Topological data analysis for high-dimensional data in biology

Katherine Benjamin

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2026

Acknowledgements

This thesis could never have been written without the steadfast mentorship of my supervisors. Heather and Ulrike, your enthusiasm, patience, insight, and candour have been an inspiration, and it has been an enormous honour to get to do research with you both.

I would also like to thank my undergraduate mentors, especially Marc Lackenby, Gavin Lowe, Tom Sanders, and Zubin Siganporia, whose dedicated support set me up in the best possible way for my future studies. A special thanks to Agnese Barbensi, who taught me not only how to do research, but also that mathematicians can be absolutely hilarious. I'm also indebted to my secondary school maths teachers. In particular Jon Priday, who snuck me out of all manner of boring lessons on Friday afternoons to teach me the much more exciting topic of differential equations; and Mike Taylor, who diligently supported my (largely unsuccessful) attempts at Olympiad maths, and who taught me how to do the many t-tests that now populate this thesis.

Thank you to all my collaborators, especially Katherine and Aneesha, who admirably dealt with my endless questions as I first jumped into the world of transcriptomics. Thank you also to the incredible Oxford applied topology community, who have been the best friends as well as colleagues.

There are three incredible women without whom this thesis would have been impossible. Jenny, I miss you dearly. Thank you for being by my side from the beginning, and for teaching me to grasp every opportunity with both hands. Anna, stop me if you've heard this one before. Thank you for always being up for an adventure, and for doing my hair before every single party. Claire, you are a ray of pure sunshine. Thank you for looking after me when I couldn't look after myself, and for not telling anyone about the duck.

Finally, I owe everything to my family. I love you all more than I can say.

Abstract

Biological and medical scientists are increasingly producing rich data sets with some high-dimensional component. In this thesis we show three new approaches to study such data.

First, in the realm of multiparameter persistent homology, we provide a new algorithm for computing the rank invariant and persistence landscapes from certain minimal presentations of 3-parameter persistence modules. In particular, this algorithm is designed to work on Gröbner bases as output by an algorithm of Bender, Gäfvert, and Lesnick. We use these advances to compute multiparameter persistence landscapes of spatiotemporal trifiltrations of dynamic metric spaces as introduced by Kim to analyse a data set arising from swarm dynamics. This work marks, to our knowledge, the first application of 3-parameter persistent homology to a problem in data science.

We then turn our attention to the problem of cell-type classification in next-generation subcellular spatial transcriptomics data. We introduce a new algorithm, called Topological Automatic Cell Types (TopACT), which uses multiscale information pooled from local neighbourhoods to produce cell-type annotations with unprecedented spatial detail. We demonstrate TopACT on both synthetic and real-world data sets, where the method shows significantly increased accuracy and, in combination with 2-parameter persistent homology, provides new insights into immune cell organisation in the mouse kidney.

Finally, we show how a recent measure in ecological diversity can be repurposed to study tissue heterogeneity in single-cell transcriptomics data. We will show how this measure provides a much-needed cell-type-agnostic tool for transcriptomics analysis, and demonstrate its utility on two single-cell data sets from embryonic development studies and a spatial transcriptomics data set of the mouse hippocampus.

Contents

List of Figures

List of Abbreviations

1PH	single-parameter persistent homology
3PH	three-parameter persistent homology
DGE	differential gene expression
DMS	dynamic metric space
DoG	difference of Gaussians
FOV	field of view
GBS	Gröbner Bases and Syzygies
H&E	haematoxylin and eosin
HVG	highly variable gene
MPH	multiparameter persistent homology
MSE	mean squared error
PCA	principal component analysis
PCR	polymerase chain reaction
PLS	partial least squares
PVM	perivascular macrophage
RCTD	Robust Cell Type Decomposition
RT	reverse transcription
sc/snRNA-seq	single-cell/single-nucleus RNA sequencing
scRNA-seq	single-cell RNA sequencing
SEM	stem-cell-based embryo model
smFISH	single-molecule fluorescent *in situ* hybridisation
snRNA-seq	single-nucleus RNA sequencing
ST	spatial transcriptomics
SVM	support vector machine
TDA	topological data analysis
TLR7	toll-like receptor 7
TopACT	Topological Automatic Cell Types

Chapter 1

Introduction

The Human Genome Project was declared complete with the publication of a single human reference genome in 2003 [IHGSC04], after a multi-year effort with an estimated sequencing cost of $100 million [Wet23]. Two decades later, sequencing a human genome takes less than a day with costs in the hundreds of dollars. This reduced cost has come with vastly increased scale, and the UK Biobank recently published a data set containing nearly half a million human genome sequences [UKB25]. The same story can be told across the medical and biological sciences, and access to large, complex data sets has exploded. Yet a crucial question remains: how should we analyse all this information?

The burgeoning field of topological data analysis (TDA), which borrows machinery from algebraic topology to construct practical tools for data analysis, offers exciting new perspectives on this challenge. Chief among its contributions has been single-parameter persistent homology (1PH) [ZC05; Ott+17], in which multiscale homological features in a data set are represented by a discrete summary known as the barcode. Single-parameter persistent homology has been successfully applied to a broad range of topics in the life sciences, including neuroscience [GGB16; Kan+18; Sto+21; Gar+22; Bee+23; Goo+24], protein structure [Gam+14; Kov+16; BW20; Ben+23; Mad+25b], and tumour pathology [Law+19; Sto+22; Chu+23; Sto+24; Yan+25].

In recent years, multiparameter persistent homology (MPH) [CZ09; BL23] has seen increased attention as a natural extension to 1PH, driven in equal parts by theoretical advances and the development of new software implementations. A key challenge here is that MPH does not enjoy a discrete invariant analogous to the barcode in 1PH, which has frustrated efforts to apply the theory in practice. Nevertheless, many alternatives have been proposed, and software implementations

for the case of 2-parameter persistent homology are now available [RIV20; LS24] and have been applied to real-world problems [CB20; Vip+21].

Extending persistent homology to the spatiotemporal setting is a hard problem. A number of approaches based on 1PH have been proposed and implemented, including zigzag persistence [Cd10] and persistence vineyards [CEM06]. Since these extensions only include a single persistence parameter, it is always necessary to set this as *either* the spatial or the temporal dimension, naturally forgoing functoriality in the other. To capture both spatial and temporal topology in a single summary, Kim has proposed the use of three parameters to simultaneously encode multiple time intervals and spatial scales [Kim20]. However, the development of downstream data analysis in this context has been obstructed by a lack of efficient algorithms to compute persistent homology in more than two parameters.

Upcoming work of Bender *et al.* [BGLa] provides a solution to this problem by efficiently computing minimal presentations of persistence modules in three or more parameters. Our first contribution in this thesis will be to extend this work by introducing an efficient algorithm to compute both the rank invariant and multiparameter persistence landscapes given the data of a minimal presentation with a certain structure. We will then apply this result to compute representations of spatiotemporal persistence modules as introduced by Kim, thus providing the first application of three-parameter persistent homology (3PH) to a problem in data science.

Transcriptomics—the measurement of the total RNA content of a tissue sample—has been at the forefront of the data revolution in biology. Next-generation single-cell sequencing methods can now produce detailed readouts of gene expression counts across tens of thousands of genes and hundreds of thousands of cells in a sample, and spatial transcriptomics (ST) methods resolve the locations of millions of individual transcripts *in situ*. As the experimental technology races ahead, the mathematical and computational tools we use to analyse and understand these data need to keep pace [Läh+20].

One of the fundamental tasks in transcriptomics analysis is the identification and classification of cell types, which roughly speaking are 'clusters' of cells in gene expression space that correspond to the same biological function or phenotype. In the context of single-cell transcriptomics this is a well-studied problem, and standard unsupervised and supervised machine learning techniques can be readily applied [Pas+21]. Cell-type identification in ST poses a greater challenge, given that the individual biological units do not correspond to single cells but rather

spatial loci. When the spatial units are greater than the size of a cell, one needs to decompose these multicellular regions to recover single-cell information, and a wide range of methods have been proposed for this task [Gas+25]. In recent years the reverse problem has emerged: gene expression information can now be resolved at a subcellular resolution, and it is therefore necessary to combine many subcellular readings to reach the level of single cells, often without *a priori* knowledge of cell boundaries. Existing approaches to this task rely on inflexible binning or cell segmentation based on complementary image data, and both of these methods tend to underdetect rare, sparsely distributed cells which are of particular clinical importance [Che+22, p. 1789]. Cell-type identification in the subcellular regime therefore remains a crucial task, and extracting single-cell level information from *in situ* data was identified as a key challenge when ST was named 'Method of the Year' by *Nature Methods* in 2021 [Mar21].

The second contribution of this thesis is the introduction of a new method, Topological Automatic Cell Types (TopACT), for supervised cell-type classification in subcellular ST [Ben+24]. In contrast to traditional methods, TopACT works directly at the subcellular level without requiring a separate segmentation step, and therefore produces finer-resolution cell-type maps with a significantly increased detection rate of 'needle-in-haystack' cells such as immune cells. We benchmark TopACT on a bespoke synthetic data set, and then demonstrate its utility on real-world Stereo-seq [Che+22] data sets on the mouse brain and mouse kidney, as well as a 10x Xenium [Jan+23] human kidney data set. In mouse kidney we also incorporate an MPH spatial analysis with TopACT predictions to generate a novel hypothesis on immune organisation in murine lupus nephritis, providing another new real-world application of 2-parameter persistent homology.

On a more general level, the conceptual basis for cell-type classification in single-cell transcriptomics is notoriously fragile [Zen22; CS17]. Even in mature organisms, cells do not settle within clearly defined boundaries of gene expression space, but rather lie on continuous trajectories. Furthermore, cell types exhibit subclusters, and it is not always clear how to choose the correct level of resolution. Beyond these philosophical issues there are also strong technical barriers to reproducibility. On an experimental level, platform effects and natural stochastic fluctuations will mean that different data sets on the same tissue type will have different levels of gene expression. Further downstream, the computational methods to determine clusters are not standardised, and it has even been shown that running the same pipeline using two different software implementations can lead to drastically different

outputs [Ric+26]. All of these confounding effects make it essentially impossible for cell-type assignments to be reproduced, and this poses real challenges for comparative analysis. There is therefore a need for tools and statistics which are robust to the choice of cell-type assignment—or, even better, do not rely on cell-type assignments at all.

Perhaps surprisingly, this problem is highly analogous to the problem of defining species in the context of ecology. The final contribution of this thesis is to exploit this analogy to adapt a recent method for measuring ecological diversity into a tool for single-cell transcriptomics analysis. The proposed method, originally introduced by Leinster and Cobbold in the ecological context [LC12], provides a measure of transcriptional heterogeneity which is in a rigorous sense continuous with respect to cell-type assignments, and in fact does not require a cell-type assignment as input at all. We will demonstrate this measure on two real-world single-cell data sets in embryonic development, and show that it (1) sets previous claims on tissue heterogeneity on a more solid quantitative footing and (2) provides new insights that are not available by simply inspecting cell-type distributions. Furthermore, we will adapt ideas of partitioned diversity [Ree+16] to reveal the *in situ* heterogeneity of a spatial transcriptomics data set of the mouse hippocampus.

1.1 Thesis structure

Chapter 2 We begin this chapter by giving a basic overview of the theory behind single-parameter and multiparameter persistent homology, with a view towards invariants and applications to time-varying data. We then state and prove a criterion for computing the rank invariant of 3-parameter persistence modules given access to a minimal presentation with a certain structure. We use this criterion to give an algorithm for fast computation of 3-parameter persistence landscapes from such a minimal presentation. We combine this with upcoming work of Bender *et al.* [BGLa] to give an application of 3-parameter persistent homology to a simulated swarm dynamics data set from [GL23], showing competitive performance in the associated regression task.

Chapter 3 Here we provide an introduction, aimed at mathematicians, to experimental and computational methods in transcriptomics. We cover single-cell transcriptomics and ST technologies, an overview of current approaches to cell-type classification, and a review of previous applications of topological methods in the

field. In particular, we highlight the need for new methods to take advantage of the subcellular resolution of next-generation array-based *in situ* technologies, and for approaches to single-cell transcriptomics that are robust to the choice of cell-type annotation.

Chapter 4 Here we introduce Topological Automatic Cell Types (TopACT), a method for cell-type classification in subcellular spatial transcriptomics. TopACT utilises multiscale local neighbourhoods around spatial locations to provide fine-scale cell-type annotations without the need for a separate cell segmentation step. We showcase TopACT on a range of synthetic and real-world data sets, and include a hypothesis on immune cell organisation in the mouse kidney which was informed by 2-parameter persistent homology analysis. This chapter is based on work published in *Nature* [Ben+24].

Chapter 5 Here we introduce a similarity-sensitive measure of ecological diversity due to Leinster and Cobbold [LC12]. We show how this measure, which we will call *LC diversity*, can be adapted for applications to single-cell transcriptomics data. We argue that LC diversity is a valuable tool for analysing such data without having to commit to a choice of cell-type partition. As a proof of concept, we demonstrate the application of LC diversity to two real-world sc/snRNA-seq data sets on embryonic development and a spatial transcriptomics data set of the mouse hippocampus.

Chapter 6 We finish by summarising the contributions of the thesis, discussing limitations, and commenting broadly on potential future work in multiparameter persistence and single-cell and spatial transcriptomics.

Chapter 2

Spatiotemporal topology and multiparameter persistence

Persistent homology, often called the 'flagship tool' in topological data analysis (TDA), is concerned with extracting homological features (connected components, loops, voids, and higher-order analogues) from data [ZC05; Ott+17]. In the most classical setting of single-parameter persistent homology (1PH), one starts with a point cloud and constructs a nested sequence of simplicial complexes called a *filtration* on top of the point cloud. By taking homology, a filtration is transformed into a *persistence module*: a sequence of vector spaces, linked by linear maps induced by the nesting inclusions. The persistence module contains all of the homological information of the filtration, and it can be represented by a discrete invariant known as a barcode, which is a collection of intervals each of which records both the *birth* and *death* indices of a single homological feature in the filtration. As well as being directly interpretable, the barcode admits numerous different vectorisations which allow it to be used as the input for statistical or machine learning pipelines [Ali+23].

Extending 1PH to spatiotemporal data is a difficult problem. Several approaches, including zigzag persistence [Cd10] and persistence vineyards [CEM06], have been proposed and applied successfully to a range of data sets (Section 2.1.4.1). For spatiotemporal data, however, it is particularly natural to encode the temporal component as two extra parameters as proposed in the thesis of Kim [Kim20]. This takes us into the realm of multiparameter persistent homology (MPH), which carries with it both theoretical and practical challenges [BL23]. As a result of these challenges, software implementations of multiparameter persistent homology (MPH) have hitherto been limited to the 2-parameter case and it has therefore been

impossible to compute the persistent homology of Kim's 3-parameter spatiotemporal filtration.

Bender *et al.* [BGLa], in upcoming work, have developed a new Gröbner basis algorithm for the computation of minimal presentations of persistence modules in an arbitrary number of parameters. Here, we will take advantage of the unique Gröbner basis structure of these presentations to propose an efficient criterion for computing the rank invariant from these presentations, and use this to give an algorithm to compute persistence landscapes in three parameters and more. As a direct application, we will compute persistence landscapes of interlevel-Rips-DMS trifiltrations built on spatiotemporal trajectories from swarm simulations. We will show that using these 3PH landscapes as features for principal component regression is competitive with signature methods as considered in [GL23] for parameter inference in this task. We believe this to be the first application of 3PH to a problem in data science.

Attribution This chapter is based on joint work with Oliver Gäfvert, Hamid Rahkooy, Silviana Amethyst, and Heather Harrington. The proofs and the applications presented are my own original work, but the statements of Theorems 2.3.5 and 2.3.8 as well as Algorithm 2.1 are joint with my coauthors. The extensions we developed to the Muphasa software package [BGLb] to enable computation of persistence landscapes are joint with Oliver Gäfvert and Silviana Amethyst.

2.1 Background: persistent homology

We begin with a brief overview of the theory of persistent homology. A basic understanding of homology is assumed (see [Hat02]).

2.1.1 Filtrations and persistence modules

The persistent homology pipeline can be summarised as follows:

$$\text{Data} \to \text{Filtration} \xrightarrow{\text{Homology}} \text{Persistence module} \to \text{Vectorisation}.$$

We begin by studying the start of this pipeline: how a filtration of simplicial complexes can be defined on a data set and how it gives rise to an algebraic object known as a persistence module which encodes the filtration's homological content.

2.1.1.1 Simplicial complexes

We briefly introduce simplicial complexes in order to fix notation.

Definition 2.1.1. An *(abstract) simplicial complex* K consists of

1. A finite set of *vertices* $V(K)$;
2. A set $\Sigma(K)$ consisting of non-empty subsets of $V(K)$ called *simplices*.

The set $\Sigma(K)$ must satisfy:

1. For every vertex $v \in V(K)$ there is a simplex $\{v\} \in \Sigma(K)$;
2. If $\sigma \subseteq V(K)$ is non-empty and $\sigma \subseteq \tau$ for some simplex $\tau \in \Sigma(K)$ then σ is also a simplex: $\sigma \in \Sigma(K)$.

If K and L are two simplicial complexes, a *simplicial map* $f\colon K \to L$ from K to L consists of a map of vertices $f\colon V(K) \to V(L)$ satisfying $f(\sigma) \in \Sigma(L)$ for every simplex $\sigma \in \Sigma(K)$. Simplicial complexes together with simplicial maps form a category which we denote by **Simp**.

Suppose K is a simplicial complex and consider a subset $\Sigma' \subseteq \Sigma(K)$ of simplices that is itself closed under subsets (i.e. $\emptyset \neq \sigma \subseteq \tau \in \Sigma' \implies \sigma \in \Sigma'$). We can form a new simplicial complex K' by setting $V(K')$ to be the set of all singletons in Σ' and $\Sigma(K') = \Sigma'$. We say that K' is a *subcomplex* of K and write $K' \subseteq K$. Note that if $K' \subseteq K$ then there is a canonical simplicial map $K' \to K$ given by the inclusion $V(K') \subseteq V(K)$.

2.1.1.2 Filtrations

Definition 2.1.2. Let $(P, \leq)$ be a poset. A *P-filtration* F consists of a collection $(F_x)_{x \in P}$ of topological spaces (or simplicial complexes) such that $F_x \subseteq F_y$ whenever $x \leq y$.

Note that, viewing $(P, \leq)$ as a category with objects the elements of P and a single morphism $x \to y$ whenever $x \leq y$, every P-filtration is also a functor $F\colon (P, \leq) \to$ **Top** (or **Simp**).[1] We also remark that it is not necessary to specify the maps between the spaces F_x, since they are required by definition to be related by inclusions.

[1] However, note that not every such functor is a filtration, since functoriality does not guarantee that the maps are inclusions.

Of particular interest in topological data analysis is the case $P = T^r$ where T is a total order (e.g. $T = \mathbb{R}$ or $T = \mathbb{Z}$) and T^r is a product of sets equipped with the following poset structure:

$$(s_1, s_2, \ldots, s_r) \leq (t_1, t_2, \ldots, t_r) \iff s_i \leq t_i \text{ for all } i \in \{1, \ldots, r\}. \tag{2.1}$$

We refer to these filtrations as *r-parameter filtrations*, or *multifiltrations*. When $r = 2$ (resp. $r = 3$) we call them *bifiltrations* (*trifiltrations*).

Definition 2.1.3 (Sublevel filtration)**.** Let P be a poset and X a topological space. If $f : X \to P$ is an arbitrary function then the *sublevel* filtration $\mathcal{S}^{\uparrow}(f) \colon P \to \mathbf{Top}$ is given by setting

$$\mathcal{S}^{\uparrow}(f)_x = f^{-1}\big(\{y \in P : y \leq x\}\big). \tag{2.2}$$

Definition 2.1.4 (Vietoris-Rips filtration)**.** Let (X, d) be a finite metric space. The *Vietoris-Rips filtration*

$$\mathrm{Rips}(X, d) \colon \mathbb{R}^{\geq 0} \to \mathbf{Simp} \tag{2.3}$$

is defined as follows. For each $r \geq 0$, a non-empty subset $\sigma \subseteq X$ is a simplex in $\mathrm{Rips}(X, d)_r$ if and only if the pairwise distances between its vertices are at most r:

$$\sigma \in \mathrm{Rips}(X, d)_r \iff d(x, y) \leq r \text{ for all pairs } x, y \in \sigma. \tag{2.4}$$

When the underlying metric d is unambiguous we simply write $\mathrm{Rips}(X)$ in place of $\mathrm{Rips}(X, d)$. Figure 2.1 shows an example of a Vietoris-Rips filtration in $\mathbb{R}^2$ at select values of the scale parameter.

Combining Definitions 2.1.3 and 2.1.4 yields our first example of a bifiltration.

Definition 2.1.5 (Function-Rips filtration)**.** Let (X, d) be a finite metric space and P a poset. Given a P-valued function $f : X \to P$, the *f-Rips filtration*

$$\mathrm{Rips}^{\uparrow}(X, f) \colon P \times \mathbb{R}^{\geq 0} \to \mathbf{Simp} \tag{2.5}$$

is defined by setting

$$\mathrm{Rips}^{\uparrow}(X, f)_{(x,r)} = \mathrm{Rips}\big(f^{-1}\big(\{y \in P : y \leq x\}\big)\big)_r. \tag{2.6}$$

In other words, for each element x of the poset P we have a Vietoris-Rips filtration built on the sublevel filtration $\mathcal{S}^{\uparrow}(f)_x$. When P is a total order this yields

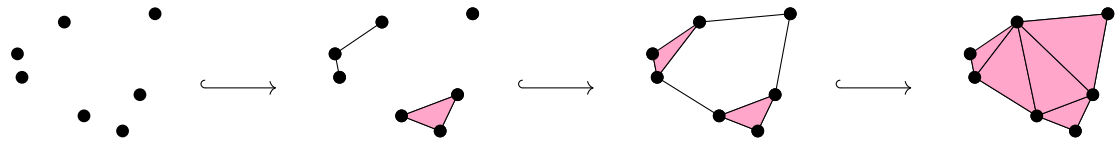


Figure 2.1. A restriction of a Vietoris-Rips filtration to four different values of the scale parameter. Only the 2-skeleton is shown.

the promised bifiltration. For an example of a function-Rips bifiltration see the codensity-Rips bifiltration of Definition 2.1.19.

In Definition 2.1.3 we made a choice to filter 'up' by considering the inverse images of sets of the form $\{y \in P : y \leq x\}$. One could also choose to filter 'down' by setting $\mathcal{S}^{\downarrow}(f)_x = f^{-1}(\{z \in P : z \geq x\})$ to obtain the *superlevel filtration*. When P is a total order, it is possible to avoid making this choice by introducing an extra parameter.

For a total order T, write $\mathrm{Int}(T)$ for the set of non-empty closed intervals in T. Note that $\mathrm{Int}(T)$ can be realised as a subposet of $T^{\mathrm{op}} \times T$ by sending an interval $[a, b] \in \mathrm{Int}(T)$ to its endpoints $(a, b) \in T^{\mathrm{op}} \times T$, since $[a, b] \subseteq [a', b']$ if and only if $a \geq a'$ and $b \leq b'$.

Definition 2.1.6 (Interlevel bifiltration)**.** Let X be a topological space and $f : X \to T$ a function taking values in the total order T. The *interlevel filtration*

$$\mathcal{S}^{\updownarrow}(f) \colon \mathrm{Int}(T) \to \mathbf{Top} \tag{2.7}$$

is given by

$$\mathcal{S}^{\updownarrow}(f)_I = f^{-1}(I). \tag{2.8}$$

Note that if f is bounded then the interlevel filtration $\mathcal{S}^{\updownarrow}(f)$ contains both the sublevel filtration $\mathcal{S}^{\uparrow}(f)$ and the superlevel filtration $\mathcal{S}^{\downarrow}(f)$ as single-parameter slices. The persistence of interlevel filtrations has been studied in detail in other work [DW07; Ben+13].

As in Definition 2.1.5, in the case that X is a finite metric space we can augment the interlevel filtration with an extra parameter by constructing a Rips filtration at each point.

Definition 2.1.7 (Interlevel-Rips trifiltration)**.** Let (X, d) be a metric space and $f : X \to T$ a function taking values in the total order T. Assume further that $f^{-1}(I)$

is finite for every interval $I \in \mathrm{Int}(T)$. The *Interlevel-Rips filtration*

$$\mathrm{Rips}^{\updownarrow}(X, f)\colon \mathrm{Int}(T) \times \mathbb{R}^{\geq 0} \to \mathbf{Simp} \tag{2.9}$$

is defined by setting

$$\mathrm{Rips}^{\updownarrow}(X, f)_{(I,r)} = \mathrm{Rips}\big(f^{-1}(I)\big)_r. \tag{2.10}$$

In the case of time-varying data, the function f can be seen as prescribing a time label for each point in X. Of interest in this work is the case of time-varying point clouds. A time-varying point cloud is a finite set Y equipped with a family $\{\gamma_t\colon Y \to M\}_{t\in T}$ of embeddings into some ambient metric space M. One can construct an Interlevel-Rips filtration from these data by considering the label f on the disjoint union $X = \bigsqcup_{t\in T} \gamma_t(Y)$ that satisfies $f(\gamma_t(x)) = t$. However, the constructed simplicial complexes will grow very large as the number of timesteps n increases.

As an alternative, we can note that such time-varying point clouds have the structure of a dynamic metric space (DMS) as considered in the thesis of Kim [Kim20]. In this context, a *dynamic metric space* consists of an underlying set X along with a family $\{d_t\}_{t\in T}$ of (pseudo) metrics on X. In the case of time-varying point clouds the metric d_t is the metric induced by the embedding γ_t. Kim also introduced a simpler version of the Interlevel-Rips filtration for the case of DMSs:

Definition 2.1.8 (Interlevel-Rips-DMS trifiltration)**.** Let $(X, \{d_t\}_{t\in T})$ be a DMS with finite underlying set X. The *Interlevel-Rips-DMS filtration*

$$\mathrm{Rips}^{\updownarrow}(X)\colon \mathrm{Int}(T) \times \mathbb{R}^{\geq 0} \to \mathbf{Simp} \tag{2.11}$$

of the space X is defined by setting

$$\mathrm{Rips}^{\updownarrow}(X)_{(I,r)} = \mathrm{Rips}(X, d_I)_r \tag{2.12}$$

where the Rips filtration is taken with the metric $d_I(x, y) \coloneqq \inf_{i\in I} d_i(x, y)$.

Kim called this filtration the 'Spatiotemporal Rips' filtration, but we will use Interlevel-Rips-DMS to make clear the distinction with Definition 2.1.7.

Remark 2.1.9. For each $t \in T$ we have $\mathrm{Rips}^{\updownarrow}(X)_{(\{t\},r)} = \mathrm{Rips}(X, d_t)_r$, so the Interlevel-Rips-DMS filtration contains all of the Vietoris-Rips filtrations at each instantaneous point in time.

Recall that in algebraic topology, degree n simplicial homology with coefficients in the field k is a functor $H_n\colon \mathbf{Simp} \to \mathbf{Vect}_k$. Given a P-filtration $F\colon (P,\leq) \to \mathbf{Simp}$, composing with H_n yields a functor $H_n \circ F\colon (P,\leq) \to \mathbf{Vect}_k$ which stores all the information about the homology of the filtration. We call this functor the *persistent homology* of F, and it is the fundamental object of study in much of topological data analysis.

2.1.1.3 Persistence modules

We will now restrict our attention to multifiltrations F over $\mathbb{Z}^r$ equipped with the product order as in (2.1). In this case, it turns out that the persistent homology $H_n \circ F\colon \mathbb{Z}^r \to \mathbf{Vect}_k$ has the structure of a graded module over the polynomial ring $\mathcal{R} \coloneqq k[x_1,\ldots,x_r]$.

Definition 2.1.10. An $\mathcal{R}$-module A is said to be $\mathbb{Z}^r$*-graded* if it is equipped with a vector space decomposition

$$A = \bigoplus_{v\in\mathbb{Z}^r} A_v \tag{2.13}$$

satisfying $x^w \cdot A_v \subseteq A_{v+w}$ for any $v \in \mathbb{Z}^r$ and $w \in \mathbb{N}^r$. A finitely generated $\mathbb{Z}^r$-graded $\mathcal{R}$-module is called an *r-parameter persistence module.*

Remark 2.1.11. Note that $\mathcal{R}$ itself has a persistence module structure given by the grading $\mathcal{R}_v \coloneqq \{\lambda x^v : \lambda \in k\}$.

A module homomorphism $\phi\colon A \to B$ between graded modules is *graded* if $\phi(A_v) \subseteq B_v$ for each $v \in \mathbb{Z}^r$. The r-parameter persistence modules equipped with graded homomorphisms form an Abelian category $\mathbf{PersMod}_r$. It can be checked that there is an equivalence of categories between the category of persistence modules and the category of finite persistent homology functors over $\mathbb{Z}^r$ [CZ09]. Therefore, to understand the persistent homology $H_n \circ F$ of an r-parameter filtration $F\colon \mathbb{Z}^r \to \mathbf{Simp}$ it suffices to understand the structure of the corresponding persistence module.

Remark 2.1.12. Note that restricting to filtrations over $\mathbb{Z}^r$ is not as restrictive as it would seem. In practice one can restrict to the critical points of an $\mathbb{R}$-parameter to obtain a $\mathbb{Z}$-parameter. A common pattern is to first restrict a filtration to its critical points, perform computations over $\mathbb{Z}^r$, and then re-parameterise to return to the original parameter space.

Let A be an r-parameter persistence module. An element $a \in A$ is called *homogeneous* if $a \in A_v$ for some $v \in \mathbb{Z}^r$, and we write $\mathrm{gr}(a) = v$ for the *grade* of

a. If $a_1, \ldots, a_n$ are homogeneous elements of A then we write $\langle a_1, \ldots, a_n \rangle$ for the submodule of A that they generate. Note that this submodule inherits the structure of a persistence module from A.

If A is a persistence module then, for any $w \in \mathbb{Z}^r$, the *shift of A by w* is the module $A(w) \coloneqq A$ with the updated grading

$$A(w)_v = A_{v+w}. \tag{2.14}$$

We say a persistence module A is *free* if there exist grades $w_1, \ldots, w_n \in \mathbb{Z}^r$ such that

$$A \cong \mathcal{R}(w_1, \ldots, w_n) \coloneqq \bigoplus_{i=1}^{n} \mathcal{R}(-w_i) \tag{2.15}$$

where each $\mathcal{R}(-w_i)$ is a shift of the ring $\mathcal{R}$ considered as a persistence module as in Remark 2.1.11. A minimal homogeneous generating set for a free persistence module is called a *basis*. It can be shown that every basis for a given free persistence module A has the same cardinality, which we call the *rank* of A. Note that $\operatorname{rank}(\mathcal{R}(w_1, \ldots, w_n)) = n$.

Let A be a free persistence module with an ordered basis $\{a_1, \ldots, a_n\}$. Every element a of A has a unique representation as a linear combination of shifts of the basis elements:

$$a = \sum_{i=1}^{n} c_i x^{\operatorname{gr}(a) - \operatorname{gr}(a_i)} a_i. \tag{2.16}$$

Note that in the case $\operatorname{gr}(a_i) \not\leq \operatorname{gr}(a)$ we will have $c_i = 0$. The *pivot* of a with respect to the basis is $\operatorname{piv}(a) \coloneqq \max\{1 \leq i \leq n : c_i \neq 0\}$.

The *coefficient vector* $c(a) \in k^n$ of a is the column vector $c(a) \coloneqq (c_1, \ldots, c_n)$. Suppose B is another free persistence module with a corresponding ordered basis $\{b_1, \ldots, b_m\}$. Consider a graded homomorphism $\varphi \colon A \to B$. We can represent φ with an $m \times n$ matrix $[\varphi]$ defined such that its jth column $[\varphi]_j$ is given by $c(\varphi(a_j))$.

2.1.2 Invariants and presentations

The aim of this section is to introduce some of the various representations for (multiparameter) persistence modules that are common in the literature. We will begin with the well-known barcode decomposition [ZC05] for single-parameter persistence, and the related rank invariant which generalises to arbitrary parameters. We will then see the persistence landscape, a vectorised invariant introduced by

Bubenik [Bub15] in the single-parameter case and generalised to multiple parameters by Vipond [Vip20]. Finally, we will define minimal presentations of persistence modules, which form the basis for our contributions in this chapter.

2.1.2.1 Barcodes and the rank invariant

Setting $r = 1$ in Definition 2.1.10 yields the special case of *single-parameter persistence*. A 1-parameter persistence module is a finitely generated graded module over the PID $k[x]$. The classical structure theorem for these modules yields a decomposition

$$A = \bigoplus_{I \in \mathcal{B}(A)} k_I \tag{2.17}$$

of A as a direct sum of interval modules parameterised by a uniquely determined multiset of intervals $\mathcal{B}(A)$ known as the *barcode* of A [ZC05]. The barcode can equivalently be represented by the set of pairs of each interval's left and right endpoints, known as the *birth* and *death* scales respectively. This set is known as the *persistence diagram* of A and can be visualised as a collection of points in $\mathbb{Z}^2 \subset \mathbb{R}^2$.

