–8]. The PDEs governing these probabilities follow a model for Turing-type pattern formation and depend on the spatial distribution of cell types as well as the availability of activators, inhibitors and feedback regulators [161].

5.2 Data Set

The authors of [161] generated a synthetic data set to verify our findings on the experimental data using their model. All but two of the parameters presented in the publication [161] are fixed. They vary $\lambda_L = 0.5$, 1, 2, the lysis parameter, and $\lambda_m^{\mathrm{SC}} = \lambda_m^{\mathrm{CP}} = 0.35$, 0.71, 1.42, cell mitosis parameters for stem cells and committed progenitor cells, respectively (all parameters are dimensionless). All other parameters are left at the default values reported in [161, Table 1]. The data set comprises the boundaries of the aforementioned organoids at times $t = 20, 40, 60, 80, 100$ (t is dimensionless).

The initial conditions are defined via

$$\begin{aligned}
\phi_T(\mathbf{x}, 0) &= 1 - \prod_{i=1}^{3} \frac{1}{2}\left(1 + \tanh\left(\frac{r_i - 2}{2\sqrt{2}\varepsilon}\right)\right) \\
r_1 &= \sqrt{(x-0.1)^2 + (y+1.2)^2 + (z+1.3)^2} \\
r_2 &= \sqrt{(x-0.2)^2 + (y-0.7)^2 + (z+1.3)^2} \\
r_3 &= \sqrt{(x+0.8)^2 + (y+0.2)^2 + (z-0.8)^2},
\end{aligned}$$

where $\varepsilon = 0.05$. We then set $\phi_{SC} = 0.1\phi_T$, $\phi_{SC} = 0.1\phi_T$, $\phi_{CP} = 0.25\phi_T$, $\phi_{TD} = 0.6\phi_T$ and $\phi_{SC} = 0.05\phi_T$ at time $t = 0$. Any remaining volume is occupied by host region and water. This setup gives an initial configuration of three overlapping spheres around the centre of the computational domain [161].

As we observe all possible combinations of death and proliferation parameters, we get time-course data for nine different computationally modelled organoids. We visualise examples of the simulated organoid development in this data set in Figure 5.1.

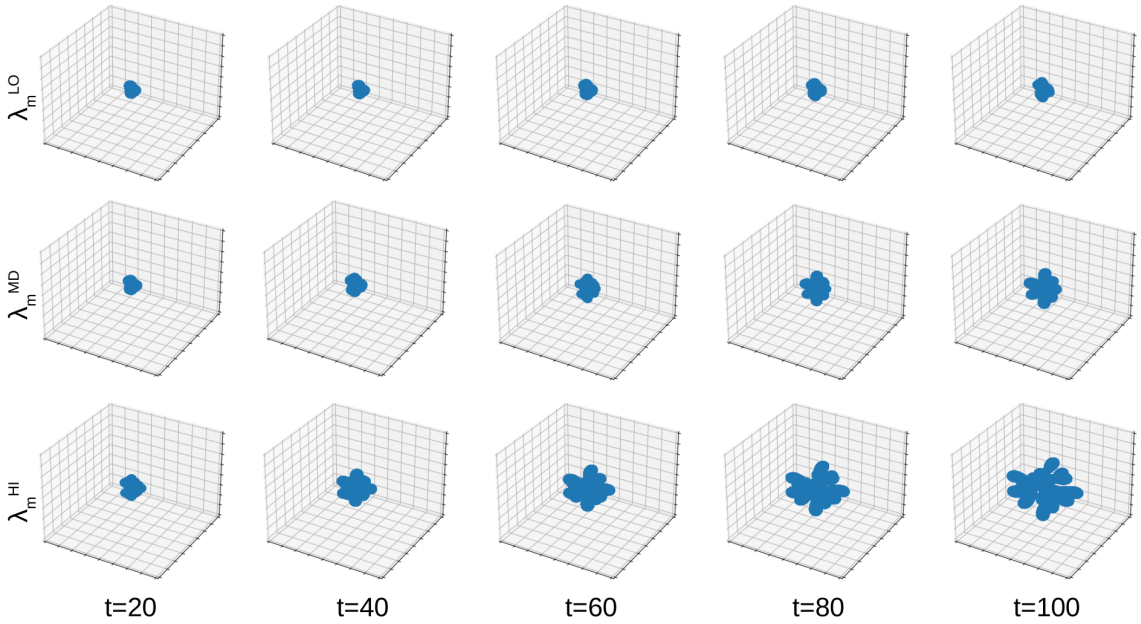


Figure 5.1: Spatio-temporal shape changes in 3D of synthetic organoids. Top column: Low (LO) mitosis rate. Middle column: Medium (MD) mitosis rate. Bottom column: High (HI) mitosis rate. In all organoids visualised, the lysis rates are at the lowest values given in the data set.

Unlike the experimental data, the number of 3D boundary points varies proportionally with the size of the simulated organoid. To ensure computational tractability, we restrict ourselves to 300 boundary points sampled uniformly at random at each time point. Different 300 samples do not lead to any notable perturbations in the downstream analyses. As some of the organoids disconnect, we first cluster the 300 points using DBSCAN [52] (with $\varepsilon = 2$, min_samples=5) to identify connected components. To triangulate the boundary surface at a given time point, we first re-centre (each connected component of) the organoid such that the mean of all boundary points is the origin. We then perform a stereographic projection into the xy-plane and perform a Delaunay triangulation and identify those points bordering an infinite area 2-cell. After projecting the finite components of the triangulations back onto the

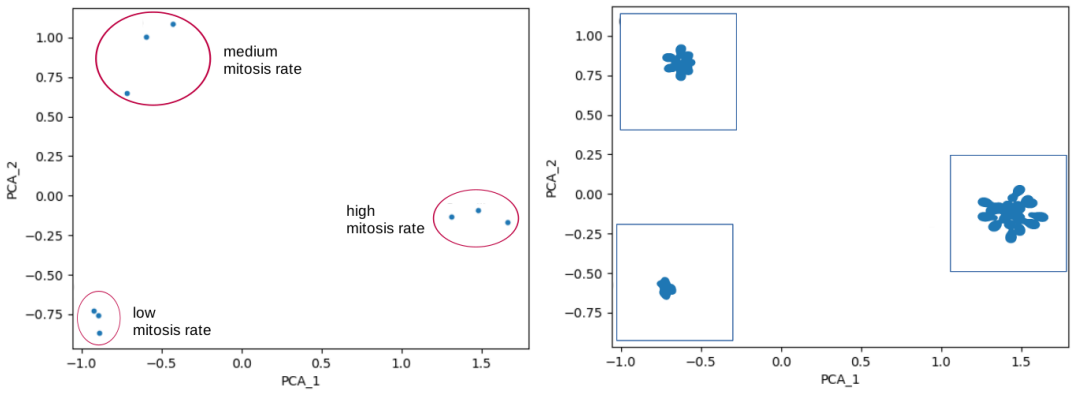
sphere, we add a further point, which is the mean of all points bordering an infinite area cell in the previous step, and insert 1 and 2-cells to fill the north-pole area of our organoid. We credit [44] with this pre-processing procedure. Unlike in the experimental data, re-scaling is not needed as all simulated organoids are identical at $t = 1$.

We perform this pre-processing step to ensure that the resulting simplicial complex has a geometric realisation homeomorphic to a sphere (or a union thereof, if an organoid disconnects). This pre-processing is a necessary step to ensure that faithful topological features (e.g. the correct number of components or holes) are present before computing Euler characteristics. Random rotations of the data ahead of pre-processing lead to negligible differences in DETECT and thus suggest this pre-processing does not introduce artificial geometric features.

5.3 Results

Finally, we apply our methodology to the synthetic data generated by the model of Yan, Kostorum and Lowengrub [161]. As described in Section 5.2, this data set contains 9 organoids and thus is too small to apply linear regression or random forest classification. We therefore only report the first two principal components of the Nystroem-transformed DETECT signatures. These outputs demonstrate that our methods generalise well to 3D shapes and identify important structures in the synthetic data, for which we know the ground truth.

This analysis, visualised in Figure 5.2, shows that organoids cluster together by their mitosis rates. This behaviour is consistent with watching the videos for these 9 organoids. We see that low-mitosis-rate organoids exhibit little growth and virtually no symmetry breaking (see Figure 5.1). Medium-mitosis-rate organoids show a little more growth than low-proliferation organoids and notable buckling of their boundary (see Figure 5.1). Finally, high-mitosis-rate organoids exhibit a large degree of growth and strong development of protrusions. In fact, the development of protrusions is so



(a) (b)

Figure 5.2: (A) The first two principal components of the Nystroem-transformed DETECT for the 3D analysis. We see that these kernelised DETECT signatures cluster by mitosis rate, which is the dominant signal in the data. (B) Example of 2D projections of an organoid with given mitosis rate at the final time-point.

pronounced that several protrusions disconnect from the main organoid at later time points (see Figure 5.1).

The principal components in Figure 5.2 therefore appear to pick up the major signal in the synthetic 3D data set. We hypothesise that the first principal component is proportional to the size of the organoid. Similarly, we conjecture that the second principal component corresponds to the geometric complexity of the organoids. In particular, low-mitosis-rate organoids have the lowest geometric complexity while medium-mitosis-rate exhibit significant symmetry breaking. The high-mitosis-rate organoids lie in between the two former groups of organoids in terms of geometric complexity, as their protrusion development is so pronounced that protrusions disconnect. The resulting connected tissues are relatively spherical.

5.4 Discussion

From the results in Figure 5.2, we see that DETECT captures biologically relevant information from organoid morphology on this data set: The organoids clearly cluster by mitosis rate, which is the parameter in our data explaining most of the variance in shape (see Figure 5.1). These findings support the use of DETECT on experimental data as described in Chapter 4. It remains future work to replicate the findings

described in this chapter on a larger synthetic data set, but unfortunately, we did not manage to obtain such data in time. Similarly, it would be desirable to apply DETECT to data generated by models describing the growth of different types of organoids. Further, the analysis pipeline of this data could be applied to other types of morphological data that do not have regularised and smooth boundaries. For example, there are no random perturbations in the synthetic data studied here. As we know the true organoid morphology in the case of synthetic data, one could thereby study and quantify the effects of random noise on DETECT.

The analysis of the data presented in this chapter also demonstrates that DETECT generalises to 3D data. First studies, such as [78], study 3D experimental segmentations of organoids. As organoids are inherently three-dimensional, such 3D experimental data is likely to become increasingly prevalent.

Chapter 6

ECT Stability and Inference

Chapter Content

6.1 Introduction 92
6.1.1 Problem Statement and Contributions 92
6.1.2 Outline 95
6.2 Background 96
6.2.1 Related Work 96
6.2.2 Topological Preliminaries 96
6.2.3 Gaussian Processes 98
6.3 ECT Stability of Non-Random Data 101
6.3.1 Stability for Smooth Curves 101
6.3.2 Stability of Piece-wise Linear Interpolation 103
6.4 ECT Stability of Random Data 104
6.5 Example 107
6.6 Discussion 108

Declaration of Authorship

The research presented in this chapter is a collaboration with David Beers. David conceptualised and proved all results in Section 6.3. I conceptualised and proved all results in Section 6.4, created all figures and created and implemented the example in

Section 6.5. All remaining sections are joint work. I present both mine and David's work in this section, as together our results yield a consistent estimator of the ECT (c.f. Theorem 6.12), which is a significant novel result.

6.1 Introduction

Classifying shapes is a ubiquitous task in data science and machine learning. A wealth of theory has been developed to distinguish different shapes and a large array of applications of these methods in the natural sciences exist [21, 48, 57, 155]. In particular the Euler characteristic transform (ECT) [146], arising from topological data analysis (TDA) and introduced in Section 2.5.5, provides a sufficient statistic for a large class of shapes [41, 59] (e.g., compact semi-algebraic sets), lies in a Hilbert space and is fast to compute. By contrast to other TDA methods, such as persistent homology [38] and the persistent homology transform [146], we are not aware of any general stability results for the ECT which are independent of the triangulation of a shape.

6.1.1 Problem Statement and Contributions

We propose a new metric on the embeddings of a finite one-dimensional CW complex that is sensitive to changes in arc length. Next, we introduce a norm on Euler characteristic transforms, in a similar vein to the norm introduced in Meng et al. [104, Equation 3.1], defined by taking first the 1-norm over the $\mathbb{R}$ component, and then the ∞-norm over the S^{d-1} component of the ECT. We then prove a novel stability result for the ECT, showing that the ECT is continuous in our metric of embedded spaces and the proposed norm (Theorem 6.3). In other words, if two embeddings of the same one-dimensional CW complex are sufficiently close in our metric, their corresponding ECTs are also close. To the best of our knowledge, our result is the first stability result for the ECT which is independent of the triangulation of a shape. Using similar ideas, we also show that the ECT of a smooth underlying shape can be approximated using sufficiently fine triangulations (Theorem 6.8). Further, we propose a smoothing

method for embeddings of one-dimensional CW complexes that were perturbed by independent Gaussian noise in ambient space. We use the two previous results to prove that our smoothing method does not only yield stability but also provides a consistent statistical estimator for the ECT of a noisy data set (Theorem 6.12), i.e. the ECT of the smoothed shape converges to the ECT of the underlying shape in probability as we increase the number of noisy observations.

The well-known stability results in applied topology for Čech and Vietoris-Rips filtrations of point clouds are stated in terms of the Hausdorff distance [34]. Proving stability results for the ECT is complicated by the fact that this metric is too coarse for the ECT to be continuous. Crucially, it is straightforward to construct examples of two shapes embedded in Euclidean space which are close in Hausdorff distance but whose ECTs are far apart. We loosely classify such instabilities into two categories.

The first type of instability arises when two shapes are close in Hausdorff distance, yet not homeomorphic to each other. Counterexamples can be constructed by adding a single point to a shape at an arbitrarily close distance, as visualised in Figure 6.1 for the case of an embedded simplicial complex. We point out that classical persistent homology and the persistent homology transform (PHT) [146] suffer from the same instability. However, extended persistence [37] and the extended persistent homology transform [147] can be used to partially overcome this type of instability. In this chapter, we resolve the described type of instability by restricting ourselves to shapes that are homeomorphic. Restricting an ECT analysis to a homeomorphism class of shapes is common in applications [3, 98, 106, 139].

Secondly, the ECT can suffer from instability through excessive curvature. For example, in the case of shapes homeomorphic to S^1 or $I = [0, 1]$, which can be parameterised as curves, this type of instability occurs when two curves are close in the embedded space, but one curve changes curvature much more rapidly. An example of such curves is given in Figure 6.2. Such instability is expected to occur if a shape is approximated based on points which are perturbed *independently* of each other by ambient noise.

Our work resolves instabilities of the second type for one-dimensional shapes by

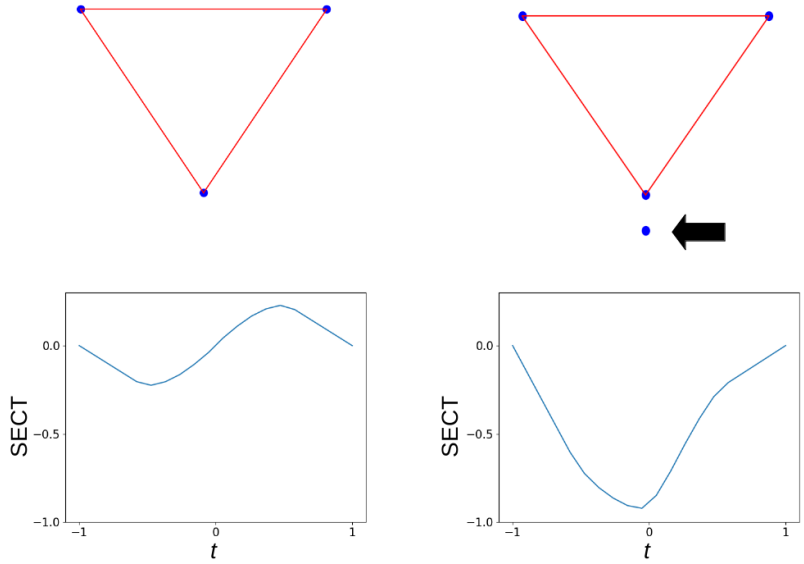


Figure 6.1: We visualise two embedded simplicial complexes (top) which differ by a single vertex. Their SECTs (bottom), visualised for a filtration in the bottom-to-top direction, are significantly different. The illustrated behaviour persists when we move the disconnected vertex in the top right panel (indicated by the arrow) arbitrarily close to the larger connected component.

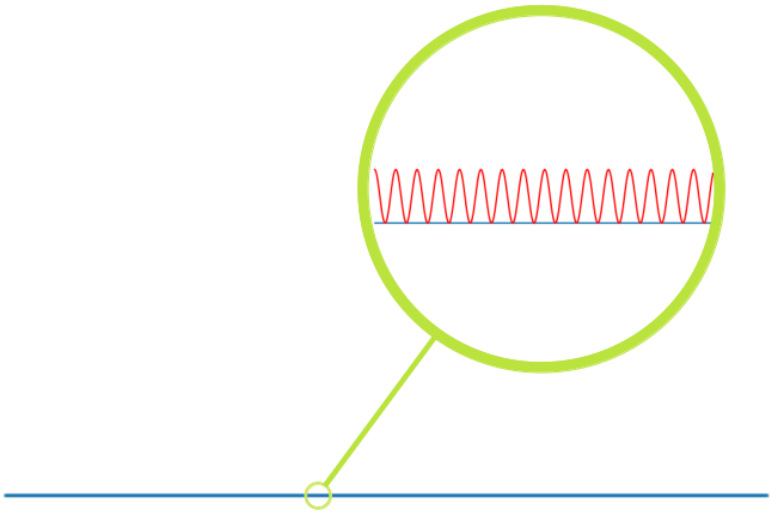


Figure 6.2: Two shapes homeomorphic to $[0,1]$ embedded into $\mathbb{R}^2$: A straight line (blue) and a wave (red) closely following the straight line with a small amplitude ε and high frequency. As long as the frequency is high enough for the wave to go through $n := \lceil 1/\varepsilon \rceil$ amplitudes, the distance of the ECTs of the two curves is at least 1 (fix the S^1-component of the ECTs to be the bottom-up direction and compute the 1-norm over $\mathbb{R}$), while the Hausdorff distance between the curves is ε.

proposing a new metric which is sensitive to curvature. We also provide a statistical estimator of the ECT which is consistent under perturbations by independently distributed Gaussian noise. These perturbations are likely to produce changes in curvature. While the PHT and extended PHT do not suffer from instabilities as illustrated in Figure 6.2, neither method provides a consistent estimator. Furthermore, the ECT arguably provides signatures more amenable to the application of further statistical and machine learning methods and are, by themselves as well as in conjunction with our new method, faster to compute.

6.1.2 Outline

This chapter is structured as follows: We start by introducing background on previous work, one-dimensional CW complexes and generalisations of the ECT and SECT in Section 6.2. Further, we introduce Gaussian processes. In Section 6.3 we propose a novel metric on the space of embeddings of a finite one-dimensional CW complex and prove that the ECT is stable against this metric for C^2-embeddings in Theorem 6.3. Then we propose a method for approximating the ECT of such an embedding by interpolating points in a finite subset in Theorem 6.8. In Theorem 6.11 of Section 6.4 we prove the probabilistic convergence of Gaussian processes on finite one-dimensional CW complexes, given a suitable kernel. Next, we construct a statistical estimator of the ECT for a shape perturbed by independent Gaussian noise. In Theorem 6.12 we combine our deterministic stability results with our probabilistic convergence result to prove that the estimator is consistent. Finally, we illustrate the power of our estimator and results on an example in Section 6.5 before concluding the chapter with a discussion in Section 6.6. The proofs of the results of this chapter can be found in Appendix A.3.

6.2 Background

6.2.1 Related Work

Already the work of Berkouk [12] shows that there is no metric on Euler curves that is stable against the interleaving distance of underlying persistence modules and satisfies a few mild, desirable conditions. We note that the stability of the Wasserstein distance proved by Skraba and Turner [135] provides a straightforward stability result for the ECT. Further, Dłotko and Gurnari [47] prove a similar result for the Euler characteristic curve. Nadimpalli et al. [106] prove a stability result for the ECT on binary image data, which is linear in the number of voxels at which two images differ. However, these stability results depend on the number of simplices in the underlying simplicial complex and the bound on the ECT becomes increasingly loose as the number of data points increases. Meng et al. [104] provide results that imply stability of the ECT when a shape is perturbed by rotations and translations but not for more general perturbations. Tameness assumptions, which are not needed in our results, are required for the stability they prove to hold. They also provide a statistical inference pipeline for shapes using the SECT. However, their pipeline considers parameterised families of shapes and random perturbations only happen in parameter space. As a result, the perturbations of points in shape space are correlated. By contrast, our results on the estimation of the ECT and SECT allow independent perturbations in ambient space.

6.2.2 Topological Preliminaries

One-Dimensional CW Complexes

A topological space Z is called a one-dimensional CW complex if it is of the form

$$Z = \left(Z_0 \sqcup \bigsqcup_{\lambda \in \Lambda} [0,1] \right) / \phi, \tag{6.1}$$

where Z_0 is a set with the discrete topology and ϕ is some map from the endpoints of the intervals in $\bigsqcup_{\lambda \in \Lambda}[0,1]$ to Z_0. We refer to the map sending the λ^{th} copy of $[0,1]$ into Z by Φ_λ (note that Φ_λ must be injective everywhere except possibly the

endpoints of the interval). The space Z is said to be a *finite* one-dimensional CW complex if it can be written as in Equation (6.1) with Z_0 and Λ finite. We refer to points in Z that are in the image of $Z_0 \to Z$ as 0-cells and subsets of Z that are the image of a map Φ_λ as 1-cells. For convenience, we denote the set of 1-cells of Z by Z_1. It may be possible for a space Z to be written as in Equation (6.1) in many different ways. For instance, the circle admits the structure of a one-dimensional CW complex with n 0-cells and n 1-cells for any natural number n. Sometimes, we need to fix a cellular decomposition of a shape. When fixing a choice of Z_0 and $\{\Phi_\lambda\}_{\lambda\in\Lambda}$, we refer to $Z^* = (Z, Z_0, \{\Phi_\lambda\}_{\lambda\in\Lambda})$ as a CW structure on Z.

We are primarily interested in shapes with this structure that are subsets of $\mathbb{R}^d$. To this end, we say that $f : Z \to X \subseteq \mathbb{R}^d$ is a C^r map under Z^* if $\Phi_\lambda \circ f$ is C^r (i.e., r-times continuously differentiable) for each $\lambda \in \Lambda$. We denote the set of such maps f by $\mathcal{F}^r(Z^*, d)$. We denote the subset of $\mathcal{F}^r(Z^*, d)$ of maps that are also homeomorphisms by $\mathcal{E}^r(Z^*, d)$ and the set of images of these homeomorphisms by $\mathcal{G}^r(Z^*, d)$.

For $r \geq 2$, we say that $f \in \mathcal{F}^r(Z^*, d)$ has curvature bounded by M if the curvature of each map $\Phi_\lambda \circ f$ is bounded by M for every $\lambda \in \Lambda$. By compactness of the unit interval, it follows that every $f \in \mathcal{G}^r(Z^*, d)$ has curvature bounded by some constant M whenever Z is a finite one-dimensional CW complex and $r \geq 2$. We say $X \in \mathcal{G}^r(Z^*, d)$ has curvature bounded by M under Z^* if the curvature of any map $h \in \mathcal{E}^r(Z^*, d)$ with image X has curvature bounded by M. It is straightforward to show that if X has curvature bounded by M under Z^*, then every map $h \in \mathcal{E}^r(Z^*, d)$ with image X has curvature bounded by M.

The Euler Characteristic Transform: Recap and Extension

Similarly to the result stated in Lemma 2.10, the Euler characteristic of a space homeomorphic to a finite one-dimensional CW complex containing c_k cells of dimension k for $k = 0, 1$, can be computed as.

$$\chi(X) = c_0 - c_1.$$

For a proof of this equation, see for example [70, Theorem 2.44]. In particular, if every path-component of X is contractible, then $\chi(X)$ is the number of path-components of X. This formula can then be used to efficiently compute the Euler characteristic of a finite one-dimensional CW complex.

The ECT and SECT of a finite one-dimensional CW complex embedded in $\mathbb{R}^d$ are then defined as in Section 2.5.5. Often one restricts to constructible families of subsets of $\mathbb{R}^d$ when studying theoretical properties of the ECT, however, for the results presented in this chapter these assumptions are unnecessary and further mention of constructible sets will be limited. However, we will focus on C^r-embeddings of finite one-dimensional CW complexes, which are always bounded subsets of $\mathbb{R}^d$. Thus, it is always possible to define the SECT of such an embedded shape.

For the remainder of this chapter, we endow the ECT (viewed as a function on $S^{d-1} \times [-a, a]$) with the norm

$$\|\mathrm{ECT}_X\| := \sup_{v \in S^{d-1}} \int_{-a}^{a} |\mathrm{ECT}_X(v,t)| \, \mathrm{d}t. \tag{6.2}$$

The norm is defined and considered analogously for the SECT.

It is also useful for us to define Euler characteristic transforms of functions f from topological spaces into $\mathbb{R}^d$. We define

$$\begin{aligned} \mathrm{ECT}_f : S^{d-1} \times \mathbb{R} &\longrightarrow \mathbb{Z} \\ (v,t) &\longmapsto \chi\big(f^{-1}\{x \in \mathbb{R}^d : \langle x, v \rangle \leq t\}\big). \end{aligned}$$

Note that if f is a homeomorphism it is immediate that $\mathrm{ECT}_f = \mathrm{ECT}_{\mathrm{im}\, f}$. We prescribe norms to Euler characteristic transforms of functions as before: restricting to functions bounded in norm by a, we let

$$\|\mathrm{ECT}_f\| := \sup_{v \in S^{d-1}} \int_{-a}^{a} |\mathrm{ECT}_f(v,t)| \, \mathrm{d}t.$$

6.2.3 Gaussian Processes

Gaussian processes (GPs) are a model for random functions. Recalling the definition of a *kernel* from Section 2.5, we define a GP as follows:

Definition 6.1. Let X be a non-empty set, $\mu : X \to \mathbb{R}$ a function and $k : X \times X \to \mathbb{R}$ be a kernel. The Gaussian process (GP) on X with mean function μ and kernel k is defined to be the random function $f : X \to \mathbb{R}$ such that for each finite set $X' = \{x'_1, ..., x'_n\} \subseteq X$ we get

$$\begin{bmatrix} f(x'_1) \\ \vdots \\ f(x'_m) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(x'_1) \\ \vdots \\ \mu(x'_n) \end{bmatrix}, K(X', X') \right). \tag{6.3}$$

The theory of GPs can be used to estimate a deterministic function $f : X \to \mathbb{R}$ given noisy observations of f at points $\{x_1, ..., x_n\} \subseteq X$. Most commonly this is done by a *Gaussian process regression* (GPR), a non-parametric Bayesian method, which models f as a random function. When performing a GPR with a given kernel k, one typically constructs a *prior distribution* by assuming

$$\begin{bmatrix} f(x'_1) \\ \vdots \\ f(x'_m) \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, K(X', X')) \tag{6.4}$$

for any finite subset $X' \subseteq X$ [120]. Assume we make n observations of the form $y_i = f(x^*_i) + \zeta_i$, where $\zeta_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d, for $i = 1, ..., n$ and $\sigma > 0$. Importantly, ζ_i does not depend on f in any way. Then, by our prior assumption and by the introduction of the shorthand $\mathbf{f}(X') := (f(x'_1), ..., f(x'_m))^T$ and $\boldsymbol{\zeta} := (\zeta_1, ..., \zeta_m)^T$ we get that

$$\begin{bmatrix} \mathbf{f}(X^*) + \boldsymbol{\zeta} \\ \mathbf{f}(X') \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X^*, X^*) + \sigma^2 I_n & K(X^*, X') \\ K(X', X^*) & K(X', X') \end{bmatrix} \right).$$

Thus, by conditioning the above multivariate normal distribution of $\mathbf{f}$ on the observations $\mathbf{y} := (y_1, ..., y_n)^T$, we get [120]

$$\begin{aligned} \mathbf{f}(X')|\mathbf{f}(X^*) + \boldsymbol{\zeta} = \mathbf{y} \sim \mathcal{N}(&K(X', X^*)(K(X^*, X^*) + \sigma^2 I)^{-1}\mathbf{y}, \\ &K(X', X') - K(X', X^*)(K(X^*, X^*) + \sigma^2 I)^{-1} K(X^*, X')). \end{aligned} \tag{6.5}$$

The above distribution, for any finite $X' \subseteq X$, is the *posterior distribution* of f at X' given observations $\mathbf{y}$. In the context of Bayesian modelling, we first summarise our knowledge in the values of f by the prior distribution: unless we gain further

information, we assume f to be mean 0 with covariance K (Equation (6.4)). For any (noisy) observation of f we make, we update our belief in the values of f by conditioning our prior distribution on our observations. The posterior density at inputs X' can then be interpreted as how strongly we believe an output value to be the true output of f at X', given our observations and modelling assumptions.