When $r > 1$, r-parameter persistence modules admit no complete discrete invariant analogous to the barcode [CZ09]. This can be viewed as a consequence of Gabriel's theorem in quiver representation theory (see [BL23]). Nevertheless, it is possible to construct incomplete invariants that capture information about the module. The most immediate of these is the rank invariant:

Definition 2.1.13. Let A be an r-parameter persistence module. The *rank invariant* of A is the function $\rho\colon \mathbb{Z}^r \times \mathbb{Z}^r \to \mathbb{N}$ given by

$$\rho(v, w) = \begin{cases} \operatorname{rank}(A_v \xrightarrow{x^{w-v}} A_w) & \text{if } v \leq w, \\ 0 & \text{otherwise.} \end{cases} \tag{2.18}$$

In other words, ρ records the ranks of the module actions between each pair of homogeneous parts of A.

When $r = 1$ the rank invariant is in fact complete. However, it is simple to construct examples for $r \geq 2$ which show that it is incomplete, i.e. that there are two non-isomorphic r-parameter persistence modules which have the same rank invariant. Figure 2.2 shows two filtrations whose homology modules provide such an example.

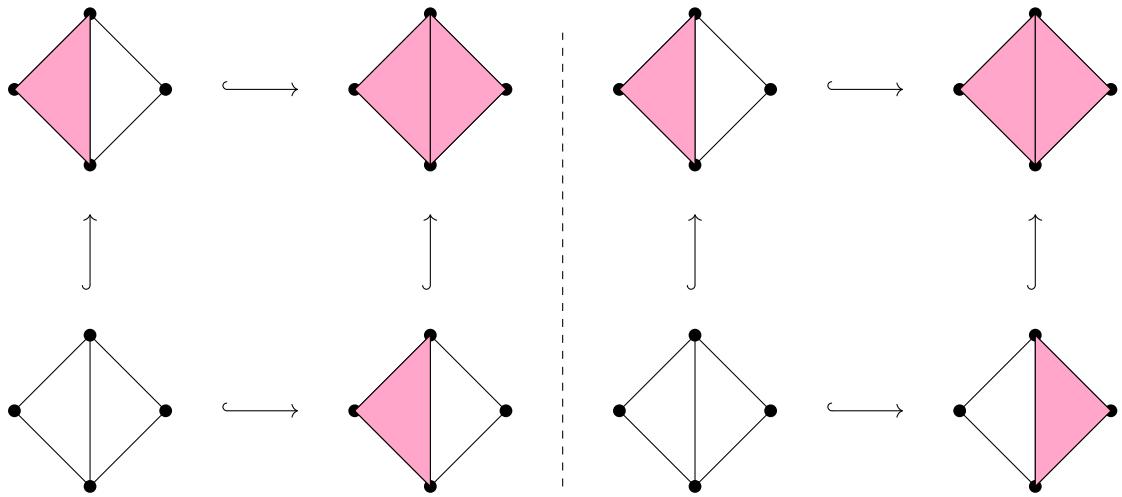


Figure 2.2. Two bifiltrations which give rise to non-isomorphic persistence modules with the same rank invariant.

2.1.2.2 Persistence landscapes

For applications in data science, and especially for machine learning, it is preferable for a topological summary to have a vector structure. There exist a number of different vector invariants of persistence modules, but the one of interest to us here is the *persistence landscape* introduced by Bubenik in the single-parameter case [Bub15] and extended to the multiparameter case by Vipond [Vip20]. We are particularly interested in persistence landscapes in this thesis for two reasons: firstly, they generalise readily to n-dimensional persistence and secondly, we are able to compute them very quickly thanks to Algorithm 2.1. We give a slightly generalised definition to Vipond's which allows for the 'direction' of the landscape to be modified.

Definition 2.1.14. Let A be an r-parameter persistence module and choose a non-zero vector $e \in \mathbb{N}^r$. The *persistence landscape of A with parameter e* is the function $\lambda^{(e)} \colon \mathbb{N} \times \mathbb{Z}^r \to \mathbb{R}$ defined by

$$\lambda^{(e)}(k, v) \coloneqq \sup\{t \in \mathbb{N} : \rho(v - te, v + te) \geq k\} \tag{2.19}$$

where the supremum of the empty set is taken to be 0.

In practice we will be working with $\mathbb{R}^r$-indexed filtrations, and so we would like a definition of persistence landscapes which shares this parameterisation:

Definition 2.1.15. Let $X\colon \mathbb{R}^r \to \mathbf{Vect}_{\mathbb{R}}$ be a finite persistent homology functor and choose a non-zero vector $e \in \mathbb{R}^r_{\geq 0}$. The *persistence landscape of X with parameter e* is the function $\lambda^{(e)}\colon \mathbb{N} \times \mathbb{R}^r \to \mathbb{R}$ defined by

$$\lambda^{(e)}(k, v) \coloneqq \sup\{t \geq 0 : \rho(v - te, v + te) \geq k\} \tag{2.20}$$

where the supremum of the empty set is taken to be 0.

Note that if we use the equivalence of categories between persistence modules and finite persistent homology functors to embed a persistence module into a functor $\mathbb{R}^r \to \mathbf{Vect}_{\mathbb{R}}$ then the two definitions agree.

Setting $e = (1, 1, \ldots, 1)$ recovers the standard landscape as defined by Vipond, which we will refer to as the *diagonal landscape.* We will take the diagonal direction to be implicit if a direction is not made explicit. In the single-parameter case, the persistence landscape (setting $e = 1$) is equivalent to the barcode, and therefore a complete invariant of the module [BBE22].

Since persistence landscapes live in the vector space of functionals $\mathbb{N} \times \mathbb{R}^r \to \mathbb{R}$ they can be summed pointwise to obtain new functionals. In particular, if $\lambda_1, \ldots, \lambda_n$ are a finite collection of persistence landscapes then we define the *average landscape* $\overline{\lambda}\colon \mathbb{N} \times \mathbb{R}^r \to \mathbb{R}$ to be the mean

$$\overline{\lambda} = \frac{1}{n} \sum_{i=1}^{n} \lambda_i. \tag{2.21}$$

Note that $\overline{\lambda}$ may not arise as the persistence landscape of a persistence module, but we will still refer to it as a persistence landscape.

The following basic properties of persistence landscapes are worth noting [Vip20, Lemma 20].

Proposition 2.1.16. *A persistence landscape λ enjoys the following properties:*

1. $\lambda(k, v) \geq 0$ *for all* $k \geq 1$ *and grades* v*;*
2. $\lambda(k, v) \geq \lambda(k + 1, v)$ *for all* $k \geq 1$ *and grades* v*;*
3. $\lambda(k, v)$ *is* 1*-Lipschitz in* v *if* λ *has real-valued parameters.*

2.1.3 Presentations

Here we summarise commutative algebra preliminaries on resolutions and presentations of persistence modules. We refer the reader to [Eis95; Pee11] for a more general

treatment of resolutions of modules in commutative algebra, and to [BL23] for more on the special case of persistence modules which we study in this thesis.

Definition 2.1.17. A *(free) resolution* of a persistence module A is an exact sequence

$$\boldsymbol{F}: \quad \cdots \to P_k \to \cdots \to P_0 \to A \to 0, \tag{2.22}$$

where P_i are free persistence modules. A *(free) presentation* of a persistence module A is an exact sequence

$$P_1 \to P_0 \to A \to 0, \tag{2.23}$$

where P_1 and P_0 are free persistence modules. A resolution of the form

$$\cdots \to 0 \to P \xrightarrow{\text{id}} P \to 0, \tag{2.24}$$

where id is the identity map on P, is called a *trivial resolution*. Similarly a presentation of the form

$$P \oplus P \xrightarrow{\text{id}\oplus 0} P \to 0, \tag{2.25}$$

is called a *trivial presentation*. A resolution $\boldsymbol{F}$ of a persistence module A is called *minimal* if every resolution of A is a direct sum of $\boldsymbol{F}$ and a trivial resolution.

Theorem 2.1.18 (Theorem 7.5, [Pee11])**.** *Every persistence module has a minimal free resolution. Moreover, the minimal resolution is unique up to isomorphism.*

Although they encode all of the information in the persistence module, minimal presentations are not necessarily unique. They therefore do not provide an analogy to the barcode decomposition given in (2.17). However, they do provide compact representations from which other invariants such as the rank invariant and the persistence landscape can be computed. In particular, in the next chapter, our starting point for the computation of a persistence landscape is a certain kind of minimal presentation for the persistence module.

In this thesis, we are interested in presentations for homology modules as opposed to persistence modules in full generality. In particular, if F is an r-parameter filtration then there is a chain complex of persistence modules

$$\cdots \to C_{n+1}F \xrightarrow{\partial_{n+1}} C_nF \xrightarrow{\partial_n} C_{n-1}F \to \dots \tag{2.26}$$

such that $H := H_nF \cong \ker \partial_n / \operatorname{im} \partial_{n+1}$. In general, the chain modules C_iF may not be free. However, a now standard construction [CSV17] can be used to compute

from this chain complex a sequence

$$C \xrightarrow{\phi} A \xrightarrow{\psi} B \tag{2.27}$$

of free persistence modules such that $\psi \circ \phi = 0$ and $H \cong \ker \psi / \operatorname{im} \phi$.

2.1.4 Existing applications

Since its introduction over two decades ago there have been a huge number of applications of persistent homology to problems ranging across all aspects of data science. Here we will briefly highlight some especially relevant applications.

2.1.4.1 Persistent homology for time-varying data

There have been a wide range of applications of persistence to time-varying data, and we provide a brief survey of the main approaches here. Most of these applications are to non-spatial data, but there are some examples where the data have a spatial component.

Zigzag persistence One of the earliest variants of persistent homology with immediate application to spatiotemporal data is *zigzag persistence* [Cd10]. Here the indexing poset is a *zigzag diagram* formed by a sequence of points $x_1, \ldots, x_n$ with arbitrary ordering between adjacent elements, i.e. $x_i \leq x_{i+1}$ *or* $x_i \geq x_{i+1}$ for each $1 \leq i < n$. Perhaps surprisingly, the persistent homology indexed by zigzag diagrams admits a barcode decomposition completely analogous to the barcode in single-parameter persistence.

To see how zigzag persistence is a useful tool for time-varying data, consider a time-varying family of topological spaces $X_1, X_2, \ldots, X_n$. In most applications you would not expect to have the inclusions $X_t \subseteq X_{t+1}$ required to apply single-parameter persistent homology. However, we can take advantage of the inclusions $X_t \supseteq X_t \cap X_{t+1} \subseteq X_{t+1}$ to form a zigzag filtration:

$$X_1 \supseteq X_1 \cap X_2 \subseteq X_2 \subseteq X_2 \cap X_3 \supseteq \cdots \subseteq X_n. \tag{2.28}$$

Symmetrically, one can make the connecting spaces the unions $X_t \cup X_{t+1}$ and reverse the directions of the inclusions. Perhaps surprisingly, it can be shown that zigzag persistence modules decompose into a barcode representation in exactly the

same way as regular single-parameter persistence modules, making them relatively straightforward to analyse.

While zigzag persistence was introduced not long after single-parameter persistence, the software implementations have not experienced the same focused optimisation as their single-parameter counterparts. As a consequence there have been relatively fewer applications of zigzag persistence to real data science problems. Nevertheless, algorithms and software have been improving [DH21; DH22], and we can expect more applications to follow.

In the context of time-varying data, there have been a number of applications of zigzag persistence to time series analysis [TMK20]. Persistence images of zigzag filtrations have been used to define a layer for graph convolutional neural networks, with an application to forecasting problems [CSG21]. Dynamic networks have also been analysed with zigzag approaches: in [Mye+23], the authors form zigzag filtrations from Vietoris-Rips complexes of temporal graphs taken at a fixed scale, and apply this technique to transport network data. In [McD+23], zigzag persistence is used to study the spatiotemporal development of coral reef structures, and in [Yan+25] it is applied to the study of tumour dynamics. Finally, in the application most relevant to this chapter, Corcoran and Jones used zigzag persistence landscapes to study fish swarm behaviour in [CJ17].

One of the limitations of basic zigzag persistence for time-varying data is that by using time as our persistence parameter we lose the multiscale quality of traditional scale-based filtrations like the Vietoris-Rips filtration. One could attempt to address this by forming a kind of 2-parameter filtration with both a zigzag and a Vietoris-Rips component and computing a modified landscape. The basic problem to overcome here is that the persistence landscape does not have an obvious definition when the rank invariant $\rho(v-t, v+t)$ may not be well defined (because the corresponding arrow may not exist in the indexing poset). One approach is to replace the rank invariant with the generalised rank invariant due to Kim and Mémoli [KM21], which has already been used in the standard 2-parameter setting to compute the 'Generalised Rank Invariant Landscape' [Xin+23]. Two groups of authors have both recently released preprints detailing this approach, which they call 'spatiotemporal persistence landscapes' [FH24] and 'Zigzag Generalised Rank Invariant Landscapes (ZZ-GRIL)' [DS25] respectively. These papers both include various applications to time series data.

Persistence vineyards The stability theorem for single-parameter persistent homology due to Cohen-Steiner *et al.* [CEH05] very roughly states that, under certain conditions, a small change in the input filtration will lead to a correspondingly small change in the output persistence module. *Persistence vineyards* utilise this idea to keep track of these movements, turning a continuously varying time series of filtrations into a continuously varying time series of persistence diagrams, where the points in the persistence diagrams trace out continuous curves in space known as *vines* [CEM06].

Despite a large body of theory having developed on the topic, persistence vineyards have not seen much application in practice.[2] Neuroscience has been a relatively popular setting, and vineyards have been used to study dynamics in both EEG [Yoo+16] and fMRI data [Sal+21]. More recently vineyards were applied to tumour immunology in [Yan+25].

Applications to swarm data In this chapter we will explore an example of 3PH computed on spatiotemporal swarm data. Such simulations are common targets for topological methods, and we will summarise previous work here. Let $(X_t)_t$ be a time series of point clouds $X_t \subseteq \mathbb{R}^d$.

Topaz *et al.* have used Betti curves to analyse simulations from the Vicsek and D'Orsogna swarm models [TZH15]. In brief, for each time t and scale r they compute the i-th Betti number $b_i(t, r) = \dim H_i(\mathrm{Rips}(X_t)_r)$ to obtain a 2-parameter family of scalars encoding the strength of topological signal across different times and scales. They represent this information in a contour plot called a *CROCKER plot*, and explore how CROCKER plots of homological degrees 0 and 1 can be used to quantify behaviour in the swarm models. In [Bha+19] this work was extended by using the CROCKER plots as the input for a machine learning pipeline for parameter inference. More recently, Giusti and Lee have developed feature maps of paths of persistence diagrams and applied them to similar simulation data from the D'Orsogna model [GL23].

2.1.4.2 Codensity-Rips bifiltrations for immune cell organisation

A notorious limitation of single-parameter persistent homology is that, while it is stable with respect to perturbations, it is not robust to outliers (see Figure 2.3). One can attempt to mitigate this problem by first performing some noise reduction

[2] On the other hand, the author is pleased to report from her literature review that scholars of biodiversity have extensively researched the *ecological* persistence of species in *grape* vineyards.

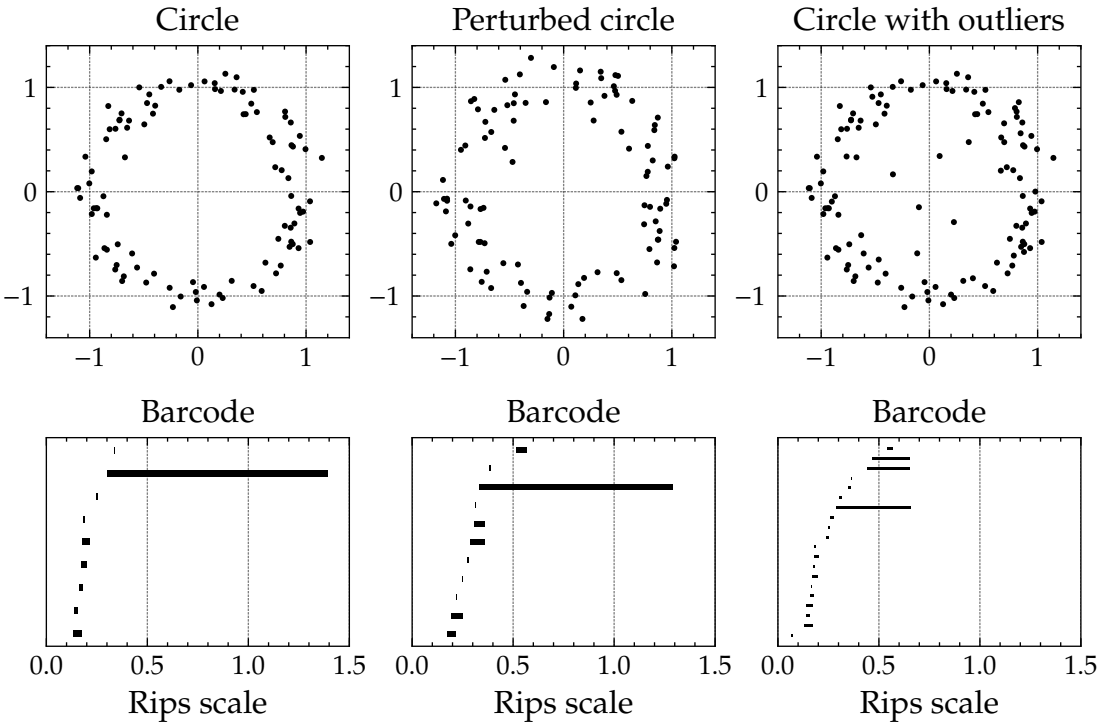


Figure 2.3. Effect of different types of noise on Vietoris-Rips persistent homology. Top row: the input point cloud. Bottom row: the corresponding Vietoris-Rips persistent homology barcode. Left: 100 point sample of a circle of radius 1 with Gaussian noise. The barcode identifies a single persistent feature. Centre: the circle data perturbed with further Gaussian noise in both the x and y coordinates. The barcode still identifies a single persistent feature. Right: the circle data with the addition of 20 uniformly sampled outliers. The outliers destroy the barcode and it fails to identify the single persistent feature. Barcodes computed with Ripser.py 0.6.14 [TSB25].

to remove outliers, however this approach raises the question of exactly which points should count as outliers. A natural resolution in the spirit of persistence is to introduce a second parameter that controls the level of noise reduction. This can be achieved with the *codensity-Rips bifiltration*, an instantiation of the function-Rips bifiltration of Definition 2.1.5.

Definition 2.1.19. Let (X, d) be a finite metric space and fix an integer $k > 0$. For a given point $x \in X$ write x_i for its i-th nearest neighbour in X (resolving ties arbitrarily). The *k-codensity* function $\rho_k \colon X \to \mathbb{R}^{\geq 0}$ is given by

$$\rho_k(x) = \frac{1}{k} \sum_{i=1}^{k} d(x, x_i). \tag{2.29}$$

The *codensity-Rips bifiltration* of X with respect to k is the function-Rips bifiltration $\mathrm{Rips}^{\uparrow}(X, \rho_k) \colon \mathbb{R}^{\geq 0} \times \mathbb{R}^{\geq 0} \to \mathbf{Simp}$.

An early application of 2-parameter persistence used the codensity-Rips bifiltration to study immune cell patterning in tumours [Vip+21]. The authors computed 2-parameter persistence landscapes of patches of immune cell locations realised as point clouds, and performed statistical tests as well as computing averages over different immune subtypes. In particular, they showed that the multiparameter approach was able to much more reliably represent immune cell spatial patterning than the single-parameter approach.

In Chapter 4 we will show a new application of the codensity-Rips bifiltration, this time to immune cell patterning emerging from the output of our proposed spatial transcriptomics cell-type classification method.

2.2 Background: slice-consistent Gröbner bases and the GBS algorithm for minimal presentations

Our contributions in this chapter are based on an upcoming algorithm for computing minimal presentations of persistence modules due to Bender, Gäfvert, and Lesnick known as the Gröbner Bases and Syzygies (GBS) algorithm [BGLa]. The output of the GBS algorithm has some special structure which we will take advantage of in the next section to more efficiently compute the rank invariant and persistence landscape. The purpose of this section is to briefly outline these properties, and we claim no ownership of any of the results presented here.

2.2.1 Presentations for homology

Here we explain the ideas developed in [BGLa] for constructing a minimal presentation of the homology $H = \ker\psi/\operatorname{im}\phi$, given the sequence $C \xrightarrow{\phi} A \xrightarrow{\psi} B$ as in (2.27).

As a first attempt, one may try to construct the following 'presentation' for H, where $\pi\colon \ker\psi \to H$ is the quotient map:

$$C \xrightarrow{\phi} \ker\psi \xrightarrow{\pi} H. \tag{2.30}$$

In fact, when $r \leq 2$ it is an immediate consequence of Hilbert's Syzygy Theorem [Eis95] that $\ker\psi$ is a free module, and so this is indeed a presentation. However, for $r > 2$ the module $\ker\psi$ may not be free, and we therefore need to take a different approach. In particular, one needs to examine the failure of $\ker\psi$ to be free by considering its *syzygies*:

Definition 2.2.1. Let A be an r-parameter persistence module with a given homogeneous generating set $\{a_1, \ldots, a_n\}$. Let $v_i := \operatorname{gr} a_i$. Write $\mathcal{R}(a_1, \ldots, a_n) := \bigoplus_{i=1}^n \mathcal{R}(-v_i)$ for the free persistence module as in (2.15), which has a canonical basis $\{e_1, \ldots, e_n\}$ satisfying $\operatorname{gr}(e_i) = v_i$. There is a map of persistence modules

$$\rho\colon \mathcal{R}(a_1, \ldots, a_n) \to A, \tag{2.31}$$

given by

$$\rho\left(\sum_{i=1}^n g_i e_i\right) = \sum_{i=1}^n g_i \rho(e_i) := \sum_{i=1}^n g_i a_i. \tag{2.32}$$

Elements of $\ker\rho$ are called *syzygies* of A with respect to the generating set $\{a_1, \ldots, a_n\}$. The set of all such syzygies forms a persistence module, denoted $\operatorname{Syz}(a_1, \ldots, a_n)$, which we call the *syzygy module* of A.

Now, let $\{f_1, \ldots, f_s\} \subseteq \ker\psi$ be a homogeneous generating set for $\ker\psi$ and construct the syzygy module $\operatorname{Syz}_0 := \operatorname{Syz}(f_1, \ldots, f_s)$ as the kernel of the map

$$\mathcal{R}(f_1, \ldots, f_s) \xrightarrow{\rho_0} \ker\psi \tag{2.33}$$

as in Definition 2.2.1.

If $\{f_1', \ldots, f_{s'}'\}$ is a generating set for Syz_0 then we have a map

$$\mathcal{R}(f_1', \ldots, f_{s'}') \xrightarrow{\rho_1} \operatorname{Syz}_0 \tag{2.34}$$

yielding the *second syzygy module* $\mathrm{Syz}_1 \coloneqq \mathrm{Syz}(f_1', \ldots, f_{s'}') = \ker \rho_1$ of $\ker \psi$. Writing $P_0 = \mathcal{R}(f_1, \ldots, f_s)$ and $P_1 = \mathcal{R}(f_1', \ldots, f_{s'}')$ we therefore have the following free presentation for $\ker \psi$:

$$P_1 \xrightarrow{\rho_1} P_0 \xrightarrow{\rho_0} \ker \psi \to 0. \tag{2.35}$$

Remark 2.2.2. One can now continue this construction inductively, by taking Syz_{k+1} to be the syzygy module of Syz_k, to arrive at the free exact sequence

$$\cdots \to P_k \xrightarrow{\rho_k} P_{k-1} \to \cdots \to P_0 \xrightarrow{\rho_0} \ker \psi \to 0, \tag{2.36}$$

where each P_i is a free persistence module.

Recall that our present goal is to construct a minimal presentation of H for arbitrary values of r. One might hope to adapt the presentation (2.35) as follows:

$$P_1 \xrightarrow{\rho_1} P_0 \xrightarrow{\pi \circ \rho_0} H \to 0. \tag{2.37}$$

However, this sequence is not exact because in general

$$\mathrm{im}(\rho_1) = \ker(\rho_0) \subsetneq \ker(\pi \circ \rho_0), \tag{2.38}$$

i.e. there are elements of P_0 that are killed by π but not by ρ_0. We therefore need to include the contribution of $\mathrm{im}\, \phi$ into our construction.

In this case, note that since C is free and ρ_0 is surjective, we can choose a (non-canonical) lift $\tilde{\phi} \colon C \to P_0$ that factors ϕ through ρ_0:

$$\begin{array}{ccc} C & & \\ {\scriptstyle \tilde{\phi}}\,\vdots & \searrow^{\phi} & \\ P_0 & \xrightarrow{\rho_0} & \ker \psi. \end{array} \tag{2.39}$$

We can now put the equations (2.35) and (2.39) together to form the following presentation of H:

$$\begin{array}{ccccc} & & \mathrm{Syz}_0 + \mathrm{im}\, \phi & & \\ & \nearrow & & \searrow & \\ P_1 \oplus C & & \xrightarrow{\rho_1 \oplus \tilde{\phi}} & & P_0 \xrightarrow{\pi \circ \rho_0} H. \end{array} \tag{2.40}$$

To see that this is exact, note that an element of $\ker(\pi \circ \rho_0)$ is either killed by ρ_0, in which case it is contained in the image of ρ_1, or it is killed by π, in which case it is contained in the image of $\tilde{\phi}$. So we have that $\operatorname{im}(\rho_1 \oplus \tilde{\phi}) = \ker(\pi \circ \rho_0)$.

Example 2.2.3. Fix $R = k[x_1, x_2, x_3]$ and consider the sequence

$$C \xrightarrow{\phi} A \xrightarrow{\psi} B \tag{2.41}$$

where $C = \mathcal{R}((1,1,1))$, $A = \mathcal{R}(e_1, e_2, e_3)$, and $B = \mathcal{R}(0)$. Let the maps ϕ and ψ, given in coefficient matrix form as in (2.16), be

$$\phi = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array} \begin{array}{c} c_1 \\ \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \end{array} \qquad \psi = c_1 \begin{array}{c} \begin{array}{ccc} a_1 & a_2 & a_3 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \end{array} \tag{2.42}$$

Our aim is to compute a presentation for the homology $H = \ker\psi / \operatorname{im}\phi$. The first step in the presentation procedure is to compute a minimal presentation for $\ker\psi$. It can be seen that $\ker\psi$ is generated by the elements $s_{ij} = x_j a_i - x_i a_j$ for the three pairs $1 \le i < j \le 3$. There is one relation $x_3 s_{12} - x_2 s_{13} + x_1 s_{23} = 0$, leading to the minimal presentation

$$\mathcal{R}((1,1,1)) \xrightarrow{\rho_1} \mathcal{R}((1,1,0),(1,0,1),(0,1,1)) \xrightarrow{\rho_0} \ker\psi \tag{2.43}$$

with maps

$$\rho_1 = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} \qquad \rho_0 = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & -1 \end{bmatrix}. \tag{2.44}$$

The next step is to compute a lift

$$\tilde{\phi} \colon C = \mathcal{R}((1,1,1)) \to P_0 = \mathcal{R}((1,1,0),(1,0,1),(0,1,1)). \tag{2.45}$$

It can be seen that $\tilde{\phi} \coloneqq (1,0,0)^T$ suffices, since

$$x_3 s_{12} = x_3(x_2 a_1 - x_1 a_2) = x_2 x_3 a_1 - x_1 x_3 a_2 = \phi(c_1). \tag{2.46}$$

Putting everything together, one arrives at the following presentation for H

$$\mathcal{R}((1,1,1),(1,1,1)) \xrightarrow{\xi_1} \mathcal{R}((1,1,0),(1,0,1),(0,1,1)) \xrightarrow{\pi\circ\xi_2} H, \tag{2.47}$$

with maps given by

$$\xi_1 = \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 1 & 0 \end{bmatrix} \qquad \xi_2 = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & -1 \end{bmatrix}. \tag{2.48}$$

While the presentation in the previous example turned out to be minimal, this is not guaranteed in general. In [BGLa], the authors describe a minimisation procedure which converts the presentation (2.40) into a minimal presentation with the same structure. In particular, the procedure finds free submodules $P \subseteq P_1$ and $Q \subseteq C$ such that

$$\begin{array}{ccccc} & & \mathrm{Syz}_0 + \mathrm{im}\,\phi & & \\ & \nearrow & & \searrow & \\ P \oplus Q & & \xrightarrow{\rho_1 \oplus \tilde{\phi}} & & P_0 \xrightarrow{\pi\circ\rho_0} H. \end{array} \tag{2.49}$$

is a minimal presentation of H factoring through $\mathrm{Syz}_0 + \mathrm{im}\,\phi$.

The key takeaway from this construction is that all you need in order to compute a minimal presentation of a persistence module is to be able to compute images and kernels of maps of free persistence modules, and this is precisely what the GBS algorithm does.

2.2.2 Gröbner bases of persistence modules and the GBS algorithm

The GBS algorithm of [BGLa] computes Gröbner bases for images and kernels of maps of free persistence modules. We give some basic definitions of Gröbner bases and the basic building blocks of GBS.

Definition 2.2.4. Let A be a free persistence module with an ordered basis $\{a_1, \ldots, a_n\}$. If B is a homogeneous submodule of A, we say that a set of homogeneous generators G for B is a *Gröbner basis* for B if for every homogeneous $b \in B$, there is an element $g \in G$ such that $\mathrm{gr}(g) \leq \mathrm{gr}(b)$ and $\mathrm{piv}(g) = \mathrm{piv}(b)$.

GBS operates by moving through parameter space in a special order which we will also need to make use of later:

Definition 2.2.5. The *colexicographical*, or *colex*, order on $\mathbb{Z}^r$ is defined as follows. We say that $v <_{\text{colex}} w$ if the rightmost non-zero entry of $w - v$ is positive. In particular $e_1 < e_2 < \cdots < e_r$.

The following definition from [BGLa] is a key building block of the GBS algorithm:

Definition 2.2.6. Let V be a vector space with a finite ordered basis S, and consider a subspace $W \subseteq V$. We say that a basis for W is *reduced* with respect to S if its elements have distinct pivots with respect to S.

Consider a map $\phi\colon B \to A$ of free persistence modules. Roughly speaking, the GBS algorithm computes a Gröbner basis for $\operatorname{im}\phi \subseteq A$ by iterating through each grade v in colexicographical order and computing a reduced basis G_v for $(\operatorname{im}\phi)_v$ as a vector space. These reduced bases are computed in such a way that the union $G = \bigcup_{v\in\mathbb{Z}^r} G_v$ forms a Gröbner basis for $\operatorname{im}\phi$ as a submodule. In fact, they show that one only needs to consider a subset of the grades called the *grid* of a generating set $\{a_1, \ldots, a_n\}$ for $\operatorname{im}\phi$.

Definition 2.2.7. Let A be a free module with some homogeneous elements $a_1, \ldots, a_n \in A$. The *grid* of $\{a_1, \ldots, a_n\}$ is defined to be

$$\operatorname{Grid}(a_1, \ldots, a_n) = \{v \in \mathbb{Z}^r : \text{for all } 1 \le i \le r \text{ there is some } j \text{ s.t. } v_i = (\operatorname{gr}(a_j))_i\}. \tag{2.50}$$

An illustration of a grid in two parameters is given in Figure 2.4.

Lemma 2.2.8. *Let $v \in \mathbb{Z}^r$ be a grade such that $v \notin \operatorname{Grid}(a_1, \ldots, a_n)$ and for each $1 \le i \le r$ there exists some $1 \le j \le n$ satisfying $(a_j)_i \le v_i$. Then there is a unique maximal grade $\lfloor v \rfloor$ in the set*

$$\{w \in \operatorname{Grid}(a_1, \ldots, a_n) : w \le v\}. \tag{2.51}$$

Proof. We construct $\lfloor v \rfloor$ component-wise. For each $1 \le i \le r$ set

$$(\lfloor v \rfloor)_i := \max\{(a_j)_i : (a_j)_i \le v_i\}. \tag{2.52}$$

It can be seen that $\lfloor v \rfloor \le v$, and $\lfloor v \rfloor \in \operatorname{Grid}(a_1, \ldots, a_n)$ by the definition of the grid. Furthermore $\lfloor v \rfloor$ is maximal over all grid elements bounded above by v, since if $w \le v$ we must have, for all $1 \le i \le r$, that $w_i \le v_i$, whence $w_i \le (\lfloor v \rfloor)_i$ by definition. □

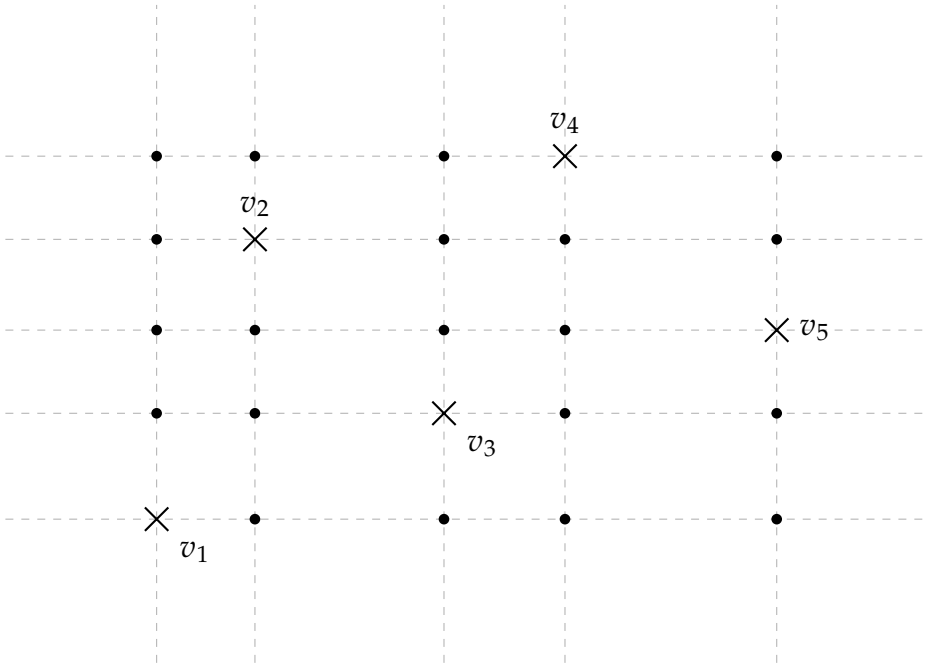


Figure 2.4. Illustration of the grid $\mathrm{Grid}(a_1, \ldots, a_5)$ of some homogeneous elements $a_1, \ldots, a_5$ with grades $\mathrm{gr}(a_i) = v_i$, as in Definition 2.2.7. The grid is composed of the five grades $v_1, \ldots, v_5$ of the elements a_i and the 20 additional grades shown.

Remark 2.2.9. Suppose $A' \subseteq A$ has a homogeneous generating set $a_1, \ldots, a_n$ and we are interested in computing a Gröbner basis for A'. It can be seen that $A'_v \cong A'_{\lfloor v \rfloor}$ for $v \notin \mathrm{Grid}(a_1, \ldots, a_n)$, since the map $x^{v-\lfloor v \rfloor}$ is both surjective and injective. So one only needs to compute reduced bases G_v at each grade v in the grid. By abuse of notation, we will therefore write $G_v \coloneqq G_{\lfloor v \rfloor}$ for non-grid elements in order to simplify later proofs.

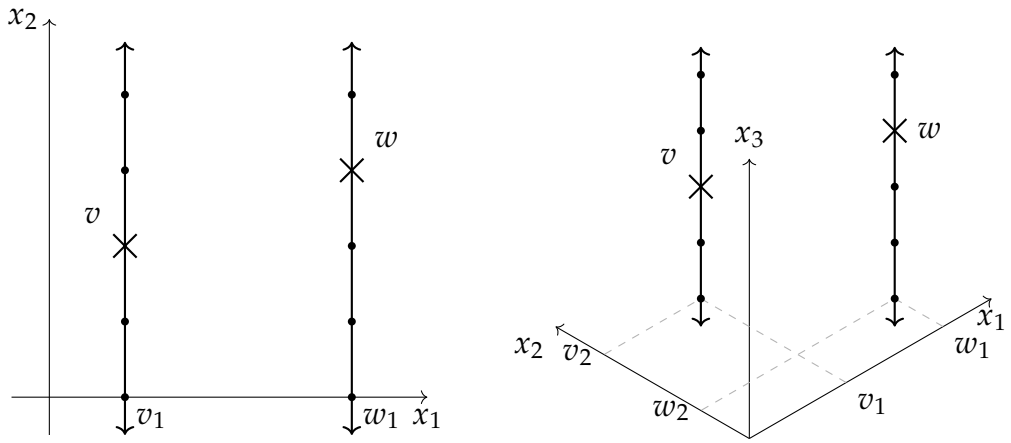
The reduced bases computed as part of GBS have a special structure which we will encapsulate in a definition for convenience. First we need to define the notion of 1-parameter slices of our parameter space.

Definition 2.2.10. Let $v = (v_1, \ldots, v_r) \in \mathbb{Z}^r$ be a grade. Then the 1-parameter *slice* of v is

$$\mathrm{Slice}(v_1, \ldots, v_{r-1}) = \{w \in \mathbb{Z}^r : w_i = v_i \text{ for all } 1 \leq i \leq r-1\}. \tag{2.53}$$

In other words, the slice of v is all elements obtained from v by only shifting the last parameter. See Figure 2.5. We now state the key property of GBS-computed reduced bases.

Definition 2.2.11 (Slice-consistency)**.** Let $\{G_v\}_{v \in \mathbb{Z}^r}$ be a collection of reduced bases for a submodule of a free persistence module. The collection is said to be *slice-consistent*



(a) Slices in $\mathbb{Z}^2$. **(b)** Slices in $\mathbb{Z}^3$.

Figure 2.5. The slices of two grades $v, w \in \mathbb{Z}^r$ when (a) $r = 2$ and (b) $r = 3$. The slice of a grade is the set of all grades obtained by varying its last parameter value (Definition 2.2.10).

if for any grades $v \leq w$ such that $w \in \text{Slice}(v_1, \ldots, v_{r-1})$ we have $G_v \subseteq G_w$ and moreover $\text{gr}(g) \not\leq v$ for any $g \in G_w \setminus G_v$.

The intermediate bases computed by GBS are slice consistent, and this is a key part of the proof of the complexity result in [BGLa]. Here we will use this fact without proof.

Fact 2.2.12. *The reduced bases used to compute Gröbner bases of images in GBS are slice-consistent.*

An implementation of GBS for computing minimal presentations of multiparameter persistence modules is available in the Muphasa software package [BGLb].

2.3 The sparse representation for computing the rank invariant

In this section, we will describe a *sparse representation* of an r-parameter persistence module which allows for quick computation of its rank invariant (in certain directions) given access to a minimal presentation with the same structure as that output by GBS. Recall from Definition 2.1.14 that the kth persistence landscape of H in the direction $e \in \mathbb{N}^r$ can be evaluated at the grade v by the following expression:

$$\lambda_k^{(e)}(v) \coloneqq \sup\{t \in \mathbb{N} : \rho(v - te, v + te) \geq k\} \tag{2.54}$$

where ρ is the rank invariant of H. We will show that given this sparse representation for H, if e_r is the standard unit vector in the r-th parameter then the rank invariant $\rho(v, v+te_r)$ can be evaluated as a set difference rather than a matrix rank computation. Furthermore, by reordering the parameter set r times and recomputing the sparse representation, it is possible to combine the sparse representation in each parameter to compute the rank invariant in any direction. The upshot of these results is that we can then compute the persistence landscape of H while sidestepping repeated matrix rank computations.

We will begin by proving some basic lemmas about presentations of persistence modules. We will then use these ideas to show that presentations with the special structure of the GBS output enable fast computation of the rank invariant in a certain direction. Then, we will present an algorithm to efficiently compute persistence landscapes. Finally, we generalise these results to arbitrary directions.

2.3.1 Algebraic preliminaries

Definition 2.3.1. Let $P_1 \xrightarrow{\rho_1} P_0 \xrightarrow{\rho_0} A$ be a free presentation of a persistence module A and consider an ordered free basis $\{z_1, \ldots, z_n\}$ for P_0. For a given generator z_i and grade $v \in \mathbb{Z}^r$ we say that z_i is *active* at v if $\mathrm{gr}(z_i) \leq v$. If z_i is active at v then we say that an element $s \in P_1$ *makes* z_i *redundant* at v if $\mathrm{gr}(s) \leq v$ and

$$\rho_1(s) \in \langle z_1, \ldots, z_i \rangle \setminus \langle z_1, \ldots, z_{i-1} \rangle. \tag{2.55}$$

We say that z_i is *essential* at v if it is active at v and not made redundant at v by any element of P_1.

Note that if z_i is made redundant by s at v then it follows that

$$x^{v-\mathrm{gr}(s)}\rho_1(s) = \sum_{j \leq i} c_j x^{v-\mathrm{gr}(z_j)} z_j \tag{2.56}$$

for some coefficients $c_1, \ldots, c_i \in k$ where $c_i \neq 0$ and $c_j = 0$ whenever z_j is not active at v. By exactness $\rho_0(\rho_1(s)) = 0$ and we find that

$$x^{v-\mathrm{gr}(z_i)}\rho_0(z_i) = -\sum_{j<i} \frac{c_j}{c_i} x^{v-\mathrm{gr}(z_j)}\rho_0(z_j) \in A_v. \tag{2.57}$$

In particular removing z_i from a set containing all of the active z_j for $j < i$ does not change the span of their images in A_v.

Lemma 2.3.2. *Let $P_1 \xrightarrow{\rho_1} P_0 \xrightarrow{\rho_0} A$ be a minimal free presentation of a persistence module A and consider an ordered free basis $\{z_1, \ldots, z_n\}$ for P_0. Fix a grade $v \in \mathbb{Z}^r$. Then a basis for the vector space A_v is given by the set*

$$\left\{x^{v-\mathrm{gr}(z_i)}\rho_0(z_i) : z_i \text{ is essential at } v\right\}. \tag{2.58}$$

Proof. Write Z_{ess} for the given set and

$$Z_{\mathrm{act}} \coloneqq \left\{x^{v-\mathrm{gr}(z_i)}\rho_0(z_i) : z_i \text{ is active at } v\right\}. \tag{2.59}$$

It can be seen by induction that

$$Z^i \coloneqq Z_{\mathrm{ess}} \cup \left\{x^{v-\mathrm{gr}(z_j)}\rho_0(z_j) \in Z_{\mathrm{act}} : j \leq i\right\} \tag{2.60}$$

spans A_v for any $0 \leq i \leq n$, with the base case being that $Z^n = Z_{\mathrm{act}}$ itself spans A_v and the inductive step following because a redundant element can be expressed in terms of later active elements. Since $Z_{\mathrm{ess}} = Z^0$ this shows that Z_{ess} spans A_v.

It remains to show that Z_{ess} is linearly independent in A_v. Suppose we have a linear relation

$$\sum_{i=1}^{n} c_i x^{v-\mathrm{gr}(z_i)}\rho_0(z_i) = 0 \in A_v \tag{2.61}$$

where $c_i = 0$ whenever $\mathrm{gr}(z_i) \not\leq v$ but not all of the c_is are 0. Since ρ_0 is a module homomorphism we therefore have that

$$\sum_{i=1}^{n} c_i x^{v-\mathrm{gr}(z_i)} z_i \in \ker \rho_0 = \mathrm{im}\, \rho_1. \tag{2.62}$$

In particular there is an $s \in P_1$ with $\rho_1(s) = \sum_{i=1}^{n} c_i x^{v-\mathrm{gr}(z_i)} z_i$. Letting k be maximal such that $c_k \neq 0$, it can be seen that s makes z_k redundant. It follows that any subset of the z_is that contains a linear dependency in A_v must contain a redundant element, and we conclude that the set Z_{ess} is linearly independent and therefore a basis for A_v. □

Having shown that the essential elements produce a basis for the graded components A_v, we will now use the results of the previous section to demonstrate an efficient criterion for determining if a basis element is essential. For notational convenience, given a map $A \xrightarrow{f} B$ of free persistence modules with some fixed

ordered basis for B, we will say that the *pivot* of an element $a \in A$ is the pivot of its image $f(a) \in B$ with respect to that basis.

Lemma 2.3.3. *Let $P_1 \xrightarrow{\rho_1} P_0 \xrightarrow{\rho_0} A$ be a free presentation of a persistence module A and consider an ordered free basis $\{z_1, \dots, z_n\}$ for P_0. Suppose that $s_1, \dots, s_m \in P_1$ are chosen such that the images $\rho_1(s_1), \dots, \rho_1(s_m)$ form a reduced basis for $(\operatorname{im} \rho_1)_v$. Then an element z_i is essential at v if and only if it is active at v and $\operatorname{piv}(\rho_1(s_j)) \neq i$ for all $1 \leq j \leq m$.*

Proof. Clearly if z_i is essential then none of the s_j can have pivot $\operatorname{piv}(s_j) = i$, as otherwise they would make z_i redundant.

In the other direction, suppose that z_i is made redundant at v by some element $s \in (P_1)_v$. Our aim is to show that there exists some k with $\operatorname{piv}(s_k) = i$. Note that $\rho_1(s)$ can be expressed uniquely in terms of the images $\rho_1(s_j)$:

$$\rho_1(s) = \sum_{j=1}^{m} c_j \rho_1(s_j) \in \langle z_1, \dots, z_i \rangle \setminus \langle z_1, \dots, z_{i-1} \rangle. \tag{2.63}$$

Choose k so that s_k has maximal pivot over all j such that $c_j \neq 0$. Note that $\operatorname{piv}(s_k) \geq i$, as otherwise $\rho_1(s) \in \langle z_1, \dots, z_{i-1} \rangle$. However by uniqueness of pivots, since the $\rho_1(s_j)$ form a reduced basis, we must also have that $\operatorname{piv}(s_k) \leq i$, since otherwise $\operatorname{piv}(s) = \operatorname{piv}(s_k) > i$ and $\rho_1(s) \notin \langle z_1, \dots, z_i \rangle$. We therefore have that $\operatorname{piv}(s_k) = i$ as required. □

Corollary 2.3.4. *Let $P_1 \xrightarrow{\rho_1} P_0 \xrightarrow{\rho_0} A$ be a free presentation of a persistence module A and consider an ordered free basis $\{z_1, \dots, z_n\}$ for P_0. Suppose that $s_1, \dots, s_m \in P_1$ are chosen such that the images $\rho_1(s_1), \dots, \rho_1(s_m)$ form a reduced basis for $(\operatorname{im} \rho_1)_v$. Then the collection of z_i that are active at v and for which $\operatorname{piv}(s_j) \neq i$ for all $1 \leq j \leq m$ is a basis for A_v.*

Proof. By Lemma 2.3.2 the essential elements of $\{z_1, \dots, z_n\}$ form a basis for A_v, and by Lemma 2.3.3 these are precisely the elements that are not pivots of any elements of S. □

2.3.2 Computing the rank invariant

Recall from Section 2.2.1 that given a sequence of free persistence modules

$$C \xrightarrow{\phi} A \xrightarrow{\psi} B \tag{2.64}$$

satisfying $\psi \circ \phi = 0$, we can construct a minimal presentation for $H = \ker \psi / \operatorname{im} \phi$ with the following structure:

$$\begin{array}{ccccc} & & \mathrm{Syz}_0 + \operatorname{im}\phi & & \\ & \nearrow & & \searrow & \\ P \oplus Q & & \xrightarrow{\rho_1 \oplus \tilde{\phi}} & & P_0 \xrightarrow{\pi \circ \rho_0} H. \end{array} \tag{2.65}$$

The GBS algorithm outputs an ordered basis $Z = \{z_1, \ldots, z_n\}$ for P_0 and a Gröbner basis G for the image of the map $\rho_1 \oplus \tilde{\phi}$ formed as a union

$$G = \bigcup_{v \in \mathrm{Grid}(z_1, \ldots, z_n)} G_v \tag{2.66}$$

of reduced bases G_v for the graded components of $\operatorname{im}(\rho_1 \oplus \tilde{\phi})$. Crucially, these reduced bases are slice-consistent (see Definition 2.2.11).

We can partition the Gröbner basis G according to pivots by setting $S_i := \{g \in G : \mathrm{piv}(g) = i\}$, as in the following example:

$$\begin{array}{c|cccccc|l} & g_1 & g_2 & g_3 & g_4 & g_5 & g_6 & \\ z_1 & * & * & * & * & * & * & S_1 = \varnothing \\ z_2 & \circledast & * & \circledast & * & * & * & S_2 = \{g_1, g_3\} \\ z_3 & 0 & * & 0 & \circledast & * & * & S_3 = \{g_4\} \\ z_4 & 0 & * & 0 & 0 & * & * & S_4 = \varnothing \\ z_5 & 0 & \circledast & 0 & 0 & \circledast & \circledast & S_5 = \{g_2, g_5, g_6\} \end{array} \tag{2.67}$$

Intuitively we will think of the elements of S_i as 'killing' the generator z_i *in the direction of the last parameter*. This is what will allow us to efficiently compute the rank invariant, as we will see in the example to follow.

Theorem 2.3.5. *Consider a Gröbner basis G obtained from a slice-consistent set as in* (2.66)*. Take $v \in \mathbb{N}^{r-1}$ and consider $w, w' \in \mathrm{Slice}(v_1, \ldots, v_{r-1})$ such that $w \leq w'$. Then the rank invariant $\rho(w, w')$ of H is equal to the number of $z_i \in Z$ satisfying:*

$$\mathrm{gr}(z_i) \leq w \text{ and } \#\{g \in S_i \mid \mathrm{gr}(g) \leq w'\} = 0. \tag{2.68}$$

Proof. There is a basis $\mathcal{B}$ for H_w given by Corollary 2.3.4:

$$\mathcal{B} = \{z_i : z_i \text{ is active at } w \text{ and } g \notin S_i \text{ for all } g \in G_w\}. \tag{2.69}$$

Similarly we can construct a basis $\mathcal{B}'$ for $H_{w'}$:

$$\mathcal{B}' = \{z_i : z_i \text{ is active at } w' \text{ and } g \notin S_i \text{ for all } g \in G_{w'}\}. \tag{2.70}$$

Let's think about what the possible differences between these two bases are. Firstly, note that if $z_i \in \mathcal{B}'$ and z_i is active at w then we must have $z_i \in \mathcal{B}$, since if $g \in G_w$ has $\text{piv}(g) = i$ it follows by slice consistency that $g \in G_{w'}$. In other words, the difference $\mathcal{B}' \setminus \mathcal{B}$ consists only of generators z_i that are active at w' but are not active at w. Note in particular that such generators will necessarily be greater than any element of $\mathcal{B}$ colexicographically, and therefore will appear later in the order of Z. Conversely, clearly any generator active at w is also active at w'. It follows that the elements of $\mathcal{B} \setminus \mathcal{B}'$ are precisely those generators $z_i \in \mathcal{B}$ for which there is some $g \in G_{w'} \setminus G_w$ such that $\text{piv}(g) = i$.

Now, consider the matrix of the module action from $H_w \to H_{w'}$ with respect to these two bases. Clearly the module action acts as the identity on $\mathcal{B} \cap \mathcal{B}'$. Moreover, if $z_i \in \mathcal{B} \setminus \mathcal{B}'$ it follows that there exists some $g \in G_{w'}$ such that $\text{piv}(g) = i$. Say $g = \sum_{j \geq 1} c_j z_j$ with $c_j = 0$ for all $j > i$ and $c_i \neq 0$. The column of the matrix corresponding to z_i will therefore have entries x_{z_j} for $z_j \in \mathcal{B}'$, where

$$x_{z_j} = \begin{cases} -c_j & \text{if } j < i, \\ 0 & \text{otherwise.} \end{cases} \tag{2.71}$$

Note that for any $z_i \in \mathcal{B}$ the column corresponding to z_i never has any non-zero entries in rows corresponding to z_j for any $j > i$. Since elements of $\mathcal{B}' \setminus \mathcal{B}$ are all colexicographically larger than elements of $\mathcal{B} \cap \mathcal{B}'$ it follows that the only non-zero rows of the matrix are those corresponding to the intersection of the two bases. Observe that the leading non-zero entry of the row corresponding to $z_i \in \mathcal{B} \cap \mathcal{B}'$ is precisely the 1 in the column corresponding to z_i. These leading entries are distinct, so it follows that all of the rows corresponding to $\mathcal{B} \cap \mathcal{B}'$ are linearly independent. In particular, the rank of the matrix is the size of this intersection.

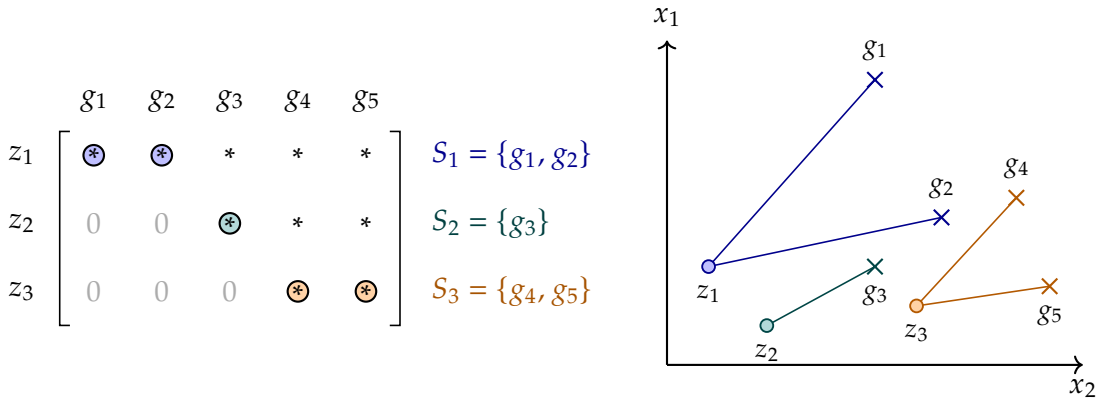
To conclude, we simply note that

$$\mathcal{B} \cap \mathcal{B}' = \{z_i \in Z : \text{gr}(z_i) \leq w \text{ and } \nexists g \in S_i \text{ s.t. } \text{gr}(g) \leq w'\} \tag{2.72}$$

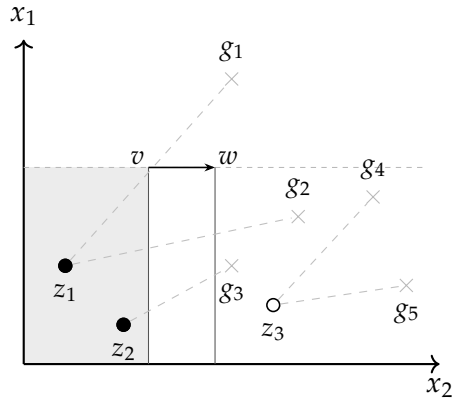
as required.

□

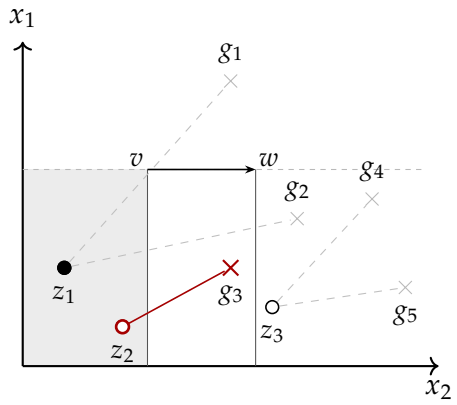
Let's see how to compute the rank invariant from a slice-consistent Gröbner basis using Theorem 2.3.5. Consider the following example generator-syzygy arrangement. Generators are represented by circles, syzygies by crosses, and $g_j \in S_i$ is represented by a line drawn from g_j to z_i:



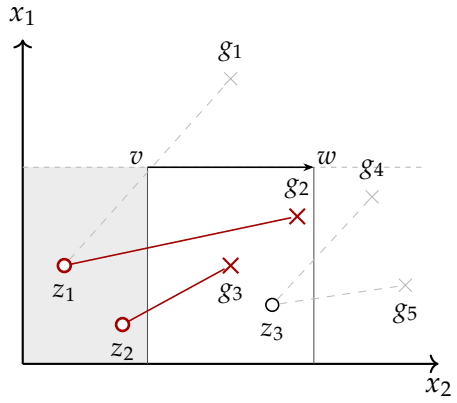
We now compute the rank invariant $\rho(v, w)$ for various choices of v and w. In each diagram, a solid black circle indicates a generator that is contributing to the rank invariant, and a red cross indicates a syzygy that has killed its assigned generator. In the following diagram, there are two active generators z_1, z_2 with grades $\leq v$ which do not have any syzygies lying below w. The generator z_3 is not active, since $\mathrm{gr}(z_3) \not\leq v$. So, the rank invariant $\rho(v, w)$ is 2:



Let's see what happens if we shift w to the right. Now we have $g_3 \leq w$, meaning that z_2 is killed. Although g_1 lies to the left of w, it does not lie below it, so g_1 does not kill z_1. We can therefore compute $\rho(v, w) = 1$.



If we shift w even further to the right, we find that $g_2 \leq w$, killing z_1. The rank invariant is now 0.



2.3.3 Computing persistence landscapes

We will now use the results in the previous section in order to compute persistence landscapes. We refer to the collection $\{(z_i, S_i)\}_{1 \leq i \leq n}$ as a *sparse representation* of the persistence landscape of H. Using Theorem 2.3.5 we can formulate an algorithm (Algorithm 2.1) to evaluate the landscape from a sparse representation.

Theorem 2.3.6. *Algorithm 2.1 is correct.*

Proof. Recall that we wish to return the largest value of t such that $\rho(v - te_r, v + te_r) \geq k$. By Theorem 2.3.5, this is the largest t such that there are at least k generators $z_i \in Z$ satisfying $\mathrm{gr}(z_i) \leq v - te_r$ and for which there are no corresponding syzygies $g \in S_i$ satisfying $\mathrm{gr}(g) \leq v + te_r$. If no such t exists then we return 0.

For each $z_i \in Z$, set $t_i \geq 0$ to be the largest value of t such that $\mathrm{gr}(z_i) \leq v - te_r$ and such that there is no $g \in S_i$ with $\mathrm{gr}(g) \leq v + te_r$. If no such t exists then we set $t_i = 0$. Write D for the set of all such t_i. Then it can be seen that $\lambda_k^{e_r}(v) = \mathrm{kmax}(D)$.

Algorithm 2.1 EvaluateLandscape

Input: A sparse representation $\{(z_i, S_i)\}$ of a persistence module H, a grade $v \in \mathbb{Z}^r$, and a positive integer $k \in \mathbb{N}$.

Output: The evaluation $\lambda^{e_r}(k, v)$ of the kth multiparameter persistence landscape of H at v in the e_r direction.

$D \leftarrow \emptyset$
for $z_i \in Z$ such that $\mathrm{gr}(z_i) \leq v$ **do**
  $\ell \leftarrow v_r - \mathrm{gr}(z_i)_r$
  $h \leftarrow \infty$ ▷ *How big can h be and have $\mathrm{gr}(g) \not\leq v + he_r$ for every $g \in S_i$?*
  for $g \in S_i$ in colexicographical order **do**
    if $\mathrm{gr}(g)_j \leq v_j$ for all $1 \leq j < r$ **then**
      $h \leftarrow \mathrm{gr}(g)_r - v_r - 1$ ▷ $\mathrm{gr}(g) \leq v + (h+1)e_r$
      break ▷ *Any later g will have a bigger $\mathrm{gr}(g)_r$*
  append $\max(\min(\ell, h), 0)$ to D
return $\mathrm{kmax}(D)$

We now claim that the outer loop of Algorithm 2.1 computes the set D. In particular we claim that for each z_i considered, the corresponding value t_i is appended to the list D. To see this, first observe that if $g \in S_i$ has $\mathrm{gr}(g)_j > v_j$ for any $1 \leq j < r$ then clearly $\mathrm{gr}(g) > v + te_r$ for all $t \geq 0$. However if $\mathrm{gr}(g)_j \leq v_j$ for all $1 \leq j < r$ then, setting $h_g = \mathrm{gr}(g)_r - v_r$ we find that $\mathrm{gr}(g) \leq v + h_g e_r$, so that $t_i \leq h_g - 1$. Moreover, if $\mathrm{gr}(g') \geq_{\mathrm{colex}} \mathrm{gr}(g)$ then $h_{g'} \geq h_g$. It therefore follows that if g_0 is colex-minimal over all $g \in S_i$ with $\mathrm{gr}(g)_j \leq v_j$ for all $1 \leq j < r$ then $\mathrm{gr}(g) \not\leq v + (h_{g_0} - 1)e_r$ for all $g \in S_i$. We conclude that t_i is indeed the value appended to D. □

Proposition 2.3.7. *The runtime of Algorithm 2.1 is $O(mn)$ where $m > 0$ and $n > 0$ are the sizes of the bases for P_1 and P_0 respectively in a minimal presentation $P_1 \to P_0 \to H$ for H.*

Proof. Within each iteration i of the outer loop there are $O(m)$ syzygies $g \in S_i$ to check, and therefore $O(m)$ iterations of the inner loop which itself takes constant time. There are n iterations of the outer loop and so the outer loop takes total time $O(mn)$. Furthermore checking for the kmax takes time $O(n)$ and it follows that the overall runtime is $O(mn + n) = O(mn)$. □

We have implemented Algorithm 2.1 in a fork of Muphasa [BGLb] available at `https://github.com/katherine-benjamin/muphasa/`. In particular, the imple-

mentation allows for the computation of persistence landscapes for the Interlevel-Rips-DMS trifiltration defined on a spatiotemporal trajectory in Definition 2.1.8. This implementation is joint work with Oliver Gäfvert and Silviana Amethyst.

2.3.4 Multi-directional sparse representations and diagonal landscapes

While Algorithm 2.1 is capable of computing the rank invariant in the e_r direction, if one reorders the parameters and recomputes the minimal presentation with GBS it is possible to compute the rank invariant in the e_j direction for any $1 \leq j \leq r$. We remark that doing so will not change the basis Z. Write $(S_i^j)_{i=1}^n$ for the sparse representation corresponding to the parameter j. Then we can combine these sparse representations into a multi-directional sparse representation:

$$\left(z_i, \left(S_i^j\right)_{j=1}^r\right)_{i=1}^n. \tag{2.73}$$

It turns out that this information is sufficient to compute the rank invariant in arbitrary directions.

Theorem 2.3.8. *Let $(z_i, (S_i^j)_j)_i$ be a multi-directional sparse representation as described. Take grades $w \leq w'$ such that $w' - w = (t_1, \ldots, t_r)$. Then the rank invariant $\rho(w, w')$ of H is equal to the number of z_i satisfying:*

1. $\operatorname{gr}(z_i) \leq w$, *and*
2. *for every $1 \leq j \leq r$ there exists no $g \in S_i^j$ satisfying $\operatorname{gr}(g) \leq w + \sum_{k=1}^j t_k e_k$.*

Proof (Sketch). The idea is to apply the proof of Theorem 2.3.5 multiple times, once each in the direction of each parameter. One can construct bases as in Lemma 2.3.2 for H_{v_j} at each of the grades $v_j := w + \sum_{k=1}^j t_k e_k$. Using matrices for the maps $H_{v_j} \to H_{v_{j+1}}$ one can track the lifespan of the basis element corresponding to each of the active z_i in H_w, and the generators that are not reduced to zero are precisely those satisfying the stated property. □

We can use this result to adapt Algorithm 2.1 to compute the diagonal landscape instead of the horizontal landscape; see Algorithm 2.2.

Algorithm 2.2 EvaluateDiagonalLandscape

Input: A multi-directional sparse representation $(z_i, (S_i^j)_j)_i$ of a persistence module H, a grade $v \in \mathbb{Z}^r$, and a positive integer $k \in \mathbb{N}$.

Output: The evaluation $\lambda(k, v)$ of the (diagonal) kth multiparameter persistence landscape of H at v.

$D \leftarrow \emptyset$
for $z_i \in Z$ such that $\mathrm{gr}(z_i) \leq v$ **do**
  $\ell \leftarrow \min_{1 \leq j \leq r}(v_j - \mathrm{gr}(z_i)_j)$
  $h \leftarrow \infty$
  for $j \in \{1, \ldots, r\}$ **do**
    for $g \in S_i^j$ **do**
      $h \leftarrow \min\big(h, \max_{k \leq j}(\mathrm{gr}(g)_k - v_k - 1)\big)$
  append $\max(\min(\ell, h), 0)$ to D
return $\mathrm{kmax}(D)$

2.3.5 Benchmarking

In this section we aim to briefly investigate the empirical efficiency of the sparse representation for *densifying* persistence landscapes, i.e. evaluating them across a densely sampled grid. We set up three different 3D spatiotemporal trajectory examples, each with 10 timesteps and a variable number of agents $8 \leq n_{\text{agents}} \leq 96$. The examples are:

1. Random: at each time point every agent samples a new position at random from the unit cube;
2. Rotating ring: initial positions are noisily sampled from a ring of radius 1 on the xy plane. The points are then rotated about the z axis.
3. Brownian motion: initial positions are sampled from a Gaussian and then each agent walks with independently sampled Gaussian increments.

For each trajectory, we used GBS to compute a minimal presentation for the homology of its Interlevel-Rips-DMS filtration as in Definition 2.1.8. We then evaluated the first landscape layer λ_1 at $10 \times 10 \times 100$ grades for each of these minimal presentations, using both Algorithm 2.1 (Sparse) and the standard algorithm based on explicit computation of the rank invariant (Naive). Experiments were run on a single core of an Apple MacBook Pro (M5 Pro chip, 24 GB memory). The results are presented in Figure 2.6. Algorithm 2.1 is able to densify every landscape in the data

set in less than 10 ms starting from a minimal presentation, representing as much as a 100,000 x speedup over the naive computation at the upper end.