If $m = 1$ (i.e., $X' = \{t\}$ for some $t \in X$), we denote the mean of the above conditional normal distribution by $\hat{f}_n(t)$ and its variance by $v_n(t)$. We henceforth call $\hat{f}_n(t)$ the *Gaussian smoothing of* f (on the set X^* of size n). When needed, we explicitly denote the dependence of $\hat{f}_n(t)$ on X^* and f by writing $\hat{f}_n(t, X^*, f)$. From Equation (6.5), it follows that $\hat{f}_n$ always lies in $\mathcal{H}_k$, the RKHS of k.

Under certain assumptions, one can show that $\hat{f}_n \to f$ in mean. In this chapter, we use results by Koepernik and Pfaff [85], which give strong probabilistic convergence results in the case of X being a compact metric space. Note that finite one-dimensional CW-complexes are always metrisable [55] and compact. Hence the results of Koepernik and Pfaff apply to all of the cases we look at.

We note that computing $\hat{f}_n$ requires the inversion of an $n \times n$-matrix and thus has a runtime of $\mathcal{O}(n^3)$. By using the ECT on $\hat{f}_n$ we thus lose some of the ECTs runtime advantage (compared to the PHT and extended PHT). However, both versions of the PHT require $\mathcal{O}(n_s^3)$ computations *per direction* [50], where $n_s \geq n$ is the number of simplices in the triangulation of a shape, which still gives a combined Gaussian process and ECT pipeline an edge in terms of runtime. More importantly, it is common practice to approximate the (inverse) Gram matrix by a low-rank matrix approximation method, such as the Nystroem method [158] or random Fourier features [119]. Such methods run in $\mathcal{O}(l^3 + l^2 n)$, where $l \ll n$ is the approximate rank of the Gram matrix and thus are significantly faster than $\mathcal{O}(n^3)$.

6.3 ECT Stability of Non-Random Data

6.3.1 Stability for Smooth Curves

As we observed in the introduction, controlling the proximity of two different one-dimensional shapes is not enough to control the difference between their ECTs. This motivates the definition of a metric between such shapes which is also concerned with perturbations to arc length.

Definition 6.2. Let Z be a finite one-dimensional CW complex with a fixed CW structure $Z^* = (Z, Z_0, \{\Phi_\lambda\}_{\lambda\in\Lambda})$. Fix $r \geq 1$. For $X, Y \in \mathcal{G}^r(Z^*, d)$, we define $d_{Z^*}(X, Y)$ to be the infimum of all ε such that there exists $h_X, h_Y \in \mathcal{E}^r(Z^*, d)$, whose images are X and Y respectively, satisfying:

1. The difference of arc lengths between $h_X \circ \Phi_\lambda$ and $h_Y \circ \Phi_\lambda$ is less than or equal to ε for each $\lambda \in \Lambda$.

2. Both $h_X \circ \Phi_\lambda$ and $h_Y \circ \Phi_\lambda$ are curves of constant velocity for each $\lambda \in \Lambda$.

3. $\|h_X - h_Y\|_\infty \leq \varepsilon$.

By using the compactness of Z it is not difficult to show that d_{Z^*} is a metric on $\mathcal{G}^r(Z^*, d)$. For the remainder of this chapter, we endow $\mathcal{G}^r(Z^*, d)$ with this metric and the topology arising from it. A key goal of this chapter is to show that the ECT is a continuous map on $\mathcal{G}^r(Z^*, d)$ for $r \geq 2$.

Theorem 6.3. *Let Z be a finite one-dimensional CW complex with a fixed CW structure $Z^* = (Z, Z_0, \{\Phi_\lambda\}_{\lambda\in\Lambda})$. The map $X \mapsto \mathrm{ECT}_X$ is continuous on $\mathcal{G}^r(Z^*, d)$ for $r \geq 2$.*

In particular, if X has curvature bounded by M, and the image of the λ^{th} 1-cell in X has arc length L_λ, then whenever $d_{Z^*}(X, Y) < \varepsilon$, we have

$$\|\mathrm{ECT}_X - \mathrm{ECT}_Y\| \leq |Z_0|\varepsilon + \sum_{\lambda\in\Lambda} G_\lambda(\varepsilon),$$

where

$$G_\lambda(\varepsilon) := \begin{cases} 8\sqrt{L_\lambda n_\lambda \varepsilon} + n_\lambda\varepsilon & L_\lambda/n_\lambda > 2\varepsilon \\ 11 n_\lambda \varepsilon & L_\lambda/n_\lambda \leq 2\varepsilon \end{cases}$$

and

$$n_\lambda := \max\left(\left\lceil\left(\frac{M^2L_\lambda^3}{24\varepsilon}\right)^{1/3}\right\rceil, \left\lceil\frac{L_\lambda M}{\pi}\right\rceil\right).$$

We prove this theorem via the following proposition, which is useful when considering functional information in Section 6.4.

Proposition 6.4. *Let Z be a finite one-dimensional CW complex with a fixed CW structure $Z^* = (Z, Z_0, \{\Phi_\lambda\}_{\lambda\in\Lambda})$. Let $f, g \in \mathcal{F}^r(Z^*, d)$, with $r \geq 2$, and suppose that:*

1. *The curves $f \circ \Phi_\lambda$ and $g \circ \Phi_\lambda$ have arc lengths that differ by at most ε for each $\lambda \in \Lambda$.*
2. *The curves $f \circ \Phi_\lambda$ and $g \circ \Phi_\lambda$ have constant velocity for each $\lambda \in \Lambda$.*
3. *$\|f - g\|_\infty \leq \varepsilon$.*

Then if f has curvature bounded by M and $f \circ \Phi_\lambda$ has arc length L_λ, we have

$$\left\|\mathrm{ECT}_f - \mathrm{ECT}_g\right\| \leq |Z_0|\varepsilon + \sum_{\lambda\in\Lambda} G_\lambda(\varepsilon),$$

where G_λ and n_λ are defined as above.

The idea of the proof of this proposition is as follows. We show that the norm of the ECT of a curve can be controlled using its differential properties. Using this observation, we can bound the difference in ECT of two curves that are nearly straight lines. We can also refine the structure of a finite one-dimensional CW complex Z^* with a map $f : Z \to \mathbb{R}^d$ into enough pieces that the image of every 1-cell is a nearly linear curve. The above proposition then follows from a glueing argument.

In this chapter, we let I denote the unit interval $[0, 1]$ and say that $f : I \to \mathbb{R}^d$ is piece-wise C^1 if it is continuous and there exists a collection $T_1, \ldots, T_k$ of closed intervals covering I, on the interiors of which f is C^1.

Proposition 6.5. *Let $\gamma : I \to \mathbb{R}^d$ piece-wise C^1 map, with X being the image of γ. Fix any $v \in S^{d-1}$. Let a and b be the minimal and maximal values of $f : x \mapsto \langle v, \gamma(x)\rangle$ on I respectively. Then*

$$\int_a^b |\mathrm{ECT}_\gamma(v, t)|\, \mathrm{d}t \leq V(f),$$

where $V(f)$ denotes the variation of f:

$$V(f) := \int_0^1 |f'| \, \mathrm{d}t.$$

For $t \geq b$ we have $\mathrm{ECT}_\gamma(v,t) = 1$. For $t < a$ we have $\mathrm{ECT}_\gamma(v,t) = 0$. In particular, $\mathrm{ECT}_\gamma(v,t)$ is defined almost everywhere.

6.3.2 Stability of Piece-wise Linear Interpolation

If we are given $X \subseteq \mathbb{R}^d$, the C^2-image of a homeomorphism h from some one-dimensional CW complex Z, it may not be easy to exactly compute ECT_X. The main goal of this section is to show that a dense subset of Z can be used to approximate $\mathrm{ECT}_h = \mathrm{ECT}_X$. First, we make precise the kind of dense subset we need to properly estimate the ECT_h.

Definition 6.6. Let $Z^* = (Z, Z_0, \{\Phi_\lambda\}_{\lambda \in \Lambda})$ be a connected finite one-dimensional CW complex with some fixed cellular decomposition and f be a C^2 map $f : Z \to \mathbb{R}^d$. We say that $\mathbf{a}_n = \{a_1, \ldots, a_n\} \subseteq Z$ is a *compatible subset* of Z^* if the following hold:

1. $Z_0 \subseteq \mathbf{a}_n$ and

2. $\mathbf{a}_n - Z_0$ contains a point in each 1-cell of Z.

These requirements ensure that $Z - \mathbf{a}_n$ is a union of disjoint open intervals. If additionally, the length of the image of each of these intervals under f is less than ε, we say that $\mathbf{a}_n$ is an *ε-dense subset* for f. An infinite subset of Z is *compatible and dense* for f if it contains a finite ε-dense subset for all positive ε.

Definition 6.7. Let f be as in the previous definition and $\mathbf{a}_n = \{a_1, \ldots, a_n\}$ be a compatible subset of Z^*. Let a_{ij} denote the line segment from a_i to a_j. We define a multiset E with elements in the set of unordered pairs in $\{1, \ldots, n\}$. E contains a copy of (i,j) for each open curve in $Z - \mathbf{a}_n$ whose endpoints are a_i and a_j. We define

$$\begin{aligned}\mathrm{ECT}_f^{\mathbf{a}_n}(v,t) =& \#\{1 \leq i \leq n : \langle f(a_i), v\rangle \leq t\} \\ & - \#\{(i,j) \in E : \max(\langle f(a_i), v\rangle, \langle f(a_j), v\rangle) \leq t\}.\end{aligned}$$

The main theorem of the section says that we can use dense subsets to approximate the Euler characteristic transform of a one-dimensional CW complex:

Theorem 6.8. *Let $Z^* = (Z, Z_0, \{\Phi_\lambda\}_{\lambda\in\Lambda})$ be a connected finite one-dimensional CW complex with some fixed cellular decomposition and f be a C^2 map $f : Z \to X \subseteq \mathbb{R}^d$. Suppose that f has curvature bounded by M and let $\mathbf{a}_n$ be an ε-dense subset of Z, where $0 < \varepsilon < \pi/M$. Let L be the sum of the arc lengths of the images of 1-cells of Z under f. Then*

$$\left\|\mathrm{ECT}_f - \mathrm{ECT}_f^{\mathbf{a}_n}\right\| \leq \frac{1}{\sqrt{12}} ML\varepsilon.$$

6.4 ECT Stability of Random Data

In this section, we consider observations taken from an embedded finite one-dimensional CW complex Z which are perturbed by ambient Gaussian noise. We show that the Gaussian smoothing of these observations converges to satisfy the assumptions of Proposition 6.4. In particular, we show that the ECT and SECT of the Gaussian smoothing give consistent estimators of the ECT and SECT of Z, respectively. To provide the theorems, we first need to introduce technical conditions on the kernel we use in the Gaussian smoothing:

Definition 6.9 (Definition 5 in [85])**.** Let Z be a topological space and $k : Z \times Z \to \mathbb{R}$ be a continuous kernel. Define

$$d_k(t, s) = \sqrt{k(t, t) + k(s, s) - 2k(t, s)}.$$

For any $\varepsilon > 0$ let $N(Z, \varepsilon, d_k)$ be the minimal numbers of d_k-balls with radius ε needed to cover Z. Then define

$$J(Z, d_k) = \int_0^\infty \sqrt{\log N(Z, \varepsilon, d_k)}\, \mathrm{d}\varepsilon.$$

Definition 6.10. Let $Z^* = (Z, Z_0, \{\Phi_\lambda\}_{\lambda\in\Lambda})$ be a connected finite one-dimensional CW complex with some fixed cellular decomposition. Let $k : Z \times Z \to \mathbb{R}$ be a continuous kernel. We say k is r-times differentiable on Z^* if

1. for each $\lambda \in \Lambda$ the map $k^\lambda : I \times I \to \mathbb{R}$ given by $(s,t) \mapsto k(\Phi_\lambda(s), \Phi_\lambda(t))$ is r-times continuously differentiable and

2. for each $\lambda \in \Lambda$ and $z \in Z$ the map $k^{\lambda,z} : I \to \mathbb{R}$ given by $s \mapsto k(\Phi_\lambda(s), z)$ is r-times continuously differentiable.

Differentiability is defined by one-sided limits at the boundaries of $I \times I$ and I.

Remark 1. For a given connected finite 1 one-dimensional CW complex $Z^* = (Z, Z_0, \{\Phi_\lambda\}_{\lambda\in\Lambda})$ with fixed cellular composition there is a straightforward way to construct an r-times differentiable kernel on Z^*: let $f : Z \to \mathbb{R}^d$ be a continuous injective function such that $f \circ \Phi_\lambda$ is r-times differentiable for each $\lambda \in \Lambda$. Then if k is an r-times differentiable kernel on $\mathbb{R}^d$, it follows that $k'(s,t) := k(f(s), f(t))$ is an r-times differentiable kernel on Z by the chain rule.

While it might be tempting to define a geodesic distance on Z and then apply a stationary kernel (such as the Gaussian kernel) to this distance, it should be noted that, even in the case of Z being a manifold, the resulting function does not give a positive-definite kernel in general [53].

We can now state the first theorem of this section:

Theorem 6.11. *Let $Z^* = (Z, Z_0, \{\Phi_\lambda\}_{\lambda\in\Lambda})$ be a connected finite one-dimensional CW complex with some fixed cellular structure. Let $k : Z \times Z \to \mathbb{R}$ be a continuous, four-times differentiable kernel on Z^*. Assume k satisfies $J(Z, d_k) < \infty$.*

Let $f : Z \to \mathbb{R}$ be a function in the RKHS of k. Let $\mathbf{a} \subset Z$ be a sequence which is dense. Denote by $\mathbf{a}_n$ the first n terms of $\mathbf{a}$ and by a_n the n-th term of $\mathbf{a}$. Let $\hat{f}_n$ denote the Gaussian smoothing of f based on observations $y_i = f(a_i) + \zeta_i$ using kernel k, where $i = 1, ..., n$ and $\zeta_i \sim \mathcal{N}(0, \sigma)$ i.i.d. for some $\sigma > 0$. Then

$$\mathbb{E}\left[\left\|\hat{f}_n(t, \mathbf{a}_n, f) - f(t)\right\|_\infty\right] \to 0$$

as $n \to \infty$. Moreover, for each $\lambda \in \Lambda$ define $\hat{f}_{n,\lambda}(t) = \hat{f}_n(\Phi_\lambda(t))$ and $f_\lambda(t) = f(\Phi_\lambda(t))$. Then

$$\mathbb{E}\left[\left|V\left(\hat{f}_{n,\lambda}\right) - V\left(f_\lambda\right)\right|^2\right] \to 0$$

for each $\lambda \in \Lambda$ as $n \to \infty$; i.e. the variation of $\hat{f}_{n,\lambda}$ converges to the variation of f_λ in mean square.

Using the above theorem, we can prove that the ECT of the Gaussian smoothing $\hat{f}_n$ of f, denoted $\mathrm{ECT}_{\hat{f}_n}$, is a consistent estimator of the ECT of X:

Theorem 6.12. *Let $Z^* = (Z, Z_0, \{\Phi_\lambda\}_{\lambda\in\Lambda})$ be a finite one-dimensional CW complex with some fixed cellular structure and $f : Z \to X \subseteq \mathbb{R}^d$ be a C^2 homeomorphism with bounded curvature. Further, assume that all components of f are functions in the RKHS of k, where k is a kernel satisfying the assumptions of Theorem 6.11. Moreover, assume that $\|f'_\lambda(t)\|_2 = L_\lambda$ is constant on all 1-cells $\lambda \in \Lambda$. Let $\mathbf{a}$ be a dense sequence in Z. Let*

$$f(t) := \left(f^1(t), ..., f^d(t)\right)^T, \qquad \hat{f}_n := \left(\hat{f}_n^1, ..., \hat{f}_n^d\right)^T,$$

where for $j = 1, ..., d$ and $i = 1, ..., n$ the function $\hat{f}_n^j$ is the Gaussian smoothing of f^j given observations $y_{ij} = f^j(a_i) + \zeta_{ij}$ using kernel k and $\zeta_{ij} \sim \mathcal{N}(0, \sigma_j)$ i.i.d for some $\sigma_j > 0$. Then for each $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(\left\|\mathrm{ECT}_{\hat{f}_n} - \mathrm{ECT}_f\right\| < \varepsilon\right) \to 1.$$

Note that as f is a homeomorphism, $\mathrm{ECT}_f = \mathrm{ECT}_{\mathrm{im}\, f} = \mathrm{ECT}_X$ and we thus have constructed a consistent estimator for ECT_X. If for given observations $\mathbf{y}_n$ the curvature of $\hat{f}_n$ is bounded on each 1-cell, $\mathbf{a}_n$ is compatible with Z^* and dense for f, we can approximate $\mathrm{ECT}_{\hat{f}_n}$ by $\mathrm{ECT}_{\hat{f}_n}^{\mathbf{a}_m}$ arbitrarily closely for a sufficiently large m by Theorem 6.8. We conjecture that for sufficiently well-behaved kernels k, $\mathrm{ECT}_{\hat{f}_n}^{\mathbf{a}_m}$ converges to $\mathrm{ECT}_{\hat{f}_n}$ in probability, where m is some function in n. Proving this conjecture will involve bounding the curvature with high probability and is beyond the scope of this thesis.

Furthermore, our consistency result extends to the SECT of X:

Lemma 6.13. *Define the* ECT *on some interval $[-a, a]$. Assume the distance between the* ECT*s of two shapes X and Y is δ. Then the distance between their* SECT*s is at most $(2a + 1)\delta$.*

The main limitation of our results is that the topology of our embedded space is assumed to be known and the results only work for a restricted class of CW complexes. Extending our statistical estimator and the related results to perturbations in the topology of the underlying shape remains future work.

6.5 Example

We now illustrate our methods by means of a simulated example. In our simulation, we focus on a single simple closed curve in $\mathbb{R}^2$ and sample different numbers of noisy points from the curve. Our curve has been constructed by judiciously choosing complex Fourier coefficients. The samples are then taken by evenly spaced evaluations of our curve and are corrupted by adding independent multivariate Gaussian noise with mean 0 and covariance $(0.002)^2 I_2$. The curve, together with the noisy samples, is visualised in Figure 6.3.

As a kernel in our Gaussian smoothing, we pick the *sine-squared exponential kernel*. Assuming our curve is parameterised by $\gamma : [0, 2\pi] \to \mathbb{R}^2$ with $\gamma(0) = \gamma(2\pi)$, it is given by

$$k(s,t) = \exp\left(-2\sin\left(\frac{s-t}{2}\right)^2\right).$$

It satisfies the conditions of Theorems 6.11 and 6.12 (see Lemma A.22; it is infinitely differentiable as it is the composition of infinitely differentiable functions). Its RKHS contains the curve we generated (see Lemma A.23).

In Figure 6.4, we visualise the SECT of our true curve (in a fixed direction) and compare it to the SECT of curves sampled from Gaussian process regression (GPR) posterior distributions based on 20, 50 and 100 noisy evaluations of our original curves, respectively. In addition, we plot the distributions of the distance (given by the norm introduced in Equation (6.2)) of the SECTs of the posterior samples with the SECT of the true curve.

In both types of plots, we see the posterior curves' mean moving closer to the true SECT as the number of samples increases, which illustrates the results of our theorems. We furthermore report that the distance between the SECT of the true

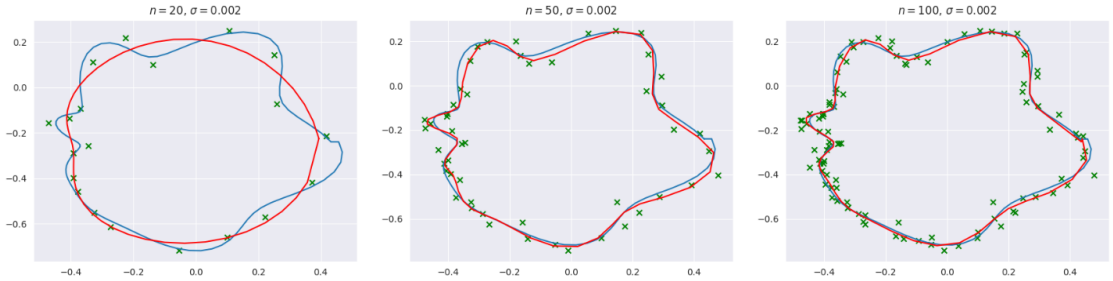


Figure 6.3: The Gaussian smoothings (red lines) of a simple closed curve (blue line) based on noisy samples (green crosses). The number of samples is 20 in the left panel, 50 in the middle panel and 100 in the right panel. All samples have been independently corrupted with mean zero Gaussian noise with standard deviation $\sigma = 0.002$ in each component.

curve and the SECTs of Gaussian smoothings are approximately 0.0627 ($n = 20$), 0.0366 ($n = 50$) and 0.0214 ($n = 100$), respectively. However, while Figure 6.4 illustrates that our results provide a consistent estimator of the SECT, the estimator need not be unbiased.

The example in this section illustrates how our estimator naturally gives rise to a posterior distribution over the space of SECT curves. We believe that there is potential to use this posterior distribution in a statistical inference or classification pipeline. Proving convergence rates for estimators like ours would help with quantifying the confidence of statistical ECT analyses.

6.6 Discussion

In this chapter, we provide stability results for the Euler characteristic transform under independent, random perturbations of underlying point cloud data. We proceed by first proving in a deterministic setting that if two embedded CW complexes are close in some well-defined sense, then their ECTs are also close. Crucially, we provide bounds on how far the two output ECTs can be apart. We also show that the ECT of such a shape can be well-approximated by simplicial complexes. Second, we propose a smoothing technique for randomly perturbed CW complexes which builds on insights from Gaussian process theory. We then prove that the smoothed shape is close to

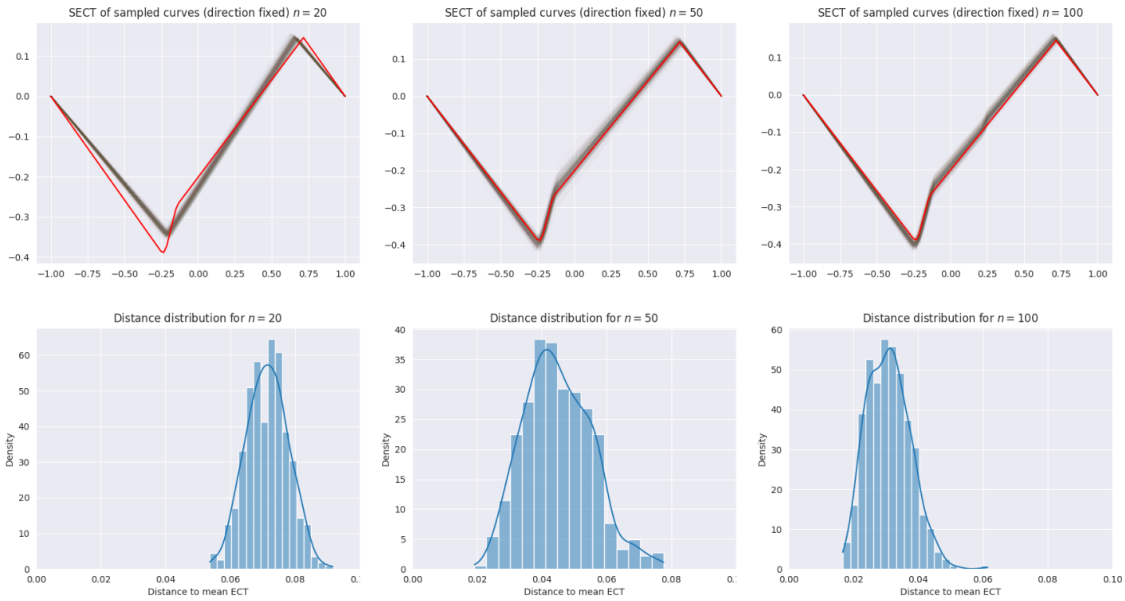


Figure 6.4: Top: The SECT in a fixed direction of the true shape (in red) compared to the SECTs of GPR posterior samples (in opaque blue) based on 20 (left), 50 (middle) and 100 (right) samples. The fixed direction corresponds to left-to-right in Figure 6.3. The SECTs are based on interpolations on the samples. Bottom: The distribution of the distance between the SECT of the true curve to SECTs of GPR posterior curves based on 20 (left), 50 (middle) and 100 (right) noisy samples from the underlying curve.

the unperturbed in the sense of our deterministic theorem with high probability, thus providing a consistent statistical estimator for the ECT. Finally, we illustrate the utility of our method and results on a synthetic data set.

The main limitation of our results is that the topology of our embedded space is assumed to be known and the results only work for a restricted class of CW complexes. Extending our statistical estimator and the related results to perturbations in the topology of the underlying shape remains future work. Further, our example in Section 6.5 illustrates how our estimator naturally gives rise to a posterior distribution over the space of SECT curves. We believe that there is potential to use this posterior distribution in a statistical inference and classification pipeline. Proving convergence rates for estimators like ours would help with quantifying the confidence of statistical ECT analyses.

Chapter 7

Discussion and Outlook

Organoids are a useful *in vitro* model for studying tissue development, drug testing and, more broadly, understanding the response of tissues to genetic and environmental changes. Organoids share favourable characteristics with *in vivo* experimental models while providing a more cost-effective and ethical alternative. Organoids have successfully aided the study of cancer development in tissues [99, 49, 56], tumour microenvironments [109, 15, 46], and precision medicine [96, 72, 141, 154, 113], which may lead to advances in drug discovery, immunotherapy and individualised cancer treatments [148].

Dynamic changes in the shape, cellular composition and gene expression of organoids can be used to understand the effect of mutations and treatments on healthy and diseased tissue. The tissue composition, closely linked to the morphology, affects the microenvironments of cells, including cell-to-cell signalling and, thus, gene expression. Conversely, the gene expression of the cells drives stem cell differentiation into distinct cell types, which rearrange in a spatially non-uniform manner. These processes lead to morphogenesis and direct dynamic changes in the organoid shape. Thus, both RNA sequencing data and organoid morphology imaging data reflect the underlying genetic composition of an organoid and its response to environmental perturbation and, therefore, yield insights into disease progression and treatment effects. In this thesis, I developed and presented new methods from TDA to analyse both scRNA-seq and morphology data of organoids.

In Chapter 3, I developed the multiscale Laplacian score, a multiscale feature

selection method on scRNA-seq data, and the UMAP diffusion cover, a heuristic that improves Mapper for use in trajectory inference. I discussed the need for new analysis methods for scRNA-seq data. First, differential expression tests use a discrete notion of cell type (i.e., they assign a unique label to each cell). Typically, cell types are assigned by applying a clustering algorithm to the data. However, the high resolution of state-of-art single-cell sequencing methods puts discrete notions of cell type into question, as it is possible to observe continuous trajectories connecting different cell types [60]. In such a setting, many cells could be described as an intermediate cell type. In particular, assigning discrete labels corresponds to determining an arbitrary cut-off along such a trajectory, leading to unstable downstream analysis. Govek et al. [60] proposed the Laplacian score as a generalised DE test which considers the topology underlying scRNA-seq data and, by extension, the notion of cell type to be continuous. The LS is inherently single-scale, while classical DE tests can easily perform multiscale analyses. To bridge this gap, I introduced a novel *multiscale Laplacian score*, which is motivated by random walk theory on simplicial complexes.