To put these results in context, the sparse representation is taking advantage of the fact that the GBS algorithm has done much of the work for us in terms of computing the rank invariant. Naively computing the rank invariant by row reduction is essentially recomputing much of the information already stored within the GBS output. By comparison, Algorithm 2.1 uses the GBS structure to simply read off the landscape values without performing any linear algebra.

2.4 Application to swarm dynamics

Swarm dynamics simulations model the spatiotemporal behaviour of a collection of self-propelled agents. The instantaneous spatial arrangements of these swarms may not fully encapsulate the subtlety of their behaviour, and so we naturally turn to methods which take into account the swarm dynamics as well. The goal of this section is to use persistence landscapes of the Interlevel-Rips-DMS trifiltration (Definition 2.1.8) to infer the parameter values of an instance of a certain swarm simulation using only its associated spatiotemporal trajectory. This work follows previous topological approaches to the same problem (see Section 2.1.4.1) [Bha+19; GL23]. Our aim here is not to outdo the state of the art for this task, but rather to provide a proof of concept for 3PH landscapes in this problem setting, and to explore some of the implementation details that arise in practice. Indeed in [GL23, Section 7.6] it is remarked that static methods, which do not take into account any of the dynamics of the simulation, perform competitively for the specific regression task in question.

Remark 2.4.1. The word 'parameter' is unfortunately overloaded in this setting. From now on, *parameter* will always refer to the parameters of the underlying swarm model. We will say *filtration indices/values* to refer to the three persistence parameters of the Interlevel-Rips-DMS trifiltration as in Definition 2.1.8.

Model description For this experiment we used a data set of swarm simulations from [GL23]. They used the 3D D'Orsogna model [DOr+06] which simulates agents interacting in $\mathbb{R}^3$, where each agent is modelled with propulsive and drag effects as well as both repulsive and attractive interactions with the other agents.

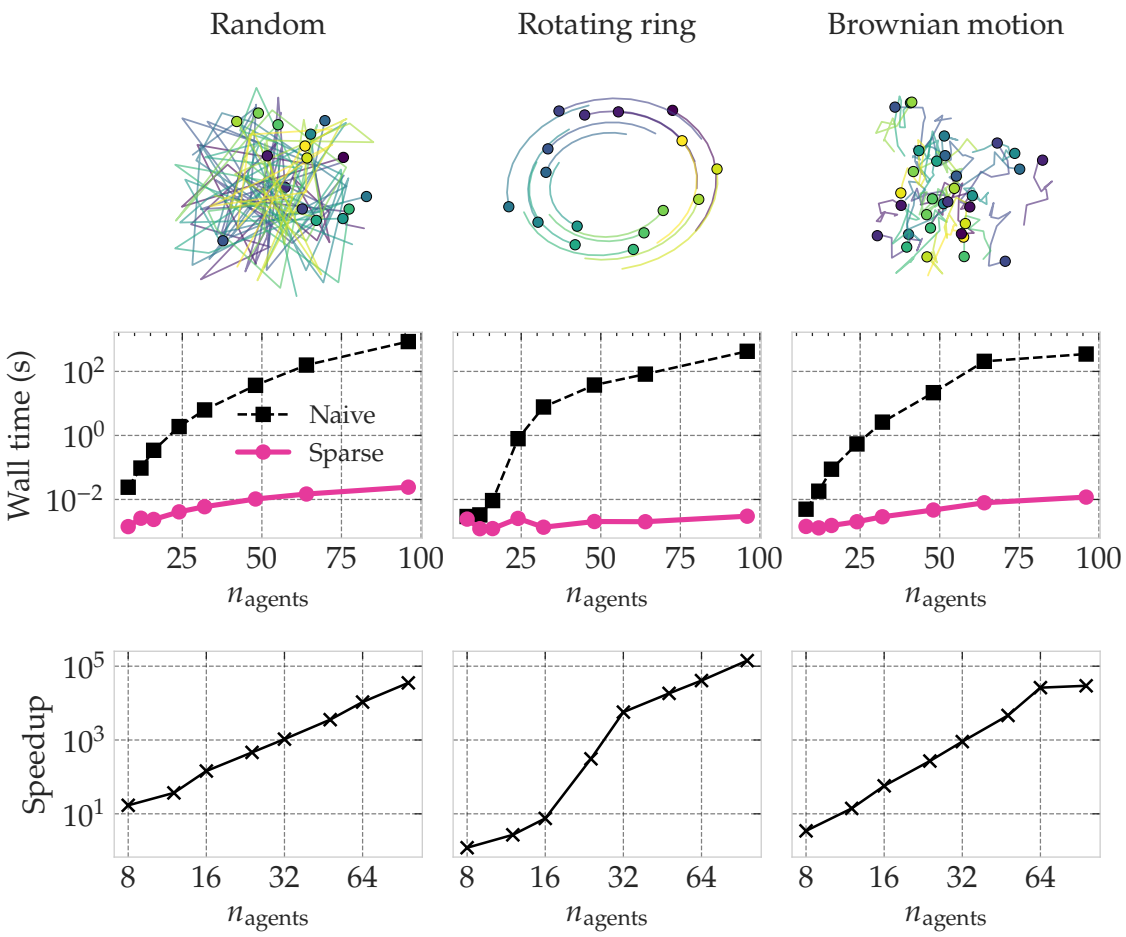


Figure 2.6. Benchmarking the sparse algorithm for computing the persistence landscape. Each column represents a different trajectory example. Top row: example trajectories. Middle row: total wall time to densify a landscape by number of agents for the naive rank algorithm (Naive; black dashed) and Algorithm 2.1 (Sparse; pink). Bottom row: Log-log plots of speedup vs number of agents.

In detail, the 3D D'Orsogna model is governed by the following fully deterministic equations of motion:

$$\frac{\partial \vec{x}_i}{\partial t} = \vec{v}_i, \tag{2.74}$$

$$m\,\frac{\partial \vec{v}_i}{\partial t} = \left(\alpha - \beta|\vec{v}_i|^2\right)\vec{v}_i - \vec{\nabla}_i U(\vec{x}_i), \tag{2.75}$$

with generalised Morse potential U given by

$$U(\vec{x}_i) = \sum_{j\neq i}\left[C_r\, e^{-|\vec{x}_i-\vec{x}_j|/\ell_r} - C_a\, e^{-|\vec{x}_i-\vec{x}_j|/\ell_a}\right]. \tag{2.76}$$

The motion of each agent is governed by the fixed per-agent mass m, a propulsion effect with strength α, a velocity-dependent drag effect with strength β, and a potential U given by the cumulative effect of pairwise interactions with each other agent. These agent-agent interactions themselves are governed by attractive and repulsive components with strengths C_a, C_r and interaction scales ℓ_a, ℓ_r respectively. The model has no boundary conditions; only initial positions are specified.

Giusti and Lee produced 500 instances of the simulation over 400 time steps as follows. They first fixed the mass $m = 1.0$, propulsion strength $\alpha = 1.0$, and drag strength $\beta = 0.5$ at standard values. The initial positions of the $N = 200$ agents were chosen uniformly at random in the unit cube $[0,1]^3$ with initial velocity sampled from a Gaussian. After nondimensionalising, the remaining two free parameters—interaction strength $C \coloneqq C_r/C_a$ and characteristic length $\ell \coloneqq \ell_r/\ell_a$—were independently chosen uniformly at random from the interval $[0.1, 2]$. Unbounded simulations, defined as a simulation where any agent has moved more than 40 units from its start position by the halfway mark, were discarded, and the simulation was repeated until 500 bounded instances were produced. The resulting distribution of parameters is shown in Figure 2.7. Note that the upper-right quadrant of the sample space did not produce any bounded simulations, because when both $C > 1$ and $\ell > 1$ the repulsive force both is stronger and acts at larger scales than the attractive force, leading to dispersion.

Persistence landscape vectorisation A persistence landscape is a sequence of functionals $\lambda_k : \mathbb{R}^r \to \mathbb{R}$ on filtration-index space. The first step in using a persistence landscape in machine learning is to convert these functionals into a finite feature vector, and this is typically achieved by evaluating the landscape at a fixed set of

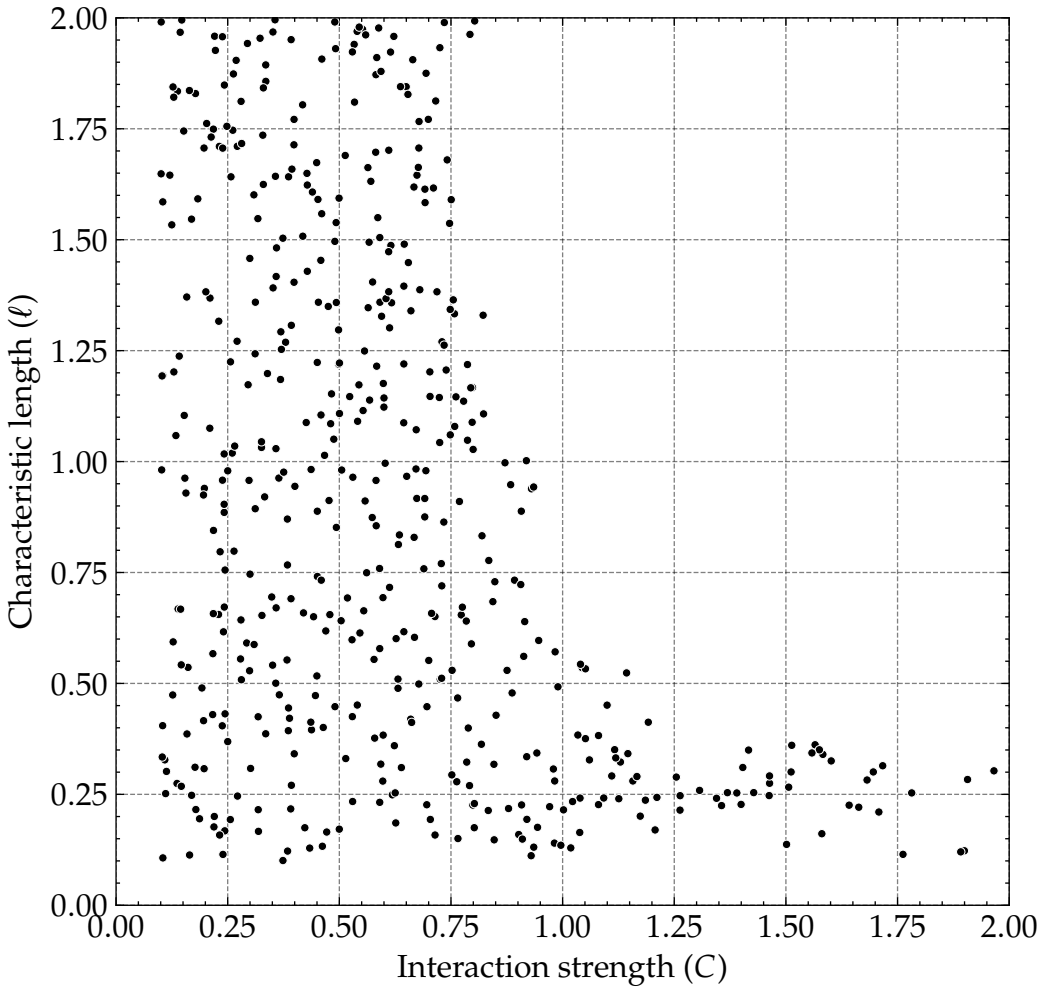


Figure 2.7. Parameter values of bounded D'Orsogna swarm simulations from [GL23]. Each point corresponds to a single instance of the simulation ($n = 500$). Note the upper-right quadrant is empty because those parameter choices lead to unbounded simulations.

discrete filtration index values. While this is relatively straightforward in the case of single and 2-parameter persistence, in r-persistence with $r \geq 3$ we begin to encounter a curse of dimensionality: if n discretised values of each filtration index are required for each of k landscape layers then the final vector will have $k \times n^r$ features. Even in the case that $n = 10$ and $k = 1$ we have only 10 features in the 1PH setting, but this explodes into 1,000 features in the 3PH setting. Furthermore, features in the discretised landscape will exhibit strong correlation effects: because persistence landscapes are continuous and Lipschitz (Proposition 2.1.16), evaluations of the landscape at neighbouring filtration-index values will be highly correlated. Both of these effects can lead to overfitting and therefore need to be resolved before any downstream analysis can be carried out.

To address these two concerns, we propose using principal component regression: we first perform principal component analysis (PCA) to reduce dimensionality and remove collinearity in the feature vectors, and then carry out a regression on the extracted principal components. This allows us to fully cover the filtration space without having to work with a high number of features or excessive correlation. To keep the model simple and interpretable, here we use a linear regression on the principal components.

Subsampling and vectorisation To keep performance tractable, we downsampled each simulation to 50 agents and 20 time points. Agents were selected uniformly at random for each instance, and we tested two different methods of temporal downsampling as suggested in [GL23]: the initial 20 time points, and a random subset of 20 time points sampled uniformly. We will focus on the random subsample for the analysis in this section; the initial subsample performed similarly. We then used our fork of Muphasa [BGLb] to compute sparse representations of H_1 of the Interlevel-Rips-DMS filtration (Definition 2.1.8) for each subsampled instance.

Choosing the discretisation of filtration space is non-trivial. For the two interlevel temporal filtration indices $1 \leq t_1 \leq t_2 \leq 20$ we choose to use the full set of $1 + \cdots + 20 = 210$ possible filtration values. For the Rips index, we took 100 values uniformly in the range $[0, 1]$. An alternative approach, with the aim of reducing the size of the feature vector, would be to follow [GL23] and use a log scale, the idea being that information is more densely packed at smaller scales. However, given that we will be performing PCA to reduce dimensionality downstream in any case, we felt this was an unnecessary optimisation. We evaluated just the first landscape

layer λ_1 in the direction of the Rips index (Definition 2.1.15), leading to a feature vector of length $210 \times 100 \times 1 = 21{,}000$ as the input for PCA.

Regression analysis After forming an 80/20 train/test split, we used 5-fold cross-validation to select the number of principal components which was fixed at $n = 30$. We then performed two linear regressions on the principal components to predict the parameters C and ℓ respectively. Figure 2.8 summarises the results of these two regressions on the test set. For C and ℓ the coefficient of determination R^2 is 0.69 and 0.64 respectively, and the mean squared error (MSE) is 0.05 and 0.13 respectively. We remark that the MSEs are very similar to those obtained from path signature methods as reported in [GL23, Figure 8]. Inspecting the squared residuals for each prediction, we can see that the model tends to underestimate both the interaction strength C and the characteristic length ℓ when their respective ground truth parameters take high values. Additionally, when C is very low and ℓ is reasonably low the model tends to overestimate ℓ.

We can pull the linear regression coefficients back to the original filtration index space to visualise the contribution of different time points and scales to the estimated simulation parameters. Figure 2.9 shows how the regression coefficient varies across choices of time interval $[t_1, t_2]$ at three different choices of scale $r \in [0.1, 0.4, 0.7]$. These visualisations allow us to determine which scales and time points are informing the inferred parameters. As an example, we can see that high 'static' persistence corresponding to diagonal points $[t_1, t_1]$ when $r = 0.1$ is indicative of a high C coefficient and a low ℓ coefficient. One interpretation of this behaviour is that swarms with a high interaction strength C tend to instantaneously exhibit more 1-dimensional homological features at the $r = 0.1$ scale. It is less straightforward to see how to interpret the off-diagonal points. One slightly weak claim that can be made is that a high ℓ coefficient causes *different* behaviour at the $r = 0.1$ scale towards the end of the simulation, for example, as indicated by the green entries at high t_1 values in the ℓ coefficient plot at this scale. It is not clear exactly *what* that difference is, however.

One potential concern is that the norm $\|\lambda_1\| := \int_{\mathbb{R}^3} \lambda_1$ is the dominant source of variation in the feature vectors, and that this is hindering the regression by drowning out more subtle variation. Indeed, it can be seen that the sum of all features, which is an approximation of $\|\lambda_1\|$, is correlated with both the parameter C and the first principal component PC1 (Figure 2.10). However, we obtained similar results even after normalising each feature vector to have unit sum. One possible explanation

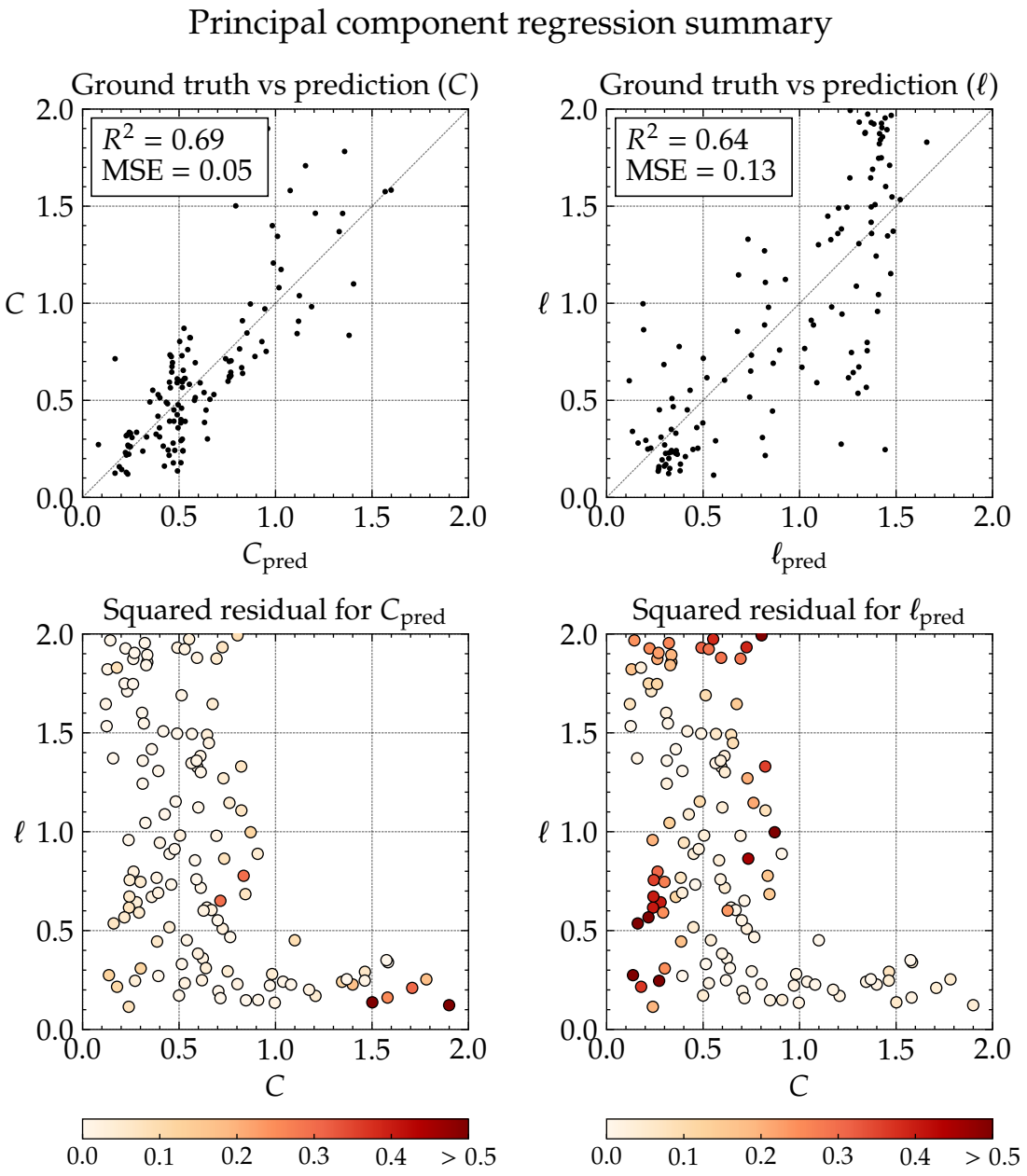


Figure 2.8. Principal component regression of D'Orsogna swarm parameters. Top row: ground truth parameter values vs predicted values for the C (left) and ℓ (right) parameters. Bottom row: test parameter pairs coloured by squared residual of predictions for C (left) and ℓ (right).

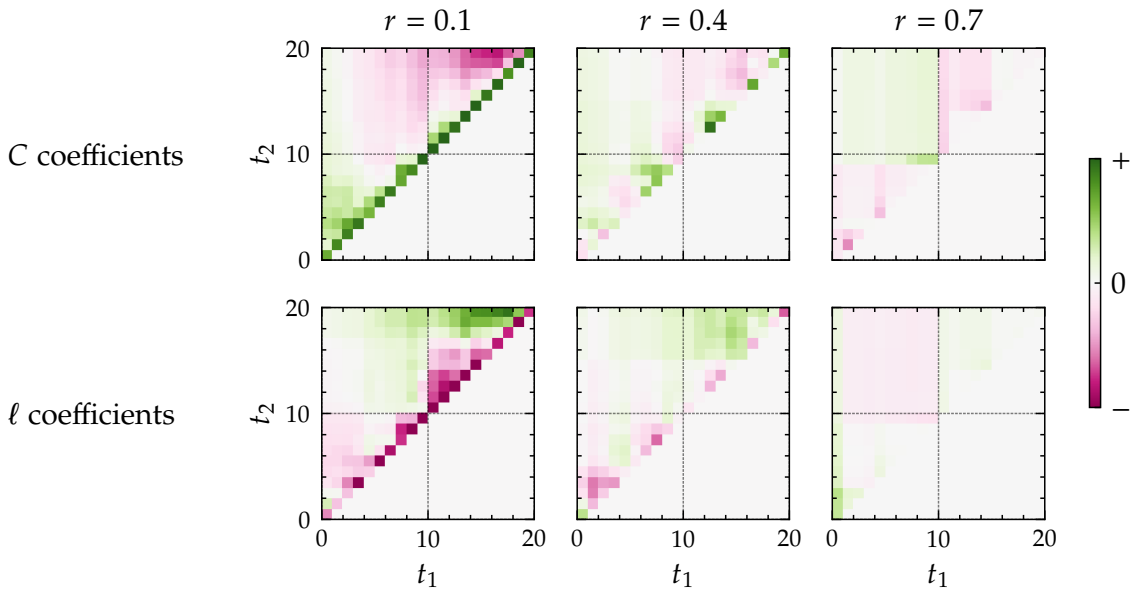


Figure 2.9. Regression coefficient visualisation for 3PH analysis of D'Orsogna swarm parameters. Each plot shows the regression coefficients for each time interval $[t_1, t_2]$ at a specific scale value r. Positive coefficient values (green) indicate filtration values that positively inform the target swarm parameter, and negative coefficient values (pink) indicate filtration values that negatively inform it. Top row: interaction strength C. Bottom row: characteristic length ℓ.

for this behaviour is that the high norms at certain choices of simulation-parameter value are caused by the existence of a few highly persistent features at a certain filtration value, and that it is the presence of this feature that is being picked up in the regression. So, while on this data set it is not necessary to normalise each feature vector, it is possible that this may be a necessary step in other contexts.

As another control for the possibility that the norm is dominating the variance in PCA, we also implemented partial least squares (PLS) regression as an alternative to principal component regression. PLS performs a similar function to principal component regression, the difference being that the learned components are informed by the dependent variables of the regression. We again used 5-fold cross-validation to select the number of components, which this time was fixed at $n = 8$. Figure E1 summarises the results of the PLS regression. We can see that the R^2 coefficients and the MSEs remain largely similar to those of the principal component regression. Moreover, the failure points for the regression are at the same ground truth parameter pairs. So, it seems to be the case that the areas of poor regression performance are not caused by the choice of dimensionality reduction.

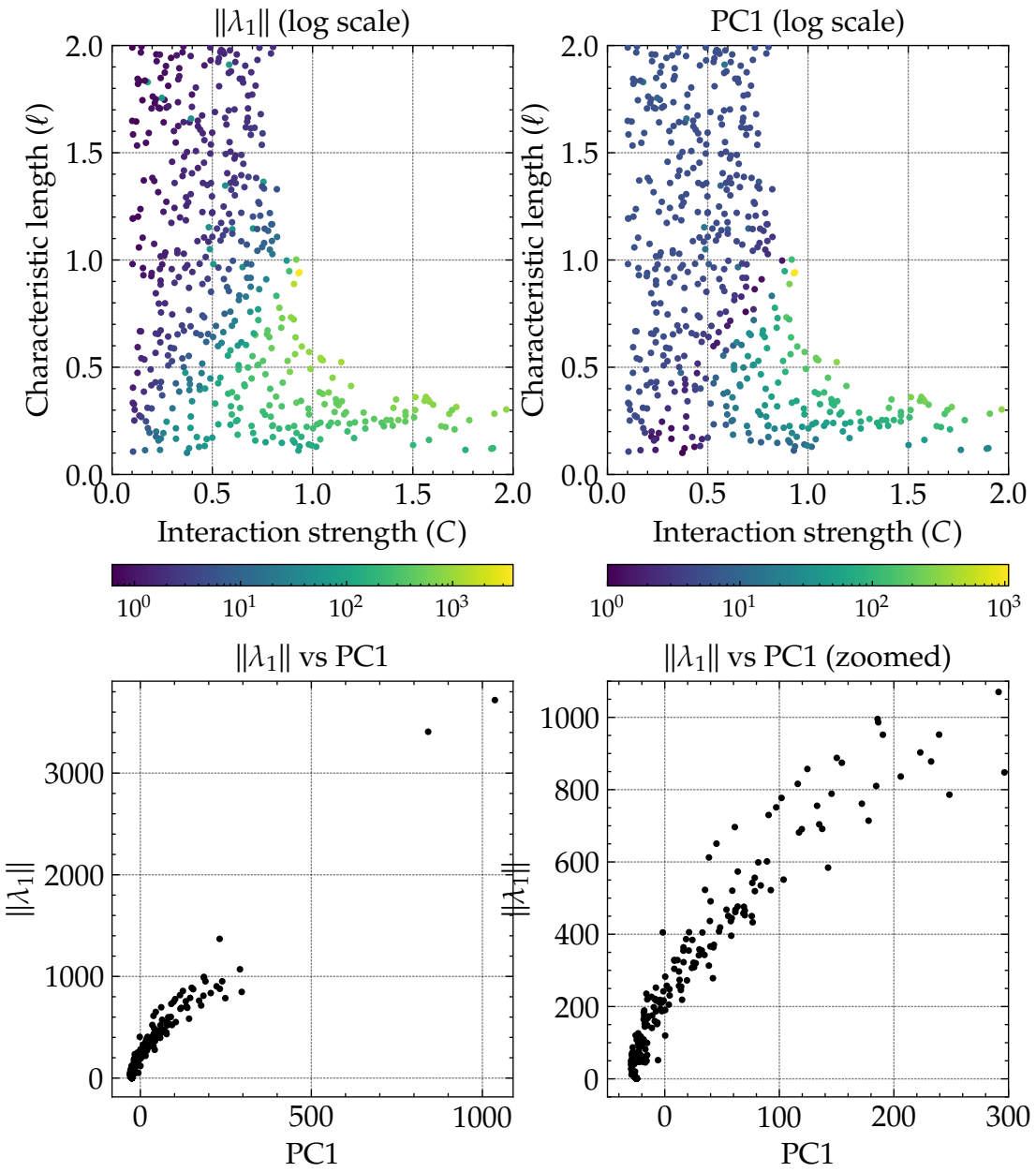


Figure 2.10. The relationship between the swarm parameters C and ℓ, the approximate landscape norm $\|\lambda_1\|$, and the first principal component PC1. Top row: the swarm model parameter space coloured by $\|\lambda_1\|$ (left) and PC1 (right). Both colourings are in a log scale, with PC1 shifted so that the minimum value is 1. Bottom row: $\|\lambda_1\|$ vs PC1. The panel on the right is a zoomed version of the panel on the left.

To better understand the relationship between the 3PH landscapes and model behaviour, we also used principal component regression to predict the late-stage time-averaged radius of each swarm [Chu+07]. To be precise, for a point cloud X we define its *radius* to be the maximal distance from any point to its centre $\overline{X}$:

$$r(X) = \max_{x \in X} \left| x - \overline{X} \right|. \tag{2.77}$$

For a swarm $(X_t)_{t=1}^{400}$ we then take its time-averaged radius $R(X) \coloneqq \frac{1}{40} \sum_{t=361}^{400} r(X_t)$ to represent the late-stage dispersion of the point cloud. We used principal component regression on $n = 6$ principal components (selected using 5-fold cross-validation) to predict R from the dense landscape vectors as before. The regression results are presented in Figure 2.11. We recover $R^2 = 0.98$, indicating that the dense landscape contains almost all of the information of the radius of the swarm.

2.5 Discussion

In this chapter we have built on work of Bender *et al.* [BGLa] to propose a new algorithm for computing the rank invariant and multiparameter persistence landscape of r-parameter persistence modules for $r \geq 3$. Starting from a slice-consistent minimal presentation of the persistence module, we form its sparse representation using the pivots of the presentation matrix. From the sparse representation we showed how to compute the rank invariant, and consequently the persistence landscape, using a simple counting criterion (Theorem 2.3.5). We showed that this approach allows for the computation of an entire dense persistence landscape in milliseconds starting from a minimal presentation, in comparison to many minutes with the naive approach. We then applied this work to compute Interlevel-Rips-DMS filtrations (Definition 2.1.8) from swarm simulations obtained from the D'Orsogna model as in [GL23]. We demonstrated that the combination of 3PH landscapes with principal component regression is competitive with path signature methods for the task of inferring the parameters of a simulation from its spatiotemporal trajectory. We also demonstrated that the 3PH landscape fully captures the information of the radius of each swarm trajectory. Moreover, by pulling the coefficients of the regression back to the spatiotemporal parameter space we can visualise the effect that varying the parameters has on the swarm dynamics. On the D'Orsogna task this visualisation shows that the long-range inputs unique to the 3PH filtration are providing useful signal for the regression.

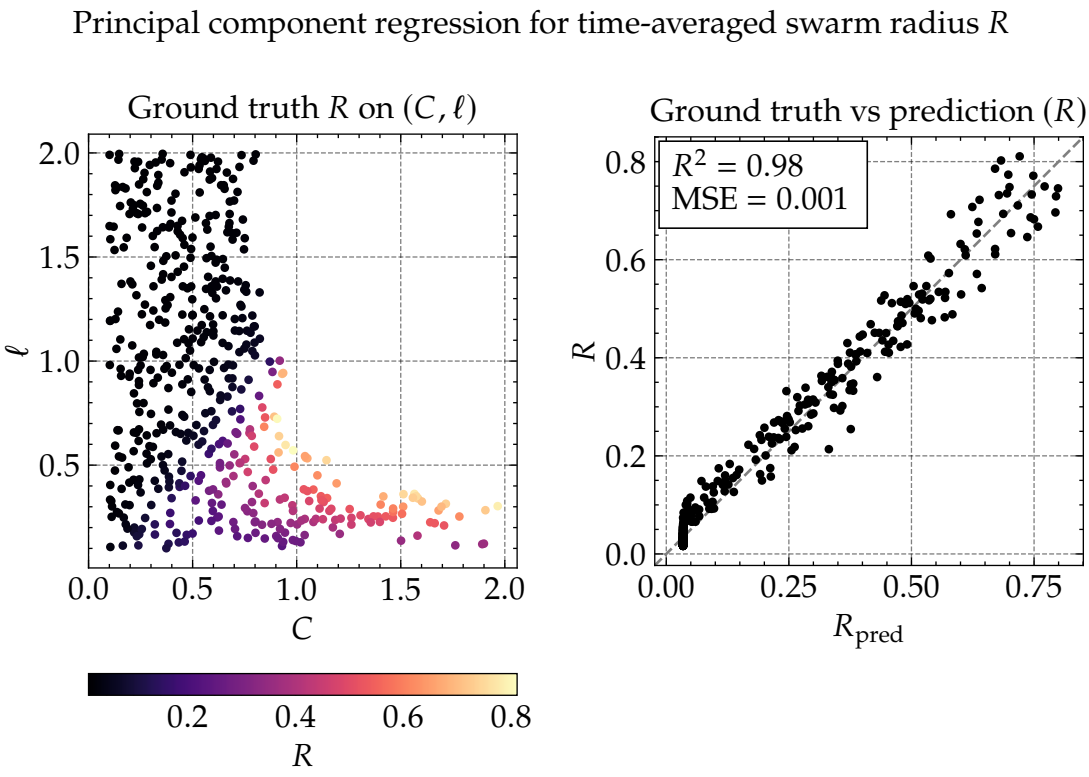


Figure 2.11. Principal component regression of time-averaged swarm radius. Left: (C, ℓ) parameter space coloured by time-averaged swarm radius. Right: ground-truth vs predicted radius values.