Unlike classical DE tests, trajectory inference methods explicitly assume a continuous notion of cell type and attempt to infer continuous trajectories along which cells change their function. Many trajectory methods may be limited by their topological modelling assumptions or biased by the large number of user-determined hyperparameters [125]. The Mapper algorithm, a TDA method, ha successfully inferred trajectories on scRNA-seq data [122, 118]. However, a key step, the so-called cover selection, is performed manually in these studies, leading to potential overfitting and bias. In Chapter 3, I proposed an unsupervised heuristic for picking the cover, called the *UMAP diffusion cover*. The UMAP diffusion cover is theoretically well-motivated and based on the UMAP graph, a method widely used in scRNA-seq analyses. I applied both the UMAP diffusion cover (in conjunction with Mapper) and the MLS to two scRNA-seq data sets: a benchmark data set of lung tumour infiltrating T cells and a mouse colon organoid data set. The MLS validated previously identified genes and detected additional biologically meaningful genes with coherent expression patterns. The UMAP diffusion cover (in conjunction with Mapper) identi-

fied complex trajectories in both data sets and outperformed the state-of-art method PAGA.

When applying the MLS on scRNA-seq data, I focused on modelling the geometry of cell space with a specific k-nn graph, the UMAP graph. The MLS is flexible for use on other graphs, such as Mapper graphs [122, 118], but the resulting analysis could be sensitive to the underlying graph structure. Studying these changes could be the subject of future research. Recently, a method for automated scale detection in multiscale community detection has been proposed [131]. This method could also be applied to the MLS to further automate the analysis pipeline. The choice of resolution(s) for the MLS is not limited to Markov stability times (e.g. graph wavelets [144] could be an alternative). It would be interesting future work to extend these signal selection approaches to other signals more generally (e.g., epigenetic factors), other complex single-cell network structures [79] or other higher-order networks [130, 14], with a view towards multi-modal data integration [88].

In Chapter 3, I also compared the output of the Mapper graph using a UMAP diffusion cover to PAGA, a state-of-art method for trajectory inference. Further benchmarking on additional data sets and against additional trajectory inference methods is important future work. The review by Saelens et al. [125] provides a general framework, several metrics and data sets for such benchmarking.

In Chapter 4, I developed DETECT, a rotationally invariant signature capturing the temporal evolution of a shape. DETECT, an extension of the Euler characteristic transform [146], is theoretically well-motivated, interpretable and fast to compute. I applied DETECT to an experimental data set of mouse small intestine organoids. For this data set, organoid experiments were filmed over a period of 80 hours and the boundary of each organoid was segmented at each video frame, yielding a sequence of shapes summarising the morphological evolution of each organoid for a given experiment. On this data set, DETECT captured information contained in multiple classical shape descriptors (e.g. diameter, area, centroid distances, major axis length). We regressed these descriptors from DETECT with high accuracy. Further, we classified organoids into treated and untreated groups using DETECT, thereby demonstrating

that the given treatment has a significant effect on the dynamic changes of organoid morphology. By contrast, the predictive accuracy of classical shape descriptors did not exceed guessing on the same task.

In future work in we could apply our methods to data sets of different types of organoids (derived from different organs, with different genetic backgrounds and/or cultured under different conditions). Future research could involve selecting features in (kernelised) DETECT space and reconstructing dynamic organoid shapes along those features. We would first identify those features which vary most across different organoid categories (e.g. experimental conditions). Second, we would then reconstruct how the temporal evolution of an organoid shape changes along that feature. While Wang et al. [155] give a blueprint for such an analysis in their SINATRA pipeline, further theoretical work is needed to extend their work to account for the temporal component and rotational invariance of DETECT. Such an inversion of DETECT could yield further information about how genetics and treatments affect organoid morphology.

In Chapter 5, I validated the DETECT method of Chapter 4 on a synthetic data set. This data set was generated from a 3D mechanistic continuum model describing the growth of mouse colon organoids [161]. DETECT clustered organoids by mitosis rate, the dominating biological signal in the data set. We concluded that DETECT is capable of extracting biologically meaningful information from organoid morphology. This analysis also demonstrated that DETECT generalises to 3D data. As such, we could apply DETECT to 3D experimental data in the future.

In the future, we could also study information loss between 3D data and their 2D projections. One possible approach to this problem would be to consider synthetic 3D data generated from mathematical models (such as the one presented in Chapter 5.2) and to compare the generated 3D data to its 2D projections. However, these models neglect certain biophysical processes (e.g. the effects of gravity, the production of extracellular matrix, mechanical stress) [161] and the analysis would be complicated by the fact that DETECT signatures of 3D and 2D organoids are markedly different, even in the absence of information loss. Other future work may include replicating the

findings described in Chapter 5 on a larger synthetic data set, generated on a larger set of parameter values and with different initial conditions. Due to time constraints, we did not manage to obtain such data in time. Similarly, applying DETECT to data generated by models describing the growth of different types of organoids or to perturb boundaries of *in silico* organoids would be desirable to test the effects of noise on DETECT in a controlled setting.

Finally, in Chapter 6, I proved that ECT can be consistently estimated from noisy data by smoothing the boundary by taking weighted averages. The smoothing method was inspired by Gaussian process theory. If a given shape is perturbed by independent Gaussian noise, it is likely that the ECT of the smoothed perturbed shape is close to the ECT of the true shape. I did not apply this smoothing method to the experimental data of Chapter 4 as it had already been regularised by a segmentation algorithm. However, the results of Chapter 6 showed that a heuristic as simple as taking weighted averages of vertices is sufficient to remove noise from the ECT in the probabilistic limit. These insights further justified applying the ECT to smoothed experimental data.

A future research direction based on the work from Chapter 6 would be to generalise the results to a larger class of CW complexes and to perturbations in the topology of the underlying shape. Further, the example in Section 6.5 illustrates how the ECT estimator naturally gives rise to a posterior distribution over the space of ECT/SECT curves. This posterior distribution could be useful, more generally, in a statistical inference and classification pipeline. Proving convergence rates for estimators, like the one proposed in Chapter 6, could assist with quantifying the confidence of statistical ECT analyses.

In the broader context of this thesis, important future work involves conducting an integrated analysis of scRNA-seq and morphology data once generated from the same organoid. The methods and approaches presented in this thesis could serve as a roadmap towards such an integrated analysis once such data becomes available in the future. One approach for such an analysis could be:

1. Identify genes that are consistently expressed at a number of scales using the MLS and VI.

2. Compute the DETECT signatures of the same organoids.

3. Perform a canonical correlation analysis between the expression levels of the genes found in 1. and the (kernelised) DETECT computed in 2.

This analysis would allow us to identify correlations between gene expression and (changes in) morphology. More research is needed to detect causality between morphology and gene expression (likely they both interact with each other in a feedback loop).

The canonical directions on the DETECT signatures could be used to detect experimental times at which a gene drives morphological changes. A possible future analysis could involve reconstructing the shape changes which correspond to the canonical directions in the DETECT signature. Wang et al. [155] provide a framework for such a reconstruction with the Euler characteristic transform, which DETECT extends. However, the rotational invariance of DETECT may demand a theoretical extension of their work.

Alternatively, a Mapper graph (or another graph generated by a trajectory inference method) could be constructed on the SECT signatures of boundaries of individual organoid boundaries in individual video frames. Such a graph would summarise the developmental trajectories of organoid morphologies. In a second step, the MLS could be used to identify genes that are consistent with these developmental trajectories, possibly identifying genes that correlate with different morphological changes. Regardless of whether such an integrated analysis follows the blueprint suggested above, it would enable genes to be associated with morphological changes with higher confidence than analyses made on distinct scRNA and morphology data sets alone.

Bibliography

[1] Abdul L, Xu J, Sotra A, Chaudary A, Gao J, Rajasekar S, Anvari N, Mahyar H and Zhang B. 'D-CryptO: deep learning-based analysis of colon organoid morphology from brightfield images'. In: *Lab on a Chip* 22.21 (2022), pp. 4118–4128.

[2] Adams TS, Schupp JC, Poli S, Ayaub EA, Neumark N, Ahangari F, Chu SG, Raby BA, DeIuliis G, Januszyk M et al. 'Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis'. In: *Science advances* 6.28 (2020), eaba1983.

[3] Amezquita EJ, Quigley MY, Ophelders T, Landis JB, Koenig D, Munch E and Chitwood DH. 'Measuring hidden phenotype: Quantifying the shape of barley seeds using the Euler Characteristic Transform'. In: *in silico Plants* (Dec. 2021). DOI: 10.1093/insilicoplants/diab033.

[4] Anders S and Huber W. 'Differential expression analysis for sequencecount data'. In: *Genome Biology* 11.R106 (2010).

[5] Barahona M. *The stability of a graph partition*. URL: https://www.ma.imperial.ac.uk/~mpbara/Partition_Stability/ (visited on 23/05/2022).

[6] Baryawno N, Przybylski D, Kowalczyk MS, Kfoury Y, Severe N, Gustafsson K, Kokkaliaris KD, Mercier F, Tabaka M, Hofree M et al. 'A cellular taxonomy of the bone marrow stroma in homeostasis and leukemia'. In: *Cell* 177.7 (2019), pp. 1915–1932.

[7] Beck LE, Lee J, Cote C, Dunagin MC, Lukonin I, Salla N, Chang MK, Hughes AJ, Mornin JD, Gartner ZJ et al. 'Systematically quantifying morphological features reveals constraints on organoid phenotypes'. In: *Cell Systems* 13.7 (2022), pp. 547–560.

[8] Beguerisse-Diaz M, Garduno-Hernandez G, Vangelov B, Yaliraki SN and Barahona M. 'Interest communities and flow roles in directed networks: the Twitter network of the UK riots'. In: *Journal of The Royal Society Interface* 11.101 (2014), p. 20140940.

[9] Beguerisse-Diaz M, Vangelov B and Barahona M. 'Finding role communities in directed networks using Role-Based Similarity, Markov Stability and the Relaxed Minimum Spanning Tree'. In: *2013 IEEE Global Conference on Signal and Information Processing* (Dec. 2013). DOI: 10.1109/globalsip.2013.6737046. URL: http://dx.doi.org/10.1109/GlobalSIP.2013.6737046.

[10] Bennstein SB, Weinhold S, Manser AR, Scherenschlich N, Noll A, Raba K, Kögler G, Walter L and Uhrberg M. 'Umbilical cord blood-derived ILC1-like cells constitute a novel precursor for mature KIR+ NKG2A-NK cells'. In: *Elife* 9 (2020), e55232.

[11] Bergmann C, Guay-Woodford LM, Harris PC, Horie S, Peters DJ and Torres VE. 'Polycystic kidney disease'. In: *Nature reviews Disease primers* 4.1 (2018), p. 50.

[12] Berkouk N. 'Persistence and the sheaf-function correspondence'. In: *arXiv preprint arXiv:2207.06335* (2022).

[13] Bernink JH, Peters CP, Munneke M, Te Velde AA, Meijer SL, Weijer K, Hreggvidsdottir HS, Heinsbroek SE, Legrand N, Buskens CJ et al. 'Human type 1 innate lymphoid cells accumulate in inflamed mucosal tissues'. In: *Nature immunology* 14.3 (2013), pp. 221–229.

[14] Bick C, Gross E, Harrington HA and Schaub MT. 'What are higher-order networks?' In: *arXiv preprint arXiv:2104.11329* (2021).

[15] Biffi G, Oni TE, Spielman B, Hao Y, Elyada E, Park Y, Preall J and Tuveson DA. 'IL1-Induced JAK/STAT Signaling Is Antagonized by TGFβ to Shape CAF Heterogeneity in Pancreatic Ductal AdenocarcinomaPathway Antagonism Shapes CAF Heterogeneity in PDAC'. In: *Cancer discovery* 9.2 (2019), pp. 282–301.

[16] Birey F, Andersen J, Makinson CD, Islam S, Wei W, Huber N, Fan HC, Metzler KRC, Panagiotakos G, Thom N et al. 'Assembly of functionally integrated human forebrain spheroids'. In: *Nature* 545.7652 (2017), pp. 54–59.

[17] Biswas R, Cultrera di Montesano S, Edelsbrunner H and Saghafian M. 'A window to the persistence of 1D maps. I: Geometric characterization of critical point pairs'. In: *LIPIcs* (2022).

[18] Blondel VD, Guillaume JL, Lambiotte R and Lefebvre E. 'Fast unfolding of communities in large networks'. In: *Journal of Statistical Mechanics: Theory and Experiment* (2008).

[19] Borten MA, Bajikar SS, Sasaki N, Clevers H and Janes KA. 'Automated brightfield morphometry of 3D organoid populations by OrganoSeg'. In: *Scientific Reports* 8.1 (2018), p. 5319. ISSN: 2045-2322. URL: `https://doi.org/10.1038/s41598-017-18815-8`.

[20] Bost P, Giladi A, Liu Y, Bendjelal Y, Xu G, David E, Blecher-Gonen R, Cohen M, Medaglia C, Li H et al. 'Host-viral infection maps reveal signatures of severe COVID-19 patients'. In: *Cell* 181.7 (2020), pp. 1475–1488.

[21] Boyer DM, Lipman Y, St. Clair E, Puente J, Patel BA, Funkhouser T, Jernvall J and Daubechies I. 'Algorithms to automatically quantify the geometric similarity of anatomical surfaces'. In: *Proceedings of the National Academy of Sciences* 108.45 (2011), pp. 18221–18226.

[22] Brazovskaja A, Treutlein B and Camp JG. 'High-throughput single-cell transcriptomics on organoids'. In: *Current opinion in biotechnology* 55 (2019), pp. 167–171.

[23] Breiman L. 'Random forests'. In: *Machine learning* 45.1 (2001), pp. 5–32.

[24] Bremaud P. 'Random Walks on Graphs'. In: *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues*. Cham: Springer International Publishing, 2020, pp. 255–287. ISBN: 978-3-030-45982-6. DOI: 10.1007/978-3-030-45982-6_8. URL: https://doi.org/10.1007/978-3-030-45982-6_8.

[25] Bremond-Martin C, Simon-Chane C, Clouchoux C and Histace A. 'TDA-Clustering Strategies for the Characterization of Brain Organoids'. In: *Ethical and Philosophical Issues in Medical Imaging, Multimodal Learning and Fusion Across Scales for Clinical Decision Support, and Topological Data Analysis for Biomedical Imaging: 1st International Workshop, EPIMI 2022, 12th International Workshop, ML-CDS 2022, 2nd International Workshop, TDA4BiomedicalImaging, Held in Conjunction with MICCAI 2022, Singapore, September 18–22, 2022, Proceedings*. Springer. 2022, pp. 113–122.

[26] Bues J, Biocanin M, Pezoldt J, Dainese R, Chrisnandy A, Rezakhani S, Saelens W, Gardeux V, Gupta R, Sarkis R et al. 'Deterministic scRNA-seq captures variation in intestinal crypt and organoid composition'. In: *Nature Methods* 19.3 (2022), pp. 323–330.

[27] Buske P, Przybilla J, Loeffler M, Sachs N, Sato T, Clevers H and Galle J. 'On the biomechanics of stem cell niche formation in the gut–modelling growing organoids'. In: *The FEBS journal* 279.18 (2012), pp. 3475–3487.

[28] Byrne H. 'Three Dimensional Biological Cultures and Organoids'. In: *Interface Focus* 10.20200014 (2020).

[29] Cagnol S and Rivard N. 'Oncogenic KRAS and BRAF Activation of the MEK/ERK Signaling Pathway Promotes Expression of Dual-Specificity Phosphatase 4 (DUSP4/MKP2) Resulting in Nuclear ERK1/2 Inhibition'. In: *Oncogene* 32 (2013), pp. 564–576.

[30] Camp JG, Badsha F, Florio M, Kanton S, Gerber T, Wilsch-Brauninger M, Lewitus E, Sykes A, Hevers W, Lancaster M et al. 'Human cerebral organoids recapitulate gene expression programs of fetal neocortex development'. In: *Proceedings of the National Academy of Sciences* 112.51 (2015), pp. 15672–15677.

[31] Camp JG, Sekine K, Gerber T, Loeffler-Wirth H, Binder H, Gac M, Kanton S, Kageyama J, Damm G, Seehofer D et al. 'Multilineage communication regulates human liver bud development from pluripotency'. In: *Nature* 546.7659 (2017), pp. 533–538.

[32] Carlsson G, Zomorodian A, Collins A and Guibas L. 'Persistence Barcodes for Shapes'. In: *International Journal of Shape Modeling* (2005).

[33] Chari T, Banerjee J and Pachter L. 'The specious art of single-cell genomics'. In: *BioRxiv* (2021), pp. 2021–08.

[34] Chazal F, De Silva V and Oudot S. 'Persistence stability for geometric complexes'. In: *Geometriae Dedicata* 173.1 (2014), pp. 193–214.

[35] Chern SS. 'Curves and surfaces in Euclidean space'. In: *Studies in global geometry and analysis* 4.1 (1967), p. 967.

[36] Chipman KC and Singh AK. 'Predicting genetic interactions with random walks on biological networks'. In: *BMC bioinformatics* 10.1 (2009), pp. 1–11.

[37] Cohen-Steiner D, Edelsbrunner H and Harer J. 'Extending Persistence Using Poincare and Lefschetz Duality'. In: *Foundations of Computational Mathematics* 9 (2008), pages79–103.

[38] Cohen-Steiner D, Edelsbrunner H and Harer J. 'Stability of persistence diagrams'. In: *Proceedings of the twenty-first annual symposium on Computational geometry*. 2005, pp. 263–271.

[39] Crawford L, Monod A, Chen AX, Mukherjee S and Rabadan R. 'Predicting Clinical Outcomes in Glioblastoma: An Application of Topological and Functional Data Analysis'. In: *Journal of the American Statistical Association* 115.531 (2020), pp. 1139–1150.

[40] Crinier A, Dumas PY, Escalière B, Piperoglou C, Gil L, Villacreces A, Vély F, Ivanovic Z, Milpied P, Narni-Mancinelli É et al. 'Single-cell profiling reveals the trajectories of natural killer cell differentiation in bone marrow and a stress signature induced by acute myeloid leukemia'. In: *Cellular & molecular immunology* 18.5 (2021), pp. 1290–1304.

[41] Curry J, Mukherjee S and Turner K. 'How Many Directions Determine A Shape and Other Sufficiency Results for Two Topological Transforms'. In: *arXiv: 1805.09782v2* (2019).

[42] Cuyx S, Santo Ramalho A, Corthout N, Fieuws S, Fuerstova E, Arnauts K, Ferrante M, Verfaillie C, Munck S, Boon M et al. 'Rectal organoid morphology analysis (ROMA) as a promising diagnostic tool in cystic fibrosis'. In: *Thorax* 76.11 (2021), pp. 1146–1149.

[43] Daitch SI, Kelner JA and Spielman DA. 'Fitting a graph to vector data'. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009, pp. 201–208.

[44] *Delaunay+Voronoi on a sphere*. https://www.redblobgames.com/x/1842-delaunay-voronoi-sphere/. Accessed: 2022-08-02.

[45] Delvenne JC, Schaub MT, Yaliraki SN and Barahona M. 'The stability of a graph partition: A dynamics-based framework for community detection'. In: *Dynamics On and Of Complex Networks, Volume 2: Applications to Time-Varying Dynamical Systems* (2013), pp. 221–242.

[46] Dijkstra KK, Cattaneo CM, Weeber F, Chalabi M, Haar J van de, Fanchi LF, Slagter M, Velden DL van der, Kaing S, Kelderman S et al. 'Generation of tumor-reactive T cells by co-culture of peripheral blood lymphocytes and tumor organoids'. In: *Cell* 174.6 (2018), pp. 1586–1598.

[47] Dłotko P and Gurnari D. 'Euler Characteristic Curves and Profiles: a stable shape invariant for big data problems'. In: *arXiv preprint arXiv:2212.01666* (2022).

[48] Donnat C, Levy A, Poitevin F, Zhong ED and Miolane N. 'Deep generative modeling for volume reconstruction in cryo-electron microscopy'. In: *Journal of Structural Biology* (2022), p. 107920.

[49] Drost J, Van Jaarsveld RH, Ponsioen B, Zimberlin C, Van Boxtel R, Buijs A, Sachs N, Overmeer RM, Offerhaus GJ, Begthel H et al. 'Sequential cancer mutations in cultured human intestinal stem cells'. In: *Nature* 521.7550 (2015), pp. 43–47.

[50] Edelsbrunner H and Harer J. *Computational topology: an introduction*. American Mathematical Soc., 2010.

[51] Emery CM, Vijayendran KG, Zipser MC, Sawyer AM, Niu L, Kim JJ, Hatton C, Chopra R, Oberholzer PA and Karpova MB. 'MEK1 Mutations Confer Resistance to MEK and B-RAF Inhibition'. In: *Proc. Natl. Acad. Sci.* 106 (2009), pp. 20411–20416.

[52] Ester M, Kriegel HP, Sander J and Xu X. 'A density-based algorithm for discovering clusters in large spatial databases with noise'. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)* (1996), pp. 226–231.

[53] Feragen A, Lauze F and Hauberg S. 'Geodesic exponential kernels: When curvature and linearity conflict'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3032–3042.

[54] Ferraro F, Celso CL and Scadden D. 'Adult stem cels and their niches'. In: *The Cell Biology of Stem Cells* (2010), pp. 155–168.

[55] Fritsch R and Piccinini RA. 'CW-complexes and Euclidean spaces'. In: *Rendiconti del circolo matematico di Palermo* 24 (1993), pp. 79–95.

[56] Fumagalli A, Drost J, Suijkerbuijk SJ, Van Boxtel R, De Ligt J, Offerhaus GJ, Begthel H, Beerling E, Tan EH, Sansom OJ et al. 'Genetic dissection of colorectal cancer progression by orthotopic transplantation of engineered cancer organoids'. In: *Proceedings of the National Academy of Sciences* 114.12 (2017), E2357–E2364.

[57] Gao T, Kovalsky SZ and Daubechies I. 'Gaussian process landmarking on manifolds'. In: *SIAM Journal on Mathematics of Data Science* 1.1 (2019), pp. 208–236.

[58] Gao Z, Long Y, Wu Y, Pu Y and Xue F. 'LncRNA LINC02253 activates KRT18/MAPK/ERK pathway by mediating N 6-methyladenosine modification of KRT18 mRNA in gastric cancer'. In: *Carcinogenesis* 43.5 (2022), pp. 419–429.

[59] Ghrist R, Levanger R and Mai H. 'Persistent homology and Euler integral transforms'. In: *Journal of Applied and Computational Topology* 2.1 (2018), pp. 55–60.

[60] Govek KW, Yamajala VS and Camara PG. 'Clustering-Independent Analysis of Genomic Data Using Spectral Simplicial Theory'. In: *PLOS Computational Biology* 15.11 (22nd Nov. 2019), e1007509. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007509. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007509 (visited on 20/10/2021).

[61] Goyal Y, Jindal GA, Pelliccia JL, Yamaya K, Yeung E, Futran AS, Burdine RD, Schüpbach T and Shvartsman SY. 'Divergent Effects of Intrinsically Active MEK Variants on Developmental Ras Signaling'. In: *Nat. Genet.* 49 (2017), pp. 465–469.

[62] Gritti N, Lim JL, Anlacs K, Pandya M, Aalderink G, Martinez-Ara G and Trivedi V. 'MOrgAna: accessible quantitative analysis of organoids with machine learning'. In: *Development* 148.18 (2021), dev199611.

[63] Grun D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H and Van Oudenaarden A. 'Single-cell messenger RNA sequencing reveals rare intestinal cell types'. In: *Nature* 525.7568 (2015), pp. 251–255.

[64] Guillen KP, Fujita M, Butterfield AJ, Scherer SD, Bailey MH, Chu Z, DeRose YS, Zhao L, Cortes-Sanchez E, Yang CH et al. 'A breast cancer patient-derived xenograft and organoid platform for drug discovery and precision oncology'. In: *bioRxiv* (2021).

[65] Haber AL, Biton M, Rogel N, Herbst RH, Shekhar K, Smillie C, Burgin G, Delorey TM, Howitt MR, Katz Y et al. 'A single-cell survey of the small intestinal epithelium'. In: *Nature* 551.7680 (2017), pp. 333–339.

[66] Habermann AC, Gutierrez AJ, Bui LT, Yahn SL, Winters NI, Calvi CL, Peter L, Chung MI, Taylor CJ, Jetter C et al. 'Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis'. In: *Science advances* 6.28 (2020), eaba1972.

[67] Hafemeister C and Satija R. 'Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression'. In: *Genome Biology* 20.296 (2019).

[68] Hao Y et al. 'Integrated analysis of multimodal single-cell data'. In: *Cell* (2021). DOI: 10.1016/j.cell.2021.04.048. URL: https://doi.org/10.1016/j.cell.2021.04.048.

[69] Hartung N, Mollard S, Barbolosi D, Benabdallah A, Chapuisat G, Henry G, Giacometti S, Iliadis A, Ciccolini J, Faivre C et al. 'Mathematical modeling of tumor growth and metastatic spreading: validation in tumor-bearing mice'. In: *Cancer research* 74.22 (2014), pp. 6397–6407.

[70] Hatcher A. 'Algebraic Topology'. In: *http://www. math. cornell. edu/˜ hatcher/AT/ATpage. html* (2002).

[71] He X, Cai D and Niyogi P. 'Laplacian score for feature selection'. In: *Advances in neural information processing systems* 18 (2005).

[72] Hill SJ, Decker B, Roberts EA, Horowitz NS, Muto MG, Worley Jr MJ, Feltmate CM, Nucci MR, Swisher EM, Nguyen H et al. 'Prediction of DNA repair inhibitor response in short-term patient-derived ovarian cancer organoids'. In: *Cancer discovery* 8.11 (2018), pp. 1404–1421.

[73] Ho TK. 'Random decision forests'. In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282.

[74] Hof L, Moreth T, Koch M, Liebisch T, Kurtz M, Tarnick J, Lissek SM, Verstegen MM, Laan LJ van der, Huch M et al. 'Long-term live imaging and multiscale analysis identify heterogeneity and core principles of epithelial organoid morphogenesis'. In: *BMC biology* 19 (2021), pp. 1–22.

[75] Huang B, Song JH, Cheng Y, Abraham JM, Ibrahim S, Sun Z, Ke X and Meltzer S. 'Long non-coding antisense RNA KRT7-AS is activated in gastric cancers and supports cancer cell progression by increasing KRT7 expression'. In: *Oncogene* 35.37 (2016), pp. 4927–4936.

[76] Huang CS, Chu J, Zhu XX, Li JH, Huang XT, Cai JP, Zhao W and Yin XY. 'The C/EBPβ-LINC01133 axis promotes cell proliferation in pancreatic ductal adenocarcinoma through upregulation of CCNG1'. In: *Cancer letters* 421 (2018), pp. 63–72.

[77] Im S, Yoo C, Jung JH, Choi HJ, Yoo J and Kang CS. 'Reduced expression of TFF1 and increased expression of TFF3 in gastric cancer: correlation with clinicopathological parameters and prognosis'. In: *International journal of medical sciences* 10.2 (2013), p. 133.

[78] Ishihara K, Mukherjee A, Gromberg E, Brugues J, Tanaka EM and Juelicher F. 'Topological morphogenesis of neuroepithelial organoids'. In: *Nature Physics* (2022), pp. 1–7.

[79] Jeitziner R, Carriere M, Rougemont J, Oudot S, Hess K and Brisken C. 'Two-tier mapper: a user-independent clustering method for global gene expression analysis based on topology'. In: *arXiv preprint arXiv:1801.01841* (2017).

[80] Johnson RA and Wichern DW. *Applied multivariate statistical analysis*. Prentice hall, Upper Saddle River, NJ, 2002.

[81] Kanton S, Boyle MJ, He Z, Santel M, Weigert A, Sanchis-Calleja F, Guijarro P, Sidow L, Fleck JS, Han D et al. 'Organoid single-cell genomic atlas uncovers human-specific features of brain development'. In: *Nature* 574.7778 (2019), pp. 418–422.

[82] Kanton S, Treutlein B and Camp JG. 'Single-cell genomic analysis of human cerebral organoids'. In: *Methods in Cell Biology* 159 (2020), pp. 229–256.

[83] Kassis T, Hernandez-Gordillo V, Langer R and Griffith LG. 'OrgaQuant: human intestinal organoid localization and quantification using deep convolutional neural networks'. In: *Scientific reports* 9.1 (2019), pp. 1–7.

[84] Kim S, Choung S, Sun RX, Ung N, Hashemi N, Fong EJ, Lau R, Spiller E, Gasho J, Foo J et al. 'Comparison of cell and organoid-level analysis of patient-derived 3D organoids to evaluate tumor cell growth dynamics and drug response'. In: *SLAS DISCOVERY: Advancing the Science of Drug Discovery* 25.7 (2020), pp. 744–754.

[85] Koepernik P and Pfaff F. 'Consistency of Gaussian Process Regression in Metric Spaces.' In: *J. Mach. Learn. Res.* 22 (2021), pp. 244–1.

[86] Kondor R and Pan H. 'The multiscale laplacian graph kernel'. In: *Advances in neural information processing systems* 29 (2016).

[87] Kretzschmar K and Clevers H. 'Organoids: modeling development and the stem cell niche in a dish'. In: *Developmental cell* 38.6 (2016), pp. 590–600.

[88] Kuchroo M, Godavarthi A, Tong A, Wolf G and Krishnaswamy S. 'Multimodal Data Visualization and Denoising with Integrated Diffusion'. In: *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2021, pp. 1–6.

[89] Kumar S, Mohri M and Talwalkar A. 'Sampling methods for the Nystroem method'. In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 981–1006.

[90] Lähnemann D et al. 'Eleven grand challenges in single-cell data science'. In: *Genome Biology* 21.1 (2020), p. 31. DOI: 10.1186/s13059-020-1926-6. URL: https://doi.org/10.1186/s13059-020-1926-6.

[91] Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, Bassez A, Decaluwe H, Pircher A, Van den Eynde K et al. 'Phenotype molding of stromal cells in the lung tumor microenvironment'. In: *Nature medicine* 24.8 (2018), pp. 1277–1289.

[92] Langlands AJ, Almet AA, Appleton PL, Newton IP, Osborne JM and Näthke IS. 'Paneth cell-rich regions separated by a cluster of Lgr5+ cells initiate crypt fission in the intestinal stem cell niche'. In: *PLoS biology* 14.6 (2016), e1002491.

[93] Lannagan TR, Lee YK, Wang T, Roper J, Bettington ML, Fennell L, Vrbanac L, Jonavicius L, Somashekar R, Gieniec K et al. 'Genetic editing of colonic organoids provides a molecularly distinct and orthotopic preclinical model of serrated carcinogenesis'. In: *Gut* 68.4 (2019), pp. 684–692.

[94] Lee JK, Wang J, Sa JK, Ladewig E, Lee HO, Lee IH, Kang HJ, Rosenbloom DS, Camara PG, Liu Z et al. 'Spatiotemporal genomic architecture informs precision oncology in glioblastoma'. In: *Nature genetics* 49.4 (2017), pp. 594–599.

[95] Lee RD, Munro SA, Knutson TP, LaRue RS, Heltemes-Harris LM and Farrar MA. 'Single-cell analysis identifies dynamic gene expression networks that govern B cell development and transformation'. In: *Nature communications* 12.1 (2021), pp. 1–16.

[96] Lee SH, Hu W, Matulay JT, Silva MV, Owczarek TB, Kim K, Chua CW, Barlow LJ, Kandoth C, Williams AB et al. 'Tumor evolution and drug response in patient-derived organoid models of bladder cancer'. In: *Cell* 173.2 (2018), pp. 515–528.

[97] Liu X, Cheng Y, Abraham JM, Wang Z, Wang Z, Ke X, Yan R, Shin EJ, Ngamruengphong S, Khashab MA et al. 'Modeling Wnt signalling by CRISPR-Cas9 genome editing recapitulates neoplasia in human Barrett epithelial organoids'. In: *Cancer letters* 436 (2018), pp. 109–118.

[98] Marsh L, Zhou FY, Qin X, Lu X, Byrne HM and Harrington HA. 'Detecting Temporal shape changes with the Euler Characteristic Transform'. In: *arXiv preprint arXiv:2212.10883* (2022).

[99] Matano M, Date S, Shimokawa M, Takano A, Fujii M, Ohta Y, Watanabe T, Kanai T and Sato T. 'Modeling colorectal cancer using CRISPR-Cas9–mediated engineering of human intestinal organoids'. In: *Nature medicine* 21.3 (2015), pp. 256–262.

[100] Matthews JM, Schuster B, Kashaf SS, Liu P, Ben-Yishay R, Ishay-Ronen D, Izumchenko E, Shen L, Weber CR, Bielski M et al. 'OrganoID: A versatile deep learning platform for tracking and analysis of single-organoid dynamics'. In: *PLOS Computational Biology* 18.11 (2022), e1010584.

[101] Maust JD, Whitehead CE and Sebolt-Leopold JS. 'Oncogenic Mutants of MEK1: A Trilogy Unfolds'. In: *Cancer Discov.* 8 (2018), pp. 534–536.

[102] McInnes L and Healy J. 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction'. In: *arXiv: 1802.03426* (2018).

[103] Meng JR, Tang HZ, Zhou KZ, Shen WH and Guo HY. 'TFF3 and survivin expressions associate with a lower survival rate in gastric cancer'. In: *Clinical and experimental medicine* 13.4 (2013), pp. 297–303.

[104] Meng K, Crawford L and Eloyan A. 'Randomness and Statistical Inference of Shapes via the Smooth Euler Characteristic Transform'. In: *arXiv preprint arXiv:2204.12699* (2022).

[105] Monticelli LA, Osborne LC, Noti M, Tran SV, Zaiss DM and Artis D. 'IL-33 promotes an innate immune pathway of intestinal tissue protection dependent on amphiregulin–EGFR interactions'. In: *Proceedings of the National Academy of Sciences* 112.34 (2015), pp. 10762–10767.

[106] Nadimpalli KV, Chattopadhyay A and Rieck B. 'Euler Characteristic Transform Based Topological Loss for Reconstructing 3D Images from Single 2D Slices'. In: (2023).

[107] Navis A and Nelson CM. 'Pulling together: Tissue-generated forces that drive lumen morphogenesis'. In: *Seminars in cell & developmental biology*. Vol. 55. Elsevier. 2016, pp. 139–147.

[108] NCSS L. *NCSS 2020 Statistical Software*. Kaysville, Utah, USA, 2020.

[109] Ohlund D, Handly-Santana A, Biffi G, Elyada E, Almeida AS, Ponz-Sarvise M, Corbo V, Oni TE, Hearn SA, Lee EJ et al. 'Distinct populations of inflammatory fibroblasts and myofibroblasts in pancreatic cancer'. In: *Journal of Experimental Medicine* 214.3 (2017), pp. 579–596.

[110] Ordan M, Pallara C, Maik-Rachline G, Hanoch T, Gervasio FL, Glaser F, Fernandez-Recio J and Seger R. 'Intrinsically Active MEK Variants are Differentially Regulated by Proteinases and Phosphatases'. In: *Sci. Rep.* 8 (2018), pp. 1–16.

[111] Oudot SY. *Persistence Theory: From Quiver Representations to Data Analysis*. American Mathematical Society, 2015.

[112] Park J, Shrestha R, Qiu C, Kondo A, Huang S, Werth M, Li M, Barasch J and Susztak K. 'Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease'. In: *Science* 360.6390 (2018), pp. 758–763.

[113] Pauli C, Hopkins BD, Prandi D, Shaw R, Fedrizzi T, Sboner A, Sailer V, Augello M, Puca L, Rosati R et al. 'Personalized In Vitro and In Vivo Cancer Models to Guide Precision MedicinePersonalized Cancer Models to Guide Precision Medicine'. In: *Cancer discovery* 7.5 (2017), pp. 462–477.

[114] Pearson K. 'LIII. On lines and planes of closest fit to systems of points in space'. In: *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2.11 (1901), pp. 559–572.

[115] Pedregosa F et al. 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[116] Pijuan-Sala B, Griffiths JA, Guibentif C, Hiscock TW, Jawaid W, Calero-Nieto FJ, Mulas C, Ibarra-Soria X, Tyser RC, Ho DLL et al. 'A single-cell molecular map of mouse gastrulation and early organogenesis'. In: *Nature* 566.7745 (2019), pp. 490–495.

[117] Quadrato G, Nguyen T, Macosko EZ, Sherwood JL, Min Yang S, Berger DR, Maria N, Scholvin J, Goldman M, Kinney JP et al. 'Cell diversity and network dynamics in photosensitive human brain organoids'. In: *Nature* 545.7652 (2017), pp. 48–53.

[118] Rabadan R, Mohamedi Y, Rubin U, Chu T, Alghalith AN, Elliott O, Arnes L, Cal S, Obaya AJ, Levine AJ et al. 'Identification of relevant genetic alterations in cancer using topological data analysis'. In: *Nature communications* 11.1 (2020), pp. 1–10.

[119] Rahimi A and Recht B. 'Random features for large-scale kernel machines'. In: *Advances in neural information processing systems* 20 (2007).

[120] Rasmussen CE. 'Gaussian processes in machine learning'. In: *Summer school on machine learning*. Springer. 2003, pp. 63–71.

[121] Riehl E. *Category Theory in Context*. Courier Dover Publications, 2017.

[122] Rizvi AH, Camara PG, Kandror EK, Roberts TJ, Schieren I, Maniatis T and Rabadan R. 'Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development.' eng. In: *Nature biotechnology* 35 (6 June 2017), pp. 551–560.

[123] Rossi R, Manfrin A and Lutolf A. 'Progress and Potential in Organoid Research'. In: *Nat. Rev. Genet.* 19 (2018), pp. 671–687.

[124] Rosvall M and Bergstrom CT. 'Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems'. In: *PloS one* 6.4 (2011), e18209.

[125] Saelens W, Cannoodt R, Todorov H and Saeys Y. 'A comparison of single-cell trajectory inference methods'. In: *Nature biotechnology* 37.5 (2019), pp. 547–554.

[126] Saggar M, Sporns O, Gonzalez-Castillo J, Bandettini PA, Carlsson G, Glover G and Reiss AL. 'Towards a new approach to reveal dynamical organization of the brain using topological data analysis'. In: *Nature communications* 9.1 (2018), pp. 1–14.

[127] Sard A. 'The measure of the critical values of differentiable maps'. In: *Bulletin of the American Mathematical Society* 48.12 (1942), pp. 883–890.

[128] Schaub MT, Benson AR, Horn P, Lippner G and Jadbabaie A. 'Random walks on simplicial complexes and the normalized Hodge 1-Laplacian'. In: *SIAM Review* 62.2 (2020), pp. 353–391.

[129] Schaub MT, Delvenne JC, Yaliraki SN and Barahona M. 'Markov Dynamics as a Zooming Lens for MultiscaleCommunity Detection: Non Clique-Like Communitiesand the Field-of-View Limit'. In: *PLoS ONE* 7.2 (2012).

[130] Schaub MT, Zhu Y, Seby JB, Roddenberry TM and Segarra S. 'Signal processing on higher-order networks: Livin'on the edge... and beyond'. In: *Signal Processing* 187 (2021), p. 108149.

[131] Schindler D, Clarke J and Barahona M. 'Multiscale mobility patterns and the restriction of human mobility under lockdown'. In: *arXiv preprint arXiv:2201.06323* (2022).

[132] Shang Y, Feng B, Zhou L, Ren G, Zhang Z, Fan X, Sun Y, Luo G, Liang J, Wu K et al. 'The miR27b-CCNG1-P53-miR-508-5p axis regulates multidrug resistance of gastric cancer'. In: *Oncotarget* 7.1 (2016), p. 538.

[133] Sheather SJ and Jones MC. 'A reliable data-based bandwidth selection method for kernel density estimation'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.3 (1991), pp. 683–690.

[134] Singh G, Memoli F and Gunnar C. 'Topological Methods for the Analysis of High DimensionalData Sets and 3D Object Recognition'. In: *Proc. Eurographics Symposium on Point-Based Graphics* (2007).

[135] Skraba P and Turner K. 'Wasserstein stability for persistence diagrams'. In: *arXiv preprint arXiv:2006.16824* (2020).

[136] Slack J. *Stem Cells: A Very Short Introduction*. Oxford University Press, 2012.

[137] Smillie CS, Biton M, Ordovas-Montanes J, Sullivan KM, Burgin G, Graham DB, Herbst RH, Rogel N, Slyper M, Waldman J et al. 'Intra-and inter-cellular rewiring of the human colon during ulcerative colitis'. In: *Cell* 178.3 (2019), pp. 714–730.

[138] Spivak DI. 'Metric Realization of Fuzzy Simplicial Sets'. `http://math.mit.edu/~dspivak/files/metric_realization.pdf`, last checked: 04/10/21.

[139] Tang WS, Silva GM da, Kirveslahti H, Skeens E, Feng B, Sudijono T, Yang KK, Mukherjee S, Rubenstein B and Crawford L. 'A topological data analytic approach for discovering biophysical signatures in protein dynamics'. In: *PLoS computational biology* 18.5 (2022), e1010045.

[140] Thalheim T, Quaas M, Herberg M, Braumann UD, Kerner C, Loeffler M, Aust G and Galle J. 'Linking stem cell function and growth pattern of intestinal organoids'. In: *Developmental Biology* 433.2 (2018), pp. 254–261. ISSN: 0012-1606. URL: `http://www.sciencedirect.com/science/article/pii/S0012160617303767`.

[141] Tiriac H, Belleau P, Engle DD, Plenker D, Deschenes A, Somerville TD, Froeling FE, Burkhart RA, Denroche RE, Jang GH et al. 'Organoid profiling identifies common responders to chemotherapy in pancreatic cancerpancreatic cancer organoids parallel patient response'. In: *Cancer discovery* 8.9 (2018), pp. 1112–1129.

[142] Torres BY, Oliveira JHM, Tate AT, Rath P, Cumnock K and Schneider DS. 'Tracking resilience to infections by mapping disease space'. In: *PLoS biology* 14.4 (2016), e1002436.

[143] Traag VA, Waltman L and Van Eck NJ. 'From Louvain to Leiden: guaranteeing well-connected communities'. In: *Scientific reports* 9.1 (2019), pp. 1–12.

[144] Tremblay N and Borgnat P. 'Graph wavelets for multiscale community mining'. In: *IEEE Transactions on Signal Processing* 62.20 (2014), pp. 5227–5239.

[145] Tsuji T, Satoyoshi R, Aiba N, Kubo T, Yanagihara K, Maeda D, Goto A, Ishikawa K, Yashiro M and Tanaka M. 'Agr2 Mediates Paracrine Effects on Stromal Fibroblasts That Promote Invasion by Gastric Signet-Ring Carcinoma CellsAgr2 Promotes Fibroblast-Associated Cancer Invasion'. In: *Cancer Research* 75.2 (2015), pp. 356–366.

[146] Turner K, Mukherjee S and Boyer DM. 'Peristent Homology Transform for Modeling Shapes and Surfaces'. In: *arxiv: 310.1030v2* (2014).

[147] Turner K, Robins V and Morgan J. 'The Extended Persistent Homology Transform of manifolds with boundary'. In: *arXiv preprint arXiv:2208.14583* (2022).

[148] Tuveson D and Clevers H. 'Cancer modeling meets human organoid technology'. In: *Science* 364.6444 (2019), pp. 952–955.

[149] Ullah H, Sajid M, Yan K, Feng J, He M, Shereen MA, Li Q, Xu T, Hao R, Guo D et al. 'Antiviral activity of interferon alpha-inducible protein 27 against hepatitis B virus gene expression and replication'. In: *Frontiers in Microbiology* 12 (2021), p. 656353.

[150] Valdeolivas A, Tichit L, Navarro C, Perrin S, Odelin G, Levy N, Cau P, Remy E and Baudot A. 'Random walk with restart on multiplex and heterogeneous biological networks'. In: *Bioinformatics* 35.3 (2019), pp. 497–505.

[151] Van den Dries L et al. *Tame topology and o-minimal structures*. Vol. 248. Cambridge university press, 1998.

[152] Van der Maaten L and Hinton G. 'Visualizing data using t-SNE.' In: *Journal of machine learning research* 9.11 (2008).

[153] Van Veen HJ, Saul N, Eargle D and Mangham SW. 'Kepler Mapper: A flexible Python implementation of the Mapper algorithm.' In: *Journal of Open Source Software* 4.42 (2019), p. 1315.

[154] Vlachogiannis G, Hedayat S, Vatsiou A, Jamin Y, Fernandez-Mateos J, Khan K, Lampis A, Eason K, Huntingford I, Burke R et al. 'Patient-derived organoids model treatment response of metastatic gastrointestinal cancers'. In: *Science* 359.6378 (2018), pp. 920–926.

[155] Wang B, Sudijono T, Kirveslahti H, Gao T, Boyer DM, Mukherjee S and Crawford L. 'A statistical pipeline for identifying physical features that differentiate classes of 3D shapes'. In: *The Annals of Applied Statistics* 15.2 (2021), pp. 638–661.

[156] Wang X, Ghareeb WM, Lu X, Huang Y, Huang S and Chi P. 'Coexpression network analysis linked H2AFJ to chemoradiation resistance in colorectal cancer'. In: *Journal of Cellular Biochemistry* 120.6 (2019), pp. 10351–10362.

[157] Wegelin JA. 'A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case'. In: (2000).

[158] Williams C and Seeger M. 'Using the Nyström method to speed up kernel machines'. In: *Advances in neural information processing systems* 13 (2000).

[159] Wolf FA, Angerer P and Theis FJ. 'SCANPY: large-scale single-cell gene expression data analysis'. In: *Genome biology* 19.1 (2018), pp. 1–5.

[160] Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, Rajewsky N, Simon L and Theis FJ. 'PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells'. In: *Genome biology* 20.1 (2019), pp. 1–9.

[161] Yan H, Konstorum A and Lowengrub JS. 'Three-dimensional spatiotemporal modeling of colon cancer organoids reveals that multimodal control of stem cell self-renewal is a critical determinant of size and shape in early stages of tumor growth'. In: *Bulletin of mathematical biology* 80.5 (2018), pp. 1404–1433.

[162] Yin Y, Liu PY, Shi Y and Li P. 'Single-cell sequencing and organoids: a powerful combination for modelling organ development and diseases'. In: *Reviews of Physiology, Biochemistry and Pharmacology* (2021), pp. 189–210.

[163] Yuan S, Norgard RJ and Stanger BZ. 'Cellular Plasticity in CancerCancer Cells Change Identity during Tumor Progression'. In: *Cancer discovery* 9.7 (2019), pp. 837–851.

[164] Zaiss DM, Gause WC, Osborne LC and Artis D. 'Emerging functions of amphiregulin in orchestrating immunity, inflammation, and tissue repair'. In: *Immunity* 42.2 (2015), pp. 216–226.

[165] Zhang Q, He Y, Luo N, Patel SJ, Han Y, Gao R, Modak M, Carotta S, Haslinger C, Kind D et al. 'Landscape and dynamics of single immune cells in hepatocellular carcinoma'. In: *Cell* 179.4 (2019), pp. 829–845.

[166] Zhu A, Srivastava A, Ibrahim JG, Patro R and Love MI. 'Nonparametric expression analysis using inferential replicate counts'. In: *Nucleic Acids Research* 47.18 (2019), e105–e105.

Appendix

Chapter Content

A.1 Differential Expression Analyses 131

A.2 UMAP Theory 133

A.3 ECT Stability Proofs 141

A.1 Differential Expression Analyses

	Cluster 0		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	Gene	P-value	Gene	P-value	Gene	P-value	Gene	P-value	Gene	P-value	Gene	P-value	Gene	P-value
1	LTB	0	GZMK	0	CCL4L2	0	HSPA6	0	CD52	0	NKG7	0	EEF1A1	0
2	RPL34	0	SPP1	0	GNLY	0	HSPA1A	0	S100A4	2.073E-283	FGFBP2	0	RPL17	0
3	RPL11	0	CXCR4	0	CCL3	0	TRDC	0	ZFP36L2	6.056E-232	PRF1	0	RPL3	0
4	SELL	0	BTG1	1.35E-293	CXCL13	0	HSPA1B	4.8E-271	ANKRD28	7.172E-199	SPON2	0	RPS2	0
5	MALAT1	0	TXNIP	3.38E-259	GZMB	0	MT-ND3	1.14E-202	CD40LG	1.149E-195	GNLY	0	RPL7	0
6	RPS6	2.59E-300	CD69	1.4E-245	CCL4	0	CCL5	1.05E-172	GLUL	9.413E-183	FCGR3A	0	NBEAL1	0
7	RPL32	4.79E-288	RPS14	1.5E-186	NKG7	0	CD7	1.52E-168	SFTPC	1.233E-162	KLRD1	0	HNRNPA1	0
8	RPS8	2.2E-277	MALAT1	3.3E-181	ISG15	0	AC092580.4	4.26E-164	RBPJ	2.597E-148	TYROBP	0	RPL6	0
9	RPS18	1.86E-251	JUNB	1.06E-164	LAG3	0	MT-ND4	1.87E-153	RGCC	2.103E-142	CST7	0	RPS3A	0
10	RPS12	1.92E-234	bRPLP2	6.24E-158	GZMA	0	KLRD1	2.77E-137	TMEM173	2.319E-134	KLRF1	0	RPL15	0

	Cluster 7		Cluster 8		Cluster 9		Cluster 10		Cluster 11		Cluster 12	
	Gene	P-value	Gene	P-value	Gene	P-value	Gene	P-value	Gene	P-value	Gene	P-value
1	CRIP1	0	CXCL13	0	TUBA1B	0	IFI27	0	HBB	0	IGHGP	0
2	FTH1	0	SPP1	8.49E-146	STMN1	0	ISG15	2.78E-186	HBA2	0	JCHAIN	3E-188
3	ANXA1	0	SRGN	1.3E-127	TUBB	0	IFI6	1.64E-175	HBA1	0	IGHG1	8E-131
4	RPL41	0	TNFRSF18	1.29E-115	HMGB2	0	MX1	1.27E-163	SFTPC	8.3411E-26	IGHG3	3E-100
5	CREM	0	RBPJ	1.92E-104	H2AFZ	0	IFI44L	1.86E-157	ZFP36	8.7624E-13	IGKC	7.13E-86
6	VIM	0	RGS1	1.301E-96	HIST1H4C	0	LY6E	3.04E-116	ZFP36L2	4.9141E-10	IGHA1	2.32E-57
7	RPS29	0	NR3C1	2.67E-86	GAPDH	0	OAS1	9.56E-105	MT-ND3	1.463E-08	MZB1	3.16E-46
8	ZFP36	0	SLA	1.342E-78	KIAA0101	0	IFIT3	7.82E-104	MT-CO1	6.9688E-08	IGLC2	1.49E-45
9	RPS27	0	CTLA4	9.869E-76	bUBE2C	0	IFIT1	4.577E-97	JUNB	1.1411E-07	IGLC3	4.29E-39
10	SFTPC	0	FKBP5	6.111E-71	MKI67	0	HERC5	5.141E-91	NFKBIA	2.0394E-06	SSR4	5.4E-16

Table A.1: The top 10 differentially expressed genes in each cluster of the T cell data set. For each cluster, 10 pairs of gene names and adjusted p-value are given. The clusters are given in Figure A.1 and are determined using a 20-nn graph on the first 30 PCs on the variance stabilised data via the Louvain algorithm. The differential expression test is a Wilcoxon Rank Sum test and the p-values are adjusted using the Benjamini-Hochberg procedure with a false-discovery rate of 25%.

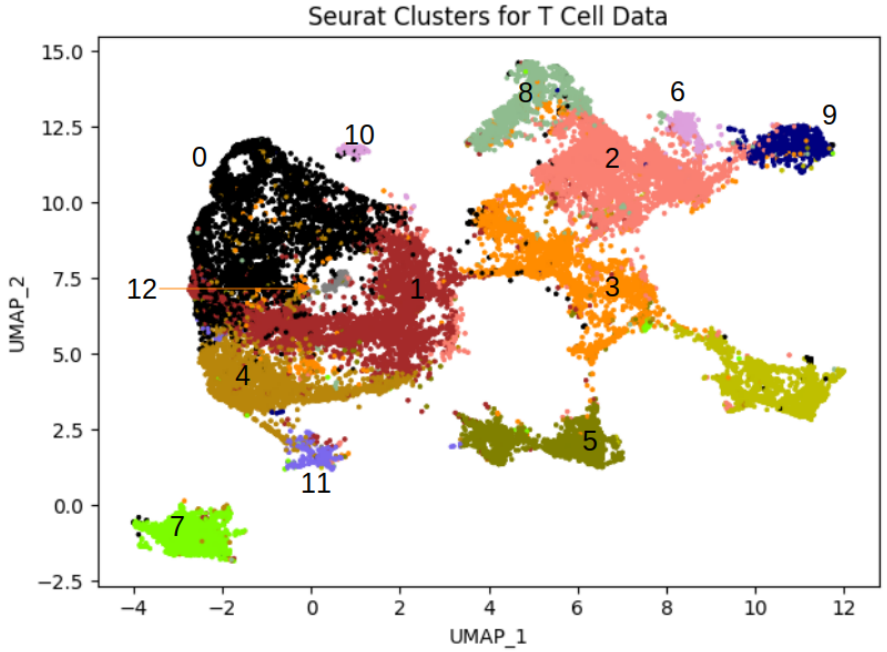


Figure A.1: Seurat clusters for differential expression testing on T cell data.

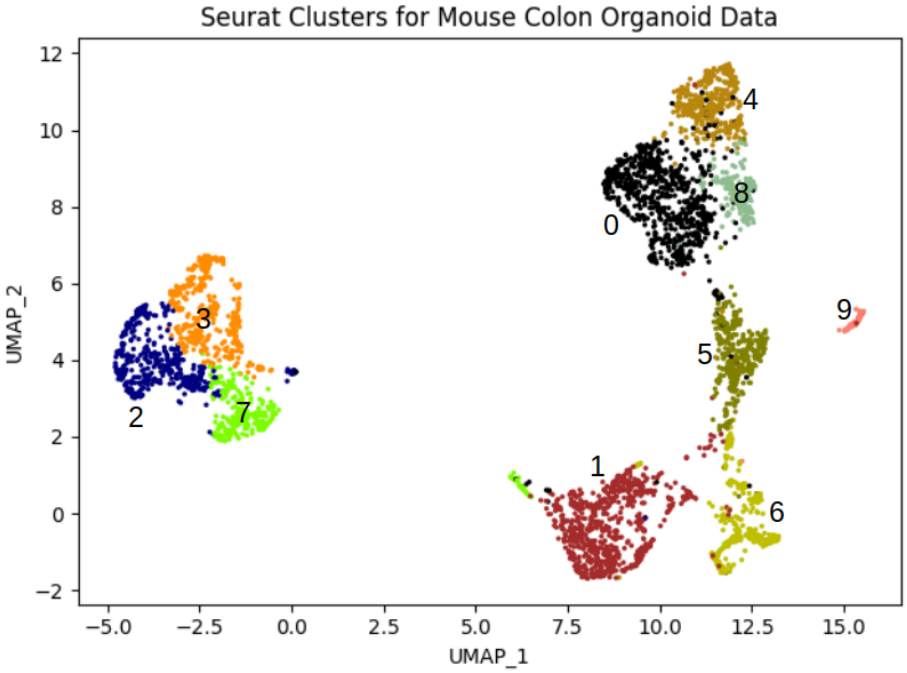


Figure A.2: Seurat clusters for differential expression testing on mouse colon organoid data.