The sparse representation $(z_i, S_i)_{i=1}^n$ has another interpretation. Note that the grades of the generators and syzygies are entirely determined by the 0th- and 1st-degree multigraded Betti numbers of the persistence module H. Let those sets of grades be β_0 and β_1 respectively. We can therefore view the sparse representation in a given direction as a choice of assignment $\beta_1 \to \beta_0$ which says to which partition each element in β_1 belongs. It is very natural to ask: is this assignment canonical in some sense, such that the (multi-directional) sparse representation represents an invariant of the module H? We are unable to answer this question in either the affirmative or negative in this thesis, but we offer some observations.

Firstly, the sparse representation as defined in this thesis depends on the output of the GBS algorithm. It is not at all obvious that this algorithm, when given two different representations of the same homology module as input, should produce a Gröbner basis with the same assignments of pivots. So it is certainly not immediate that the sparse representation should be an invariant. To tackle this problem, then, we propose attempting to define the sparse representation by some property that uniquely determines it among all possible assignments $\beta_1 \to \beta_0$, and then show that the representation output by GBS satisfies this property. In this chapter we have already proven one candidate property, namely that the criterion (2.68) computes the rank invariant. However, it is fairly straightforward to come up with an example where two different assignments $\beta_1 \to \beta_0$ both allow one to compute the rank invariant. Whether the sparse representation as output by GBS is an invariant or not therefore remains an open problem.

We briefly comment on why it would be desirable to show that the sparse representation is indeed an invariant. Recall Figure 2.2, which shows two 2-parameter persistence modules with the same rank invariant. These two modules are also indistinguishable from their multigraded Betti numbers. However, their sparse representations are different, because they keep track of *which* generator is killed, not only *that* a generator is killed. If the sparse representation is an invariant, then, it would be strictly stronger than both the rank invariant and the combination of the 0th and 1st Betti numbers.

Another adjacent question to the question of invariance is whether there exists a metric, or other suitable notion of distance, directly computable on sparse representations. At present, to compare two persistence modules from their sparse representations, one must first compute the dense persistence landscape of each module to compare as vectors. This densification gives up the advantages of storing the module in a sparse form, and incurs a not-insignificant computational cost.

Comparing modules directly via their sparse representations would therefore be a significant computational benefit. Finding such a metric is an open problem, and indeed for it to pull back to a metric on the underlying modules it first needs to be shown that the sparse representation is itself an invariant.

On the application side, there are a wide range of additional data sets which could be amenable to 3PH analysis. We have previously used 1PH to analyse biomolecular structure in the context of knotted proteins [Ben+23], where we showed that the persistence landscape of the Vietoris-Rips filtration defined on a protein's C_α backbone is a rich isometry invariant of the protein's primary structure. A natural extension to this work is therefore to apply 3PH landscapes as a fingerprint of a protein's molecular dynamics trajectory, and this is the subject of ongoing work. Outside the realm of spatiotemporal data, another natural source of applications is multiplex imaging data, where multiple markers (e.g. protein markers for cell types) in a tissue are imaged simultaneously. For such a data set with n markers, one can construct an $(n + 1)$-parameter filtration where each of the first n parameters is a superlevel filtration on the level of intensity of the respective marker and the final parameter is a Rips-like scale parameter. It would be interesting to compare this approach to other recent PH extensions which have been applied to this type of data set, which include PWDS visualisation [Tor+25] and Dowker/Witness complexes [Sto+24].

Chapter 3

Background: methods in transcriptomics

In its simplest form, the Central Dogma of Molecular Biology states that genetic information within eukaryotic cells flows through the following pipeline:

$$\text{Genome} \xrightarrow{\text{Transcription}} \text{Transcriptome} \xrightarrow{\text{Translation}} \text{Proteome.}$$

As mathematicians, we can think of the genome as being a set of genes $\mathcal{G}$, fixed for each organism, which are recorded in DNA. Through the process of transcription, genes are *expressed*, many times each, as individual RNA molecules called *transcripts*. The multiset of these transcripts is called the cell's *transcriptome*, and this will vary per cell. Each gene is in effect a code for the production of a specific protein, and in the final stage the genes expressed as RNA are *translated* into proteins, which then go on to perform functions in the cell. In this sense, we can view a cell's transcriptome as being a proxy for its biological function. *Transcriptomics* is precisely the measurement of the transcriptome.

The process of transcription is intricate and highly sensitive to environmental conditions. *Transcription factors* are special proteins that modify the rate of transcription of certain target genes, by either promoting their transcription (*activators*) or inhibiting it (*repressors*). The mutual effects of these transcription factors, along with other information in the cell, cause the cell's transcriptome to constantly vary. The collection of all genes along with their activation/repression relations is called the *gene regulatory network*, and is determined by an organism's genotype. The viewpoint of epigenetics is that the transcriptomic state of a cell can be modelled as a dynamical system which evolves according to the aggregate pressures on transcription encoded

by the organism's gene regulatory network [GW09]. It is by this mechanism that a cell's phenotype can develop from one state to another without any change to its genotype.

The dynamical system model of epigenetics makes natural the notion of a *cell type* as an attractor of the transcriptomic dynamical system. Remarkably, this model developed far earlier than the experimental tools that have later allowed us to confirm it. The British biologist Conrad Waddington introduced his concept of the 'epigenetic landscape', where cells are likened to balls rolling down a surface and settling in 'valleys', as early as 1957 [Wad57, pp. 29–30]. These valleys, the attractor states in our dynamical system model, should be thought of not only as the classical cell types identified by functional specialisation (immune, neuronal, muscle, etc.) but also as more subtle transcriptional subpopulations.

While the Waddington landscape offers a powerful mental model for cell differentiation, it does have some major deficiencies.[1] Chiefly, it is not the case that cells neatly organise in transcriptome space into separable cell types once an organism has matured. Instead, as biologist Michael Elowitz put it, 'cellular properties vary continuously as well as discretely, may not follow rigid hierarchies, and are highly dynamic' [Elo17]. In other words, cell states lie on continuous trajectories in gene expression space, and transitions from one cell type to another are commonplace. Nevertheless, partitioning of cell types remains a crucial tool for practitioners aiming to reduce the complexity of high-dimensional transcriptomics data sets.

Modern experimental technology allows us to measure the transcriptome of a tissue sample with incredible precision and efficiency. As we will see in this chapter, transcripts can be localised to their cell of origin (*single-cell* transcriptomics) or resolved *in situ* to recover their distribution in space (*spatial* transcriptomics). The end result of these experiments is the production of a *gene expression vector* associated to each biological unit (cell or spatial location), which identifies the unit in D-dimensional gene expression space where D is the size of the genome. Our task as mathematicians is to make sense of these vectors and turn them into useful insights into the underlying biology.

Having now established some intuition for the field of transcriptomics, the goal of this chapter is to provide a survey, for a mathematical audience, of the

[1] Of course, the theoretical frontier has been pushed forward greatly since Waddington's contribution, but a history of this work is outside the scope of this thesis. We refer the reader to the historical survey [AOM19], the review [Hua12], and for an especially mathematical flavour the textbook [Alo19].

various experimental approaches to transcriptomics of the modern day. As well as introducing the experimental platforms which measure the transcriptome, we will also discuss approaches to the analysis of the resulting data. The aim is that, by the end of this chapter, the mathematical reader will be equipped with all the prerequisite knowledge on the current state of the art in experimental transcriptomics to enable them to follow the contributions in the following chapters.

3.1 Experimental methods

The aim of this section is to establish the current state of the art in transcriptomics technologies, with a view towards data analysis questions. The aim of these methods is, simply put, to measure the genes expressed as RNA in a tissue sample—in other words, its transcriptome. We start with single-cell transcriptomics before discussing spatial methods. We refer the reader to the survey [MP22] for a much more comprehensive review of ST technologies.

3.1.1 Single-cell RNA sequencing

Single-cell/single-nucleus RNA sequencing (sc/snRNA-seq) broadly refers to a family of related technologies concerned with measuring the gene expression of individual cells. In brief, these technologies work by separating cells into droplets and then barcoding the RNA in each droplet. Next, following reverse transcription (RT) into cDNA and polymerase chain reaction (PCR) amplification, each fragment of barcoded cDNA is sequenced (i.e. its sequence of base pairs is read off) and the sequence is matched back to the barcode. The RT and PCR steps need to be carried out in a way that maximally preserves the relative abundance of the fragments of each gene. Once each cDNA fragment is recovered with its barcode, its sequence can be mapped to a reference genome to determine the gene of the original RNA molecule. The end result is that for each cell we have a record of the (relative) abundance of each of its genes expressed as RNA.

The difference between single-*cell* and single-*nucleus* RNA-seq is that in single-nucleus RNA sequencing (snRNA-seq) the cell nuclei are first isolated before suspension into droplets, so that only nuclear (as opposed to cytoplasmic) RNA is measured. There are various technical reasons why a practitioner may or may not want to do this, but they lie far outside the scope of this thesis.

As a more mathematical abstraction, one can think of the output of an sc/snRNA-seq experiment as being a *count matrix*: a matrix with rows indexed by cells and columns indexed by genes, whose entries record the detected RNA transcript count of a given gene in a given cell. We will work directly from this abstraction in this thesis, and not concern ourselves with the (very important) initial preparation and preprocessing steps.

3.1.2 Imaging-based spatial methods

The most well-established approach to ST profiling is to construct fluorescent probes which bind to specific gene targets *in situ*. Roughly speaking, to detect a molecule of a given gene one can construct a complementary strand of RNA which hybridises to molecules of that gene. By dyeing these probes with a fluorescent label, it is possible to resolve the spatial locations of individual RNA molecules of the gene by taking images of the sample under a microscope. In practice the situation is much more complicated, for example to combat noise one actually needs to produce multiple distinct probes which bind to subsequences of the target gene, but this is the general idea. These techniques are broadly referred to as single-molecule fluorescent *in situ* hybridisation (smFISH) [Fem+98; Raj+08]. Multi-round versions of smFISH, where each gene is assigned a sequence of colours to be detected over multiple imaging rounds with different sets of probes, are commonly used to increase the number of genes which can be detected.

Imaging-based ST methods are typically characterised by very high detection efficiency ('near 100 %' [LBL17, p. 2]) and true single-molecule resolution. This comes at the cost of transcriptome depth, with panels typically limited to hundreds of genes, although some techniques such as MERFISH [Xia+19; Che+15] and seqFISH+ [Eng+19] are able to approach 10,000 genes.

The imaging-based method of the greatest relevance to this thesis is the Xenium platform developed by 10x Genomics [Jan+23]. In Xenium, a circular probe design which binds twice to each RNA target is used to increase the specificity of the detected transcripts. Xenium typically operates with a panel of around 300 genes [Mar+25], although 10x advertises functionality for 5,000 gene panels. Other features include the option for same-section proteomics and morphology staining.

3.1.3 Array-based spatial methods

Array-based methods take a spatial barcoding approach to sequencing. In essence, these technologies proceed very similarly to sc/snRNA-seq, except that barcodes are assigned to *spots* lying in a spatial array, as opposed to droplets. The upshot is that processed transcripts can be matched downstream to spots in the array, thus recovering the spatial location of each transcript. The exact mechanism of the array can vary, but that is beyond the scope of this thesis. A key point is that, because transcripts are sequenced in a similar fashion to sc/snRNA-seq, array-based transcriptomics platforms typically measure the whole transcriptome.

Table 3.1 shows the progression in spatial resolution of array-based ST technologies. Early technologies localised transcripts at multicellular (>10 µm) resolution. However, in the last four years platforms operating at subcellular (<5 µm) resolution have become increasingly common. For a more comprehensive comparison of array-based methods released up to 2024 we refer the reader to the survey [You+24].

The array-based platform of most relevance in this thesis is Stereo-seq [Che+22] by BGI, which achieves 500 nm resolution on capture arrays far larger than any of its competitors. Stereo-seq V2 was recently released [Zha+25] with the same spatial resolution but promising improved robustness to certain technical effects. We do not consider Stereo-seq V2 in this thesis, but all of the methods developed for ST would work equally well on the updated platform.

3.2 Cell-type classification

There are many different questions one can ask about a transcriptomics data set. In Chapter 4, we are chiefly interested in the problem of *cell-type assignment*. That is to say, given some biological unit (e.g. a cell or a subcellular region), can we identify the cell type to which it belongs? Embedded within this question are two key subproblems. Firstly, one must *discover* the cell types of interest. Only after the cell types have been defined can we *classify* the cell type of each unit. It turns out that each of these problems is highly non-trivial, and dependent on the underlying experimental modality.

Cell-type classification of spatial data sets is much less standardised than its single-cell counterpart. This is a result of the increased complexity of data with a spatial component, the more varied landscape of spatial experimental modalities, and the field's relative lack of maturity. Typically one first discovers cell types on a

Table 3.1. Comparison of resolution and capture area for array-based ST technologies. For platforms with an associated journal publication, the year is the date of that publication. For commercial platforms without an associated journal publication, the year is the date of first announcement. Note that alongside progress in spatial resolution and size, technologies have also progressed by improving capture efficiency, cost, and other parameters not listed here.

Method	Year	Spot diameter (µm)	Spot-to-spot distance (µm)	Capture area (mm^2)
ST [Stå+16]	2,016	100	200	6.2×6.6
Slide-seq [Rod+19]	2,019	10	10	3×3
Visium v1 [10x19]	2,019	55	100	6.5×6.5
DBiT-seq [Liu+20]	2,020	10	20	1×1
Visium v2 [10x22]	2,022	55	100	11×11
Stereo-seq [Che+22]	2,022	0.22	0.5	132×132
Slide-tags [Rus+24]	2,024	10	10	10×10
Nova-ST [Poo+24]	2,024	0.3	0.625	10×8
Visium HD [Oli+25]	2,025	2	2	6.5×6.5

single-cell reference data set and then tries to use these learned cell types to classify the spatial data.

3.2.1 Single-cell classification

Recall from Section 3.1.1 that the output of a sc/snRNA-seq experiment is a count matrix containing, for each cell in the sample and each gene in the genome, the number of RNA transcripts of the gene counted in the cell. The Waddington landscape model implies that the rows of this matrix (which are indexed by the cells) should cluster around the accumulation points of the landscape. A cell-type discovery pipeline therefore aims to identify clusters of cells in gene space.

With this model in mind, the following cell-type discovery pipeline is typical [Sat+23]:

1. Quality control: cells with low quality (e.g. low read count) are filtered out.
2. Feature selection: genes with low information (e.g. low overall abundance or low variability across samples) are filtered out.
3. Dimensionality reduction: a linear dimensionality reduction technique such as PCA is employed.

4. Unsupervised clustering: a clustering method is used to identify clusters in PCA space.

5. Labelling: a domain expert uses differential gene expression (DGE) analysis to assign cell-type labels to each cluster.

We note that each of the steps in this cell-type discovery pipeline involves a number of choices. Of particular importance is the resolution of the clustering method, which will for example determine whether an 'immune' cluster is split into subclusters corresponding to different immune subtypes. So, it cannot be expected that two different practitioners, when presented with the same data set, will produce the same list of cell types. In fact, even when attempting to follow the same pipeline but using different software packages or versions it has been shown that variability in output can be enormous [Ric+26].

Once cell-type discovery has been carried out on some initial data set, the discovered cell types can then be used to classify further single-cell samples. This can either be by repeating the same unsupervised pipeline on the new data, or by treating the newly discovered cell-type labels as training data for supervised classification.

3.2.2 Classification in imaging-based data

The ideal starting point for any transcriptomics data analysis is a count matrix, with columns indexed by genes and rows indexed by some kind of biological unit. Imaging-based transcriptomics methods, however, tend to output stacks of images whose intensities correlate with the locations of specific genes. When working with imaging-based data, then, it is necessary to first process these images to arrive at a count matrix suitable for further downstream analysis. This preprocessing follows roughly three stages:

1. (Image alignment) Imaging-based methods typically involve several rounds of imaging to enable detection of more than a handful of genes. So the first preprocessing step is to spatially align all of the images at each field of view (FOV) to recover a multiplexed image.

2. (Transcript localisation) Transcripts correspond to bright 'spots' in the multiplexed images, and so need to be detected by some form of spot detection. This typically involves a filtering step to remove background noise followed

by a blob detection algorithm and finally a decoding step to map the spot and its intensity data to a reference gene.

3. (Cell segmentation) After transcripts have been localised, their data along with inferred nuclei loci are used to detect cell boundaries and form a cell segmentation.

Following cell segmentation, a count matrix can be computed by counting the detected transcripts in each segmented cell. For commercial platforms, these pre-processing steps are usually bundled into the provided software. There also exist open-source libraries for these tasks such as Starfish [Axe+18]. Once a count matrix has been produced, cell-type classification can continue as with single-cell data. Even after segmentation, however, cell-type classification for imaging-based data sets can be difficult because of the typically low number of genes profiled per experiment.

Of particular interest in this thesis is the cell segmentation procedure provided with the 10x Xenium platform (see Section 3.1.2). This software first identifies cell nuclei from DAPI staining, and then expands boundaries outwards until another cell boundary is reached, up to a maximum of 15 µm [Jan+23, p. 4].

3.2.3 Classification in traditional array-based data

For array-based platforms, spatial resolution has until recently been limited to multiple cells per bin (Table 3.1). A number of approaches to deconvolve such data sets have been proposed.

3.2.3.1 Cell-type decomposition

Decomposition methods aim to recover the cell-type proportions in each multicellular bin. While a wide range of these methods have been proposed, they most commonly fall into one of four broad categories: matrix factorisation, probabilistic inference, regression, and deep learning. We give a brief overview of these methods here, and refer the reader to the survey [Gas+25] for a more thorough treatment.

Matrix factorisation With matrix factorisation methods [Rod+19; Elo+21; MZ22], the aim is to first decompose an annotated single-cell matrix, often with non-negative matrix factorisation, to express cells in terms of learned 'metagenes' encoding gene programs. The idea is that different metagenes correspond to different cell types,

and so after the spatial data are expressed in terms of the metagene basis this information can be used to estimate cell-type proportions.

Probabilistic inference The probabilistic, or Bayesian, approach [And+20; Cab+22; Kle+22; Lop+22; Che+23; Dan+22] is to set up a statistical model for gene expression, typically informed by annotated sc/snRNA-seq data, and then to infer cell-type proportions via maximum likelihood estimation.

Regression Regression models have a simple idea: estimate the cell-type proportions as the coefficients of a linear or more general regression. These typically begin by first identifying marker genes for cell types in a single-cell reference. Examples include SpatialDWLS [DY21], NLSDeconv [CRW24], SpatialDecon [Dan+22], and SONAR [Liu+23].

Deep learning Deep learning approaches are the most varied. A common idea is to use graph neural networks to model the spatial component of the data [SS21; Li+22; Xu+23; YWZ24; LL24]. CellDART [Bae+22] and SpatialDDLS [Mañ+24] both use a neural network trained on pseudospots aggregated from single-cell data to directly estimate cell-type proportions on array spots. A recent trend more broadly has been to produce large foundation models for ST data [Bla+25; Mad+25a]. UCDBase is a foundation model specifically designed for the task of cell-type deconvolution in spatial data [CBS23].

3.2.3.2 Spatial reconstruction

Another approach is to take matching sc/snRNA-seq data and attempt to map single cells to spatial locations using the low-resolution spatial data as 'lampposts'. An implementation of this idea has been offered as part of Seurat [Sat+21] since version 3 [Hao+21]. A popular deep learning approach to spatial reconstruction is Tangram [Bia+21], which optimises cell placements such that they approximate the gene-gene correlation observed in the spatial reference. Other deep learning techniques for the same task include DEEPsc [MCN21] and GraphST [Lon+23].

3.2.4 Classification in subcellular array-based data

Subcellular array-based methods are still in their infancy, and so there has been little development of bespoke methods for cell-type classification in this paradigm.

These data are incredibly sparse—in the Stereo-seq mouse kidney data we consider in Chapter 4 the modal number of transcripts per 715 nm spot is 0 and the mean is ~1.7—and so it is necessary to perform some aggregation to determine cell types. In the original Stereo-seq publication [Che+22] a segment-then-classify approach is taken, where cell boundaries are determined first, followed by classification into cell types. They considered two different approaches to this problem:

1. (Fixed-window binning) In the fixed-window approach the array is segmented into a square grid with cells of size approximating the size of a single cell. We describe the grid as 'Bin n' if the cells have size n spots $\times$ n spots.

2. (Cell segmentation) In this approach, complementary imaging data are used to define cell boundaries.

Both of these methods are somewhat problematic. The inflexible nature of the fixed-window approach leads to two major issues: (1) many of the bins are likely to cover more than one cell, and (2) individual sparsely dispersed cells can get caught between the grid boundaries, leading to their signal being lost. The cell segmentation approach is also tricky to get correct in practice, and it is very easy for smaller cells to be incorrectly segmented. As a result, both of these methods can lead to underdetection of small and rare cells, as was noted in the original Stereo-seq study [Che+22, p. 1789].

3.3 Topological methods in transcriptomics

By far the most ubiquitous application of topological methods to transcriptomics is the use of UMAP [MHM18] to produce low-dimensional embeddings of gene expression data for visualisation [Bec+19], where it has emerged as a competitor to the previously dominant t-SNE [MH08]. Indeed, UMAP is now included alongside t-SNE as an example visualisation tool in the introductory guide to Seurat, one of the most popular scRNA-seq analysis toolkits [Sat+23]. However, it should be noted that the use of UMAP (and indeed all non-linear low-dimensional embeddings) for single-cell transcriptomics analysis has become contentious in recent years. There is evidence that the reported benefits of UMAP over t-SNE are nullified if t-SNE is appropriately initialised [KL21]. It has further been argued that the distortions introduced by reducing from thousands of dimensions to just two are so great that

the use of methods such as UMAP and t-SNE is actually counter-productive in practice [CP23].

Given such questions about the suitability of manifold-learning techniques for sc/snRNA-seq data analysis, a natural line of research has been to probe the geometry of gene expression space. It has been shown that for some single-cell data sets the PCA-reduced space exhibits signs of nonzero curvature [SWH21].

Another prominent use of topological methods in transcriptomics is the application of MAPPER [SMC07] by Nicolau *et al.* to breast cancer microarray data [NLC11], where the authors identified a previously unidentified subgroup of Oestrogen Receptor-positive breast cancers. A more recent application of MAPPER in single-cell transcriptomics is scTDA, a tool for identifying cell differentiation pathways in scRNA-seq data [Riz+17].

There have been a number of applications of spectral theory to single-cell transcriptomics. Spectral simplicial theory has been used to perform DGE analysis in scenarios where replicates cannot be easily sorted into classes [GYC19]. A trio of spectral and persistence methods were used in [Hoe+22] for feature selection and DGE analysis.

For cell-type identification in single-cell data, persistent homology forms the basis for the HiDef multiscale community detection method [Zhe+21], which outputs a hierarchical tree of cell types. A combination of curvature and persistent homology analysis was used in [HC24] to extract properties of cells including their developmental trajectories.

In spatial transcriptomics, the recent PHD-MS method uses persistent homology on a multiscale graph of expression clusters to identify spatial domains [BC26]. The persistence machinery here enables the authors to identify hierarchical features across multiple scales simultaneously. MCIST [CW25] uses persistent spectral graph theory [WNW20] along with deep learning to produce embeddings informed by multiscale cell-cell interaction information.

3.4 Discussion

In this chapter we have gone on a whistle-stop tour of the world of single-cell and spatial transcriptomics. Moving forward into the rest of the thesis, we would invite the reader to keep in mind two open problems raised in this chapter.

The first problem is concrete. We have seen how transcriptomics technologies separate into two broad classes: single-cell, and spatial. In 2021, *Nature Methods*

interviewed researchers about the future of spatial transcriptomics [Mar21]. Overwhelmingly, they dreamt of merging the single-cell and spatial worlds: 'one day people will send out tissue and get back spatially resolved single-cell genetic information', said Hongkui Zeng of the Allen Institute for Brain Science. As we have now seen, the underlying technology has now reached that point, and whole-transcriptome methods such as Stereo-seq are measuring at subcellular resolution. The challenge, then, rests with the mathematicians: how do we turn this high-resolution data into true *in situ* single-cell transcriptomics? This is the question we will answer in Chapter 4.

The second problem is more subtle. In Section 3.2.1 we saw that cell-type classification in single-cell transcriptomics is a notoriously ill-defined task. Indeed, two different practitioners given access to the same data set are likely to produce considerably different cell-type classifications from it. This presents a real challenge for downstream analysis, where results are almost always dependent on the choice of cell-type assignment. However, it also presents an interesting frontier for mathematicians: is there a way we can analyse single-cell data without exposing ourselves too much to these issues? In Chapter 5 we will propose one new method which does exactly that, by leveraging an insight from the field of ecological diversity to provide an analysis tool for single-cell and spatial transcriptomics which is robust to the underlying choice of cell-type assignment.

Chapter 4

TopACT: a novel method for cell-type classification in subcellular spatial transcriptomics

Spatial transcriptomics (ST) was named Method of the Year by *Nature Methods* in 2021 [Mar21], where it was noted that an open problem is the inference of information 'at the level of single cells'. Historically this has been a limitation of experimental technologies, where the spatial resolution of array-based ST platforms has been limited to >10 µm, above the typical size of a single mammalian cell. Computational approaches for data analysis in this regime have therefore largely been focused on decomposing information gathered from multicellular 'bins'. However, in recent years there has been a rapid trend towards subcellular-resolution array-based ST platforms, which necessitate an entirely novel computational approach.

In this chapter, which is based on results published in [Ben+24], we will consider the problem of cell-type identification in subcellular ST data sets. As presented in Chapter 3, experimental technologies have been converging on whole transcriptome reads at subcellular spatial resolution. The Stereo-seq technology [Che+22] in particular achieves a 0.5 µm to 0.715 µm resolution with a field of view of up to 13 cm × 13 cm. In the original publication the authors used a 'segment-then-classify' approach (see Section 3.2.4) to identify cell types. However, this approach failed to identify rare, sparsely dispersed cells, which are often of critical clinical importance [Che+22, p. 1789].

Here, we present TopACT, a new method for cell-type identification in subcellular ST data which takes a spot-level approach to classification. That is, instead of segmenting and then classifying, TopACT directly classifies at the subcellular level

by pooling information from local neighbourhoods around each spot. This allows for transcriptomic information to directly inform the resulting cell segmentation. In particular, because TopACT does not rely on any prior cell segmentation, it is much better suited to identifying those sparsely dispersed cells which evade detection under the traditional segment-then-classify regime.

We will showcase TopACT on a number of data sets. One of the difficulties in developing methods for subcellular spatial transcriptomics is the lack of appropriate ground truth labels for cell segmentation. So, first we will develop a synthetic data set based on a random Voronoi model to examine the accuracy of TopACT in a setting with established ground truth. We will see that TopACT achieves an accuracy well above the best-case scenario for the traditional fixed-window regime. Next, we will validate TopACT on a Stereo-seq mouse brain data set [Che+22], and use the method to identify a population of perivascular macrophage cells which evaded detection under the original authors' segment-then-classify approach.

We will then use TopACT to analyse a new Stereo-seq data set of the mouse kidney under treatment with a lupus-like condition, and show that the method is able to identify glomerular immune infiltration characteristic to the disease. Finally, we will examine the adaptability of TopACT to imaging-based ST platforms by analysing a new 10x Xenium [Jan+23] data set of the human kidney. We will show that TopACT is able to achieve much more biologically plausible cell segmentations than the standard commercial tool.

Overall, these results demonstrate that TopACT is able to extract true single-cell-level information from next-generation subcellular ST technologies, providing a resolution to the problem posed in [Mar21].

Attribution This chapter is based on the paper 'Multiscale topology classifies cells in subcellular spatial transcriptomics' published in *Nature* [Ben+24]. It is therefore based on joint work with the coauthors of that paper. Here we present only the parts of the work that were directly carried out by the author of this thesis. Some parts of the applications sections, especially the biological context and interpretation, were written in collaboration with Katherine Bull and Aneesha Bhandari.

4.1 Method description

TopACT is a method for cell-type identification on subcellular ST data. In contrast to ubiquitous segment-then-classify approaches, TopACT produces an independent

cell-type classification for every subcellular spot in the underlying data set. To achieve this, the method makes use of a local classifier, which is a cell type classifier trained on a sc/snRNA-seq reference. For a given spot, TopACT pools together the gene expression in a local neighbourhood about the spot. The local classifier is used to determine both a cell-type classification for the neighbourhood as well as a confidence level in the classification. If the confidence level does not meet a user-determined threshold, then TopACT increases the size of the neighbourhood in order to gain more information. By repeating this process until a confident classification is produced, the method is able to produce a cell-type classification for each spot using the minimal necessary spatial context, thus avoiding contamination from neighbouring cells.

In this section we will describe the TopACT method in full technicality. We will begin by establishing a mathematical model for array-based ST experiments which is agnostic to the platform chosen.

4.1.1 Spatial transcriptomics model

We begin by abstracting the notion of an ST experiment. A general (non-spatial) transcriptomics experiment can be thought of as a collection of objects (for example, in single-cell transcriptomics the objects are cells), each equipped with a cell type t and an expression vector $v \in \mathbb{R}^D$. The vector v measures the number of reads in each of the D genes in the genome $\mathcal{G}$, and it is assumed that these are sampled from random variables corresponding to the cell type t. The key difference in an ST experiment is that the objects are now equipped with a notion of distance, giving rise to a metric space. The present section formalises this notion.

4.1.1.1 Experimental setup

We begin with the following preliminary objects:

1. A metric space X called a *sample*;
2. A finite subset $\mathcal{S} \subset X$ of *spots*;
3. A finite ordered set $\mathcal{G} = \{g_1, \ldots, g_D\}$ of D *genes*;
4. A finite ordered set $\mathcal{T} = \{t_1, \ldots, t_K\}$ of K *cell types*.

These items together can be seen as a mathematical abstraction of a typical array-based ST experimental setup: we aim to measure the expression of each gene in $\mathcal{G}$

across the sample X, by taking readings from each spot in $\mathcal{S}$. These readings are determined by the underlying cell type in $\mathcal{T}$ associated to each spot.

In this setting, an experimental reading can be thought of as an assignment of an expression $v_{sg} \in \mathbb{R}$ for each spot $s \in \mathcal{S}$ and gene $g \in \mathcal{G}$. Equivalently, making use of the ordering on $\mathcal{G}$, we have a map

$$v \colon \mathcal{S} \to \mathbb{R}^D, \tag{4.1}$$

where $v(s)_i = v_{sg_i}$ for each $s \in \mathcal{S}$ and $1 \leq i \leq D$.

4.1.1.2 Expression model

We now describe our model for how these expression assignments arise in practice. Underlying each experiment, we assume there is a set $\mathcal{T}$ of disjoint *cell types*, and that for each cell type $t \in \mathcal{T}$ and gene $g \in \mathcal{G}$ there is a corresponding random variable V_{tg} giving the count of the gene g measured in a cell of type t.

A subcellular ST experiment can be seen as a partial assignment

$$\tau \colon X \rightharpoonup \mathcal{T} \tag{4.2}$$

of a cell type to some of the points in X. Given such an assignment, for each point $x \in \operatorname{dom} \tau$ we model

$$v_{xg} \sim \alpha V_{\tau(x)g}, \tag{4.3}$$

where α is a scalar corresponding to the technical efficiency of the underlying experimental technology being modelled, and we can assign $v_{xg} = 0$ whenever $x \notin \operatorname{dom} \tau$. Restricting these values to spots in $\mathcal{S}$, we recover an experimental reading.

We emphasise that in this subcellular model, each spot is assigned *at most* one cell type. In the case of multicellular ST, as seen with data produced by e.g. ST/Visium [Stå+16] and Slide-seq(v2) [Rod+19; Sti+21], this assumption will not hold, as each spot records transcripts from multiple distinct cells.

4.1.2 Cell-type classification

Given that the expression vector $v(x) \in \mathbb{R}^D$ assigned to a point $x \in X$ depends on its cell type $\tau(x)$, a natural objective is to deduce the cell-type map τ given the expression map v. In the case of single-cell transcriptomics, where each expression

vector contains sufficient information to deduce a cell type, this is a relatively straightforward task. In contrast, subcellular spatial data typically have very low read counts, and it is therefore necessary to aggregate readings from neighbouring spots in order to recover enough information to reliably predict a cell type. However, it is not clear how best to perform this aggregation without prior knowledge of cell boundaries.

Our approach is to assume that there exists a local neighbourhood around each spot that belongs entirely to a single cell type. By combining the expression readings from this neighbourhood, one obtains a pseudo-single-cell reading that can be classified by existing techniques. This yields a classification for each individual spot. The challenge now is to identify the correct scale at which to draw the neighbourhood, and we resolve this by taking a 'multiscale' approach.

4.1.2.1 Local classifier definition

Let T be a $\mathcal{T}$-valued random variable and Q a positive-integer-valued random variable. We are going to study a random variable describing the aggregated gene expression of Q spots that are all assigned the cell type T.