	Cluster 0		Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	Gene	P-Value	Gene	P-Value	Gene	P-Value	Gene	P-Value	Gene	P-Value
1	mBanana	1.569E-192	EYFP	0	Aldh3a1	0	Spp1	7.515E-238	Aunip	2.92E-114
2	Gsta4	1.833E-156	Car2	1.04E-225	Ptgs1	0	Notum	1.179E-220	Smc2	2.288E-110
3	Eif2s2	1.536E-152	Khdc1a	2.269E-211	Krt90	0	Ephx1	1.118E-209	Cebpb	2.108E-108
4	Areg	5.383E-148	Crip1	1.509E-184	Stra6	1.682E-285	Sftpd	1.206E-209	Rpl36a	1.529E-107
5	Nupr1	9.651E-145	Gpx2	1.724E-176	Gabrp	6.281E-275	mCherry	1.167E-195	Hmgb2	3.327E-107
6	Eif4ebp1	5.692E-129	Cyba	3.834E-151	Gm42047	6.619E-239	Glipr1	2.718E-192	Fbxo5	1.378E-104
7	Ly6g	6.309E-126	H2afz	3.811E-143	Spp1	1.367E-235	Itm2b	4.517E-179	Hmgb1	7.518E-104
8	Gpx2	2.252E-123	Ptma	1.426E-141	mCherry	1.888E-232	Dcxr	8.186E-176	Tubb4b	1.04E-103
9	H2afv	1.045E-122	Tm4sf20	7.44E-129	Klk10	1.389E-220	Cyp2f2	8.253E-167	Aurkb	1.858E-103
10	Il1rn	1.013E-120	Gstp2	3.909E-119	Krt4	3.034E-210	Rpl19	9.274E-165	Lonp1	6.577E-100

	Cluster 5		Cluster 6		Cluster 7		Cluster 8		Cluster9	
	Gene	P-Value	Gene	P-Value	Gene	P-Value	Gene	P-Value	Gene	P-Value
1	Muc2	0	Pycard	0	Serpinb9b	5.088E-286	Hsd17b2	1.399E-144	Neurod1	0
2	Tpsg1	0	Osr2	3.326E-237	Psca	9.981E-255	Maf	1.6429E-85	Fev	5.114E-302
3	Tff3	1.98E-293	Clca3b	5.12E-236	Sprr1a	3.014E-240	Rfx3	1.7948E-84	Cck	2.113E-279
4	Ang4	5.969E-291	Insc	2.007E-197	S100a7a	2.519E-233	Stbd1	2.3731E-74	Chga	5.553E-277
5	Spdef	1.803E-270	Mtus2	1.731E-193	Mab21l4	2.549E-175	Gsdmc2	1.5539E-73	Tubb3	1.563E-267
6	Spink4	2.592E-264	Hoxb2	1.335E-191	Gm4610	1.399E-172	Sult1d1	1.3186E-70	Rimbp2	2.664E-260
7	Creb3l1	6.838E-261	Ms4a10	2.584E-181	Mal	6.844E-164	Gpc6	1.1688E-66	Scn3a	6.957E-255
8	Hpd	1.991E-253	Hoxb8	5.704E-180	Serpinb2	6.203E-161	Esx1	1.161E-64	Rfx6	2.395E-246
9	Agr2	8.054E-239	Wfdc2	2.151E-169	Clic3	1.78E-151	Rarb	4.982E-64	Gfra3	3.715E-242
10	Atoh1	2.264E-213	Hsd17b14	6.722E-143	Cryab	1.154E-150	Tstd1	1.4381E-58	Gdap1l1	1.084E-237

Table A.2: The top 10 differentially expressed genes in each cluster of the mouse colon organoid data set. For each cluster, 10 pairs of gene names and adjusted p-value are given. The clusters are given in Figure A.2 and are determined using a 20-nn graph on the first 30 PCs on the variance stabilised data via the Louvain algorithm. The differential expression test is a Wilcoxon Rank Sum test and the p-values are adjusted using the Benjamini-Hochberg procedure with a false-discovery rate of 25%.

A.2 UMAP Theory

In this section of the appendix, I provide more theoretical motivation for the UMAP algorithm, as well as a generalisation of the algorithm which uses filtered simplicial complexes instead of weighted graphs. This motivation follows the theoretical part of the UMAP paper [102]. However, I present their motivation in terms of Vietoris-Rips filtrations, a central construct in topological data analysis. This presentation in terms of Vietoris-Rips filtrations is novel (the paper [102] presents the motivation in terms of fuzzy simplicial sets, a related construction [138]). Further, I provide a concrete generalisation of the UMAP algorithm that uses higher-order interaction between data points to construct a dimension reduction. Such a generalisation is not provided by McInnes et al. [102] nor, to the best of my knowledge, elsewhere.

Extended Pseudometric Spaces and Neighbourhood Graphs

As mentioned in Section 2.5.2, the Euclidean metric is often a good approximation of the Riemannian metric on an embedded manifold $(\mathcal{M}, g)$ (up to multiplication by a positive scalar) for points that are close to each other. By contrast, the Euclidean metric yields no conclusive insight on the value g takes for points that are far apart. McInnes et al. use extended pseudometric spaces to formalise the intuition of a metric space with an 'inconclusive' option.

Definition A.1. Let X be a set and $d : X \times X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ a function. If d satisfies

1. $d(x, x) = 0$,
2. $d(x, y) = d(y, x)$, and
3. $d(x, z) \leq d(x, y) + d(x, z)$ or $d(x, z) = \infty$

for all $x, y, z \in X$, then the pair (X, d) is called an *extended pseudometric space*.

Given two extended pseudometric spaces (X, d_X) and (Y, d_Y), a map $f : X \to Y$ is called *non-expansive* if $d_Y(f(x_1), f(x_2)) \leq d_X(x_1, x_2)$ for all $x_1, x_2 \in X$. The collection of all extended pseudometric spaces (considered as objects) and all non-expansive maps between them (considered as morphisms) define a category, denoted **EPMet**.

If (X, d) is an extended pseudometric space and X is finite, we call it a *finite extended pseudometric space*. Analogously, all finite extended pseudometric spaces and the non-expansive maps between them form a category, denoted **fEPMet**.

All finite extended pseudometric spaces considered by UMAP are constructed from a finite metric space (X, d). The metric $d(x, y)$ is then updated by being scaled (and possibly translated, hence the need for the possibility $d(x, y) = 0$ for $x \neq y$) for points x and y which are 'close' and set to $d(x, y) = \infty$ for points which are distant to obtain a pseudometric. This extended pseudometric can then be constructed from a k-nearest-neighbour (k-nn) graph with weights in $(0, 1]$ (such as in Section 2.5.2):

For an $x \in X$, let E_x denote its k-nearest-neighbours (again, excluding x itself). Given $\tilde{w}$ as in Section 2.5.2, we define $c : X \times X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ (mnemonic for cost) by

$$c(x, y) = -\log\left(\tilde{w}(x, y)\right).$$

Subsequently, define $d : X \times X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ by

$$d(x, y) = \begin{cases} 0 & \text{if } x = y, \\ \min\left\{\sum_{i=1}^{n} c(z_{i-1}, z_i) \,|\, z_i \in X, z_0 = x, z_n = y\right\} & \text{if } y \in E_x \text{ or } x \in E_y, \\ \infty & \text{otherwise.} \end{cases}$$

Lemma A.2. *The function d defines an extended pseudometric on X.*

We will prove the above lemma later in this section.

Filtrations of Simplicial Complexes Revisited

As noted before, UMAP assumes a discrete sample from a continuous manifold. To approximate and describe the topology of the underlying manifold, it uses *simplicial complexes* and *filtrations*, which were introduced in Section 2.5.4. In this section, we endow filtered simplicial complexes with the structure of a category:

Definition A.3. We call a map $f : \mathcal{K} \to \mathcal{K}'$ between two simplicial complexes on vertex sets X and X' a *simplicial map* if there exists a map $g : X \to X'$ such that $f(\sigma) = \{g(x) \,|\, x \in \sigma\}$ for all $\sigma \in \mathcal{K}$.

The collection of all simplicial complexes (objects), together with all simplicial maps between them (morphisms), form a category, denoted **SC**.

Given two filtrations of simplicial complexes $(\mathcal{K}, \phi)$ and $(\mathcal{K}', \phi')$ and a simplicial map $f : \mathcal{K} \to \mathcal{K}'$, we call f a morphism of filtrations if $\phi(\sigma) \geq \phi'(f(\sigma))$ for all $\sigma \in \mathcal{K}$.

The collection of all filtrations of simplicial complexes and all morphisms between them form a category, denoted **fSC**. Moreover, we write $\mathbf{fSC}_{[0,1)}$ for the full subcategory containing filtrations $(\mathcal{K}, \phi)$ such that $0 \leq \phi(\sigma) < 1$ for all $\sigma \in \mathcal{K}$.

An Adjunction As An Optimisation Problem

Finite extended pseudometric spaces can be related to filtrations of simplicial complexes by constructing two functors between **fEPMet** and $\mathbf{fSC}_{[0,1)}$:

Definition A.4. We define the functor Sing : **fEPMet** $\to$ **fSC**$_{[0,1)}$ by action on finite extended pseudometric spaces (X, d) as

$$\mathrm{Sing}((X,d)) = \Big(\{\{x_0, ..., x_n\} \subseteq X \,|\, d(x_i, x_j) < \infty \,\forall 0 \leq i, j \leq n\}, \\ \phi(\sigma) = \max_{x,y \in \sigma} \left\{1 - e^{-d(x,y)}\right\} \Big).$$

Given a non-expansive map $f : (X, d_X) \to (Y, d_Y)$, we define

$$\mathrm{Sing}\,(f)\,(\sigma) = \{f(x) \,|\, x \in X\}.$$

Additionally, for $\lambda \geq 0$, we define Sing_λ : **fEPMet** $\to$ **fSC**$_{[0,1)}$ as $\mathrm{Sing}_\lambda((X,d)) = \mathrm{Sing}((X,d'))$, where $d'(x,y) = \max\{0, d(x,y) - \lambda\}$ for all $x, y \in X$.

UMAP uses (constructions equivalent to) Sing_λ for performance reasons [102]. Note that while d' need not satisfy the triangle inequality and thus need not be an extended pseudometric, Sing_λ is still well-defined.

Lemma A.5. *The maps* $\mathrm{Sing}\,(f)$ *defined in Definition A.4 are morphisms of filtrations.*

Proof. Let $f : (X, d_X) \to (Y, d_Y)$ be non-expansive and let $\sigma \in \mathcal{K}_X$ where $\mathrm{Sing}((X, d_X)) = (\mathcal{K}_X, \phi_X)$ and $\mathrm{Sing}((Y, d_Y)) = (\mathcal{K}_Y, \phi_Y)$.

Then for any $x, y \in \sigma$ we have $d_Y(f(x), f(y)) \leq d_X(x, y) < \infty$. Hence, Sing maps f into $\mathcal{K}_Y$ and is a well-defined simplicial map. Furthermore, $x \mapsto 1 - e^{-x}$ is monotonically increasing. Thus, for the $x', y' \in \sigma$ which maximise $1 - e^{-d_Y(f(x'),f(y'))}$ we have

$$\phi_Y(f(\sigma)) = 1 - e^{-d_Y(f(x'),f(y'))} \leq 1 - e^{-d_X(x',y')} \leq \max_{x,y \in \sigma} \left\{1 - e^{-d_X(x,y)}\right\} = \phi_X(\sigma).$$

Hence, $\mathrm{Sing}\,(f)$ is a morphism of filtrations of simplicial complexes. □

The above proof readily generalises to Sing_λ. We note that $\mathrm{Sing}((X,d))$ is also called the Vietoris-Rips filtration of X with (updated) extended pseudometric $d'(x,y) = 1 - e^{-d(x,y)}$.[1] In Euclidean space, Vietoris-Rips filtrations approximate

[1]Note that d' is again an extended pseudometric as $x \mapsto 1 - e^{-x}$ is monotonically increasing, concave, and maps 0 to 0.

Čech filtrations, which are homotopy equivalent to a thickening of the underlying discrete space X [111].

Definition A.6. Let $\mathcal{K}$ be a simplicial complex. We call an ordered sequence of elements of $\mathcal{K}$, $(\sigma_0, \sigma_1, \ldots, \sigma_n)$, a *path* in $\mathcal{K}$ if, for all i, we have $\sigma_i \in \mathcal{K}$, $|\sigma_i| = 2$, $|\sigma_{i-1} \cap \sigma_i| = 1$ and $|\sigma_{i-1} \cap \sigma_i \cap \sigma_{i+1}| = 0$. Denote the set of all paths in $\mathcal{K}$ by $\Pi(\mathcal{K})$. If x and y are in the vertex set of $\mathcal{K}$, we denote by $\Pi(\mathcal{K}, x, y)$ all paths of $\mathcal{K}$ such that $x \in \sigma_0$, $x \notin \sigma_1$, $y \in \sigma_n$ and $y \notin \sigma_{n-1}$.

We define the functor Real : $\mathbf{fSC}_{[0,1)} \to \mathbf{fEPMet}$ by mapping a filtration $(\mathcal{K}, \phi) \in \mathbf{fSC}_{[0,1)}$ to an extended pseudometric space (X, d) with $X = \{x \in \sigma \,|\, \sigma \in \mathcal{K}\}$ and

$$d(x,y) = \begin{cases} 0 & \text{if } x = y, \\ \min\left\{\sum_{i=0}^{n} -\log\left(1-\phi(\sigma_i)\right) \,|\, (\sigma_0, \ldots, \sigma_n) \in \Pi(\mathcal{K}, x, y)\right\} & \text{if } \{x,y\} \in \mathcal{K}, \\ \infty & \text{otherwise.} \end{cases}$$

Let $f : (\mathcal{K}, \phi) \to (\mathcal{K}', \phi')$ be a morphism of filtrations of simplicial complexes. Recall that f is a simplicial map, thus there exists a map g between the vertex sets of $\mathcal{K}$ and $\mathcal{K}'$ respectively which extends to f. We define $\mathrm{Real}(f)$ to be g.

Lemma A.7. *For any morphism of filtrations of simplicial complexes $f : (\mathcal{K}, \phi) \to (\mathcal{K}', \phi')$ the map* $\mathrm{Real}\,(f)$ *as defined in Definition A.6 is non-expansive.*

Proof. Let X be the vertex set of $\mathcal{K}$ and thus also the set of the extended pseudometric space $(\mathcal{K}, \phi)$ gets mapped to by Real. Let $x, y \in X$. The cases $x = y$ and $\{x, y\} \notin \mathcal{K}$ are trivial. We, therefore, focus on the remaining case, in which $x \neq y$ but $\{x, y\} \in \mathcal{K}$.

Let $(\sigma_0, ..., \sigma_n) \in \Pi(\mathcal{K}, x, y)$ a the path that minimises $\sum_{i=0}^{n} -\log\left(1-\phi(\sigma_i)\right)$ in the definition of $d(x, y)$. Note that $(f(\sigma_0), ..., f(\sigma_n)) \in \Pi(\mathcal{K}', \mathrm{Real}(f)(x), \mathrm{Real}(f)(y))$ (after removing all $f(\sigma_i)$ with cardinality 1).

As $x \mapsto -\log(1-x)$ is monotonically increasing, we find

$$\sum_{i=0}^{n} -\log\left(1-\phi(\sigma_i)\right) \geq \sum_{i=0}^{n} -\log\left(1-\phi'(f(\sigma_i))\right).$$

Thus, $d'(\mathrm{Real}(f)(x), \mathrm{Real}(f)(y)) \leq d(x, y)$. □

Lemma A.8. *The function d defined in Definition A.6 is an extended pseudometric.*

Proof. Condition 1 of Definition A.1 is satisfied by construction. To see that condition 2 is met, note that for any path $(\sigma_0, ..., \sigma_n)$ in $\mathcal{K}$ the reverse sequence of simplices, $(\sigma_n, ..., \sigma_0)$ is also a path in $\mathcal{K}$. Hence, the definition is symmetric.

Assume that $x, y, z \in X$ are such that $\{x, z\}, \{x, y\}, \{y, z\} \in \mathcal{K}$. Let $(\sigma_0, ..., \sigma_n) \in \Pi(\mathcal{K}, x, y)$ a the path that minimises $\sum_{i=0}^{n} -\log(1 - \phi(\sigma_i))$ in the definition of $d(x, y)$. Define $(\sigma'_0, ..., \sigma'_{n'}) \in \Pi(\mathcal{K}, y, z)$ analogously for $d(y, z)$. Note that $(\sigma_0, ..., \sigma_n, \sigma'_0, ..., \sigma'_{n'}) \in \Pi(\mathcal{K}, x, z)$. This implies $d(x, z) \leq d(x, y) + d(y, z)$. □

Theorem A.9. *The functors* Sing *and* Real *form an adjunction* Real $\dashv$ Sing.

One can view Theorem A.9 as an insight on an optimisation problem: if a functor F is left-adjoint to functor G, then F can be interpreted as the most efficient solution to the problem posed by G. Conversely, G represents the hardest problem F can solve [121]. To put this in the context of UMAP, we now return to Lemma A.2 and give its proof:

Proof of Lemma A.2. Define a simplicial complex $\mathcal{K}$ on vertex set X such that $\{x\} \in \mathcal{K}$ for all $x \in X$ and $\{x, y\} \in \mathcal{K}$ whenever $\tilde{w}(x, y) > 0$. Define $\phi : \mathcal{K} \to [0, 1)$ by $\phi(\{x\}) = 0$ for all $x \in X$ and $\phi(\{x, y\}) = 1 - \tilde{w}(x, y)$ for all $\{x, y\} \in \mathcal{K}$.

Then the metric space described in Lemma A.2 is Real$((\mathcal{K}, \phi))$. It is an extended pseudometric space by Lemma A.8. □

The construction of the filtration in the above proof can be viewed as applying Sing_{ρ_x} to the neighbourhood of each point x (defined by the k-nn graph), rescaling the extended pseudometric by σ_x and patching all of these local filtrations together in a compatible way [102]. More concretely, for a fixed point x and a point y in its neighbourhood, $\tilde{w}(x, y)$ represents the probability of an edge (i.e., the simplex $\{x, y\}$) existing between the two points. The probability is modelled in a neighbourhood of a point x and, in general, the probability that the same edge exists is, in general, different if the probability is modelled by y. These inconsistencies are handled by the definition of $\tilde{w}(x, y)$, which is the probability that the edge $\{x, y\}$ is modelled by

the neighbourhood of x or that of y. A higher-order simplex is added whenever all 1-simplices it contains are contained in the k-nn graph.

Theorem A.9 states that there exists a most efficient way of turning such a filtration into an extended pseudometric space, which uses the functor Real. The UMAP algorithm approximates this extended pseudometric space by embedding its points into low-dimensional Euclidean space.

Remark. Given $\mathcal{K}$ as in the above proof, $\mathrm{Sing}(\mathrm{Real}(\mathcal{K}, \phi))$ should be approximately the same as $\mathcal{K}$ for a well-behaved data set. Large discrepancies in filtration values between these two filtrations are typically an indication of outlier points or rapidly changing density in the original embedding in $\mathbb{R}^N$. Both phenomena should usually be investigated separately.

Proof of Theorem A.9. To prove this theorem, we need to show that

$$\mathbf{hom}(\mathrm{Real}((\mathcal{K}, \phi)), (X', d')) \simeq \mathbf{hom}((\mathcal{K}, \phi), \mathrm{Sing}((X', d'))$$

naturally for all $(\mathcal{K}, \phi) \in \mathbf{fSC}_{[0,1)}$ and $(X', d') \in \mathbf{fEPMet}$. We do this by explicit construction:

Assume $f : \mathrm{Real}((\mathcal{K}, \phi)) \to (X', d')$ is non-expansive. We map it to $g : (\mathcal{K}, \phi) \to \mathrm{Sing}((X', d'))$ given by $\sigma \mapsto \{f(x) \,|\, x \in X\}$ where X is the vertex set underlying $\mathcal{K}$. Fix σ and let $x, y \in \sigma$ be such that $1 - e^{-d'(f(x), f(y))}$ is maximised. As f is non-expansive and $d(x, y) \leq -\log(1 - \phi(\{x, y\}))$, we find $\phi'(f(\sigma)) \leq \phi(\sigma)$. Thus, the mapping is well-defined.

Conversely, assume that $g : (\mathcal{K}, \phi) \to \mathrm{Sing}((X', d'))$, a morphism of simplicial complexes, is given. We map it to $f : \mathrm{Real}((\mathcal{K}, \phi)) \to (X', d')$ defined by $f : x \mapsto *g(\{x\})$, where $*$ is the operator that replaces a one-element set with the element it contains. Let $\{x, y\} \in \mathcal{K}$ and assume for contradiction that $d(x, y) < d'(f(x), f(y))$. Then

$$\phi(\{x, y\}) \leq 1 - e^{-d(x,y)} < 1 - e^{-d'(f(x), f(y))} \leq \phi'(g(\{x, y\}),$$

a contradiction.

To see that the two mappings described are inverses to each other, observe that they are entirely determined by the action of f on points and of g on singletons. Applying the two mappings in order fixes points and singletons. By the same reasoning, they satisfy the naturality axioms of an adjunction too. □

The UMAP Algorithm Revisited

The UMAP algorithm first approximates the geodesic distance between samples $X = \{x_i\}_{i=1,\dots,n}$ *locally*. The distance between two close points (i.e., points that are k-nearest neighbours of each other) is rescaled and then symmetrised (see Equation (2.2)). In particular, the metric d in Lemma A.2 yields an extended pseudometric which (up to a scalar constant) approximates the geodesic distance of two close points on $\mathcal{M}$. Equivalently, the functor Real is applied to a k-nn graph weighted by $\tilde{w}$. To retain computational tractability, UMAP makes the simplifying assumption that $n = 1$ always in the definition of the extended pseudometric used in Lemma A.2.

To embed the sample $X = \{x_i\}_{i=1,\dots,n}$, UMAP turns the above extended pseudometric space into a filtered simplicial complex $\mathrm{Sing}((X, d))$. UMAP then aims to find coordinates $\{y_i\} \subset \mathbb{R}^d$ (with x_i corresponding to y_i) such that an extended pseudometric space constructed from $\{y_i\}$ would yield a similar realisation to $\mathrm{Sing}((X, d))$. In particular, if we define

$$d'(x_i, x_j) = \begin{cases} \|y_i - y_j\|_2 & \text{if } d(x_i, x_j) < \infty, \\ \infty & \text{otherwise,} \end{cases}$$

the filtered simplicial complexes $\mathrm{Sing}((X, d))$ and $\mathrm{Sing}_\lambda((X, d'))$ for user-defined $\lambda :=$ min-dist should be 'close'. The measure UMAP uses to define the closeness of filtered simplicial complexes is the Kullback-Leibler divergence:

Definition A.10. Let $\mathcal{K}$ be a simplicial complex and let ϕ and ψ be two filtrations of $\mathcal{K}$ in $\mathbf{fSC}_{[0,1)}$. Their *Kullback-Leibler divergence* is defined as

$$D_{SC}((\mathcal{K}, \phi), (\mathcal{K}, \psi)) = \sum_{\sigma \in \mathcal{K}} \phi(\sigma) \log\left(\frac{\phi(\sigma)}{\psi(\sigma)}\right) + (1 - \phi(\sigma)) \log\left(\frac{1 - \phi(\sigma)}{1 - \psi(\sigma)}\right). \quad \text{(A.1)}$$

Similarly to D defined in Section 2.5.2, we note that while D_{SC} is differentiable for changes in ψ, but ψ is not differentiable for small changes in its inputs in the

low-dimensional space $\mathbb{R}^d$ as Sing_λ is not differentiable. For a one-simplex connecting $y, y' \in \mathbb{R}^d$, we have that its filtration value is

$$v(y, y') = 1 - \exp\left(-\max\{0, \|y - y'\| - \text{min-dist}\}\right).$$

The filtration value of any higher-order simplex is again the maximal filtration value of the one-simplices it contains.

UMAP approximates v by the smooth function

$$\Phi(y, y') = 1 - \left(1 + a\|y - y'\|^{2b}\right)^{-1},$$

where a and b are determined by least-squares fitting against v. To generalise Φ to an n-simplex $\sigma \in \mathcal{K}$, I propose to define the C_2^n-dimensional vector P_σ which contains the Euclidean distance of all pairs of embedded vertices of σ in its coordinates. Then

$$\Phi(\sigma) := 1 - \left(1 + a\|P_\sigma\|_q^{2b}\right)^{-1}$$

for a user-defined q (q should be large in order to approximate the ∞-norm).

Then Algorithm 1 generalises to simplicial complexes to give Algorithm 4. By extension, UMAP generalises to filtered simplicial complexes. In this case, only the one-skeleton $\mathcal{K}_1$ of $\mathcal{K}$ is used for the initialisation of $\{y_i\}$ by the spectral embedding. To the best of my knowledge, no implementation of UMAP in the public domain currently implements the version of UMAP generalised to simplicial complexes, which is presented in this section. The Python UMAP implementation `umap-learn` justifies its use of only the 1-skeleton of $\mathcal{K}$ by improved runtime [102]. While Algorithm 4 uses the rigorous motivation of McInnes et al. in full, due to the lack of an implementation there is currently no evidence that it performs better in practice than the algorithm only using $\mathcal{K}_1$.

A.3 ECT Stability Proofs

Proofs of ECT Stability

Proof of Proposition 6.5. The last two statements follow from the fact that I has Euler characteristic one and the empty set has Euler characteristic zero. We now prove the remainder of the proposition.

Algorithm 4 Stochastic gradient descent on filtered simplicial complexes

procedure OPTIMISEEMBEDDING($(\mathcal{K}, \phi)$, $\{y_i\}$, min-dist, n-epochs, n-neg-samples)
  $\alpha \leftarrow 1.0$
  Φ is fitted from min-dist
  for $i \leftarrow 1, ...,$ n-epochs **do**
    for $\sigma \in \mathcal{K} \setminus \mathcal{K}_0$ **do**
      for $a \in \sigma$ **do**
        if Random() $\leq 1 - \phi(\sigma)$ **then**
          $y_a \leftarrow y_a + \alpha \cdot \nabla(\log(1 - \Phi))(\sigma)$
          for $j \leftarrow 1, ...,$ n-neg-samples **do**
            $c \leftarrow$ random vertex in $\mathcal{K}_1$
            $y_a \leftarrow y_a + \alpha \cdot \nabla(\log(\Phi))(y_a, y_c)$
          end for
        end if
      end for
    end for
    $\alpha \leftarrow 1.0 - i/$n-epochs
  end for
  return $\{y_i\}$
end procedure

The main goal of the proof is to establish the following two equalities:

$$\int_a^b |\mathrm{ECT}_\gamma(v,t)| \, \mathrm{d}t = \int_a^b \left| \pi_0\big[f^{-1}(-\infty, t]\big] \right| \, \mathrm{d}t$$
$$V(f) = \int_a^b \left| \pi_0\big[f^{-1}(t)\big] \right| \, \mathrm{d}t.$$

The first of these equalities follows from the fact that every subset of the unit interval is component-wise contractible. Hence, the Euler characteristic of any subset of I is the number of path-components it has. The second of these equalities is more difficult to show, and its establishment is the bulk of the proof. Once both equalities are shown, demonstrating that the first integrand on the right is less than or equal to the second integrand on the right completes the proof.

To begin, notice that f is piece-wise C^1 since γ is. For the proof, we let $G(f)$ be the set of points where f' is defined and positive, $D(f)$ be the set of points where f' is defined and negative, and $C(f)$ be the set of points in I neither in $G(f)$ or $D(f)$. Since f is piece-wise C^1, $G(f)$ and $D(f)$ are both open. Meanwhile, all but finitely

many points of $C(f)$ satisfy $f'(x) = 0$. Clearly, $G(f)$, $D(f)$, and $C(f)$ partition I. Hence,

$$\begin{aligned} V(f) &= \int_{G(f)} f'(x)\,\mathrm{d}x + \int_{D(f)} -f'(x)\,\mathrm{d}x + \int_{C(f)} |f'(x)|\,\mathrm{d}x \\ &= \int_{G(f)} f'(x)\,\mathrm{d}x + \int_{D(f)} -f'(x)\,\mathrm{d}x, \end{aligned}$$

since $f' = 0$ almost everywhere on $C(f)$. Since both $G(f)$ and $D(f)$ are open, each is a countable union of open sub-intervals of I, which we denote by $\{I_k\}_{k\in\Xi}$ and $\{J_l\}_{l\in\Theta}$ respectively. On each I_k f is increasing and on each J_l, f is decreasing. Hence, we get

$$\begin{aligned} V(f) &= \int_{G(f)} f'(x)\,\mathrm{d}x + \int_{D(f)} -f'(x)\,\mathrm{d}x \\ &= \sum_{k\in\Xi} \int_{I_k} f'(x)\,\mathrm{d}x + \sum_{l\in\Theta} \int_{J_l} -f'(x)\,\mathrm{d}x \\ &= \sum_{k\in\Xi} \int_{\mathbb{R}} \left|\pi_0\big[(f|_{I_k})^{-1}(t)\big]\right| \mathrm{d}t + \sum_{l\in\Theta} \int_{\mathbb{R}} \left|\pi_0\big[(f|_{J_l})^{-1}(t)\big]\right| \mathrm{d}t \\ &= \int_{\mathbb{R}} \left|\pi_0\big[(f|_{G(f)})^{-1}(t)\big]\right| \mathrm{d}t + \int_{\mathbb{R}} \left|\pi_0\big[(f|_{D(f)})^{-1}(t)\big]\right| \mathrm{d}t \\ &= \int_{\mathbb{R}} \left|\pi_0\big[(f|_{G(f)\cup D(f)})^{-1}(t)\big]\right| \mathrm{d}t. \end{aligned}$$

Here, the last line follows from the fact that if x and y have the same f value, each point in $G(f)$ or $D(f)$, then by the definition of these sets there must be a point between them that obtains either a smaller or larger value of f. The line before follows from similar reasoning. If it is granted that $f(C(f))$ has measure zero we then have

$$\begin{aligned} V(f) &= \int_{\mathbb{R}} \left|\pi_0\big[(f|_{G(f)\cup D(f)})^{-1}(t)\big]\right| \mathrm{d}t \\ &= \int_{\mathbb{R}} \left|\pi_0\big[(f|_{G(f)\cup D(f)})^{-1}(t)\big]\right| \mathrm{d}t + \int_{\mathbb{R}} \left|\pi_0\big[(f|_{C(f)})^{-1}(t)\big]\right| \mathrm{d}t \\ &= \int_{\mathbb{R}} \left|\pi_0\big[f^{-1}(t)\big]\right| \mathrm{d}t \\ &= \int_{a}^{b} \left|\pi_0\big[f^{-1}(t)\big]\right| \mathrm{d}t. \end{aligned}$$

Here, the second to last equality follows from the fact that if x is in $C(f)$ and y is in $G(f)$ or $D(f)$, then by definition of $G(f)$ and $D(f)$ there is a point between x and

y with an f value not equal to $f(y)$. In particular, the integrand at the end of this equation must be finite almost everywhere.

Now we show that indeed $f(C(f))$ has measure zero. Since f is piece-wise C^1, let $T_1, \ldots, T_k$ be closed intervals covering I on the interiors of which f is C^1. Consider Z_i, the subset of the interior of I_i with $f' = 0$. By Sard's Theorem (see for example [127, Theorem 7.2]), $f(Z_i)$ has measure zero. We have that $C(f)$ is a subset of $Z_1 \cup \ldots \cup Z_k \cup \partial T_1 \cup \ldots \cup \partial T_k$, whose image under f has measure zero. Thus $f(C(f))$ has measure zero.

Hence, the desired result follows once we establish that for any $t \in [a, b]$,

$$\left|\pi_0\big[f^{-1}(-\infty, t]\big]\right| \leq \left|\pi_0\big[f^{-1}(t)\big]\right|.$$

This inequality implies that $\mathrm{ECT}_f(v, t)$ is defined for almost all t since $\mathrm{ECT}_f(v, t)$ is just the left-hand side of this inequality, which is positive-valued, and bounded by a function that is finite for almost all t.

To prove this statement, it suffices to show that any path-component of $f^{-1}(-\infty, t]$ contains a point with f value t. Suppose otherwise, that there is a path-component C of $f^{-1}(-\infty, t]$ with $f(C) < t$. By continuity of f, C must also be a path-component of $f^{-1}(-\infty, b]$. Indeed, suppose α is a path from the complement of C to C. Thus, every neighborhood of $\alpha^{-1}(C)$ must intersect $(f \circ \alpha)^{-1}(t, b]$. But this produces a contradiction of the intermediate value theorem. So C is a path-component of $I = f^{-1}(-\infty, b]$ and hence is I. The fact that $f(C) < t \leq b$ thus contradicts the definition of b as the maximum of f, completing the proof. □

Remark 2. Suppose $\gamma : I \to \mathbb{R}^d$ is continuous and definable with respect to an o-minimal structure on $\mathbb{R}$. By [151, Chapter 7, Theorem 3.2]), γ is piece-wise C^1 and hence the above result applies.

Remark 3. In the case where f is tame, this result is implied by a stronger result of [17, Corollary 4.6] for tame functions. The main contribution of this proposition is that the stated bound still holds when f is not tame.

With the previous result in mind, we establish a bound on the variation of a curve that is approximately straight.

Lemma A.11. *Suppose $\gamma : I \to \mathbb{R}^d$ is a piece-wise differentiable path with length L and the first coordinate γ_1 satisfies $|\gamma_1(1) - \gamma_1(0)| = L_x$. Then the variation of any other coordinate function γ_n of γ is bounded by*

$$V(\gamma_n) \leq \sqrt{L^2 - L_x^2}. \tag{A.2}$$

Proof. Without loss of generality, we can assume that $\gamma(0) = 0$, $\gamma_1(1) = L_x$, and we can show this bound holds for γ_2 only.

From γ we can construct another function $\bar{\gamma}$ by

$$\bar{\gamma}(t) := \int_0^t (\gamma_1'(t),\ |\gamma_2'(t)|,\ \gamma_3'(t), \ldots,\ \gamma_d'(t))\ \mathrm{d}t.$$

Put differently, $\bar{\gamma}$ has the same coordinate functions as γ except in the second coordinate, where $\bar{\gamma}_2$ has the same absolute value of its derivative as γ, but is never decreasing. It is immediate that γ and $\bar{\gamma}$ have the same length, the same value of $\gamma_1(1)$, and variation in the second coordinate. Hence, it suffices to show that the lemma holds for γ_2 for curves γ with length L, $\gamma(0) = 0$, $\gamma_1(1) = L_x$, and $\gamma_2'(t) \geq 0$ for all t.

The fact that γ has length L implies that $\gamma(1)$ lies in the closed d-disk of radius L centred at the origin. The fact that $\gamma_1(x) = L_x$ implies that $\gamma(1)$ lies on the hyperplane of points with the first coordinate L_x. Elementary geometry shows that the intersection of the disk and the hyperplane is

$$\{(y_1, y_2, \ldots, y_n) \in \mathbb{R}^d : y_1 = L_x,\ y_2^2 + \ldots + y_d^2 \leq L^2 - L_x^2\}.$$

It is easily seen that the greatest value of y_2 on this set is $\sqrt{L^2 - L_x^2}$. So $\gamma_2(1)$ is bounded above by this value. Hence,

$$V(\gamma_2) = \int_0^1 \left|\gamma_2'(t)\right| \mathrm{d}t = \int_0^1 \gamma_2'(t)\ \mathrm{d}t = \gamma_2(1) - \gamma_2(0) = \gamma_2(1) \leq \sqrt{L^2 - L_x^2}.$$

□

We can now bound the L_1 distance between Euler characteristic transforms of nearby curves, assuming one of them is approximately straight.

Proposition A.12. *Let $\alpha : I \to \mathbb{R}^d$ be a piece-wise C^1 map such that the distance of $\alpha(0)$ to $\alpha(1)$ is L, with arc length no greater than $L + \varepsilon$. Let $\beta : I \to \mathbb{R}^d$ be another piece-wise C^1 map with arc length no greater than $L + 2\varepsilon$, and endpoints within ε of the corresponding endpoints of α. Then*

$$\left\|\mathrm{ECT}_\alpha - \mathrm{ECT}_\beta\right\| \leq \begin{cases} 8\sqrt{L\varepsilon} & L > 2\varepsilon \\ 10\varepsilon & L \leq 2\varepsilon. \end{cases}$$

Proof. Let v be an arbitrary unit vector, $w = \alpha(1) - \alpha(0)$, and θ be the angle between w and the hyperplane normal to v. After potentially applying a rotation to α and β, we may assume that $v = (0, 1, 0, \ldots, 0)$. By applying another rotation we may assume also that w is only non-zero in the first two coordinates.

Let f denote the inner product with v and let a and b be the minimum and maximum of $f \circ \alpha$ and c and d be the minimum and maximum of $f \circ \beta$. Throughout the proof, we use the fact that if both $\mathrm{ECT}_\alpha(v, t)$ and $\mathrm{ECT}_\beta(v, t)$ are non-zero, then

$$|\mathrm{ECT}_\alpha(v, t) - \mathrm{ECT}_\alpha(v, t)| \leq \mathrm{ECT}_\alpha(v, t) + \mathrm{ECT}_\beta(v, t) - 2,$$

as both Euler characteristic transforms must be positive (since subsets of the interval are component-wise contractible) and greater than 1.

First, suppose $\max(a, c) \leq \min(b, d)$. Then

$$\begin{aligned} \int_{\mathbb{R}} |\mathrm{ECT}_\alpha(v, t) - \mathrm{ECT}_\beta(v, t)| \, \mathrm{d}t = &\int_{\min(a,c)}^{\max(a,c)} |\mathrm{ECT}_\alpha(v, t) - \mathrm{ECT}_\beta(v, t)| \, \mathrm{d}t \\ &+ \int_{\max(a,c)}^{\min(b,d)} |\mathrm{ECT}_\alpha(v, t) - \mathrm{ECT}_\beta(v, t)| \, \mathrm{d}t \\ &+ \int_{\min(b,d)}^{\max(b,d)} |\mathrm{ECT}_\alpha(v, t) - \mathrm{ECT}_\beta(v, t)| \, \mathrm{d}t. \end{aligned}$$

Suppose also that $b \leq d$. Then the above expression is bounded by

$$\begin{aligned} \int_{\min(a,c)}^{\max(a,c)} \mathrm{ECT}_\alpha(v, t) + \mathrm{ECT}_\beta(v, t) \, \mathrm{d}t + \int_{\max(a,c)}^{b} &\mathrm{ECT}_\alpha(v, t) \\ + \mathrm{ECT}_\beta(v, t) - 2 \, \mathrm{d}t + \int_b^d &\mathrm{ECT}_\beta(v, t) - 1 \, \mathrm{d}t, \end{aligned}$$

where we have only had to approximate the middle term. If we additionally suppose $a \leq c$, then our bound is equal to

$$\int_a^c \mathrm{ECT}_\alpha(v,t)\,\mathrm{d}t + \int_c^b \mathrm{ECT}_\alpha(v,t) + \mathrm{ECT}_\beta(v,t) - 2\,\mathrm{d}t + \int_b^d \mathrm{ECT}_\beta(v,t) - 1\,\mathrm{d}t.$$

Rearranging and by the linearity of the integral, the above is equal to

$$\int_c^d \mathrm{ECT}_\beta(v,t)\,\mathrm{d}t + \int_a^b \mathrm{ECT}_\alpha(v,t)\mathrm{d}t - 2(b-c) - (d-b).$$

Similar analysis when $a > c$ and/or $b > d$ shows that when $\max(a,c) \leq \min(b,d)$,

$$\begin{aligned}\int_{\mathbb{R}} |\mathrm{ECT}_\alpha(v,t) - \mathrm{ECT}_\beta(v,t)|\,\mathrm{d}t \leq \int_c^d \mathrm{ECT}_\beta(v,t)\,\mathrm{d}t + \int_a^b \mathrm{ECT}_\alpha(v,t)\,\mathrm{d}t \\ - 2|\min(b,d) - \max(a,c)| - |\max(b,d) - \min(b,d)|.\end{aligned} \tag{A.3}$$

Note that $|b-a| = L|\sin\theta|$. By hypothesis $\|\alpha(0) - \beta(0)\| \leq \varepsilon$, so $|a-c| \leq \varepsilon$. Similarly $|b-d| \leq \varepsilon$. Hence,

$$|\min(b,d) - \max(a,c)| \geq L|\sin\theta| - 2\varepsilon$$

by the triangle inequality. Of course, the quantity on the left is also positive, so

$$|\min(b,d) - \max(a,c)| \geq \max(0, L|\sin\theta| - 2\varepsilon).$$

Trivially, we also have $|\max(b,d) - \min(b,d)| \geq 0$. Applying these inequalities, Proposition 6.5, and Lemma A.11 to Equation (A.3), we have

$$\begin{aligned}\int_{\mathbb{R}} |\mathrm{ECT}_\alpha(v,t) - \mathrm{ECT}_\beta(v,t)|\,\mathrm{d}t \leq \sqrt{(L+2\varepsilon)^2 - \max(0, L|\cos\theta| - 2\varepsilon)^2} \\ + \sqrt{(L+\varepsilon)^2 - L^2|\cos^2\theta|} \\ - 2\max(0, L|\sin\theta| - 2\varepsilon).\end{aligned} \tag{A.4}$$

In the application of Lemma A.11 for the first term, we use that $|\beta_1(1) - \beta_1(0)| \leq \max(0, L|\cos\theta| - 2\varepsilon)$.

Otherwise, $\max(a,c) \geq \min(b,d)$, so either $a \leq b \leq c \leq d$ or $c \leq d \leq a \leq b$. In either of these cases, we must have that $L|\sin\theta| \leq \varepsilon$. Consider the first of these cases. We observe

$$\begin{aligned}\int_{\mathbb{R}} |\mathrm{ECT}_\alpha(v,t) - \mathrm{ECT}_\beta(v,t)|\,\mathrm{d}t = &\int_a^b |\mathrm{ECT}_\alpha(v,t) - \mathrm{ECT}_\beta(v,t)|\,\mathrm{d}t \\ &+ \int_b^c |\mathrm{ECT}_\alpha(v,t) - \mathrm{ECT}_\beta(v,t)|\,\mathrm{d}t \\ &+ \int_c^d |\mathrm{ECT}_\alpha(v,t) - \mathrm{ECT}_\beta(v,t)|\,\mathrm{d}t.\end{aligned}$$

This quantity is equal to

$$\int_a^b \mathrm{ECT}_\alpha(v,t)\,\mathrm{d}t + \int_b^c 1\,\mathrm{d}t + \int_c^d \mathrm{ECT}_\beta(v,t) - 1\,\mathrm{d}t,$$

Bounding from above, we have

$$\begin{aligned}\int_{\mathbb{R}} |\mathrm{ECT}_\alpha(v,t) - \mathrm{ECT}_\beta(v,t)|\mathrm{d}t \leq &\int_a^b \mathrm{ECT}_\alpha(v,t)\,\mathrm{d}t + \int_c^d \mathrm{ECT}_\beta(v,t)\,\mathrm{d}t \\ &+ (c-b) - (d-c) \\ \leq &\int_a^b \mathrm{ECT}_\alpha(v,t)\,\mathrm{d}t + \int_c^d \mathrm{ECT}_\beta(v,t)\,\mathrm{d}t + (c-b).\end{aligned}$$

Similar analysis when $c \leq d \leq a \leq b$ shows that in general, if $\max(a,c) \leq \min(b,d)$, then

$$\begin{aligned}\int_{\mathbb{R}} |\mathrm{ECT}_\alpha(v,t) - \mathrm{ECT}_\beta(v,t)|\,\mathrm{d}t \leq &\int_a^b \mathrm{ECT}_\alpha(v,t)\,\mathrm{d}t + \int_c^d \mathrm{ECT}_\beta(v,t)\,\mathrm{d}t \\ &+ \min(|c-b|, |d-a|).\end{aligned}$$

By the triangle inequality, $\min(|c-b|, |d-a|) \leq \varepsilon - L\sin\theta$. Applying Proposition 6.5 and Lemma A.11 once again, we see

$$\begin{aligned}\int_{\mathbb{R}} |\mathrm{ECT}_\alpha(v,t) - \mathrm{ECT}_\beta(v,t)|\,\mathrm{d}t \leq &\sqrt{(L+2\varepsilon)^2 - \max(0, L|\cos\theta| - 2\varepsilon)^2} \\ &+ \sqrt{(L+\varepsilon)^2 - L^2|\cos^2\theta|} \\ &+ \max(0, \varepsilon - L|\sin\theta|).\end{aligned}$$

In summary, $\int_{\mathbb{R}} |\mathrm{ECT}_\alpha(v,t) - \mathrm{ECT}_\beta(v,t)|\,\mathrm{d}t$ is bounded by

$$\sqrt{(L+2\varepsilon)^2 - \max(0, L|\cos\theta| - 2\varepsilon)^2} + \sqrt{(L+\varepsilon)^2 - L^2|\cos^2\theta|} - 2\max(0, L|\sin\theta| - 2\varepsilon)$$

whenever $L|\sin\theta| \geq \varepsilon$. Otherwise, either we still have $\max(a,c) \leq \min(b,d)$ and the above bound still holds or $\max(a,c) \geq \min(b,d)$ and we instead have the bound

$$\sqrt{(L+2\varepsilon)^2 - \max(0, L|\cos\theta| - 2\varepsilon)^2} + \sqrt{(L+\varepsilon)^2 - L^2|\cos^2\theta|} + \max(0, \varepsilon - L|\sin\theta|).$$

Hence, in general,

$$\begin{aligned}\int_{\mathbb{R}} |\mathrm{ECT}_\alpha(v,t) - \mathrm{ECT}_\beta(v,t)|\,\mathrm{d}t = &\sqrt{(L+2\varepsilon)^2 - \max(0, L|\cos\theta| - 2\varepsilon)^2} \\ &+ \sqrt{(L+\varepsilon)^2 - L^2|\cos^2\theta|} \\ &- 2\max(0, L|\sin\theta| - 2\varepsilon) + \max(0, \varepsilon - L|\sin\theta|).\end{aligned}$$

The proof is complete once we have established the following tedious lemma. □

Lemma A.13. *The function*

$$\begin{aligned}f(\theta) = &\sqrt{(L+2\varepsilon)^2 - \max(0, L|\cos\theta| - 2\varepsilon)^2} + \sqrt{(L+\varepsilon)^2 - L^2|\cos^2\theta|} \\ &- 2\max(0, L|\sin\theta| - 2\varepsilon) + \max(0, \varepsilon - L|\sin\theta|)\end{aligned}$$

is bounded above by

$$f(\theta) \leq \begin{cases} 8\sqrt{L\varepsilon} & L > 2\varepsilon \\ 10\varepsilon & L \leq 2\varepsilon. \end{cases}$$

Proof. Thanks to the symmetries of the sine and cosine functions and the absolute values present in the formula for f, we have that $f(-\theta) = f(\theta)$ and $f(\pi/2 - \theta) = f(\theta)$. So $f(\theta) = f(\pi/2 + \theta)$. Therefore it suffices to bound f on the interval $[0, \pi/2]$. On this interval, we can remove the absolute values in the formula for f, giving

$$\begin{aligned}f(\theta) = &\sqrt{(L+2\varepsilon)^2 - \max(0, L\cos\theta - 2\varepsilon)^2} + \sqrt{(L+\varepsilon)^2 - L^2\cos^2\theta} \\ &- 2\max(0, L\sin\theta - 2\varepsilon) + \max(0, \varepsilon - L\sin\theta).\end{aligned}$$

With the exception of finitely many values of θ, the derivative of f exists and is equal to

$$\begin{aligned}&\mathbb{I}(L\cos\theta > 2\varepsilon)\frac{L\sin\theta(L\cos\theta - 2\varepsilon)}{\sqrt{(L+2\varepsilon)^2 - (L\cos\theta - 2\varepsilon)^2}} + \frac{L^2\sin\theta\cos\theta}{\sqrt{(L+\varepsilon)^2 - L^2\cos^2\theta}} \\ &-\mathbb{I}(L\sin\theta > 2\varepsilon)2L\cos(\theta) - \mathbb{I}(L\sin\theta < \varepsilon)L\cos\theta,\end{aligned}$$

where $\mathbb{I}$ denotes the indicator function.

Notice that

$$\frac{L^2\sin\theta\cos\theta}{\sqrt{(L+\varepsilon)^2 - L^2\cos^2\theta}} \leq \frac{L^2\sin\theta\cos\theta}{\sqrt{L^2 - L^2\cos^2\theta}} = L\cos\theta,$$

and similarly

$$\frac{L\sin\theta(L\cos\theta-2\varepsilon)}{\sqrt{(L+2\varepsilon)^2-(L\cos\theta-2\varepsilon)^2}} \leq \frac{L^2\sin\theta\cos\theta}{\sqrt{L^2-L^2\cos^2\theta}} = L\cos\theta.$$

Using these identities, we get that

$$\begin{aligned} f'(\theta) \leq & \mathbb{I}(L\cos\theta > 2\varepsilon)L\cos\theta + L\cos\theta \\ & - \mathbb{I}(L\sin\theta > 2\varepsilon)2L\cos(\theta) - \mathbb{I}(L\sin\theta < \varepsilon)L\cos\theta. \end{aligned}$$

Hence f is weakly decreasing whenever $L\sin\theta > 2\varepsilon$. When $\varepsilon < L\sin\theta < 2\varepsilon$, every non-zero term in f' is positive, and so f is increasing. We now bound f' in absolute value when $L\sin\theta < \varepsilon$. In this case,

$$\begin{aligned} |f'(\theta)| &\leq \mathbb{I}(L\cos\theta > 2\varepsilon)\frac{L\sin\theta(L\cos\theta-2\varepsilon)}{\sqrt{(L+2\varepsilon)^2-(L\cos\theta-2\varepsilon)^2}} + \frac{L^2\sin\theta\cos\theta}{\sqrt{(L+\varepsilon)^2-L^2\cos^2\theta}} + L\cos\theta \\ &\leq 3L\cos\theta \\ &\leq 3L, \end{aligned}$$

using our approximations for the first and second terms from earlier.

Further, if $L > 2\varepsilon$,

$$\begin{aligned} f(0) &= \sqrt{(L+2\varepsilon)^2-\max(0,L-2\varepsilon)^2} + \sqrt{(L+\varepsilon)^2-L^2} - \varepsilon \\ &= \sqrt{8L\varepsilon} + \sqrt{2L\varepsilon+\varepsilon^2} - \varepsilon \\ &\leq (\sqrt{8}+\sqrt{3})\sqrt{L\varepsilon} - \varepsilon. \end{aligned}$$

Otherwise $L \leq 2\varepsilon$ and,

$$\begin{aligned} f(0) &= \sqrt{(L+2\varepsilon)^2-\max(0,L-2\varepsilon)^2} + \sqrt{(L+\varepsilon)^2-L^2} - \varepsilon \\ &= L + 2\varepsilon + \sqrt{2L\varepsilon+\varepsilon^2} - \varepsilon \\ &= L + \varepsilon + \sqrt{2L\varepsilon+\varepsilon^2} \\ &\leq (3+\sqrt{5})\varepsilon. \end{aligned}$$

Hence, we can bound $f(\theta)$ for θ on the interval $[0, \sin^{-1}(\varepsilon/L)]$ (or $[0, \pi/2]$ if $\varepsilon > L$) by using our upper bounds for $f(0)$ and $|f'(\theta)|$ on this interval. By additionally using the inequality $\sin^{-1}(x) \leq \pi x/2$ for positive x, we obtain that when $L \sin\theta \leq \varepsilon$,

$$\begin{aligned}
f(\theta) &\leq \begin{cases} (\sqrt{8}+\sqrt{3})\sqrt{L\varepsilon} + (3\pi/2 - 1)\varepsilon & L > 2\varepsilon \\ (3+\sqrt{5}+3\pi/2)\varepsilon & \varepsilon \leq L \leq 2\varepsilon \\ (3+\sqrt{5})\varepsilon + (3\pi/2)L & L < \varepsilon \end{cases} \\
&\leq \begin{cases} (\sqrt{8}+\sqrt{3})\sqrt{L\varepsilon} + (3\pi/2 - 1)\varepsilon & L > 2\varepsilon \\ (3+\sqrt{5}+3\pi/2)\varepsilon & L \leq 2\varepsilon \end{cases} \\
&\leq \begin{cases} (\sqrt{8}+\sqrt{3}+(3\pi/4 - 1/2)\sqrt{2})\sqrt{L\varepsilon} & L > 2\varepsilon \\ (3+\sqrt{5}+3\pi/2)\varepsilon & L \leq 2\varepsilon \end{cases} \\
&\leq \begin{cases} 8\sqrt{L\varepsilon} & L > 2\varepsilon \\ 10\varepsilon & L \leq 2\varepsilon. \end{cases}
\end{aligned}$$

Otherwise, we know f is weakly increasing until $L \sin\theta > 2\varepsilon$, after which point it is weakly decreasing. Hence, if f is not maximised where $L \sin\theta \leq \varepsilon$, it must attain its maximum when $L \sin\theta = 2\varepsilon$, or equivalently $\theta = \sin^{-1}(2\varepsilon/L)$. Note that this implies $2\varepsilon \leq L$. We compute

$$\begin{aligned}
f(\sin^{-1}(2\varepsilon/L)) &= \sqrt{(L+2\varepsilon)^2 - \max(0, \sqrt{L^2 - 4\varepsilon^2} - 2\varepsilon)^2} + \sqrt{(L+\varepsilon)^2 - L^2 - 4\varepsilon^2} \\
&= \sqrt{(L+2\varepsilon)^2 - \max(0, \sqrt{L^2 - 4\varepsilon^2} - 2\varepsilon)^2} + \sqrt{2L\varepsilon - 3\varepsilon^2} \\
&= \begin{cases} \sqrt{4L\varepsilon + 4\varepsilon^2 + 2\varepsilon\sqrt{L^2 - 4\varepsilon^2}} + \sqrt{2L\varepsilon - 3\varepsilon^2} & L > 2\sqrt{2}\varepsilon \\ L + 2\varepsilon + \sqrt{2L\varepsilon - 3\varepsilon^2} & 2\varepsilon \leq L \leq 2\sqrt{2}\varepsilon \end{cases} \\
&\leq \begin{cases} \sqrt{6L\varepsilon + 4\varepsilon^2} + \sqrt{2L\varepsilon} & L > 2\sqrt{2}\varepsilon \\ (2+2\sqrt{2})\varepsilon + \sqrt{2L\varepsilon} & 2\varepsilon \leq L \leq 2\sqrt{2}\varepsilon \end{cases} \\
&\leq \begin{cases} (\sqrt{6+\sqrt{2}} + \sqrt{2})\sqrt{L\varepsilon} & L > 2\sqrt{2}\varepsilon \\ (2+2\sqrt{2})\sqrt{L\varepsilon} & 2\varepsilon \leq L \leq 2\sqrt{2}\varepsilon \end{cases} \\
&\leq 5\sqrt{L\varepsilon}.
\end{aligned}$$

Hence, to totally bound $f(\theta)$ on the interval $[0, \pi/2]$ we need only use our earlier bound, namely

$$f(\theta) \leq \begin{cases} 8\sqrt{L\varepsilon} & L > 2\varepsilon \\ 10\varepsilon & L \leq 2\varepsilon. \end{cases}$$

□

The goal of the following proposition is to bound from below the chord length of a short segment of a curve given that it has bounded curvature.