For any cell type $t \in \mathcal{T}$ let

$$V_t := (V_{t,g_1}, \ldots, V_{t,g_D}) \tag{4.4}$$

be the $\mathbb{R}^D$-valued random variable describing the total expression over all genes of the cell type t. Then, if $V_t^1, \ldots, V_t^Q$ are i.i.d copies of V_t, set $\Sigma V_t := \sum_{i=1}^{Q} V_t^i$ and

$$Z := \Sigma V_t / \|\Sigma V_t\|_1. \tag{4.5}$$

Z is therefore the normalised sum of Q expression readings independently drawn from the cell type t.

We say that a *local classifier* is any method that estimates the probability of each cell type given an observed normalised expression reading. More specifically, recalling that the cell types have an ordering $\mathcal{T} = \{t_1, \ldots, t_K\}$, we say that a local classifier is a function

$$f : \mathcal{Z} \to [0,1]^K, \tag{4.6}$$

where $\mathcal{Z} := \{z \in [0,1]^D : \|z\|_1 = 1\}$, such that

$$f(z)_i \approx \mathbb{P}(T = t_i \mid Z = z) \tag{4.7}$$

for all $1 \leq i \leq K$.

4.1.2.2 Producing a local classifier from sc/snRNA-seq data

We can use single-cell or single-nucleus reference data sets to estimate the effect of the different cell types on the gene expression behaviour and produce a local classifier. In detail, we take a collection $\mathcal{C}$ of single cell samples along with a gene expression map

$$v^{\mathrm{sc}} \colon \mathcal{C} \to \mathbb{R}^D$$

and a cell-type map

$$\tau^{\mathrm{sc}} \colon \mathcal{C} \to \mathcal{T}.$$

From this information, we seek a classifier that takes as input normalised expression vectors and outputs probability distributions over the cell types in $\mathcal{T}$. To do this, we normalise each expression vector:

$$z^{\mathrm{sc}}(c) \coloneqq \frac{v^{\mathrm{sc}}(c)}{\|v^{\mathrm{sc}}(c)\|_1}. \tag{4.8}$$

The input-output pairs $(z^{\mathrm{sc}}(c), \tau^{\mathrm{sc}}(c))$ then form training data for any standard supervised learning platform. In our case, we use a linear support vector machine (SVM) [BGV92; CV95] and estimate probabilities with Platt scaling [Pla+99] to produce a local classifier f.

4.1.2.3 Multiscale confidence matrix

Let f be a local classifier. In order to classify a point $x \in X$ it may be necessary to aggregate expression readings around x. Write $B(x, r) = \{y \in X : d(x, y) \leq r\}$ for the closed ball of radius r in X centred on x. We define the aggregated gene expression

$$v(x, r) \coloneqq \sum_{s \in B(x,r) \cap \mathcal{S}} v(s) \in \mathbb{R}^D, \tag{4.9}$$

and, if this is non-zero, set

$$z(x, r) \coloneqq v(x, r)/\|v(x, r)\|_1 \in \mathcal{Z}. \tag{4.10}$$

In words, $z(x, r)$ describes the normalised gene expression about x at radius r. Then, if f is a local classifier as defined in (4.7), one obtains a probability vector

$$f(z(x, r)) \tag{4.11}$$

which can be interpreted as a cell-type classification at the scale r.

Given an ordered collection $R = (r_1 \leq \cdots \leq r_L)$ of radii one then obtains a sequence of corresponding probability vectors, which can be combined into an $L \times K$ matrix $\mathfrak{M}^x$ defined by

$$\mathfrak{M}^x_{ij} \coloneqq f(z(x, r_i))_j \tag{4.12}$$

which we call a *multiscale confidence matrix*. Here $\mathfrak{M}^x_{ij}$ records the confidence in cell type t_j at scale r_i around the point x.

4.1.2.4 Extracting cell-type annotations

Given a multiscale confidence matrix $\mathfrak{M}^x$, we would like to extract a cell-type annotation for the spot x. The general principle followed by TopACT is that one should use the smallest scale possible to classify a point, because this minimises the chance that the aggregated expression has been taken from surrounding cells of a different type.

Let $\theta \in [0, 1]$ be a *confidence hyperparameter*. The *classification index* $i_\theta = i_\theta(x)$ of x is

$$i_\theta \coloneqq \inf\big\{i \in \{1, \ldots, L\} : \|\mathfrak{M}^x_i\|_1 \geq \theta\big\}. \tag{4.13}$$

In other words, $r_{i_\theta(x)}$ is the lowest scale at which a cell type was predicted with confidence at least θ at the point x.

We now define the *TopACT predicted cell types* with respect to the collection R and confidence threshold θ:

$$\mathfrak{T}_{R,\theta} \colon X \rightharpoonup \mathcal{T}. \tag{4.14}$$

If $i_\theta(x) < \infty$ then we set $\mathfrak{T}_{R,\theta}(x)$ to be the cell type with the highest predicted probability at scale $r_{i_\theta(x)}$. Precisely, it is $\mathfrak{T}_{R,\theta}(x) = t_j$ where j maximises the value of $\mathfrak{M}^x_{i_\theta(x)j}$.[1] If $i_\theta(x) = \infty$, i.e. if no scale produced sufficient confidence, then we do not specify a cell type. In other words, we have that $\mathrm{dom}(\mathfrak{T}_{R,\theta}) = \{x \in X : i_\theta(x) < \infty\}$.

[1] A tie between multiple cell types can be resolved by equipping the cell types with an order of precedence. Note that if $\theta > 0.5$ then a tie can never occur.

4.1.2.5 Restricting TopACT to a square grid

In the experiments considered in the manuscript, we almost always work with either simulated or real-world Stereo-seq [Che+22] data. For Stereo-seq experiments, we assume that spots are evenly spaced on a 2D square lattice. More precisely, we assume that the metric space X is a subspace of $\mathbb{R}^2$ and the set of spots is

$$\mathcal{S} = ([I] \times [J]) \cap X \tag{4.15}$$

for some $I, J \in \mathbb{N}$, where $[k] = \{1, \ldots, k\}$ for any $k \in \mathbb{N}$.

By further equipping $\mathbb{R}^2$, and therefore X, with the ℓ_∞ norm, it follows that the neighbourhoods $B(x, r)$ are squares in X. In particular, for a spot $s \in \mathcal{S}$ the critical values $r_0 \leq r_1 \leq \ldots$ for which $B(x, r_i) \cap \mathcal{S}$ changes are precisely $r_i = i \in \mathbb{N}$. In this setting, then, we set $R = (0, 1, 2, \ldots, r_{\max})$ for some maximal radius parameter $r_{\max} \in \mathbb{N}$. Algorithm 4.3 demonstrates how to produce TopACT cell-type annotations from these assumptions.

Algorithm 4.3 TopACT (Square grid)

Input: M, N: the dimensions of the spot grid;
V: an $M \times N \times D$ array: V_{ijk} is the expression of gene g_k at the spot (i, j);
$f : \mathcal{Z} \to [0, 1]^K$: a local classifier;
θ: a confidence hyperparameter;
$r_{\max}$: the maximum radius.

Output: The TopACT cell-type assignment $\mathfrak{T}_{(0,\ldots,r_{\max}),\theta} : [M] \times [N] \rightharpoonup \mathcal{T}$.

$\tau \leftarrow \emptyset$ ▷ *An empty cell-type assignment*
for $s = (i, j) \in [M] \times [N]$ **do**
  $r \leftarrow 0 \in \mathbb{N}$
  $v \leftarrow 0 \in \mathbb{R}^D$
  while $r \leq r_{\max}$ and $s \notin \operatorname{dom} \tau$ **do**
    for all $s' = (i', j') \in [M] \times [N]$ such that $\|s - s'\|_\infty = r$ **do**
      $v \leftarrow v + V_{i'j'}$
    if $v \neq 0$ **then**
      $z \leftarrow v / \|v\|_1$ ▷ *Normalise expression for input to local classifier*
      $k^* \leftarrow \operatorname{argmax}_{1 \leq k \leq K} f(z)_k$
      if $f(z)_{k^*} \geq \theta$ **then** $\tau(s) \leftarrow t_{k^*}$ ▷ *Sufficient confidence to classify spot*
    $r \leftarrow r + 1$ ▷ *Increment radius to the next critical value*
return τ

We remark that this setup may differ for different ST technologies. For example, the spots may lie on a hexagonal grid as in HDST [Vic+19] or be randomly distributed as in Seq-Scope [Cho+21]. Our method is general and applies equally to any such specification, including 3D or spatiotemporal data. TopACT only requires some notion of distance between the spots in $\mathcal{S}$.

4.2 Validation

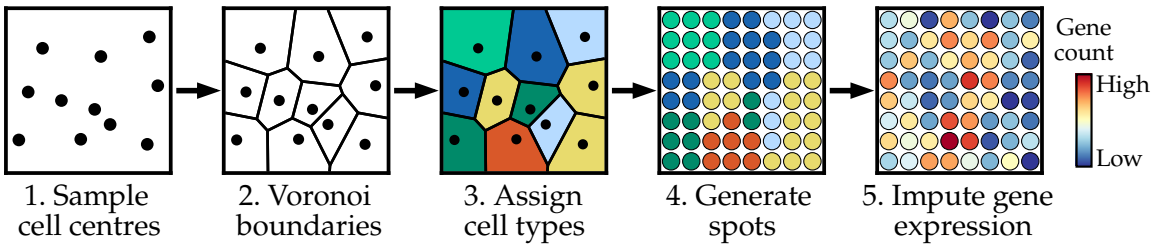
The goal of this section is to demonstrate that the TopACT method is both reliable and suitably more powerful than pre-existing methods for classifying cell types in subcellular transcriptomics data. First, we will generate a synthetic Stereo-seq data set based on a random Voronoi spatial model with gene expression imputed from an snRNA-seq reference data set of the mouse kidney. We will show with this data set that TopACT achieves significantly higher accuracy than is possible with Bin 20 methods, and that our new method is able to reliably identify sparsely dispersed cells which elude traditional methods.

While synthetic data are a useful benchmark given that they provide ground truth cell types for directly assessing model performance, real-world data will always have additional complexities and nuances which are not captured by any synthetic model [Cro+23]. To validate TopACT on real-world data we will consider a mouse brain Stereo-seq data set which was published alongside the Stereo-seq method [Che+22]. In particular, we will show that TopACT is able to extract rare perivascular macrophage cells which were unable to be resolved by cell segmentation approaches in the original Stereo-seq manuscript.

4.2.1 *In silico*

4.2.1.1 Synthetic data generation

We use a multi-stage process to generate synthetic benchmark data, first generating a synthetic cell-type map and then imputing gene expression from an snRNA-seq reference data set. The process is outlined in the following schematic:



Firstly, a synthetic grid of spots with cell-type annotations is produced. We sample 625 points uniformly at random from the unit square $[0,1] \times [0,1]$, taking these points to be cell centres. We draw a Voronoi diagram based on these points to simulate cell boundaries. Cell types are then assigned at random to each Voronoi region, in proportion to the cell type abundances in the snRNA-seq data. These cell types are then applied to a 500×500 grid of spots overlaid on the unit square. The end result is a grid of spots, each annotated with a cell type.

Next, we impute the gene expression at each spot using a Poisson process with parameters inferred from an snRNA-seq data set (the same snRNA-seq data set used in Section 4.3). This process is based on a simplified version of the model described in [Cab+22]. In detail, for a cell type T and gene g let λ_{Tg} denote the mean expression of gene g over all cells in the snRNA-seq data set with cell type T. If a spot s is assigned the cell type T, we then model the expression v_{sg} of gene g at s by

$$v_{sg} \sim \text{Poisson}(\alpha \lambda_{Tg}) \tag{4.16}$$

where $\alpha = \exp(-7.3)$ is a fixed parameter determining the transcriptional abundance. To model zero-inflation, we then select 20 % of spots uniformly at random to be assigned zero reads, regardless of the Poisson-modelled expression.

4.2.1.2 Classification methodology

We ran TopACT directly on synthetic data, with an SVM local classifier trained on the same snRNA-seq reference data set used for generation. For fixed-window Bin 20 analysis, we split the 500×500 synthetic grid into square bins, each covering a 20×20 region of spots. Bin 20 was chosen so that each bin matches the mean area of a synthetic cell. Moreover, at Bin 20 the resulting grid approximates 10 µm resolution, which is considered the 'sweet spot' for single-cell analysis [Mar21]. We then summed the expression over all spots in each region. Robust Cell Type Decomposition (RCTD) [Cab+22], a decomposition method for multicellular data (see Section 3.2.3), was run on couplet mode with default settings, using the same

snRNA-seq reference data set, and we assigned each bin the RCTD predicted 'first type'.

It is possible that another method would outperform RCTD at the Bin 20 level. To check this, and provide an upper bound on the performance of Bin 20 methods, we computed a 'Modal' cell-type assignment which assigned to each bin its most frequent ground truth cell type. This assignment would necessarily have the highest possible classification accuracy on a Bin 20 grid. This allows us to compare TopACT's accuracy to a theoretically optimal Bin 20 method, and to examine the gap between practical and best-case performance at the Bin 20 level.

4.2.1.3 Classification accuracy

Figure 4.1 shows a representative synthetic data patch with ground truth cell-type annotations along with both TopACT and RCTD cell classification outputs. It can be seen that TopACT's Bin 1 resolution enables much finer resolution of spatial detail in the resulting cell-type assignment compared with the Bin 20 approach. Moreover, many of the sparsely dispersed cells that are missed by the Bin 20 approaches are faithfully recovered in the TopACT classification. One such cell is circled in Figure 4.1.

We define the accuracy of a classification as the proportion of spots which are assigned their correct ground truth cell type. Figure 4.2 shows accuracy distributions of TopACT against both RCTD and the Modal assignment at Bin 20. We find that TopACT's accuracy (M = 0.808, SD = 0.006) surpasses the accuracy of both RCTD (M = 0.668, SD = 0.010) and the Modal assignment (M = 0.693, SD = 0.009). These results show that TopACT not only outperforms existing software packages, but also that its spot-level approach allows it to outperform even the best case for assignments produced with a fixed-window approach.

4.2.1.4 Sparse cell detection

A notable limitation of previous experimental and computational methods for array-based transcriptomics is the inability to reliably resolve rare, sparsely dispersed cells. These cells are of key clinical importance in a wide range of settings; in this chapter we will examine the distribution of immune cells in autoimmune disease. One of the major promises of the proposed method for cell-type classification is therefore its ability to resolve these 'needle in haystack' cells.

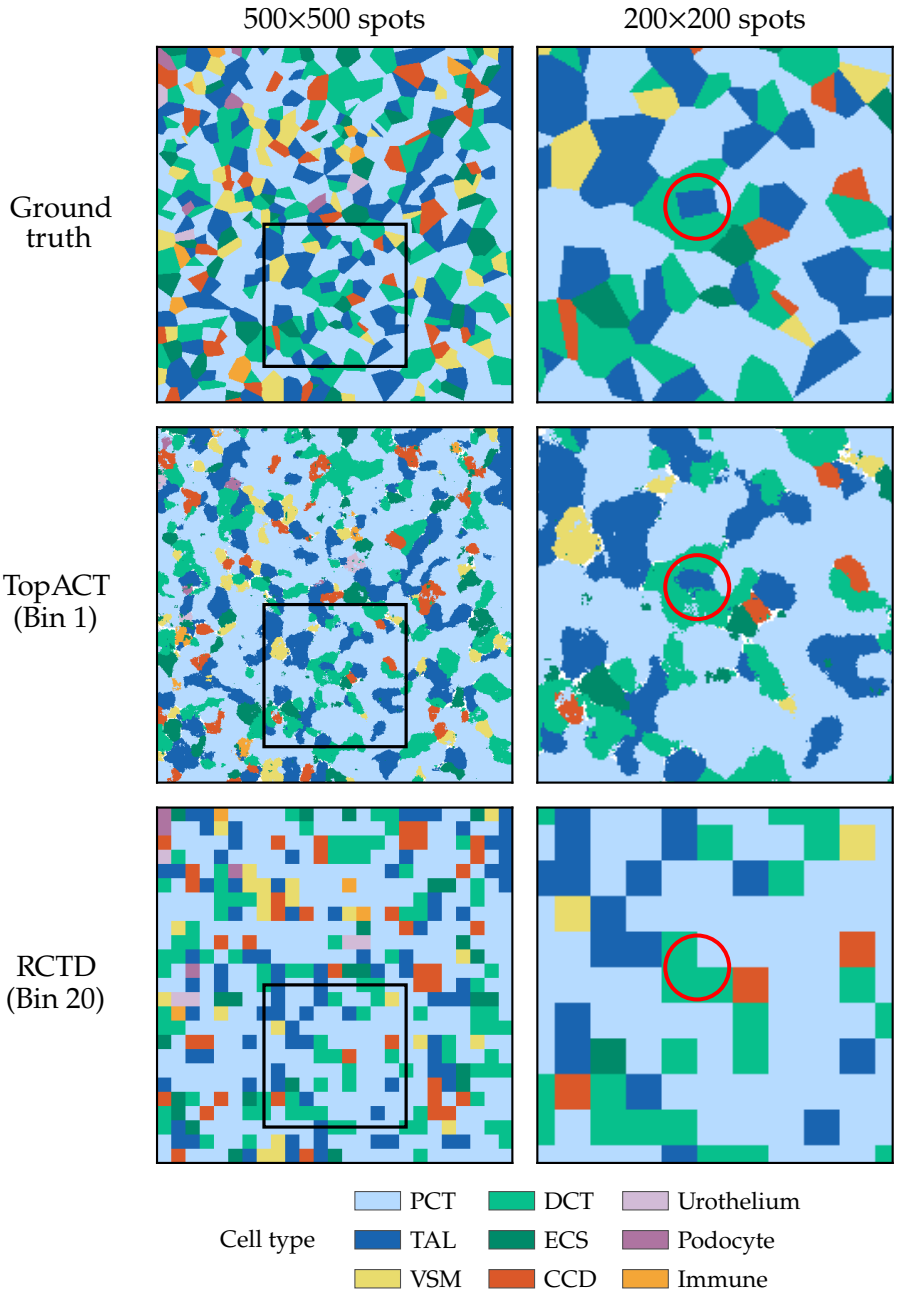


Figure 4.1. Sample output of cell-type identification algorithms on synthetic data. Top row: synthetic ground truth. Middle row: TopACT at bin 1. Bottom row: RCTD at bin 20. Second column is a magnification of a patch in the first column. Red circles highlight a ground truth cell identified by TopACT but not bin 20 methods.

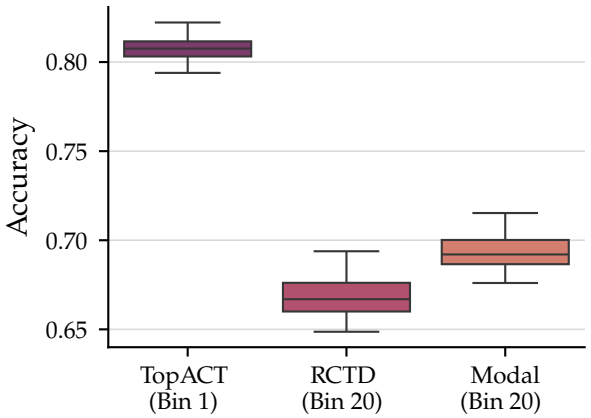


Figure 4.2. Box plots of per-iteration accuracy of cell-type classification methods on synthetic data (n = 100 iterations). Centre line shows median; box limits show interquartile range; whiskers show full range. Modal is the optimal Bin 20 classification assigning to each bin its most common ground truth cell type.

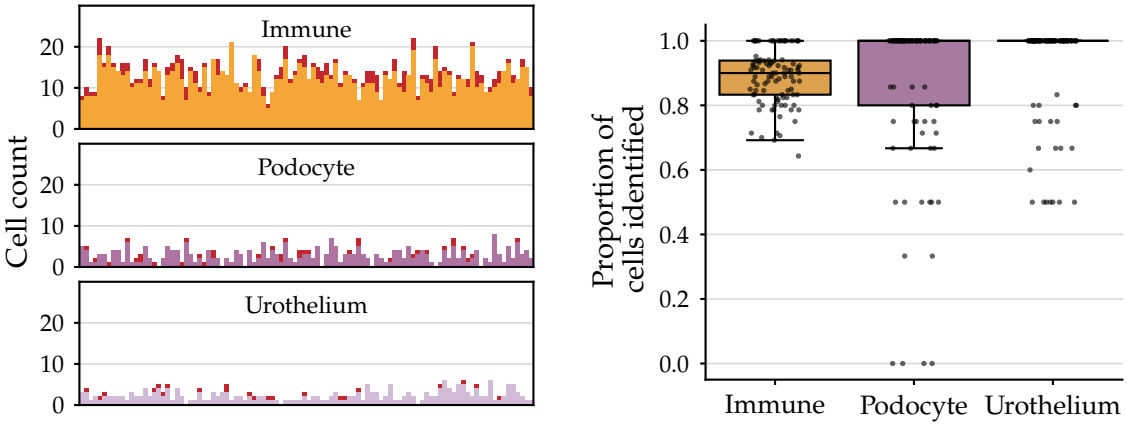


Figure 4.3. TopACT performance on rare cell types. Left: number of cells of each type detected per iteration. Coloured regions denote cells detected by TopACT, red regions denote cells not detected by TopACT. Right: box plots showing recall of sparsely dispersed cells (n = 100 iterations). Centre lines show median; box limits show interquartile range; whiskers show full range of non-outlier points. Outliers are points more than 1.5 interquartile ranges from the upper or lower quartiles. Full distributions overlaid.

For this analysis, we took rare cell types to be those making up less than 5 % of the total samples in the snRNA-seq reference data. These were namely immune cells, podocyte cells, and urothelium cells. From spot-level TopACT cell-type classifications, we used an image analysis pipeline (see Section 4.3.1) to isolate individual cells of these types. We say that a ground truth cell has been correctly identified by TopACT if it is within a 20 spot centre-to-centre distance from a TopACT-predicted cell of the same type. Across 100 iterations, we find that TopACT consistently identified all but a few of the cells of each rare cell type and often identified all of them, validating the ability of the method to pinpoint rare cells (Figure 4.3).

4.2.1.5 Molecular diffusion

Molecular diffusion (the lateral displacement of RNA molecules between the permeabilisation and sequencing stages) can have a significant effect on the spatial accuracy of array-based ST methods. Estimates of molecular diffusion strength are difficult to produce without access to ground truth, however there is evidence that in some tissue types the Stereo-seq method exhibits a more pronounced diffusion effect than other methods [You+24]. In the mouse brain, average Stereo-seq diffusion was estimated at 6.84 µm based on localisation of *Vip* transcripts relative to cell centroids [Che+22]. However, this is likely to be an overestimate as not all *Vip* transcripts will be localised precisely to cell centroids. In Seq-Scope, comparison with haematoxylin and eosin (H&E) stains suggested a mean diffusion distance of 1.7 µm [Stå+16]. The true lateral diffusion effect in Stereo-seq is likely to lie between these two figures. Moreover, the recent Stereo-seq V2 technology seems to exhibit lower levels of molecular diffusion than the original Stereo-seq technology [Zha+25].

To test the robustness of TopACT to molecular diffusion, we added diffusion effects of increasing intensity to the simulated spatial data. For a given expected diffusion distance λ_{diff}, we modelled the diffusion effect separately for each synthetic transcript. Given a transcript a, we sampled a diffusion magnitude

$$D_a \sim \exp(\lambda_{\text{diff}}) \tag{4.17}$$

from an exponential distribution. Independently, a diffusion direction θ_a was sampled from a uniform distribution over $[0, 2\pi)$. This yielded coordinate-wise displacements

$$d_a^x = D_a \cos\theta_a, \quad d_a^y = D_a \sin\theta_a. \tag{4.18}$$

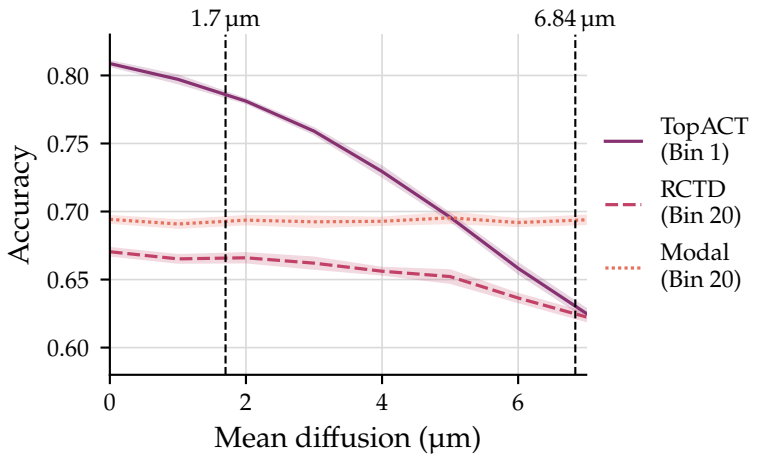


Figure 4.4. Accuracy of methods under simulated molecular diffusion. Methods run for $n = 10$ iterations each on mean diffusion magnitudes of λ µm for $\lambda = 0, 1, 2, \ldots, 7$. Lines show mean, bands show standard error of the mean. Vertical dashed lines refer to previous diffusion estimates in the literature [Stå+16; Che+22].

The original spot coordinates x_a, y_a for the transcript were then revised accordingly to displaced coordinates

$$x_a^{\text{diff}} = x_a + \lfloor d_a^x / d_{\text{spot}} \rfloor, \quad y_a^{\text{diff}} = y_a + \lfloor d_a^y / d_{\text{spot}} \rfloor. \tag{4.19}$$

The rescaling by $d_{\text{spot}} = 0.715$ accounts for the assumed inter-spot distance of 0.715 µm in Stereo-seq.

We applied this diffusion model on 70 synthetic samples, 10 each with diffusion magnitudes from 1 µm to 7 µm at 1 µm intervals. Figure 4.4 shows the accuracy of TopACT compared with RCTD and the Modal Bin 20 assignment as the simulated diffusion effect becomes stronger. The Modal assignment is unaffected by the added diffusion as it is based directly on ground truth and ignores the transcript information. We see that both TopACT and RCTD lose accuracy as the diffusion effect becomes stronger, with TopACT being more sensitive to the diffusion effect. This is unsurprising: the higher resolution method will naturally be more sensitive to smaller perturbations of the transcript data. In particular, the pooling step of Bin 20 methods already essentially 'diffuses' transcripts by the radius of the bins (here about 7 µm). We can see however that TopACT remains more accurate than RCTD even at the pessimistic upper bound of 6.84 µm on diffusion strength given in [Che+22].

4.2.2 *In vitro*

In the original Stereo-seq study of adult mouse brain data [Che+22], the authors considered both a Bin 50 analysis as well as an image-based cell segmentation

based on a nucleic acid stain. Both of these approaches failed to detect any immune populations other than microglia. This behaviour was attributed to issues with these approaches when 'multiple cell types are close to each other, particularly smaller cell types like immune cells'. Given that TopACT was designed to detect precisely these sparse, rare cells, we decided to test the method's ability to pick up immune cells in these data.

We used TopACT with a local classifier trained on the same scRNA-seq reference data set used for classification in the original Stereo-seq study [Zei+18]. We focused on the perivascular macrophage (PVM) subpopulation of immune cells. After spot-level classification, we performed a binary dilation and called as PVM cells any connected components of more than 60 spots. This pipeline detected 66 PVM cells across the entire sample, shown in Figure 4.5. We validated these findings by confirming high expression of PVM marker genes (as defined in [Zei+18]) in TopACT-predicted PVM cells compared with background (Figure 4.6). TopACT-predicted PVM cells expressed 3.4 ± 4.9 PVM mean marker gene transcripts per cell compared with 0.1 ± 0.5 marker gene transcripts in background. Across all transcripts TopACT-predicted cells expressed 12,840 ± 6,197 transcripts compared with 16,757 ± 6,429 transcripts in background, so increased marker gene expression does not simply reflect higher read depth.

These results show that in real-world data TopACT is able to pinpoint rare and sparsely distributed cell populations which are inaccessible with the standard binning and cell segmentation methodologies. In particular, it validates TopACT's ability to extract true single-cell level information from subcellular array-based ST data.

4.3 Application: immune cell identification in array-based data

Autoimmune disease is driven by abnormal immune activity. In autoimmune diseases of the kidney such as glomerulonephritis, immune cells do not aggregate but rather are sparsely dispersed throughout the tissue. To understand these diseases it is therefore essential to be able to pinpoint individual sparsely dispersed cells. Detection of dispersed immune cells is especially difficult in this context, as they are likely to be found next to highly metabolically active tubular cells which tend to dominate any neighbouring immune signal.

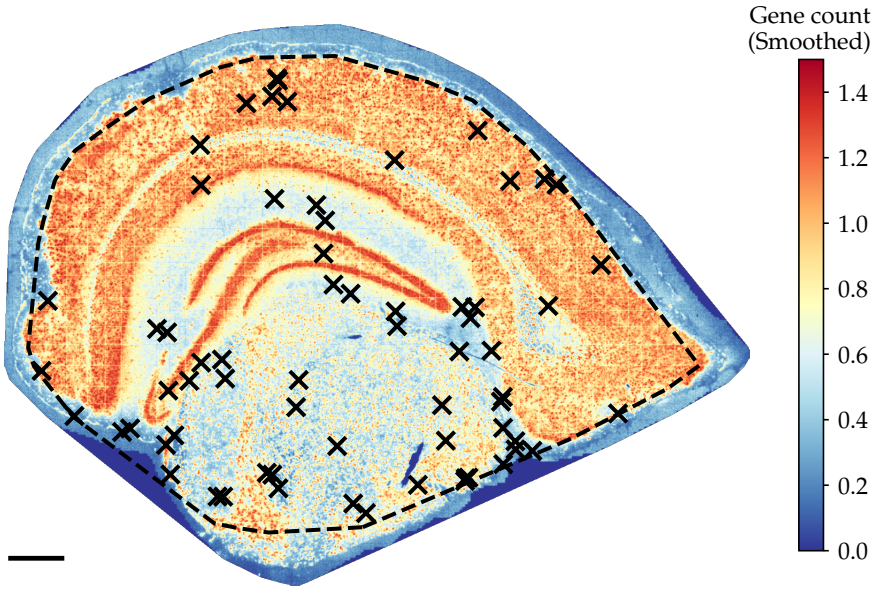


Figure 4.5. PVM cells (black crosses) localised by TopACT in adult mouse brain data profiled by Stereo-seq [Che+22]. Background heatmap shows smoothed transcript count. Black dashed line shows convex hull of high-density regions, to which analysis is restricted. Scale bar: 0.5 mm.

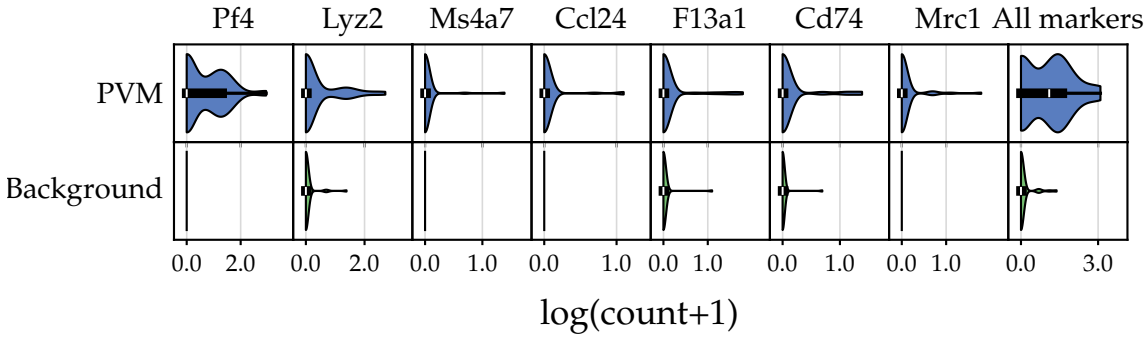


Figure 4.6. Violin plots for expression of common markers of PVM cells, for TopACT predicted PVM cells (blue, $n = 66$ cells), and randomly sampled background cells (green, $n = 66$ cells), across the entire mouse brain sample. Each plot corresponds to the expression counts of a single given marker gene in cells labelled with the given cell type. Violins show kernel density estimate of data distribution. Inner box-and-whisker shows summary statistics as follows: white centre line shows median; box limits show interquartile range; whiskers show full range. Log scale.

In this section we will use TopACT to detect immune cells in mouse kidneys treated with toll-like receptor 7 (TLR7), which promotes lupus-like autoimmune behaviour [Yok+14]. This behaviour includes mild renal immune infiltration, where immune cells infiltrate filtering structures in the kidney known as glomeruli. Detection of this behaviour was found to be challenging with segment-then-classify approaches to cell-type classification, and so the ultimate goal of this section is to demonstrate that TopACT is able to detect the presence of immune infiltration in samples treated with TLR7.

For this experiment we analysed a new Stereo-seq [Che+22] data set of the mouse kidney with an inter-spot distance of 0.715 µm. The data set contained sections from three mice: four slices from a control sample, and two and four slices respectively from two samples treated with TLR7. To build a local classifier for TopACT we used a paired snRNA-seq data set of aggregated control and treated samples with cell types annotated using Seurat [Sat+21].

4.3.1 Cell localisation pipeline

We used a standard image analysis pipeline to extract single-cell loci from TopACT output. In detail, for a given cell type and sample we produced a binary image representing spots that were classified by TopACT with the given cell type. We then performed a difference of Gaussians (DoG) blob detection [Low04; vdW+14] on a Gaussian smoothing of this binary image to extract single-cell loci (see Figure 4.7).

4.3.2 Method validation

To validate the performance of TopACT on kidney data, we first tested its ability to detect podocyte cells. Podocyte cells localise exclusively within glomeruli, which are large enough that they can be easily detected by existing methods at Bin 20. This provides an ideal ground truth for validation. We used the described cell localisation pipeline to extract single podocyte cell loci from TopACT output. Figure E2 shows that these predicted podocyte cells strongly colocalise with the ground truth glomeruli, validating the use of TopACT on these data. Furthermore, violin plots show that TopACT-predicted podocyte cells are enriched in key podocyte marker genes (Figure E3). The same figure also shows that TopACT-predicted tubular cells are similarly enriched in their respective marker genes, although it is harder to establish ground truth in these cell types. However, one can observe strong

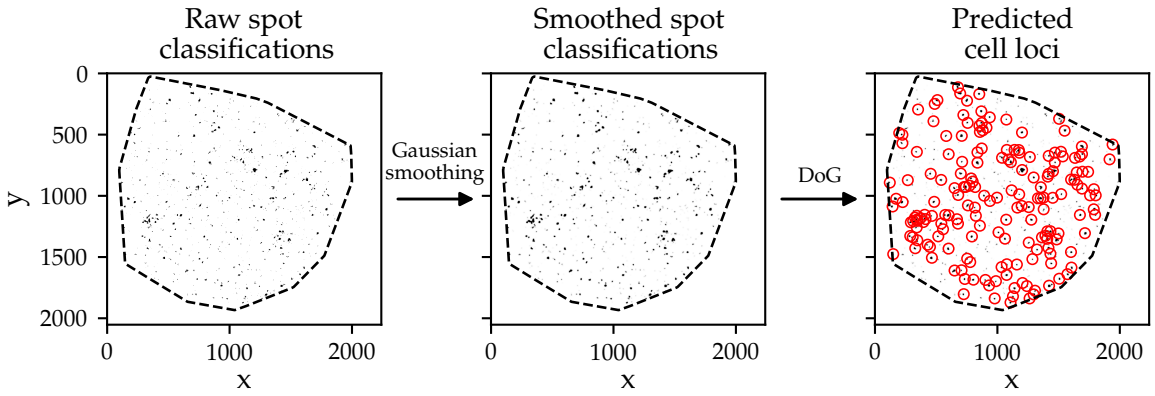


Figure 4.7. Pipeline for extracting single-cell loci from spot-level cell type predictions. First, a binary image is produced indicating spots assigned the given cell type. Then, Gaussian smoothing is applied to produce a greyscale image. Finally, DoG blob detection [Low04; vdW+14] is used to detect regions of high density of the given cell type. These regions are taken as predicted cell loci. In this example, immune cell loci are detected in a representative sample of mouse kidney Stereo-seq data.

colocalisation of TopACT-predicted tubular cells with the relevant marker genes, providing further validation (Figures E4 and E5).

For further validation, we compared the locations of TopACT-called podocyte and immune cells with a cell segmentation produced from ssDNA images, as described in [Che+22], on a single representative sample. We found that 110 out of 137 (80 %) TopACT-predicted immune cells and 46 out of 50 (92 %) TopACT-predicted podocyte cells coincided with an ssDNA-based cell bin. Figure E6 shows cell bins annotated according to the assigned TopACT cell type. Only three ssDNA bins were found to coincide with more than one TopACT-predicted cell, providing further evidence that TopACT predictions correspond to ground truth cells. Visual inspection of these three examples is suggestive of the underlying ssDNA bins being doublets.

4.3.3 Characterising immune infiltration

With the efficacy of TopACT on this data set established, we next sought to characterise the immune activity of the lupus-like condition in treated samples. We used TopACT to localise immune cells across each control and treated sample (Figure 4.8a). We remark that Bin 20 methods failed to recover any significant immune populations in these samples. We split each sample into glomerular and non-glomerular regions (Figure 4.8b) and compared the level of TopACT-detected immune activity in each region type across control and treated samples (Figure 4.8c). In control samples, we

found no significant difference in immune activity between the two region types. In contrast, we found a statistically significant increase (one-sided Welch's t-test, $t = 3.988$, $p = 4.2 \times 10^{-5}$) in immune cell counts per patch in glomerular regions (M = 1.56, SEM = 0.12, $n = 161$ patches) compared with non-glomerular regions (M = 0.95, SEM = 0.09, $n = 180$ patches) in treated samples. Thus, TopACT was able to identify the increased glomerular immune activity in treated samples consistent with lupus-like immune infiltration.

Having established TopACT's ability to extract the locations of individual sparsely distributed cells, it is natural to ask more complex questions about their spatial organisation. In this direction, we applied multiparameter persistent homology to study the spatial distribution of infiltrating immune cells in samples treated with the lupus-like condition. This analysis will be discussed in Section 4.4.

4.4 2-parameter persistence of immune cell organisation in the mouse kidney

Having produced a fine-scale mapping of individual immune cells, and successfully identified the presence of glomerular immune infiltration in treated kidney samples, we next sought to identify more complex spatial patterns in glomerular immune distribution. In particular, we wanted to compute the persistent homology of the predicted immune cell loci around each glomerulus. However, TopACT predictions are noisy, with a high number of outliers, rendering the application of 1PH difficult.

So, given the presence of outliers, we took inspiration from [Vip+21] and sought to apply multiparameter persistence landscapes to a codensity-Rips bifiltration (see Definition 2.1.19) of our immune patches. We were especially interested in the immune activity in glomeruli, so we took the TopACT output on small patches surrounding each glomerulus in each sample. For each patch, we produced a point cloud where points correspond to spots classified as Immune by TopACT. For each of these, we computed the first MPH landscape λ_1 (see Section 2.1.2.2) of the first persistent homology H_1 of each codensity-Rips bifiltration. MPH was computed with RIVET [RIV20] and converted to persistence landscapes using the code from [Vip+21]. For the filtration, we used ρ_5 for codensity and set the maximum Rips radius to 100 spots. In RIVET we set the resolution parameter to 30.

This process yielded a persistence landscape for each glomerulus, which we then averaged over both treated and control samples to arrive at an average summary

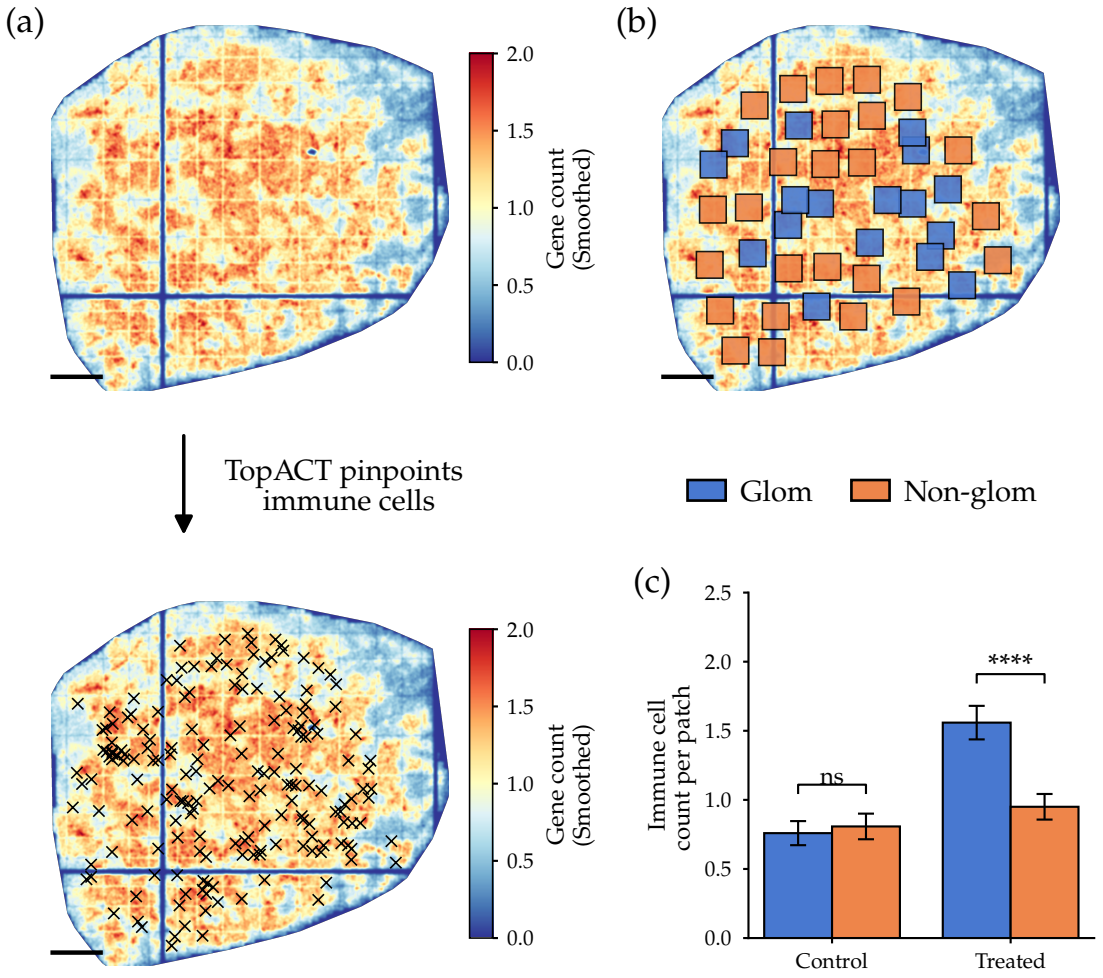


Figure 4.8. TopACT analysis of Stereo-seq kidney sections (4 control, 6 treated). (a) Example TopACT-predicted immune cells. Background: transcript density. (b) Example glomerular (blue) and non-glomerular (orange) patch distribution. (c) Mean TopACT-predicted immune count per patch, by condition and patch type. Error bars show standard error of the mean. Increased immune cell levels (one-sided Welch's t-test; $p = 4.2 \times 10^{-5}$) observed in glomerular ($n = 161$) vs non-glomerular ($n = 180$) patches in treated samples. Scale bars: 0.2 mm.

of the multiscale immune cell topology around each glomerulus (Figure 4.9). We observed that the average persistence landscape corresponding to treated kidneys is activated at high radius parameters, indicating the presence of large loops of immune cells. This led to the hypothesis that this signal is caused by the presence of a peripheral ring structure in immune cells infiltrating glomeruli in treated kidneys.

In [Ben+24], our collaborators performed additional experiments to investigate this proposed hypothesis. They performed multiplex immunofluorescence imaging with markers for T-cells (CD4/CD8), B-cells (CD19) and myeloids (CD11b, CD68, GR1) in both control ($n = 3$) and treated ($n = 3$) kidney sections (Figure 4.10). By comparing the intensity of these markers in the periphery of the glomeruli compared with the centre, they were able to show a significant increase in immune activity in the periphery. This experiment therefore confirms the prediction we made based on MPH analysis.

4.5 Application: state-of-the-art cell segmentation on imaging-based data

Recall that imaging-based spatial platforms trade sample size and transcriptome depth for transcript-level resolution (Section 3.1.2). After initial image preprocessing, the standard approach is to perform cell segmentation based on supplementary data such as nucleus or cell boundary staining, followed by a standard single-cell classification pipeline (Section 3.2.2). In this section we will instead explore the use of TopACT as a cell classifier on imaging-based data.

For this experiment, colleagues at the Oxford Centre for Histopathology Research generated a new human kidney data set on the 10x Xenium platform from an IgA nephropathy biopsy core. They used a standard panel of 377 genes not specific to the kidney [Coo+23]. A matching snRNA-seq data set based on four healthy control kidneys was also generated and annotated using Seurat [Sat+21]. Colleagues performed cell segmentation using the Xenium Onboard Analysis pipeline provided by 10x Genomics, and then used Seurat again to produce a supervised cell-type classification based on the snRNA-seq data set.

Recall that the Xenium platform localises individual RNA transcripts to sub-micron resolution, rather than reading aggregated RNA counts in a fixed grid. To simulate an array-based data set, we binned the Xenium data to 1 µm resolution. We then ran TopACT with a local classifier based on the matching snRNA-seq data

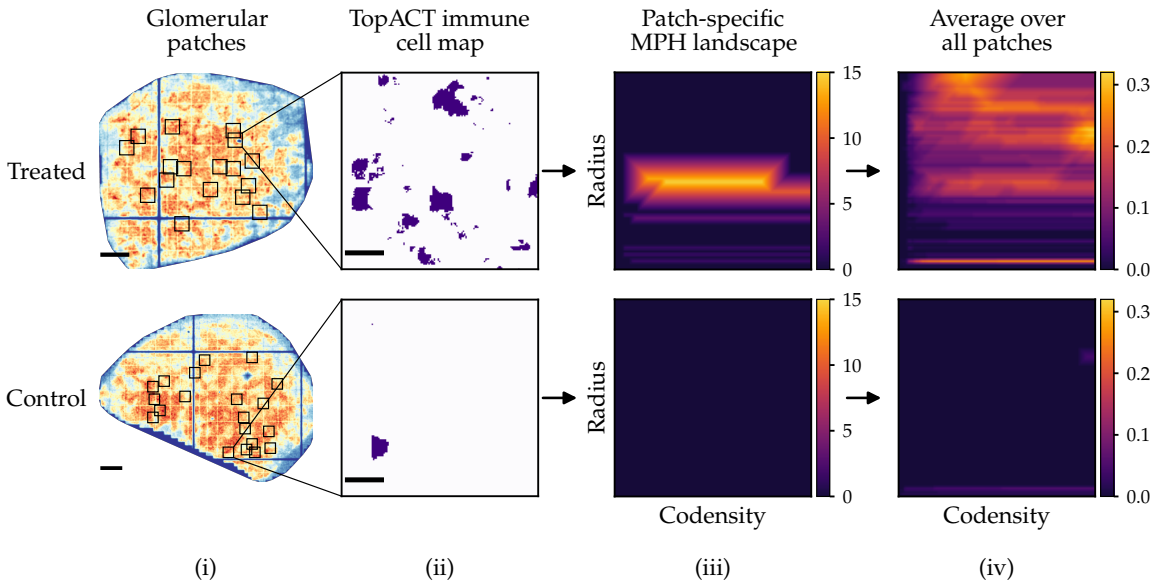


Figure 4.9. MPH analysis of TopACT-predicted immune cell patterns. (i) Glomerular patches. Scale bars: 0.2 mm. (ii) TopACT spot-level immune annotations. Scale bars: 20 µm. (iii) Single-patch MPH landscapes λ_1. (iv) Average MPH landscapes $\overline{\lambda_1}$ over all patches. Treated average indicates large peripheral loop structures.

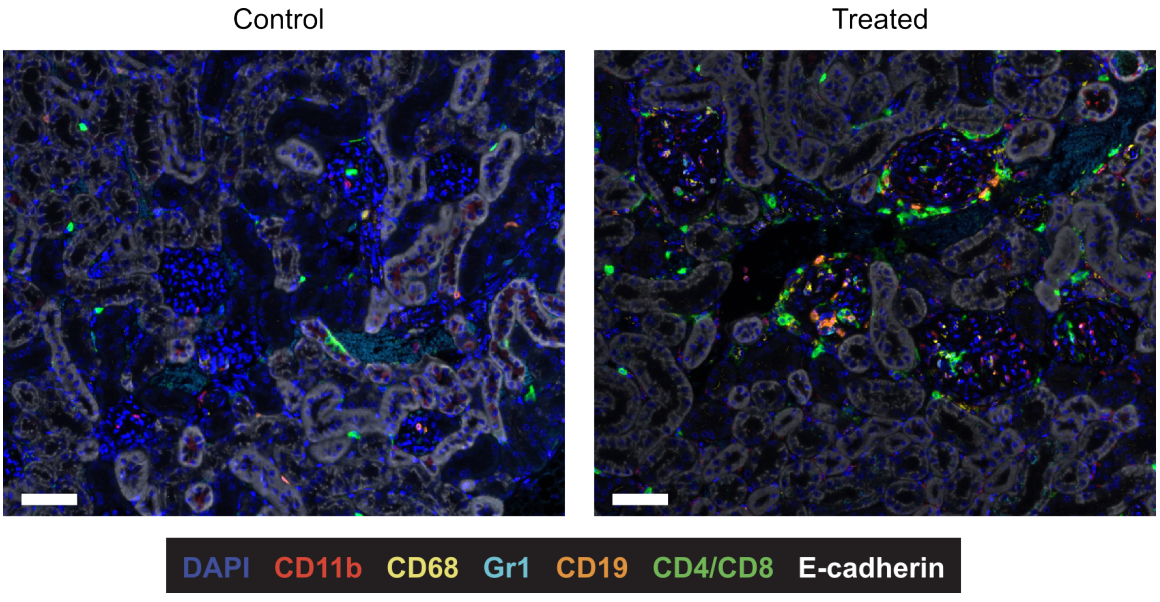


Figure 4.10. Multiplex immunofluorescence imaging informed by TopACT MPH predictions. Colours correspond to DAPI nuclear staining (blue) and T cell (CD4/CD8; green), B cell (CD19; orange) and myeloid (CD11b, CD68 and GR1; red, yellow and teal) immune subtype markers. Scale bars: 100 µm. Reproduced from [Ben+24].

set. For this analysis we focused on nine cell types corresponding to tubular and glomerular structures in the kidney.

Figure 4.11 shows a representative region with a number of tubular structures present. The top row shows H&E staining of the section, which is a standard visualisation of tissue structure. In the middle row is the output of the Xenium Onboard Analysis pipeline restricted to three tubular cell types: proximal tubule (PT), thick ascending limb (TAL), and distal tubular principal cell (PC). Finally, below this is the TopACT cell-type assignment restricted to the same three cell types. Notice that both methods are able to detect the presence of all three types of tubular structure. However, the morphology of the Xenium-predicted tubule segments is in general less consistent with the H&E reference than the TopACT-predicted tubule segments. In particular, Xenium fails to detect the empty lumen space within each tubule segment, and consistently predicts individual cells with distended boundaries. TopACT also performs much better at identifying distal tubular structures, especially the two in the top right of the example region which are only partially resolved by the Xenium pipeline.

Figure 4.12 shows a representative region containing two glomeruli. Again, the top row shows a stain of the section with the glomeruli outlined with red circles, and a magnification of a single glomerulus is shown in the second column. The middle and the bottom rows show the Xenium Onboard Analysis and TopACT classifications respectively, restricted to two cell types which localise to glomeruli: endothelial cells (ECS) and podocytes (Pod). Podocytes in particular are highly specialised cells which are found exclusively in the glomeruli. However, we can see that the Xenium segmentation identifies a spurious family of podocyte cells external to the glomeruli, highlighted with red circles in Figure 4.12. The TopACT-predicted podocyte cells are correctly localised exclusively to the glomeruli. Moreover, similar to the tubular structures, the TopACT segmentation more faithfully recovers the empty space around the periphery of each glomerulus.

4.6 Discussion

In this chapter, we have introduced, implemented and applied TopACT, a multiscale method for topological automatic cell-type classification. The proposed approach resolves cell-type information at subcellular resolution and zeros in on the location of elusive sparsely dispersed cells. By replacing the fixed-window view with a flexible, multiscale lens, TopACT achieves significantly higher accuracy in subcellular spatial

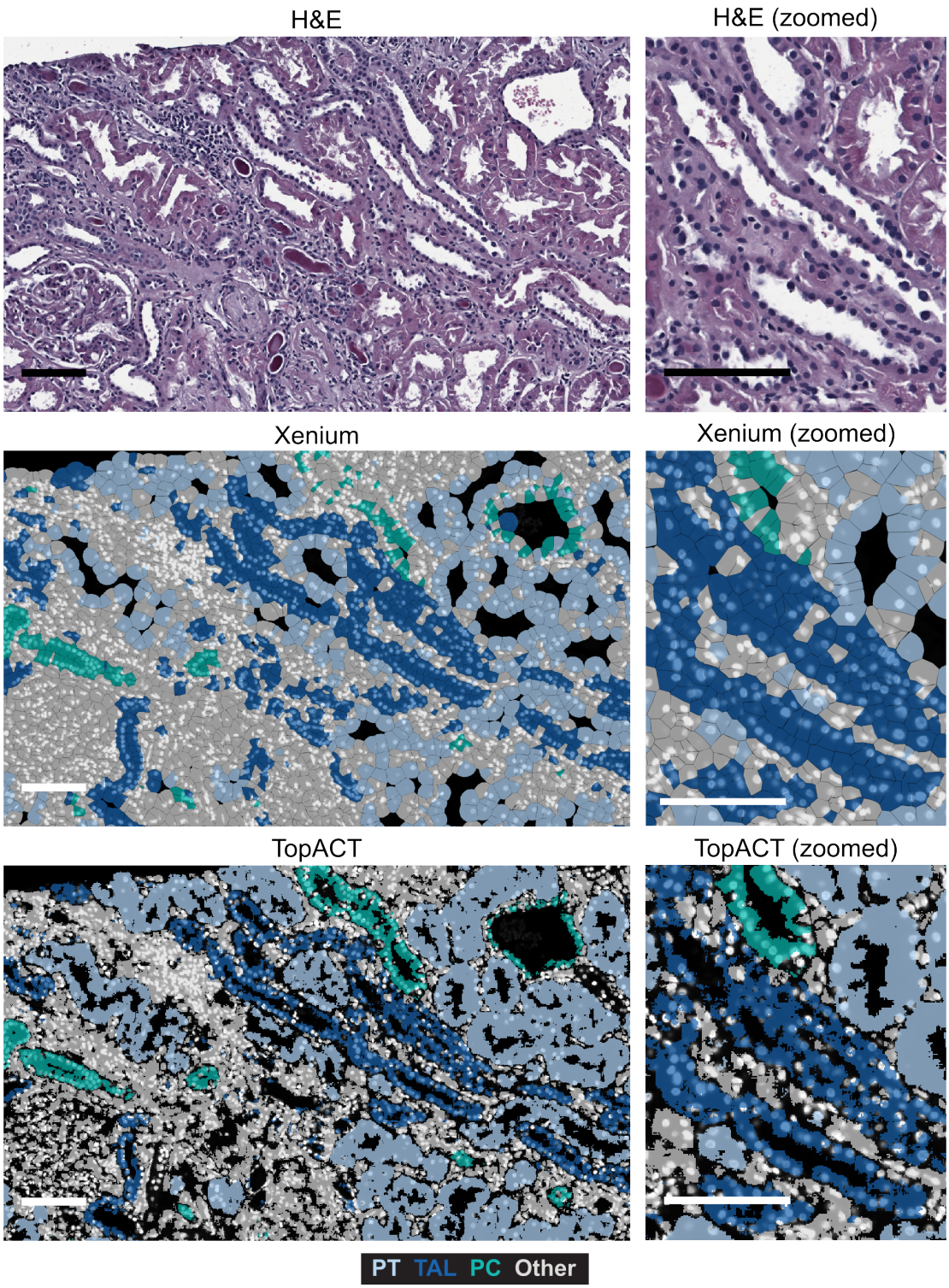


Figure 4.11. Analysis of tubular structures in human kidney profiled on the Xenium platform. Top row: H&E staining of an exemplar tubular region. Middle row: Standard Xenium cell segmentation and supervised labelling. Bottom row: TopACT cell-type assignment. Scale bars: 100 µm. Cell types: Proximal tubule (PT, pale blue); thick ascending limb (TAL, dark blue); principal cell (PC, turquoise).

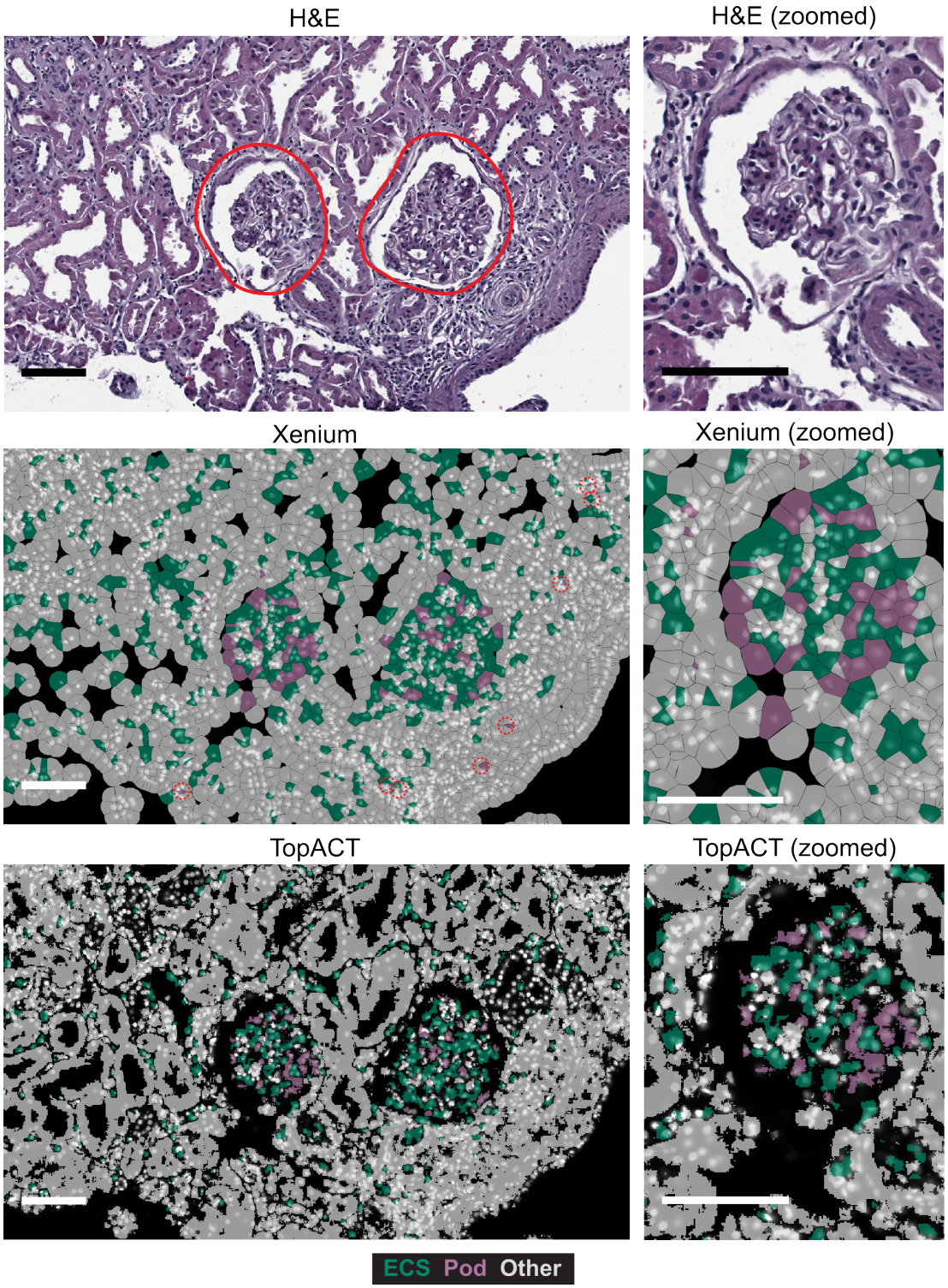


Figure 4.12. Analysis of glomerular structures in human kidney profiled on the Xenium platform. Top row: H&E staining of an exemplar glomerular region. Red circles: glomeruli. Middle row: Standard Xenium cell segmentation and supervised labelling. Red circles: spurious podocytes. Bottom row: TopACT cell-type assignment. Scale bars: 100 µm. Cell types: Endothelial (ECS, green); podocyte (Pod, purple).

cell-type identification than does the naive fixed-window approach. Crucially, our method achieves single-cell resolution by individually resolving rare, sparsely dispersed cells in array-based data which evaded detection by the traditional fixed-window binning approach. On imaging-based platforms, by forgoing the need for cell segmentation, TopACT simplifies the cell-type classification process and yields qualitatively more detailed and biologically plausible cell-type maps.

A key limitation of the present TopACT implementation is performance, given that a cell-type classifier is required to be executed at multiple scales over millions of subcellular spots across each sample. In this direction, the recent Sainsc method [Mül+25] has achieved promising results with a similar approach to TopACT by applying a per-spot classification on a smoothed gene expression map, and selecting cell types by cosine nearest neighbour. A key difference is that Sainsc does not consider multiple bandwidths for its smoothing (which would be analogous to TopACT's multiple scales of local neighbourhood), so it is possible that a future version of TopACT could save performance by restricting to a preselected scale.

Another area for improvement is in the quality of the local classifier (see Section 4.1.2.1). In the experiments in this thesis we have only considered a very simple SVM-based classifier, and while this has performed well, the provided modular implementation of TopACT means that different local classifiers can be readily integrated. The choice of classifier could even be partly avoided by using a consensus method such as popV which aggregates the classifications from several different methods [Erg+24].

One fundamental limitation of the TopACT method is that it is unable to see cell boundaries between cells of the same type. This means that, for example, TopACT is unable to predict the total number of cells present in a sample. To overcome this, it could be prudent to use TopACT annotations in tandem with a complementary segmentation step based on paired imaging data. TopACT would be used to ensure that rare cells are not lost by the segmentation, and the segmentation step would be used to provide further clarity on areas of contiguous cells such as tubular structures.

Looking further ahead, TopACT is a flexible framework which can be easily adapted to new technologies. ST in 3D is quickly becoming the new frontier in the field [Sch+24; Sui+25; Pre+25], and TopACT is readily applicable to such data sets by simply changing the underlying metric space X. Identifying cell boundaries is much harder in the 3D setting, so a segmentation-free approach such as TopACT is likely to be an invaluable tool for the next generation of *in situ* transcriptomics.

Chapter 5

Diversity measures in single-cell transcriptomics

In single-cell transcriptomics, almost all data analysis begins with a sequence of transformations aimed at partitioning cells into distinct cell types. This approach is natural for a wide range of reasons. However, it also presents some challenges in both practice and theory. One way to deal with this problem is to develop methods for single-cell analysis which are robust to the choice of cell-type classification or, even better, do not rely on it at all. The aim of this chapter is to introduce one such method, by drawing inspiration from recent developments in the study of ecological diversity.

The problems associated with cell-type classification are numerous. On a practical level, it is not obvious exactly which preprocessing steps should be taken, and moreover each step typically has a number of hyperparameters which must be set. Even if a preprocessing pipeline has been fixed, it has been shown that even the choice of software library and its specific version can result in highly non-negligible differences in downstream analysis [Ric+26]. On a theoretical level, there are further problems. Chief among them is the fact that cell states lie on continuous trajectories, so there are no strict cut-offs at which to define the boundaries between cell types. Furthermore, there is always a choice of resolution to be made with respect to cell subtypes: practitioners often choose to further subclassify cell types of particular interest, and these choices are again highly non-canonical. The consequence of all of these choices is that it is almost impossible to guarantee reproducibility of cell-type classifications. This presents a problem: how do we design methods which are robust to different equally acceptable partitions into cell types? In other words,

we would like the output of our methods to be stable when given any biologically plausible cell-type assignment as input.

In this chapter, we will present a surprising connection between this problem and recent advances in the study of ecological diversity. A paramount question for ecologists is to quantify the diversity of a community of species. A traditional measure of diversity in this context is the information-theoretic entropy of the distribution of species, however this can have some unsatisfactory properties. As an example, consider two ecosystems: A and B. Ecosystem A consists of three different species of pine tree, in equal abundance. Ecosystem B consists of a species of tree, a species of fox, and a species of butterfly, all in equal abundance. It seems sensible that ecosystem B should be given a higher diversity score than ecosystem A, but both species distributions have the same entropy. The problem here is that the entropy-based diversity measure does not take into account the relative similarity of the species in each ecosystem. While this might seem a contrived example, in practice this issue can arise in quite subtle ways. For example, sub-partitioning a species into subspecies will completely change the entropy of an ecosystem, despite the only change having been to the labels we have chosen to assign.

To address this problem, Leinster and Cobbold have proposed a family of *similarity-sensitive* diversity measures [LC12]. These *LC diversity* measures incorporate information about the relative similarity of the species in the ecosystem as a way to make them more robust to the choice of species assignment. In particular, the existence of certain continuity properties (Remark 5.2.8) ensures that LC diversity is robust to the kinds of relabelling that cause problems for entropy-based measures.