Proposition A.14. *Suppose $\gamma : [0, L] \to \mathbb{R}^d$ is a twice differentiable curve parameterised by arc length with curvature κ bounded in norm by M. Let $0 < \varepsilon < \pi/M$. Then for any $t \in [0, L - \varepsilon]$,*

$$\varepsilon \geq \|\gamma(t+\varepsilon) - \gamma(t)\|_2 \geq \frac{2}{M} \sin\left(\frac{M}{2}\varepsilon\right).$$

In particular,

$$\|\gamma(t+\varepsilon) - \gamma(t)\|_2 \geq \varepsilon - \frac{M^2}{24}\varepsilon^3.$$

To prove this we make use of the following theorem of Schwarz, which we cite from [35]:

Theorem A.15 (Schwarz)**.** *Let C be an arc joining two given points A and B with curvature $\kappa(s) \leq 1/R$, such that $R \geq \frac{1}{2}\delta$, where δ is the distance between A and B. Let S be a circle of radius R through A and B. Then the length of C is either less than, or equal to, the shorter arc AB or greater than, or equal to, the longer arc AB on S.*

Proof of Proposition A.14. The first inequality is clear since γ is parameterised by arc length. Now fix t. For the second inequality, consider the optimisation problem of minimising $\|\alpha(t+\varepsilon)-\alpha(t)\|_2$ subject to the constraints that $\|\alpha'\|_2 = 1$ and $\|\alpha''\|_2 \leq M$. Consider an arc of length ε on the circle of curvature M, which has radius $1/M$. Elementary geometry shows that the distance between the endpoints of such an arc is $\frac{2}{M}\sin(M\varepsilon/2)$. We claim that this arc provides an optimal solution. Indeed, let γ be any curve that performs at least as well as this arc in the sense that

$$\|\gamma(t+\varepsilon) - \gamma(t)\|_2 \leq \frac{2}{M} \sin\Big(\frac{M}{2}\varepsilon\Big),$$

while $\|\gamma'\|_2 = 1$, $\|\gamma''\|_2 \leq M$.

Let S be a circle of radius $1/M$ crossing both $\gamma(t)$ and $\gamma(t+\varepsilon)$. Such a circle must exist since $\|\gamma(t+\varepsilon)-\gamma(t)\|_2 \leq 2/M$. The curve γ on the interval $[t, t+\varepsilon]$ is of length $\varepsilon < \pi/M$ while the longer arc on S connecting $\gamma(t)$ and $\gamma(t+\varepsilon)$ has length greater than π/M. Hence, by the theorem of Schwarz, ε is less than or equal to the length of the shorter arc on S from $\gamma(t)$ to $\gamma(t+\varepsilon)$. If this inequality is strict, we may take a shorter portion of the circular arc with length ε, which has a shorter distance between endpoints. This proves that an arc of length ε on a circle of radius $1/M$ is an optimal solution of the optimisation problem.

Thus for potentially suboptimal γ, we have

$$\|\gamma(t+\varepsilon)-\gamma(t)\|_2 \geq \frac{2}{M}\sin\left(\frac{M}{2}\varepsilon\right).$$

For the last statement of the proposition, by the Lagrange remainder theorem

$$\sin x - x + x^3/6 = \int_0^x \cos t \frac{(x-t)^5}{5!}\,\mathrm{d}t.$$

The right side is clearly positive provided that $0 < x \leq \pi/2$. Since $0 < \frac{M}{2}\varepsilon < \pi/2$, we have

$$\frac{2}{M}\sin\left(\frac{M}{2}\varepsilon\right) \geq \frac{2}{M}\left[\frac{M}{2}\varepsilon - \frac{M^3}{48}\varepsilon^3\right] = \varepsilon - \frac{M^2}{24}\varepsilon^3.$$

□

We now use the results we have already proven about curves that are approximately straight to obtain a stability result for the Euler characteristic transform of more general shapes. To do this, we prove a lemma that allows us to glue together Euler characteristic transforms of functions restricted to different regions of a domain.

Definition A.16. Let $V^* = (V, V_0, \{\Phi_\lambda\}_{\lambda\in\Lambda_V})$ and $W^* = (W, W_0, \{\Phi_\lambda\}_{\lambda\in\Lambda_W})$ be finite one-dimensional CW complexes, each with a fixed CW structure. Suppose there exist maps $f_V \in \mathcal{F}^r(V^*, d)$ and $f_W \in \mathcal{F}^r(W^*, d)$, a subset $S \subseteq V_0$ and an injective map $m : S \to W_0$ such that $f_V = f_W \circ m$ on S. We define the glue of V^* and W^* under m to be a finite complex with structure:

$$Z^* = (Z, Z_0, \{\Phi_\lambda\}_{\lambda\in\Lambda_Z}) := \big((V \sqcup W)/m,\ \ (V_0 \sqcup W_0)/m,\ \ \{\Phi_\lambda\}_{\lambda\in\Lambda_V\sqcup\Lambda_W}\big).$$

We define the glue of f_V and f_W under m to be the map $f_Z : Z \to \mathbb{R}^d$ which restricts to f_V on V and f_W of W. This map is well defined (since $f_V = f_W \circ m$ on S) and is an element of $\mathcal{F}^r(Z^*, d)$.

Lemma A.17. *Using the notation of the previous definition, suppose* $\mathrm{ECT}_{f_V}(v,t)$ *and* $\mathrm{ECT}_{f_W}(v,t)$ *are defined for almost all* t *for any fixed* v*. Then*

$$\mathrm{ECT}_{f_Z}(v,t) = \mathrm{ECT}_{f_V}(v,t) + \mathrm{ECT}_{f_W}(v,t) - \mathrm{ECT}_{f_S}(v,t) \tag{A.5}$$

for almost all t *when* v *is fixed, where* f_S *is the restriction of* f_V *to* S*.*

Proof. Fix a unit vector v in $\mathbb{R}^d$. Let $p_1, \ldots, p_k$ be the points of S. We denote by $V(v,t)$ the subset of points x in V satisfying that $\langle v, f_V(x)\rangle \leq t$. We define $W(v,t)$ and $Z(v,t)$ analogously. We let $S(v,t)$ denote the intersection of S and $V(v,t)$.

Via the inclusions of V and W into Z, we can view Z as the union of V and W, with V and W intersecting in Z at S. Similarly, we can view $Z(v,t)$ as the union of $V(v,t)$ and $W(v,t)$, with these two subsets intersecting at $S(v,t)$. For almost all $t \in \mathbb{R}$, $\langle v, f_V(p_i)\rangle \neq t$ for all i. Fix any such t. Hence, we have that the interiors of $V(v,t)$ and $W(v,t)$ cover their intersection $S(v,t)$, by continuity of f_V and f_W. Therefore, we have a Mayer-Vietoris exact sequence of homology groups [70, p. 149]:

$$\ldots \to H_i\big(S(v,t)\big) \to H_i\big(V(v,t)\big) \oplus H_i\big(W(v,t)\big) \to H_i\big(Z(v,t)\big) \to \ldots.$$

A routine argument then deduces the identity

$$\chi\big(Z(v,t)\big) = \chi\big(V(v,t)\big) + \chi\big(W(v,t)\big) - \chi\big(S(v,t)\big),$$

whenever all Euler characteristics on the right-hand side are defined. This happens for almost all t and is another way of writing the identity of Equation (A.5). □

We now have the prerequisites to prove Proposition 6.4.

Proof of Proposition 6.4. Let $\alpha_\lambda := f \circ \Phi_\lambda$ and $\beta_\lambda := g \circ \Phi_\lambda$. Since the index set Λ is finite, we let $\Lambda = \{1, \ldots, k\}$. We define $Z^0 := Z_0$, and inductively, $Z^\lambda = Z^{\lambda-1} \cup \mathrm{im}\ \Phi_\lambda$ for $\lambda \in \Lambda$. We then let $f_\lambda = f|_{Z^\lambda}$ and $g_\lambda = g|_{Z^\lambda}$.

Inductively, we assume that

$$\left\|\mathrm{ECT}_{f_{\lambda-1}} - \mathrm{ECT}_{g_{\lambda-1}}\right\| \leq |Z_0|\varepsilon + \sum_{k=1}^{\lambda-1} G_k(\varepsilon).$$

Indeed, as a base case, it is easily observed that

$$\left\|\mathrm{ECT}_{f_0} - \mathrm{ECT}_{g_0}\right\| \leq |Z_0|\varepsilon.$$

We can split α_λ into n pieces by restricting $\alpha_{\lambda,i} : [\frac{i-1}{n}, \frac{i}{n}] \to \mathbb{R}^d$. Analogously we can split β_λ into curves $\beta_{\lambda,i}$. By Proposition A.14, the arc length of each $\alpha_{\lambda,i}$ is at most $M^2L_\lambda^3/24n^3$ greater than the distance between its endpoints if $n > L_\lambda M/\pi$. We now apply Proposition A.12. Thus, provided

$$\frac{M^2L_\lambda^3}{24n^3} \leq \varepsilon, \text{ or equivalently, } n \geq \left(\frac{M^2L_\lambda^3}{24\varepsilon}\right)^{1/3},$$

we observe

$$\left\|\mathrm{ECT}_{\alpha_{\lambda,i}} - \mathrm{ECT}_{\beta_{\lambda,i}}\right\| \leq \begin{cases} 8\sqrt{L_\lambda\varepsilon/n} & L_\lambda/n > 2\varepsilon \\ 10\varepsilon & L_\lambda/n \leq 2\varepsilon. \end{cases}$$

Let $m_\lambda \in \{1, 2\}$ be the number of 0-cells (i.e. elements of Z_0) in the image of Φ_λ. By repeatedly applying Lemma A.17 we have that

$$\mathrm{ECT}_{\alpha_\lambda}(v,t) = (m_\lambda - 2)\mathrm{ECT}_{\alpha_\lambda(0)}(v,t) + \sum_{i=1}^{n} \mathrm{ECT}_{\alpha_{\lambda,i}}(v,t) - \sum_{i=1}^{n-1} \mathrm{ECT}_{\alpha_{\lambda,i}(i/n)}(v,t),$$

for almost all t when v is fixed. By the same argument, a similar equality holds for $\mathrm{ECT}_{\beta_\lambda}$. Hence, by the triangle inequality, we deduce that $\|\mathrm{ECT}_{\alpha_\lambda} - \mathrm{ECT}_{\beta_\lambda}\|$ is bounded above by

$$\begin{aligned}(m_\lambda - 2)\left\|\mathrm{ECT}_{\alpha_\lambda(0)} - \mathrm{ECT}_{\beta_\lambda(0)}\right\| + \sum_{i=1}^{n} \left\|\mathrm{ECT}_{\alpha_{\lambda,i}} - \mathrm{ECT}_{\beta_{\lambda,i}}\right\| \\ + \sum_{i=1}^{n-1} \left\|\mathrm{ECT}_{\alpha_{\lambda,i}(i/n)} - \mathrm{ECT}_{\beta_{\lambda,i}(i/n)}\right\| \\ \leq \begin{cases} 8\sqrt{L_\lambda n\varepsilon} + (n + m_\lambda - 3)\varepsilon & L_\lambda/n > 2\varepsilon \\ (11n + m_\lambda - 3)\varepsilon & L_\lambda/n \leq 2\varepsilon. \end{cases}\end{aligned}$$

In particular, this bound hold when we let

$$n = n_\lambda := \max\left(\left\lceil\left(\frac{M^2L_\lambda^3}{24\varepsilon}\right)^{1/3}\right\rceil, \left\lceil\frac{L_\lambda M}{\pi}\right\rceil\right).$$

Applying Lemma A.17 again, we have

$$\mathrm{ECT}_{f_\lambda}(v,t) = \mathrm{ECT}_{f_{\lambda-1}}(v,t) + \mathrm{ECT}_{\alpha_\lambda}(v,t) - \mathrm{ECT}_{\alpha_\lambda(0)}(v,t) - (m_\lambda - 1)\mathrm{ECT}_{\alpha_\lambda(1)}(v,t),$$

for almost all t when v is fixed. Similarly, such an equation holds involving g_λ, $g_{\lambda-1}$, and β_λ.

Applying the triangle inequality as before, along with our bound for $\|\mathrm{ECT}_{\alpha_\lambda} - \mathrm{ECT}_{\beta_\lambda}\|$, we deduce

$$\left\|\mathrm{ECT}_{f_\lambda} - \mathrm{ECT}_{g_\lambda}\right\| \leq \left\|\mathrm{ECT}_{f_{\lambda-1}} - \mathrm{ECT}_{g_{\lambda-1}}\right\| + \left\|\mathrm{ECT}_{\alpha_\lambda} - \mathrm{ECT}_{\beta_\lambda}\right\| + (2 - m_\lambda)\varepsilon.$$

The last two terms sum to $G_\lambda(\varepsilon) - \varepsilon$, so in particular we have the bound

$$\left\|\mathrm{ECT}_{f_\lambda} - \mathrm{ECT}_{g_\lambda}\right\| \leq |Z_0|\varepsilon + \sum_{k=1}^{\lambda} G_k(\varepsilon).$$

Induction then proves the proposition. □

From Proposition 6.4, Theorem 6.3 follows easily.

Proof of Theorem 6.3. Fix some $X, Y \in \mathcal{G}^r(Z^*, d)$ and suppose $d_{Z^*}(X,Y) < \varepsilon$. Hence, we may choose $h_X, h_Y \in \mathcal{E}^r(Z^*, d)$ with the properties given in Definition 6.2. Suppose that X has curvature bounded by M under Z^*. It follows that h_X also has curvature bounded by M. Proposition 6.4 gives that

$$\|\mathrm{ECT}_{h_X} - \mathrm{ECT}_{h_Y}\| \leq |Z_0|\varepsilon + \sum_{\lambda\in\Lambda} G_\lambda(\varepsilon),$$

but $\mathrm{ECT}_X = \mathrm{ECT}_{h_X}$ and $\mathrm{ECT}_Y = \mathrm{ECT}_{h_Y}$ since h_X and h_Y are homeomorphisms. The second statement of the theorem follows.

For the first statement, note that every $X \in \mathcal{G}^r(Z^*, d)$ has a bound M on its curvature and that $G_\lambda(\varepsilon) \to 0$ as $\varepsilon \to 0$ for all $\lambda \in \Lambda$. □

Proofs of ECT Stability under Piece-wise Linear Interpolations

In practice, the Theorem 6.8 implies that the ECT of a function on a one-dimensional CW complex can be computed approximately via a dense subset. The proof of this theorem is similar to the proof of Theorem 6.3, but requires two additional lemmas.

Lemma A.18. *Using the notation Definition 6.7, let $b_i := f(a_i)$, and b_{ij} denote the line segment from b_i to b_j, and let c_i denote the number of pairs in E containing i. Then*

$$\mathrm{ECT}_f^A = \sum_{(i,j)\in E} \mathrm{ECT}_{b_{ij}} - \sum_{i=1}^{n} (c_i - 1)\mathrm{ECT}_{b_i}.$$

Proof. Fix some v and t. If $\max(\langle b_i, v\rangle, \langle b_j, v\rangle) \le t$, then $\mathrm{ECT}_{b_{ij}}(v,t) = \chi(b_{ij}) = 1$. If instead $\min(\langle b_i, v\rangle, \langle b_j, v\rangle) > t$, then $\mathrm{ECT}_{b_{ij}}(v,t) = \chi(\emptyset) = 0$. Otherwise, without loss of generality, suppose $\langle b_i, v\rangle \le t$ and $\langle b_j, v\rangle > t$. Again, we have that $\mathrm{ECT}_{b_{ij}(v,t)}$ is equal to the Euler characteristic of a line segment, which is equal to 1.

Define the submultisets

$$\begin{aligned}
E_{\mathrm{up}} &= \{(i,j) \in E : \min(\langle b_i, v\rangle, \langle b_j, v\rangle) > t\},\\
E_{\mathrm{down}} &= \{(i,j) \in E : \max(\langle b_i, v\rangle, \langle b_j, v\rangle) \le t\},\\
E_{\mathrm{mid}} &= \{(i,j) \in E : \max(\langle b_i, v\rangle, \langle b_j, v\rangle) > t,\ \min(\langle b_i, v\rangle, \langle b_j, v\rangle) \le t\}.
\end{aligned}$$

Note $E = E_{\mathrm{up}} \sqcup E_{\mathrm{down}} \sqcup E_{\mathrm{mid}}$. Therefore,

$$\begin{aligned}
&\sum_{(i,j)\in E} \mathrm{ECT}_{b_{ij}}(v,t) - \sum_{i=1}^{n} (c_i - 1)\mathrm{ECT}_{b_i}(v,t)\\
&= \sum_{(i,j)\in E_{\mathrm{up}}} \mathrm{ECT}_{b_{ij}}(v,t) + \sum_{(i,j)\in E_{\mathrm{down}}} \mathrm{ECT}_{b_{ij}}(v,t)\\
&\qquad + \sum_{(i,j)\in E_{\mathrm{mid}}} \mathrm{ECT}_{b_{ij}}(v,t) - \sum_{i=1}^{n} (c_i - 1)\mathrm{ECT}_{b_i}(v,t)\\
&= \sum_{(i,j)\in E_{\mathrm{down}}} 1 + \sum_{(i,j)\in E_{\mathrm{mid}}} 1 - \sum_{i=1}^{n} (c_i - 1)\mathrm{ECT}_{b_i}(v,t)\\
&= \sum_{(i,j)\in E_{\mathrm{down}}} (2-1) + \sum_{(i,j)\in E_{\mathrm{mid}}} 1 - \sum_{i=1}^{n} (c_i - 1)\mathrm{ECT}_{b_i}(v,t)\\
&= \sum_{(i,j)\in E_{\mathrm{down}}} 2 + \sum_{(i,j)\in E_{\mathrm{mid}}} 1 - \sum_{i=1}^{n} (c_i - 1)\mathrm{ECT}_{b_i}(v,t) - \sum_{(i,j)\in E_{\mathrm{down}}} 1\\
&= \sum_{i=1}^{n} c_i \mathrm{ECT}_{b_i}(v,t) - \sum_{i=1}^{n} (c_i - 1)\mathrm{ECT}_{b_i}(v,t) - \sum_{(i,j)\in E_{\mathrm{down}}} 1\\
&= \sum_{i=1}^{n} \mathrm{ECT}_{b_i}(v,t) - \sum_{(i,j)\in E_{\mathrm{down}}} 1 = \mathrm{ECT}_f^A(v,t).
\end{aligned}$$

□

Lemma A.19. *Let $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be any differentiable function with increasing positive derivative satisfying $f(0) = 0$. For positive numbers L and ε consider the set*

$$S(L) = \left\{ (x_1, \ldots, x_k) \in \mathbb{R}^k : 0 \leq x_i \leq \varepsilon,\ \sum_{i=0}^{k} x_i = L \right\}.$$

If $S(L)$ is non-empty, then

$$\sum_{i=1}^{k} f(x_i) \leq L f(\varepsilon)/\varepsilon$$

on $S(L)$ (note that $S(L)$ is always non-empty if $k > L/\varepsilon$).

Proof. Let $a \leq c$ and $b \geq 0$. We have

$$f(b+c) - f(a+b) - \big(f(c) - f(a)\big) = \int_a^c f'(b+t) - f'(t)\, \mathrm{d}t \geq 0$$

and so

$$f(a+b) + f(c) \leq f(a) + f(b+c) \quad \text{when} \quad a \leq c \text{ and } b \geq 0. \tag{A.6}$$

Suppose $S(L)$ is non-empty. Since $S(L)$ is compact, f must attain a maximum on $S(L)$. Pick any such maximiser $x = (x_1, \ldots, x_k) \in S(L)$. By potentially reordering entries, we may assume the x_i are in decreasing order without affecting the value of $\sum_i f(x_i)$. Let j be the smallest index with $x_j \neq \varepsilon$ and l be the largest index with x_l not equal to zero.

If $l > j$, let $m = \min(x_l, \varepsilon - x_j)$. Equation (A.6) shows that if we replace x_j with $x_j + m$ and replace x_l with $x_l - m$, the value of $\sum_i f(x_i)$ does not decrease. Therefore, by applying this replacement procedure several times, we can always find a maximiser of $\sum_i f(x_i)$ on $S(L)$ with $j \geq l$. This condition forces the value of $\sum_i f(x_i)$ to be

$$\lfloor L/\varepsilon \rfloor f(\varepsilon) + f\big(L - \lfloor L/\varepsilon \rfloor \varepsilon\big).$$

If L is divisible by ε, the result is immediate. Otherwise, since f is convex,

$$\begin{aligned} f(L - \lfloor L/\varepsilon \rfloor \varepsilon) &= f\Big(\big(\lceil L/\varepsilon \rceil - L/\varepsilon\big) 0 + \big(L/\varepsilon - \lfloor L/\varepsilon \rfloor\big)\varepsilon \Big) \\ &\leq \big(\lceil L/\varepsilon \rceil - L/\varepsilon\big) f(0) + \big(L/\varepsilon - \lfloor L/\varepsilon \rfloor\big) f(\varepsilon) \\ &= \big(L/\varepsilon - \lfloor L/\varepsilon \rfloor\big) f(\varepsilon). \end{aligned}$$

Thus

$$\lfloor L/\varepsilon \rfloor f(\varepsilon) + f\big(L - \lfloor L/\varepsilon \rfloor \varepsilon\big) \leq \lfloor L/\varepsilon \rfloor f(\varepsilon) + \big(L/\varepsilon - \lfloor L/\varepsilon \rfloor\big) f(\varepsilon) = Lf(\varepsilon)/\varepsilon.$$

Since the value on the left is the maximum of $\sum_i f(x_i)$ on $S(L)$, we are done. □

Proof of Theorem 6.8. For each $e = (i, j) \in E$, we let $b_e = b_{ij}$. Each $e \in E$ corresponds to some open interval in $Z - A$. We can always fix a finer CW structure $Z^\dagger = (Z, A, \{\Phi_e\}_{e \in E})$ of Z, and still have that $f \in \mathcal{F}^r(Z^\dagger, d)$. Let $\gamma_e = f \circ \Phi_e$. Adopting the notation of Lemma A.18, induction on the number of elements in E with Lemma A.17 applied to $Z^\dagger$ gives that for fixed v,

$$\mathrm{ECT}_f(v, t) = \sum_{e \in E} \mathrm{ECT}_{\gamma_e}(v, t) - \sum_{i=1}^{n} (c_i - 1)\mathrm{ECT}_{b_i}(v, t),$$

for almost all t.

Therefore, again fixing v and using Lemma A.18,

$$\begin{aligned}
&\int_{\mathbb{R}} \big|\mathrm{ECT}_f(v, t) - \mathrm{ECT}_f^A(v, t)\big| \, \mathrm{d}t \\
=&\int_{\mathbb{R}} \Big| \sum_{e \in E} \mathrm{ECT}_{\gamma_e}(v, t) - \sum_{i=1}^{n} (c_i - 1)\mathrm{ECT}_{b_i}(v, t) - \\
&\quad \Big[\sum_{e \in E} \mathrm{ECT}_{b_e}(v, t) - \sum_{i=1}^{n} (c_i - 1)\mathrm{ECT}_{b_i}(v, t) \Big] \Big| \, \mathrm{d}t \\
=&\int_{\mathbb{R}} \Big| \sum_{e \in E} \mathrm{ECT}_{\gamma_e}(v, t) - \sum_{e \in E} \mathrm{ECT}_{b_e}(v, t) \Big| \, \mathrm{d}t \\
\leq&\sum_{e \in E} \int_{\mathbb{R}} \Big| \mathrm{ECT}_{\gamma_e}(v, t) - \mathrm{ECT}_{b_e}(v, t) \Big| \, \mathrm{d}t.
\end{aligned} \tag{A.7}$$

Focusing on any particular $e = (i, j) \in E$, let d_1 be the minimum of $\langle \gamma_e(s), v \rangle$ over s, and d_4 be the maximum of the same function over s. Let $d_2 = \min(\langle b_i, v \rangle, \langle b_j, v \rangle))$ and $d_3 = \max(\langle b_i, v \rangle), \langle b_j, v \rangle))$. It follows that $d_1 \leq d_2 \leq d_3 \leq d_4$.

Since subsets of I and b_e always consist of contractible components, ECT_{γ_e} and ECT_{b_e} never have negative values. We have

$$\begin{aligned} t \geq d_1 \quad &\Longrightarrow \quad \mathrm{ECT}_{\gamma_e}(v,t) \geq 1, \\ t \geq d_4 \quad &\Longrightarrow \quad \mathrm{ECT}_{\gamma_e}(v,t) = 1, \\ t < d_1 \quad &\Longrightarrow \quad \mathrm{ECT}_{\gamma_e}(v,t) = 0, \\ t \geq d_2 \quad &\Longrightarrow \quad \mathrm{ECT}_{b_e}(v,t) = 1, \\ t < d_2 \quad &\Longrightarrow \quad \mathrm{ECT}_{b_e}(v,t) = 0. \end{aligned}$$

Combining these observations, we see

$$\begin{aligned} \int_{\mathbb{R}} \left| \mathrm{ECT}_{\gamma_e}(v,t) - \mathrm{ECT}_{b_e}(v,t) \right| \, \mathrm{d}t &= \int_{d_1}^{d_4} \mathrm{ECT}_{\gamma_e}(v,t) - \mathrm{ECT}_{b_e}(v,t) \, \mathrm{d}t \\ &\leq \int_{d_1}^{d_4} \mathrm{ECT}_{\gamma_e}(v,t) \, \mathrm{d}t - (d_3 - d_2). \end{aligned}$$

After applying a rotation, we may assume that $v = (0, 1, 0, \ldots, 0)$. After applying another rotation about v we may assume that b_e is parallel to the plane spanned by the first two coordinates. Let l_e be the arc length of γ_e. By Proposition A.14, the length of b_e is at least $l_e - M^2 l_e^3/24$. Suppose that the line segment b_e meets the hyperplane perpendicular to v at an angle $\theta \in [0, \pi/2]$.

Applying Proposition 6.5 and Lemma A.11 to this scenario, we observe

$$\int_{d_1}^{d_4} \mathrm{ECT}_{\gamma_e}(v,t) \, \mathrm{d}t - (d_3 - d_2) \leq \sqrt{l_e^2 - \left(l_e - \frac{M^2}{24} l_e^3 \right)^2 \cos^2\theta} - \left(l_e - \frac{M^2}{24} l_e^3 \right) \sin\theta.$$

We refer to the right side of this inequality as $f(\theta)$. Let $G = l_e - M^2 l_e^3/24$. G is positive since $l_e < \varepsilon < \pi/M < \sqrt{24}/M$. We have

$$f'(\theta) = \frac{G^2 \sin\theta \cos\theta}{\sqrt{l_e^2 - G^2 \cos^2\theta}} - G\cos\theta.$$

A routine calculation shows that f' is either zero only when $\theta = \pi/2$ or for every θ. Meanwhile $f'(0) = -G$. Since this value is negative, f must be maximised at $\theta = 0$.

Hence,

$$\begin{aligned}\int_{\mathbb{R}} |\mathrm{ECT}_{\gamma_e}(v,t) - \mathrm{ECT}_{b_e}(v,t)| \ \mathrm{d}t &\leq \int_{d_1}^{d_4} \mathrm{ECT}_{\gamma_e}(v,t) \ \mathrm{d}t - (d_3 - d_2) \\ &\leq \sqrt{l_e^2 - \left(l_e - \frac{M^2}{24} l_e^3\right)^2} \\ &= \sqrt{\frac{M^2}{12} l_e^4 - \frac{M^4}{24^2} l_e^6} \\ &\leq \frac{M}{\sqrt{12}} l_e^2.\end{aligned} \tag{A.8}$$

For $\lambda \in \Lambda$, let L_λ denote the arc length of $f \circ \Phi_\lambda$. Now for $\lambda \in \Lambda$, denote by $\Gamma(\lambda)$ the submultiset of $e \in E$ such that im Φ_e is a subset of im Φ_λ. By Equations (A.7) and (A.8), along with Lemma A.19, we get that

$$\begin{aligned}\int_{\mathbb{R}} \left|\mathrm{ECT}_f(v,t) - \mathrm{ECT}_f^A(v,t)\right| \ \mathrm{d}t &\leq \frac{M}{\sqrt{12}} \sum_{e \in E} l_e^2 \\ &= \frac{M}{\sqrt{12}} \sum_{\lambda \in \Lambda} \sum_{e \in \Gamma(\lambda)} l_e^2 \\ &\leq \frac{M}{\sqrt{12}} \sum_{\lambda \in \Lambda} L_\lambda \varepsilon^2 / \varepsilon \\ &= \frac{ML\varepsilon}{\sqrt{12}}.\end{aligned}$$

Since this bound holds for any v, we are done. □

Proof ECT Stability for Random Data

When proving Theorem 6.11, we write k_x and k_y for the partial derivatives in the first and second components, respectively, and K_x and K_y for their corresponding Gram matrices. In particular, for fixed $t \in I$, $\lambda \in \Lambda$, and $\mathbf{a}$, we write $K_x^\lambda(t, \mathbf{a}_n) = [k_x^{\lambda,a_1}(t), ..., k_x^{\lambda,a_n}(t)]$ and K_y^λ for its transpose. A repeated subscript indicates repeated differentiation in that variable.

Let $g : X \to \mathbb{R}$ be a GP with kernel k and a deterministic function $h : X' \to X$. Then $g \circ h$ is a GP with kernel $k'(x,y) := k(h(x), h(y))$ for all $x, y \in X'$. This insight immediately follows from the definition of a GP in Definition 6.1. In particular, for the GP f in the statement of this theorem and any $\lambda \in \Lambda$, the composition $f \circ \Phi_\lambda$ is a GP for any number of observations n.

The derivative of a Gaussian process on I with a differentiable kernel is almost surely differentiable. As differentiation is a linear operator, the derivative of a Gaussian process is again a Gaussian process in such a case [120]. In particular, this derivative GP has kernel k_{xy} and for any $t \in I$ we have the joint distribution

$$\begin{bmatrix} g(t) \\ g'(t) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(t) \\ \mu'(t) \end{bmatrix}, \begin{bmatrix} k(t,t) & k_y(t,t) \\ k_x(t,t) & k_{xy}(t,t) \end{bmatrix} \right). \tag{A.9}$$

In Theorem 6.11, we consider the GP regression of f_λ based on observations at $\mathbf{a}$ for fixed λ. Even if not all elements in the sequence $\mathbf{a}$ need to be in the image of Φ_λ, the GP posterior pre-composed with Φ_λ defines a GP on I. We are interested in the convergence of the derivative of this GP.

For fixed $\lambda \in \Lambda$, we denote the variance of this derivative GP at $t \in I$ by $v'_{n,\lambda}(t)$ (which is not the same as the derivative of $v_{n,\lambda}(t)$ in t). In particular, we have

$$v'_{n,\lambda}(t) = k^\lambda_{xy}(t,t) - K^\lambda_x(t, \mathbf{a}_n)(K(\mathbf{a}_n, \mathbf{a}_n) + \sigma^2 I)^{-1} K^\lambda_y(\mathbf{a}_n, t).$$

Lemma A.20. *Given the Gaussian processes of Theorem 6.11, we get that for each $\lambda \in \Lambda$ the $v'_{n,\lambda}$ satisfy*

$$v'_{n,\lambda}(t) = \mathbb{E}\left[\left| \hat{f}'_{n,\lambda}(t, \mathbf{a}_n, f) - f'_\lambda(t) \right|^2 \right].$$

Furthermore, $v'_{n,\lambda}(t)$ is monotonically decreasing in n for all $t \in I$.

Proof. The first statement follows from Lemma 11 of [85]. In particular,

$$\begin{aligned}
&\mathbb{E}_{f'_\lambda}\left[\left| \hat{f}'_{n,\lambda}(t) - f'_\lambda(t) \right|^2 \right] \\
&= \mathbb{E}_{\boldsymbol{\zeta}_n}\left[\mathbb{E}_{f'_\lambda}\left[\left| \hat{f}'_{n,\lambda}(t) - f'_\lambda(t) \right|^2 \right] \middle| \mathbf{f}(\mathbf{a}_n) + \boldsymbol{\zeta}_n = \mathbf{y}_n \right] \\
&= \mathbb{E}_{\boldsymbol{\zeta}_n}\left[\mathbb{E}_{f'_\lambda}\left[\left| f'_\lambda(t) - \mathbb{E}_{f'_\lambda}[f'_\lambda(t) | \mathbf{f}(\mathbf{a}_n) + \boldsymbol{\zeta}_n] \right|^2 \right] \middle| \mathbf{f}(\mathbf{a}_n) + \boldsymbol{\zeta}_n = \mathbf{y}_n \right] \\
&= \mathbb{E}_{\boldsymbol{\zeta}_n}[\mathrm{Var}(f'_\lambda(t) | \mathbf{f}(\mathbf{a}_n) + \boldsymbol{\zeta}_n)] \\
&= \mathbb{E}_{\boldsymbol{\zeta}_n}[v'_{n,\lambda}(t)] = v'_{n,\lambda}(t).
\end{aligned}$$

For the second statement, we can write

$$v'_{n,\lambda}(t) = k^\lambda_{xy}(t,t) - K^\lambda_x(t, \mathbf{a}_{n+1}) \begin{bmatrix} B_n^{-1} & \mathbf{0}_{n\times 1} \\ \mathbf{0}_{1\times n} & 0 \end{bmatrix} K^\lambda_y(\mathbf{a}_{n+1}, t),$$

where $B_n := (K(\mathbf{a}_n, \mathbf{a}_n) + \sigma^2 I_n)$. Using the bordering method to obtain an expression for B_{n+1}^{-1} in terms of B_n, we get

$$v'_{n,\lambda}(t) - v'_{n+1,\lambda}(t)$$

$$= \nu K_x^\lambda(t, \mathbf{a}_{n+1}) \begin{bmatrix} B_n^{-1}K(\mathbf{a}_n, a_{n+1})K(a_{n+1}, \mathbf{a}_n)B_n^{-1} & -B_n^{-1}K(\mathbf{a}_n, a_{n+1}) \\ -K(a_{n+1}, \mathbf{a}_n)B_n^{-1} & 1 \end{bmatrix} K_y^\lambda(\mathbf{a}_{n+1}, t)$$

$$= \nu((K_x^\lambda(t, \mathbf{a}_n)B_n^{-1}K(\mathbf{a}_n, a_{n+1}))^2 - 2K_x^\lambda(t, a_{n+1})K_x^\lambda(t, \mathbf{a}_n)B_n^{-1}K(\mathbf{a}_n, a_{n+1}) + k_x^\lambda(t, a_{n+1})^2)$$

$$= \nu(K_x^\lambda(t, \mathbf{a}_n)B_n^{-1}K(\mathbf{a}_n, a_{n+1}) - k_x^\lambda(t, a_{n+1}))^2,$$

where $\nu := 1/(k(a_{n+1}, a_{n+1}) + \sigma - K(a_{n+1}, \mathbf{a}_n)B_n^{-1}K(\mathbf{a}_n, a_{n+1}))$ is the reciprocal of the Schur complement of B_n inside B_{n+1}. As ν^{-1} is the Schur complement of a positive-definite matrix inside a positive-definite matrix, it is positive. As the second factor in the final line above is a square and thus positive too, we conclude that the sequence of functions $v'_{n,\lambda}(t)$ is monotonically decreasing. □

Proof of Theorem 6.11. The first statement follows from Theorem 8 in [85].

To prove the remainder of the theorem, we recall from Equation (A.9) that the covariance matrix of the distribution of $(f_\lambda(t), f'_\lambda(t))^T$ given n noisy observations of f is

$$\begin{bmatrix} k(t,t) - K(t, \mathbf{a}_n)B_n^{-1}K(\mathbf{a}_n, t) & k_y^\lambda(t,t) - K(t, \mathbf{a}_n)B_n^{-1}K_y^\lambda(\mathbf{a}_n, t) \\ k_x^\lambda(t,t) - K_x^\lambda(t, \mathbf{a}_n)B_n^{-1}K(\mathbf{a}_n, t) & k_{xy}^\lambda(t,t) - K_x^\lambda(t, \mathbf{a}_n)B_n^{-1}K_y^\lambda(\mathbf{a}_n, t) \end{bmatrix}.$$

As this matrix needs to be positive-definite, by taking the determinant and using the symmetry of k we get

$$(k(t,t) - K(t, \mathbf{a}_n)B_n^{-1}K(\mathbf{a}_n, t))(k_{xy}^\lambda(t,t) - K_x^\lambda(t, \mathbf{a}_n)B_n^{-1}K_y^\lambda(\mathbf{a}_n, t)) \tag{A.10}$$

$$\geq (k_x^\lambda(t,t) - K_x^\lambda(t, \mathbf{a}_n)B_n^{-1}K(\mathbf{a}_n, t))^2 \geq 0. \tag{A.11}$$

Thus, $k_x^\lambda(t,t) - K_x^\lambda(t, \mathbf{a}_n)B_n^{-1}K(\mathbf{a}_n, t) \to 0$ uniformly on I as the first factor of (A.10) converges uniformly by Proposition 10 of [85] and the second factor of (A.10) is bounded by the monotonicity established in Lemma A.20 and the compactness

of I. Repeating the same procedure with $\hat{f}''_{n,\lambda}$ in place of $\hat{f}'_{n,\lambda}$ gives $k^{\lambda}_{xx}(t,t) - K^{\lambda}_{xx}(t,\mathbf{a}_n)B_n^{-1}K(\mathbf{a}_n,t) \to 0$ uniformly: in this case, the second factor is $k^{\lambda}_{xxyy}(t,t) - K^{\lambda}_{xx}(t,\mathbf{a}_n)B_n^{-1}K^{\lambda}_{yy}(\mathbf{a}_n,t)$, which equals $v''_{n,\lambda}(t)$, the variance of the second derivative of the GP f_λ. We can show that $v''_{n,\lambda}(t)$ monotonically decreases by a proof analogous to the case $v'_{n,\lambda}(t)$ given in Lemma A.20. For $v''_{n,\lambda}(t)$ to be well-defined we require k to be four times differentiable.

Then, by Jensen's inequality and Lemma A.20, we can bound the expected value of the squared difference $V\left(\hat{f}_{n,\lambda}\right) - V(f_\lambda)$:

$$\mathbb{E}\left[\left|\int_0^1 |f'_\lambda(t)| - \left|\hat{f}'_{n,\lambda}(t,\mathbf{a}_n,f)\right| \,\mathrm{d}t\right|^2\right] \leq \mathbb{E}\left[\left(\int_0^1 \left||f'_\lambda(t)| - \left|\hat{f}'_{n,\lambda}(t,\mathbf{a}_n,f)\right|\right| \,\mathrm{d}t\right)^2\right]$$

$$\leq \mathbb{E}\left[\int_0^1 \left||f'_\lambda(t)| - \left|\hat{f}'_{n,\lambda}(t,\mathbf{a}_n,f)\right|\right|^2 \,\mathrm{d}t\right] = \int_0^1 v'_{n,\lambda}(t)\,\mathrm{d}t$$

$$= \left[k^{\lambda}_x(t,t) - K^{\lambda}_x(t,\mathbf{a}_n)B_n^{-1}K(\mathbf{a}_n,t) - \int_0^t k^{\lambda}_{xx}(s,s) - K^{\lambda}_{xx}(s,\mathbf{a}_n)B_n^{-1}K(\mathbf{a}_n,s)\,\mathrm{d}s\right]_0^1.$$

The final equation above converges to 0 as $n \to \infty$, as both the left-hand term and the function under in the integral of the right-hand term in the above difference converge uniformly to 0 by Equation (A.11) and its analogue for $v''_{n,\lambda}(t)$. □

Lemma A.21. *Let f and $\hat{f}_n$ be as in the statement of Theorem 6.12. Denote the arc-lengths of $\hat{f}_{\lambda,n} := \hat{f}_n \circ \Phi_\lambda$ and $f_\lambda := f \circ \Phi_\lambda$ by $L_{n,\lambda}$ and L_λ respectively for each $\lambda \in \Lambda$. Then $L_{n,\lambda} \to L_\lambda$ and*

$$\int_0^1 \left|\left\|f'_{n,\lambda}(t)\right\|_2 - \|f'_\lambda(t)\|_2\right| \,\mathrm{d}t \to 0$$

in probability.

Proof. First, note that $\sqrt{x+y} \leq \sqrt{x}+\sqrt{y}$ and $|\sqrt{x}-\sqrt{y}| \leq \sqrt{|x-y|}$ for all $x, y \geq 0$. We have

$$\begin{aligned}
\left| \int_0^1 \left\| f'_{n,\lambda}(t) \right\|_2 - \| f'_\lambda(t) \|_2 \, \mathrm{d}t \right| &\leq \int_0^1 \left| \left\| f'_{n,\lambda}(t) \right\|_2 - \| f'_\lambda(t) \|_2 \right| \, \mathrm{d}t \\
&\leq \int_0^1 \sqrt{\left| \left\| f'_{n,\lambda}(t) \right\|_2^2 - \| f'_\lambda(t) \|_2^2 \right|} \, \mathrm{d}t \\
&\leq \sum_{j=1}^d \int_0^1 \sqrt{\left| |(f^j_{n,\lambda})'(t)|^2 - |(f^j_\lambda)'(t)|^2 \right|} \, \mathrm{d}t \\
&\leq \sum_{j=1}^d \left[\int_0^1 |(f^j_{n,\lambda})'(t)| + |(f^j_\lambda)'(t)| \mathrm{d}t \right] \times \\
&\qquad \left[\int_0^1 \left| |(f^j_{n,\lambda})'(t)| - |(f^j_\lambda)'(t)| \right| \mathrm{d}t \right].
\end{aligned}$$

The first inequality follows from the triangle inequality for integrals. The second and third inequalities follow from the inequalities for square roots introduced at the start of the proof. The final inequality is the Cauchy-Schwarz inequality for integrals.

The first factor in the final line converges to $2V(f_\lambda)$ and the second factor converges to 0 for each $j = 1, ..., d$ by Theorem 6.11 in probability. □

Proof of Theorem 6.12. We recall that convergence in mean implies convergence in probability. By applying Theorem 6.11 to each component of f, we find that $\hat{f}_n$ converges to f in mean in the ∞-norm. Note that $\hat{f}_{n,\lambda}$ need not be parameterised to constant velocity. Denote the arc length of $\hat{f}_{n,\lambda}$ by $L_{n,\lambda}$ and the arc length of f_λ by L_λ. By Lemma A.21, $L_{n,\lambda} \to L_\lambda$ in probability for each $\lambda \in \Lambda$. Let $s_{n,\lambda}$ be the re-parameterised of $\hat{f}_{n,\lambda}$ to constant-velocity on I, which is given by

$$s_{n,\lambda}(t) = \frac{1}{L_{n,\lambda}} \int_0^t \left\| \hat{f}'_{n,\lambda}(x) \right\|_2 \, \mathrm{d}x$$

and satisfies $\left\| \left(\hat{f}_{n,\lambda} \circ s^{-1} \right)'(x) \right\|_2 = 1$ on I. Thus,

$$|s_{n,\lambda}(t) - t| = \left| \int_0^t \left\| \hat{f}'_{n,\lambda}(x) \right\|_2 / L_{n,\lambda} - 1 \, \mathrm{d}x \right| \leq \left| \int_0^t \left| \left\| \hat{f}'_{n,\lambda}(x) \right\|_2 / L_{n,\lambda} - \| f'_\lambda(x) \|_2 / L_\lambda \right| \mathrm{d}x \right|$$

$$\leq \frac{1}{L_{n,\lambda}} \int_0^1 \left| \left\| f'_{n,\lambda}(x) \right\|_2 - \| f'_\lambda(x) \|_2 \right| \, \mathrm{d}x + \| f'_\lambda(x) \|_2 \left| \frac{1}{L_{\lambda,n}} - \frac{1}{L_\lambda} \right|. \tag{A.12}$$

As both terms in Equation (A.12) converge to 0 in probability independently of t by Lemma A.21, we get $\|s_{n,\lambda}(t) - t\|_\infty \xrightarrow{p.} 0$.

We then define $s_n : Z \to Z$ as

$$s_n(z) = \begin{cases} z & \text{if } z \in Z_0, \\ \left(\Phi_\lambda \circ s_{n,\lambda}^{-1} \circ \Phi_\lambda^{-1}\right)(z) & \text{if } z \in \Phi_\lambda((0,1)). \end{cases}$$

The map s_n is continuous as each $s_{n,\lambda}^{-1}$ is continuous, $s_{n,\lambda}^{-1}(0) = 0$ and $s_{n,\lambda}^{-1}(1) = 1$.

The result of the theorem then follows from Proposition 6.4:

$$\left\|\mathrm{ECT}_{\hat{f}_n} - \mathrm{ECT}_f\right\| \leq \left\|\mathrm{ECT}_{\hat{f}_n} - \mathrm{ECT}_{\hat{f}_n \circ s_n}\right\| + \left\|\mathrm{ECT}_{\hat{f}_n \circ s_n} - \mathrm{ECT}_f\right\|. \tag{A.13}$$

Note that the first term is 0 as re-parameterisation does not change the image of a function. For the second term, we find that $\hat{f}_n \circ s_n$ converges to satisfy the conditions such that Proposition 6.4 yields increasingly tight bounds: the arc lengths of $\hat{f}_n \circ s_n \circ \Phi_\lambda = \hat{f}_{n,\lambda} \circ s_{n,\lambda}^{-1}$ converge to those of $f \circ \Phi_\lambda = f_\lambda$ by Lemma A.21 (the composition of f_λ with $s_{n,\lambda}^{-1}$ does not change its arc length). Further, both aforementioned functions have constant velocity and

$$\left\|\hat{f}_n \circ s_n - f\right\|_\infty \leq \left\|\hat{f}_n \circ s_n - \hat{f}_n\right\|_\infty + \left\|\hat{f}_n - f\right\|_\infty \xrightarrow{p.} 0.$$

In the above, the second term converges in probability by Theorem 6.11. The first term converges in probability as

$$\left\|\hat{f}_{n,\lambda} \circ s_{n,\lambda}^{-1} - \hat{f}_{n,\lambda}\right\|_\infty \leq \left\|\hat{f}_{n,\lambda} \circ s_{n,\lambda}^{-1} - f_\lambda \circ s_{n,\lambda}^{-1}\right\|_\infty + \left\|f_\lambda \circ s_{n,\lambda}^{-1} - f_\lambda\right\|_\infty + \left\|f_\lambda - \hat{f}_{n,\lambda}\right\|_\infty \xrightarrow{p.} 0$$

on each 1-cell $\lambda \in \Lambda$. The first term converges in probability by Theorem 6.11 (as re-parameterisation does not change the ∞-norm). Note that f_λ is continuous on I, which is compact, and therefore uniformly continuous. The second term equals $\left\|f_\lambda \circ s_{n,\lambda} - f_\lambda\right\|_\infty$ by pre-composition with $s_{n,\lambda}(t)$ and thus converges by Equation (A.12) and the uniform continuity of f_λ. The last term converges in probability by Theorem 6.11. □

Proof of Lemma 6.13. Fix $v \in S^{d-1}$. Then

$$\|\mathrm{SECT}_X(v, \cdot) - \mathrm{SECT}_Y(v, \cdot)\|_1$$

$$= \int_{-a}^{a} \left| \int_{-a}^{t} \mathrm{ECT}_X(v,x) - \mathrm{ECT}_Y(v,x)\,\mathrm{d}x - \frac{t+a}{2a} \int_{-a}^{a} \mathrm{ECT}_X(v,x) - \mathrm{ECT}_Y(v,x)\,\mathrm{d}x \right| \mathrm{d}t$$

$$\leq \int_{-a}^{a} \int_{-a}^{t} |\mathrm{ECT}_X(v,x) - \mathrm{ECT}_Y(v,x)|\,\mathrm{d}x + \frac{t+a}{2a} \int_{-a}^{a} |\mathrm{ECT}_X(v,x) - \mathrm{ECT}_Y(v,x)|\,\mathrm{d}x\,\mathrm{d}t$$

$$\leq 2a\delta + \delta = (2a+1)\delta.$$

Since the above is independent of v, we are done. □

Characterisation of the sine-squared exponential kernel

Lemma A.22. *For the sine-squared kernel, we have $J(S^1, d_k) < \infty$.*

Proof. For the sine squared kernel k, the metric d_k is given by

$$d_k(s,t) = \sqrt{2 - 2\exp(-2\sin^2((s-t)/2)}.$$

It can be shown that d_k is strongly equivalent to the angular metric d: let

$$f(x) = \sqrt{2 - 2\exp(-2\sin^2(x))}.$$

Then

$$f'(x) = \frac{4e^{-2\sin^2(x)}\cos(x)\sin(x)}{\sqrt{2 - 2\exp\left(-2\sin^2(x)\right)}}.$$

In particular, $f'(x) \geq 0$ on $0 \leq x \leq \pi/2$ (we can show that $\lim_{x\to 0^+} f'(x) = 2$ by L'Hopital's rule). Further, f is concave as it is the composition of non-decreasing concave functions. Thus, $d_k(s,t) = f(d(s,t)/2)$, we get $d \geq d_k \geq (2\sqrt{2-2e^{-2}}/\pi)d$ on S^1, where the first factor is $f'(0)/2$ and the second factor is the difference quotient of f between 0 and $\pi/2$.

As S^1 is bounded and d and d_k are strongly equivalent, it is thus sufficient to show that $J(S^1, d) < \infty$. For d and $\varepsilon > 0$, we get

$$N(S^1, d, \varepsilon) = \left\lceil \frac{\pi}{\varepsilon} \right\rceil \leq \frac{\pi}{\varepsilon} + 1.$$

Hence,

$$\begin{aligned} J(S^1, d) &= \int_0^\infty \sqrt{\log N(S^1, d, \varepsilon)}\, \mathrm{d}\varepsilon \\ &\leq \int_0^\pi \sqrt{\log\left(\frac{\pi}{\varepsilon} + 1\right)}\, \mathrm{d}\varepsilon \\ &= \pi \int_1^\infty \frac{\sqrt{\log(x+1)}}{x^2}\, \mathrm{d}x \\ &\leq \pi \int_1^\infty \frac{1}{x^{\frac{3}{2}}}\, \mathrm{d}x = \pi \left[-2x^{-\frac{1}{2}}\right]_1^\infty = 2\pi < \infty. \end{aligned}$$

□

Lemma A.23. *Define the Hilbert space $\mathcal{H}'$ of sequences $w_{ab} \in \mathbb{R}$, $a, b \in \mathbb{N}_0$, satisfying*

$$\sum_{n=0}^{\infty} n! \sum_{\substack{a \geq 0, b \geq 0 \\ a+b=n}} \frac{w_{ab}^2}{C_a^n} < \infty,$$

where C_a^n denotes n choose a. For $\{w_{ab}\}, \{v_{ab}\} \in \mathcal{H}'$, the inner product of $\mathcal{H}'$ is given by

$$\langle \{w_{ab}\}, \{v_{ab}\} \rangle_{\mathcal{H}'} := \gamma \sum_{n=0}^{\infty} n! \sum_{\substack{a \geq 0, b \geq 0 \\ a+b=n}} \frac{w_{ab} v_{ab}}{C_a^n}$$

where $\gamma > 0$ is a constant. Define V to be the closed subspace of sequences $\{w_{ab}\} \in \mathcal{H}'$ such that

$$\sum_{(a,b) \in \mathbb{N}_0^2} w_{ab} \cos^a(t) \sin^b(t) = 0 \tag{A.14}$$

for all $t \in [0, 2\pi)$. Then the Hilbert space $\mathcal{H}$ given by the functions

$$f(t) = \sum_{(a,b) \in \mathbb{N}^2} w_{ab} \cos^a(t) \sin^b(t) \tag{A.15}$$

with $\{w_{ab}\} \in V^\perp$ and inner product induced from $\mathcal{H}'$ is isometrically isomorphic to the RKHS of the sine-squared-exponential kernel, denoted by $\mathcal{H}_k$.

Proof. By using standard trigonometric identities, we see that the sine-squared kernel is proportional (by a positive constant) to the kernel

$$k(s, t) = \exp\left(\cos(s)\cos(t) + \sin(s)\sin(t)\right).$$

Thus, by using the Taylor expansion of exp, $k(\,\cdot\,,t) \in \mathcal{H}$ for all t with coefficients

$$w_{ab} = \frac{C_a^{a+b}}{(a+b)!} \cos^a(t) \sin^b(t).$$

Moreover, for $f \in \mathcal{H}$ with coefficients v_{ab} and fixed t, we get

$$\begin{aligned} \langle k(\,\cdot\,,t), f\rangle_{\mathcal{H}} &= \sum_{n=0}^{\infty} n! \sum_{\substack{a\geq 0, b\geq 0 \\ a+b=n}} \frac{C_a^n \cos^a(t) \sin^b(t) v_{ab}}{n! C_a^n} \\ &= \sum_{(a,b)\in\mathbb{N}^2} \cos^a(t) \sin^b(t) v_{ab} = f(t). \end{aligned} \tag{A.16}$$

Thus, the inner product $\langle\,\cdot\,,\,\cdot\,\rangle_{\mathcal{H}}$ has the reproducing property and coincides with the inner product induced by the kernel k (i.e. the inner product of $\mathcal{H}_k$) given in Equation (2.5). Further, the coefficients of $k(\,\cdot\,,t)$ lie in $V^\perp$: let $\{v_{ab}\} \in V$ and let $\{w_{ab}\}$ be the coefficients of $k(\,\cdot\,,t)$. Then by Equation (A.16), $\langle\{v_{ab}\},\{w_{ab}\}\rangle_{\mathcal{H}'} = 0$. As $V^\perp$ is closed as it is perpendicular to V, so is $\mathcal{H}$, implying that $\mathcal{H}$ is a Hilbert space. We have that $\mathcal{H}_k \subseteq \mathcal{H}$. As $\mathcal{H}_k$ is complete by definition, we get $\mathcal{H} = \mathcal{H}_k \oplus W$ for some closed subspace W. Let $f \in W$. Then $\langle g, f\rangle_{\mathcal{H}} = 0$ for all $g \in \mathcal{H}_k$ and in particular $f(t) = \langle k(\,\cdot\,,t), f\rangle_{\mathcal{H}} = 0$ for all $t \in [0, 2\pi)$. Thus, $W = 0$ and $\mathcal{H} \cong \mathcal{H}_k$. □

Lemma A.24. *Every $f \in \mathcal{H}$ is continuous and the inclusion $\mathcal{H} \hookrightarrow C(S^1, d_\infty)$ is continuous, where $C(S^1, d_\infty)$ is the space of continuous real-valued functions on S^1 endowed with the ∞-norm. Further, $\cos(nt)$ and $\sin(nt)$ are elements of $\mathcal{H}$ for all $n \in \mathbb{N}$.*

Proof. Note that $\|k(\,\cdot\,,t)\|_{\mathcal{H}} = 1$ for all t. Thus, by the reproducing property of k and the Cauchy-Schwarz inequality, for all $f \in \mathcal{H}$ and any $t \in S^1$ we get

$$|f(t)| = |\langle k(\,\cdot\,,t), f\rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}}.$$

Hence, convergence in the $\mathcal{H}$-norm implies convergence in the ∞-norm. As f can be written as a series of continuous functions converging in the $\mathcal{H}$-norm (c.f. Equation (A.15)), it follows that f is continuous. As for any $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leq \varepsilon$ and any $t \in S^1$ we have $|f(t)| \leq \varepsilon$, we get that the inclusion $\mathcal{H} \hookrightarrow C(S^1, d_\infty)$ is continuous.

Further, we can expand

$$\cos(nt) = \sum_{k \text{ even}}^{n} (-1)^{\frac{k}{2}} \binom{n}{k} \cos^{n-k}(t) \sin^{k}(t),$$

$$\sin(nt) = \sum_{k \text{ odd}}^{n} (-1)^{\frac{k-1}{2}} \binom{n}{k} \cos^{n-k}(t) \sin^{k}(t).$$

Thus, $\cos(nt)$ and $\sin(nt)$ can be expanded as powers of cos and sin with coefficients in $\{w_{ab}\} \in \mathcal{H}'$ (as in Lemma A.23). We can project these coefficients into $V^{\perp}$ without changing the value of our series at any $t \in S^1$: the difference in the series we observe by subtracting from elements of V from $\{w_{ab}\}$ is 0 for all t (c.f. Equation (A.14)). Thus, $\cos(nt), \sin(nt) \in \mathcal{H}$ for all $n \in \mathbb{N}$. $\square$