Translating these ideas into the context of single-cell transcriptomics, we propose repurposing LC diversity as a measure of gene-expression heterogeneity in tissue. The robustness to species assignment in the ecological setting translates to robustness to cell-type assignment—including subtyping—in the single-cell setting. In this chapter we give two applications of LC diversity to single-cell data sets in the context of embryonic development, where we can track the process of cell-type differentiation and maturation. In the past, authors have relied on *ad hoc* cell-type-based measures of heterogeneity, and here we show that LC diversity provides a measure that is both more robust and more powerful than these pre-existing approaches.

Cell-type-agnostic analysis does not come without trade-offs. Our gains in reproducibility are offset by losses in interpretability: cell type labels give practitioners useful a vocabulary for biological reasoning. The methods proposed in this chapter

should be seen as complementary to cell-type-based approaches, and they provide scientists with sanity checks to ensure that the coarseness of a cell type label is not obscuring important fine-grained structure.

This chapter is structured as follows. We will begin by giving an overview of both entropy-based diversity and LC diversity as they apply to ecology. Then, we will rephrase these ideas in terms of single-cell transcriptomics, and provide a pipeline for producing the prerequisite similarity data from an sc/snRNA-seq count matrix. We will then demonstrate this new method on two real-world sc/snRNA-seq data sets from studies on embryonic development in mouse and human tissue. In both cases we will show that LC diversity provides a natural quantification of properties of the data which are only discussed based on unreliable cell-type analysis in the original studies. To finish, we will see how partitioned diversity measures as proposed by Reeve *et al.* [Ree+16] can be used to visualise *in situ* diversity from mouse hippocampus tissue profiled by the Slide-seqV2 spatial transcriptomics platform [Sti+21].

Attribution This chapter is based on initial conversations with Emily Roff. All applications presented here are my own original work. I thank Jesse Veenvliet for highly illuminating discussions on the SEM data set studied in Section 5.2.3.

5.1 Diversity in ecology

The aim of this section is to give an overview of measures of ecological diversity, building towards the similarity-sensitive measure introduced by Leinster and Cobbold [LC12]. For a much more thorough treatment we refer the reader to the textbook of Leinster on the subject [Lei21], on which this section is heavily based.

Let us fix some notation. We write

$$\Delta_N \coloneqq \left\{ p \in [0,1]^N : \sum_{i=1}^{N} p_i = 1 \right\} \tag{5.1}$$

for the standard simplex on N vertices, whose elements we interpret as probability distributions. Given $p \in \Delta_N$ we write

$$\operatorname{supp}(p) \coloneqq \{1 \leq i \leq N : p_i \neq 0\} \tag{5.2}$$

for the support of p. We will write $u \in \Delta_N$ for the uniform distribution $u_i := 1/N$ for all $1 \leq i \leq N$.

Now, in the ecological setting we begin with an ecosystem populated by some set of species $\mathcal{S} = \{S_1, \ldots, S_N\}$. We assume that we have two pieces of additional data on the set of species:

1. A probability distribution $p \in \Delta_N$ which we interpret as a distribution on $\mathcal{S}$ by reading p_i as the relative abundance of the species S_i for each $1 \leq i \leq N$;

2. and an $N \times N$ similarity matrix Z such that for each $1 \leq i, j \leq N$ the entry $Z_{ij} \in [0, 1]$ measures the similarity of the two species S_i and S_j.

In this thesis we will assume that $Z_{ii} = 1$ for all $1 \leq i \leq N$, that is to say that a species is always completely similar to itself.

Remark 5.1.1. In the ecological setting, there is no prescribed choice of similarity matrix. Common choices include genetic and phylogenetic similarity, and a more complete list is given in [Lei21, Examples 6.1.1].

Example 5.1.2. Given a pseudo-metric $d\colon \mathcal{S} \times \mathcal{S} \to [0, \infty)$ one can construct a similarity matrix Z on $\mathcal{S}$ by

$$Z_{ij} = \exp\big(-d(S_i, S_j)\big). \tag{5.3}$$

Remark 5.1.3. In [LC12, p. 479] the authors note that it is not necessary to assume that Z is symmetric, and give some examples of applications where one might want to take an asymmetric similarity matrix. However, in this thesis we will always be working with symmetric Z.

With the setting established, we now turn to the more simple of our two families of diversity measures.

5.1.1 Entropy-based diversity

As a first attempt at quantifying the diversity of the ecosystem $(\mathcal{S}, p, Z)$, we will introduce a family of measures based on the information-theoretic entropy of the distribution p. These measures, called *Hill numbers*, were first introduced in [Hil73] and have been widely used by ecologists ever since.

We start with the simplest case of the *Hill number of order* 1 (Definition 5.1.7) which is based on the *Shannon entropy* of p:

Definition 5.1.4. Let $p \in \Delta_N$ be a probability distribution. The *Shannon entropy* $H(p)$ of p is given by

$$H(p) \coloneqq -\sum_{i \in \mathrm{supp}(p)} p_i \log p_i. \tag{5.4}$$

We state two properties of Shannon entropy which align with our intuition of how a measure of ecological diversity should behave [Lei21, Lemmas 2.2.4 and 2.2.5].

Proposition 5.1.5. *Let $p \in \Delta_N$ be a probability distribution. Then*

$$0 \leq H(p) \leq \log N. \tag{5.5}$$

Moreover, the lower bound is attained if and only if $p_i = 1$ for some $1 \leq i \leq N$, and the upper bound is attained if and only if $p = u$ is the uniform distribution.

Proposition 5.1.6. *Equip Δ_N with the Euclidean metric. Then the Shannon entropy $H \colon \Delta_N \to \mathbb{R}$ is continuous.*

Let us frame these two statements in ecological terms. Roughly speaking, they are saying that the Shannon entropy is recording how similar a distribution is to the uniform distribution, which we take to be the maximally diverse distribution on N species. Conversely, the minimal entropy is attained when only a single species is present, which agrees with our intuition that this is the least diverse possible distribution on N species.

We can now give our first definition of a Hill number of the ecosystem $(\mathcal{S}, p, Z)$ as the exponential of the Shannon entropy:

Definition 5.1.7. Let $p \in \Delta_N$ be a probability distribution. The *Hill number of order 1* is the exponential of the Shannon entropy:

$$D_1(p) \coloneqq \exp\big(H(p)\big) = \prod_{i \in \mathrm{supp}(p)} p_i^{-p_i}. \tag{5.6}$$

Why use the Hill number in place of entropy? There are a number of subtle reasons, and we refer the reader to [Lei21, pp. 54–55] for a discussion. The following immediate corollary of Proposition 5.1.5 provides one small motivation.

Corollary 5.1.8. *Let $p \in \Delta_N$ be a probability distribution. Then*

$$1 \leq D_1(p) \leq N. \tag{5.7}$$

Moreover, the lower bound is attained if and only if $p_i = 1$ for some $1 \leq i \leq N$, and the upper bound is attained if and only if $p = u$ is the uniform distribution.

Remark 5.1.9. In this sense, we can interpret D_1 as measuring the *effective number of species* in the ecosystem. In the uniform case, there are precisely N species present. As diversity decreases, this number decreases until we arrive at only a single species present, corresponding to $D_1(p) = 1$.

Now, at the start of this section we promised a family of diversity measures. Just as the Hill number of order 1 was attained as the exponential of the Shannon entropy, the Hill numbers of order q for $q \in [0, \infty]$ are obtained as exponentials of the *Rényi entropies* of order q [Rén61]. We refer the reader to [Lei21, Section 4.3] for a detailed exposition of Rényi entropies, and here we instead give direct expressions for the Hill numbers of arbitrary order.

Definition 5.1.10. Let $p \in \Delta_N$ be a probability distribution. The *Hill number of order q* is defined piecewise for $q \in [0, \infty]$. Firstly, if $q \neq 1, \infty$ then

$$D_q(p) := \left(\sum_{i \in \mathrm{supp}(p)} p_i^q \right)^{1/(1-q)}. \tag{5.8}$$

Additionally, we have

$$D_1(p) = \prod_{i \in \mathrm{supp}(p)} p_i^{-p_i}, \tag{5.9}$$

$$D_\infty(p) = \frac{1}{\max_{i \in \mathrm{supp}(p)} p_i}. \tag{5.10}$$

Note in particular that Definition 5.1.10 generalises Definition 5.1.7. We also have an exact generalisation of Corollary 5.1.8 [Lei21, Lemma 4.4.3]:

Proposition 5.1.11. *Let $p \in \Delta_N$ be a probability distribution. Then, for any $q \in [0, \infty]$, we have*

$$1 \leq D_q(p) \leq N. \tag{5.11}$$

Moreover, the lower bound is attained if and only if $p_i = 1$ for some $1 \leq i \leq N$, and the upper bound is attained if and only if $p = u$ is the uniform distribution.

Example 5.1.12. In the case $q = 0$ we have that $D_0(p) = \#\,\mathrm{supp}(p)$, so the Hill number of order 0 is simply measuring the number of observed species with no heed to

their distribution. It follows that D_0 is not continuous at the boundary of Δ_N, since increasing the abundance of a species from 0 to any $\varepsilon > 0$ will discontinuously increase D_0 by 1.

Nevertheless, D_q is continuous in p at every point [Lei21, Lemma 4.4.6]:

Proposition 5.1.13. *The Hill number $D_q \colon \Delta_N \to \mathbb{R}$ is continuous everywhere when $q > 0$. When $q = 0$ it is continuous on the open simplex Δ_N°.*

How should we interpret the parameter q? Let's first state some basic properties of Hill numbers, the proofs of which can be found in [Lei21, Section 4.4].

Proposition 5.1.14. *Fix a probability distribution $p \in \Delta_N$. Then*

1. *the Hill number $D_q(p)$ is continuous in q,*
2. *the Hill number $D_q(p)$ is decreasing in q, i.e. $q' \geq q \implies D_{q'}(p) \leq D_q(p)$.*

Combining Proposition 5.1.14 with our intuition for the values of D_q in the extreme cases $q \in \{0, \infty\}$ provides us with the interpretation that q is controlling the level of sensitivity to rare species. To be precise, as q increases we become more insensitive to the rare species in our ecosystem. The parameter q is therefore referred to as the *sensitivity parameter.*[1]

5.1.2 Similarity-sensitive diversity

A key issue with the Hill number is that it does not take into account the similarity matrix Z, which leads to the types of error discussed in the introduction. In this section we aim to rectify this limitation, by introducing a family of *similarity-sensitive* measures as first defined by Leinster and Cobbold [LC12]. The idea here is that a pair of species with a high mutual similarity should have a lower joint contribution to the overall diversity of the system than two species with a low mutual similarity. As we will see, these *LC diversities* generalise the Hill numbers defined in the previous section.

Take an ecosystem $(\mathcal{S}, p, Z)$. Let's consider the product $Zp \in [0,1]^N$, whose elements are given by

$$(Zp)_i = \sum_{j=1}^{N} Z_{ij} p_j. \tag{5.12}$$

[1] Perhaps *insensitivity* would be a better label, given that increasing q decreases the sensitivity to rare species, but we'll stick with the established convention.

We can think of $(Zp)_i$ as recording the expected similarity between a given member of S_i and another member of the ecosystem chosen uniformly at random. In other words, $(Zp)_i$ records how ordinary the species S_i is. It follows that the reciprocal $1/Zp$ in some sense records the uniqueness of each species in $\mathcal{S}$. The idea behind the *LC diversity*[2] measures of [LC12] is to use $1/Zp$ as a weighting when computing the geometric mean of our abundance vector p.

Definition 5.1.15. Let $(\mathcal{S}, p, Z)$ be an ecosystem on N species. The *LC diversity of order* q is defined piecewise for $q \in [0, \infty]$. Firstly, if $q \neq 1, \infty$ then

$$D_q(p; Z) \coloneqq \left(\sum_{i \in \mathrm{supp}(p)} p_i (Zp)_i^{q-1} \right)^{1/(1-q)}. \tag{5.13}$$

Additionally, we have

$$D_1(p; Z) = \prod_{i \in \mathrm{supp}(p)} (Zp)_i^{-p_i}, \tag{5.14}$$

$$D_\infty(p; Z) = \frac{1}{\max\limits_{i \in \mathrm{supp}(p)} (Zp)_i}. \tag{5.15}$$

Remark 5.1.16. It is easy to see that $D_q(\cdot; I_N) \equiv D_q(\cdot)$, so the LC diversity of order q generalises the Hill number of order q.

Remark 5.1.17. Recall that we assumed our similarity matrices satisfy $Z_{ii} = 1$ for all $1 \leq i \leq N$. We therefore have

$$(Zp)_i = \sum_{j=1}^{N} Z_{ij} p_j \geq Z_{ii} p_i = p_i. \tag{5.16}$$

It follows that incorporating similarity information into our diversity measure increases the ordinariness of each species, and consequently decreases the overall ecosystem diversity. To be precise, $D_q(p; Z) \leq D_q(p; I_N)$ for any similarity matrix Z.

We now state a key continuity result which will be useful when discussing cell subtypes in the next section [Lei21, Lemma 6.2.4(ii)].

Proposition 5.1.18. *Fix a* $q \in [0, \infty]$ *and* $p \in \Delta_N$. *Then* $D_q(p; Z)$ *is continuous in* Z.

[2] 'LC' is not standard nomenclature, but we use it in this thesis to provide further clarity when we turn to applications in transcriptomics.

There are similar continuity results in q and p, with some subtleties to consider in the extreme cases $q \in \{0, \infty\}$, but we do not state these results here.

5.2 Diversity in transcriptomics

Now that we have covered the basics of LC diversity as it applies to ecology, let's rephrase these ideas in terms of transcriptomics. Instead of an *ecosystem* given by a set of *species* and their relative abundance and similarity data, we instead have a *tissue sample* given by a set of *cell types* and their relative abundance and similarity data. In particular, the similarity data will be based on gene expression.

To make this idea precise, take a single-cell transcriptomics sample. We make no restriction about what 'sample' means here, and it could be a single experimental replicate, multiple such replicates pooled together, or a subset of such a replicate. First, we write $C = \{c_1, \ldots, c_M\}$ for the collection of cells in our sample, and $\mathcal{G} = \{g_1, \ldots, g_D\}$ for the collection of genes. An sc/snRNA-seq experiment produces a count matrix X of shape $M \times D$ where X_{ij} records the total expressed transcript count of the gene g_j in the cell c_i. We further assume that C has been partitioned into a set of disjoint cell types $\mathcal{T} = \{T_1, \ldots, T_N\}$ (see Section 3.2.1 for how this might be done in practice). We do not preclude the possibility that each cell defines its own type, i.e. $T_i = \{c_i\}$ for all $1 \leq i \leq M$.

The set $\mathcal{T}$ is the first piece of data needed to compute LC diversity. The second, the relative abundance distribution, is easily computed: $p_i = \#T_i/M$. The final piece of data, the similarity matrix, is not as obvious. There are many possible options to take, but in this thesis we offer the following procedure:

1. First, identify highly variable genes (HVGs). The idea here is that genes that do not vary much between cells will artificially inflate the similarity measure, in turn decreasing the sensitivity of the diversity score. For notational ease we will assume that the gene space has already been so filtered.

2. For a given cell type T, form its average expression vector X_T with entries:

$$(X_T)_j \coloneqq \frac{1}{\#T} \sum_{c_i \in T} X_{ij}. \tag{5.17}$$

3. Now, define the similarity between cell types T_i and T_j to be the cosine similarity between their two average expression vectors:

$$Z_{ij} \coloneqq \frac{X_{T_i} \cdot X_{T_j}}{\|X_{T_i}\|\|X_{T_j}\|} \in [0, 1]. \tag{5.18}$$

We note that $Z_{ij} \geq 0$ since all the entries of the count matrix are non-negative.

Definition 5.2.1. We say that the *diversity of order q* of the cell-type partition $\mathcal{T}$ is the diversity of the similarity matrix Z and relative abundance distribution p obtained from the above procedure:

$$D_q(\mathcal{T}) \coloneqq D_q(p; Z). \tag{5.19}$$

Remark 5.2.2. It is natural, given the discussion on ecological diversity, to partition the sample into coarse cell types. However, in the setting of transcriptomics where we have a feature vector describing each individual cell, it is possible to avoid this step. In particular, we can form singleton cell types $T_i = \{c_i\}$ for each $1 \leq i \leq M$. The resulting abundance distribution is the uniform distribution u, and we can then carry out the same procedure to arrive at a similarity matrix indexed by individual cells. In particular, this allows for diversity to be measured even in cases where it does not make sense to partition into cell types.

Remark 5.2.3. Here we have computed the similarity between cell types as the cosine similarity between their average expression vectors. It is far from obvious that this is the 'correct' choice. Other choices could include other correlation measures, or the inverse exponential $\exp(-d)$ for some distance d computed on the average expression vectors.

To shed some additional light on this notion of similarity, we can derive an alternative form for the similarity matrix Z in terms of the sub-expression-matrices for each cell type. For a cell type $T \subseteq C$ we write $X|_T$ for the submatrix of X restricted to those rows corresponding to elements of T. Given two matrices A and B of shape $M_1 \times D$ and $M_2 \times D$ respectively we write

$$\rho(A, B) = \sum_{\substack{1 \leq i \leq M_1 \\ 1 \leq j \leq M_2}} A_i \cdot B_j \in \mathbb{R} \tag{5.20}$$

for the sum of all elements of the matrix AB^{t} of shape $M_1 \times M_2$.

Proposition 5.2.4. *The similarity matrix Z corresponding to a cell-type partition $\{T_1, \ldots, T_N\}$ has entries*

$$Z_{ij} = \frac{\rho(X|_{T_i}, X|_{T_j})}{\sqrt{\rho(X|_{T_i}, X|_{T_i})\rho(X|_{T_j}, X|_{T_j})}}. \tag{5.21}$$

Proof. We will consider the numerator and denominator in (5.18) separately. First, we have that

$$\begin{aligned}
X_{T_i} \cdot X_{T_j} &= \sum_{k=1}^{D} (X_{T_i})_k (X_{T_j})_k && (5.22)\\
&= \sum_{k=1}^{D} \Big(\frac{1}{\#T_i} \sum_{c_s \in T_i} X_{sk}\Big)\Big(\frac{1}{\#T_j} \sum_{c_t \in T_j} X_{tk}\Big) && (5.23)\\
&= \frac{1}{\#T_i \#T_j} \sum_{k=1}^{D} \sum_{\substack{c_s \in T_i \\ c_t \in T_j}} X_{sk} X_{tk} && (5.24)\\
&= \frac{1}{\#T_i \#T_j} \sum_{\substack{c_s \in T_i \\ c_t \in T_j}} \sum_{k=1}^{D} X_{sk} X_{tk} && (5.25)\\
&= \frac{1}{\#T_i \#T_j} \sum_{\substack{c_s \in T_i \\ c_t \in T_j}} X_s \cdot X_t && (5.26)\\
&= \frac{1}{\#T_i \#T_j} \rho(X|_{T_i}, X|_{T_j}). && (5.27)
\end{aligned}$$

Setting $i = j$ we recover

$$\|X_{T_i}\|^2 = X_{T_i} \cdot X_{T_i} = \frac{1}{(\#T_i)^2} \rho(X|_{T_i}, X|_{T_i}), \tag{5.28}$$

from which the result follows. □

Remark 5.2.5. Note that (5.21) has an analogous form to the definition of the cosine similarity between two vectors. Of course, ρ is not akin to an inner product in any sensible way, so this is nothing more than a passing resemblance.

Consider two cell-type partitions $\mathcal{T} = \{T_1, \ldots, T_N\}$ and $\mathcal{T}' = \{T'_1, \ldots, T'_{N'},\}$. A *relabelling* is a function $\theta : \mathcal{T} \to \mathcal{T}'$. For notational convenience we will sometimes write $\theta(i) = j$ to mean $\theta(T_i) = T'_j$. Given a similarity matrix Z on $\mathcal{T}'$ and a distribution

$p \in \Delta_N$, a relabelling θ induces a similarity matrix $Z\theta$ on $\mathcal{T}$ and a distribution $\theta p \in \Delta_{N'}$ as follows [Lei21, p. 186]:

$$(Z\theta)_{ij} \coloneqq Z_{\theta(i)\theta(j)}, \tag{5.29}$$

$$(\theta p)_i \coloneqq \sum_{j:\theta(j)=i} p_j. \tag{5.30}$$

The following naturality property will be useful in showing that LC diversity is robust with respect to sensible reclassifications [Lei21, Lemma 6.2.6].

Lemma 5.2.6. *For p, Z and θ as above and any $q \in [0, \infty]$ we have*

$$D_q(p; Z\theta) = D_q(\theta p; Z). \tag{5.31}$$

Naturality implies that splitting a cell type up into two subtypes with the same average expression vectors does not change the diversity:

Corollary 5.2.7. *Let $\mathcal{T} = \{T_1, \ldots, T_{N+1}\}$ be a cell-type partition. Write $\mathcal{T}' = \{T_1, \ldots, T_{N-1}, T_N \cup T_{N+1}\}$ for the cell-type partition obtained by merging T_N and T_{N+1}. If $X_{T_N} = X_{T_{N+1}}$ then $D_q(\mathcal{T}) = D_q(\mathcal{T}')$.*

Proof. Write Z, p for the similarity matrix and relative abundance distribution on $\mathcal{T}$ and likewise Z', p' for the same quantities on $\mathcal{T}'$. Note that Z is obtained from Z' by duplicating its last row and column, with $Z_{N+1,N+1} = 1$.

Consider the relabelling $\theta\colon \{T_1, \ldots, T_{N+1}\} \to \{T_1, \ldots, T_N\}$ given by

$$\theta(T_i) = \begin{cases} T_i & \text{if } i < N+1, \\ T_N & \text{if } i = N+1. \end{cases} \tag{5.32}$$

By Lemma 5.2.6 we therefore have $D_q(p; Z'\theta) = D_q(\theta p; Z')$.

Now, we can see that $Z'\theta$ is obtained from Z' by duplicating its last row and column, with $(Z'\theta)_{N+1,N+1} = Z'_{\theta(N+1),\theta(N+1)} = Z'_{N,N} = 1$, so $Z'\theta = Z$. Similarly $\theta p = p'$. So we can conclude that $D_q(p; Z) = D_q(p'; Z')$ as required. □

Remark 5.2.8. The consequence of Corollary 5.2.7, in combination with continuity in the similarity matrix Z (Proposition 5.1.18), is that LC diversity is robust with respect to splitting a cell type into subtypes as long as the average expression of each of the subtypes is not too dissimilar to the average expression of their parent. One can also view this idea the other way round: we do not distort diversity too

much by combining cells into a cell type as long as the cells are reasonably close together in gene expression space. So, as long as two cell-type assignments have reasonably tight clusters, they should yield similar LC diversities.

5.2.1 Existing work

While questions about cell-type composition, heterogeneity, and diversity arise very frequently in single-cell transcriptomics studies, there has been little research dedicated to directly quantifying transcriptional diversity. Karagiannis *et al.* [KMS22] have proposed an adjusted entropy statistic defined as

$$E(p) \coloneqq \frac{H(p) - \log N}{\log N} \in [-1, 0] \tag{5.33}$$

where p is the cell-type relative abundance distribution and $H(p)$ is the Shannon entropy as in Definition 5.1.4. They used this statistic to quantify the effect of extreme old age on diversity in peripheral blood mononuclear cells. This entropy-based analysis showed an increase in cell-type diversity in the extreme old age group. However, this approach suffers from the problems we have already identified with the entropy-based approach, namely that entropy does not take into account cell-type similarity. Unfortunately not all of the data for this study are available online, so we were unable to compare these results to an analysis based on LC diversity.

5.2.2 Application: human adrenal gland development

Our first application is on an snRNA-seq data set of human adrenal gland development originally published in [Jan+21]. The data set consists of 17 human adrenal glands covering 7 different developmental time points ranging from 7 to 17 weeks postconception. It is known that as the adrenal gland develops the population of neuroblasts increases sharply, meaning that the cellular diversity of the adrenal medulla in turn decreases. This setting therefore provides an ideal proving ground for our proposed diversity measure: we aim to show quantitatively that the diversity of the adrenal medulla decreases over time, mirroring the claims of Jansky et al. [Jan+21].

The snRNA-seq data set consisted of 28,422 genes sequenced over 100,337 cells. We preprocessed the data broadly as in the original study [Jan+21]. In particular, we filtered out any cells expressing fewer than 1,000 counts or 500 genes, and any cells

with more than 2.5 % reads mapping to mitochondrial genes. We also filtered out doublets as detected in the original study. This left 98,138 (97.81 %) cells for further processing (Figure 5.1). The cell-type partition $\mathcal{T}$ was as annotated previously (see Table 5.1). For feature selection, we took the 2,000 most differentially expressed genes as defined in the original study.

Before restricting to medullary cells, we considered the diversity of the entire adrenal gland. We computed heterogeneity scores for each sample in three different ways. First, we formed the triple $(\mathcal{T}, p, Z)$ as outlined at the start of Section 5.2. We then computed the Hill number $D_2(p)$ and the LC diversity $D_2(p; Z)$ as normal. Finally, we also computed the population-level LC diversity as discussed in Remark 5.2.2.

The three diversity measures grouped by developmental stage are shown in Figure 5.2. The Hill number shows increased diversity in the early developmental stages and lower diversity in the later stages. We can see a similar pattern in both the cell-type-informed diversity and the population-level diversity, validating the efficacy of both methods. Importantly, the population-level diversity correctly identifies the decrease in diversity without requiring any classification into cell types.

The comparison between the LC diversity with and without a cell-type partition is instructive when taking into consideration Remark 5.2.8. In particular note that the diversity scores with the cell-type partition are roughly half the value of the diversity scores without the partition. Given that diversity is robust with respect to splitting up 'tight' cell types into their constituent cells, this discrepancy suggests that there is a high level of gene expression variability within each of the cell types.

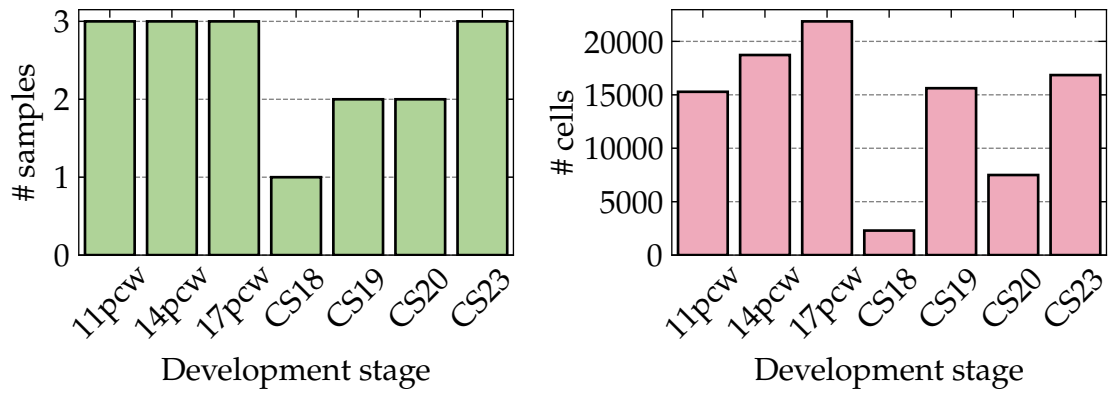


Figure 5.1. Summary statistics over time for human adrenal gland data. Left: number of samples per time point. Right: number of cells per time point.

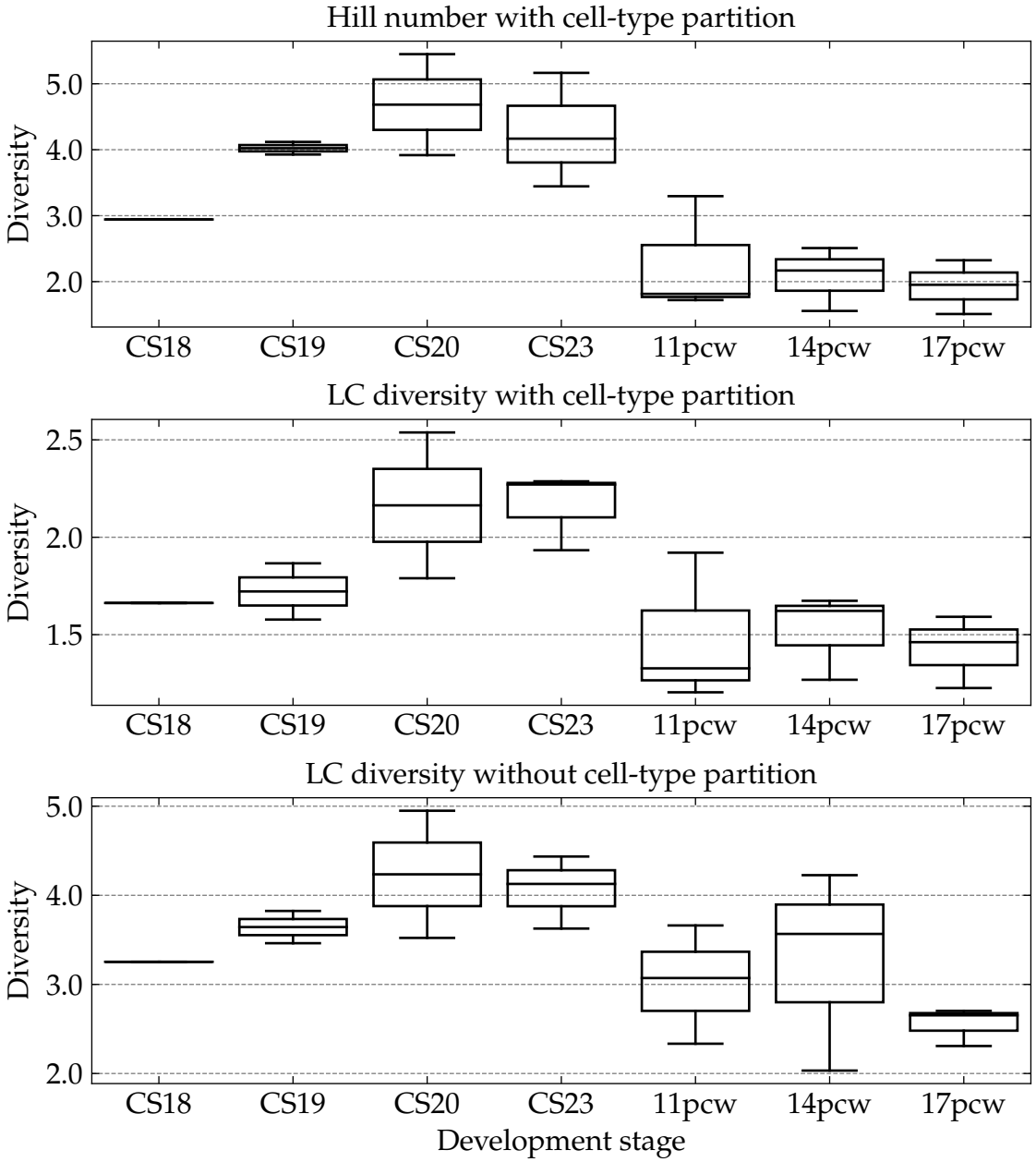


Figure 5.2. Three measures of transcriptional diversity in the human adrenal gland across developmental stages. Top: Hill number of cell-type distributions. Middle: LC diversity based on cell-type partition. Bottom: LC diversity computed without cell-type partition. Centre lines show median; box limits show interquartile range; whiskers show range.

Table 5.1. Cell-type composition of the human adrenal gland data set [Jan+21].

Cell type	n cells	%	Cumulative %
Adrenal cortex	52,507	53.50	53.50
Mesenchymal cells	16,496	16.81	70.31
Endothelial cells	13,730	13.99	84.30
Immune cells	2,167	2.21	86.51
Late neuroblasts	2,098	2.14	88.65
Connecting chromaffin cells	1,471	1.50	90.15
Erythrocytes	1,392	1.42	91.57
Neuroblasts	1,242	1.27	92.83
Chromaffin cells	1,216	1.24	94.07
Late chromaffin cells	952	0.97	95.04
Cycling neuroblasts	941	0.96	96.00
SCPs	777	0.79	96.79
Muscle progenitor cells	725	0.74	97.53
Late SCPs	614	0.63	98.16
Bridge	605	0.62	98.77
Hepatocytes	380	0.39	99.16
Myocytes	342	0.35	99.51
Cycling SCPs	257	0.26	99.77
Myofibroblasts	226	0.23	100.00
Total	98,138	100.00	

Another advantage of the cell-type-independent approach to diversity is that we can probe the diversity of subpopulations of each sample. Figure 5.3 shows the LC diversity of five cell types which are present in both early and late stages of development of the gland. To be precise, we took only cell types that have at least ten cells present in at least five samples in both the 'Early' (CS18–23) and 'Late' (11–17pcw) development stages, and diversities were only computed using those samples. We can see that, despite the overall diversity of the gland going down due to an increasing prevalence of neuroblasts, the transcriptional diversity of mesenchymal (two-sided Welch's t-test, $p = 0.022$, $n = 17$), endothelial ($p = 0.044$, $n = 17$), and cycling neuroblast ($p = 0.048$, $n = 14$) cells actually increases as the tissue develops. This is an insight into transcriptional development that is not accessible by simply inspecting distributions of cell types.

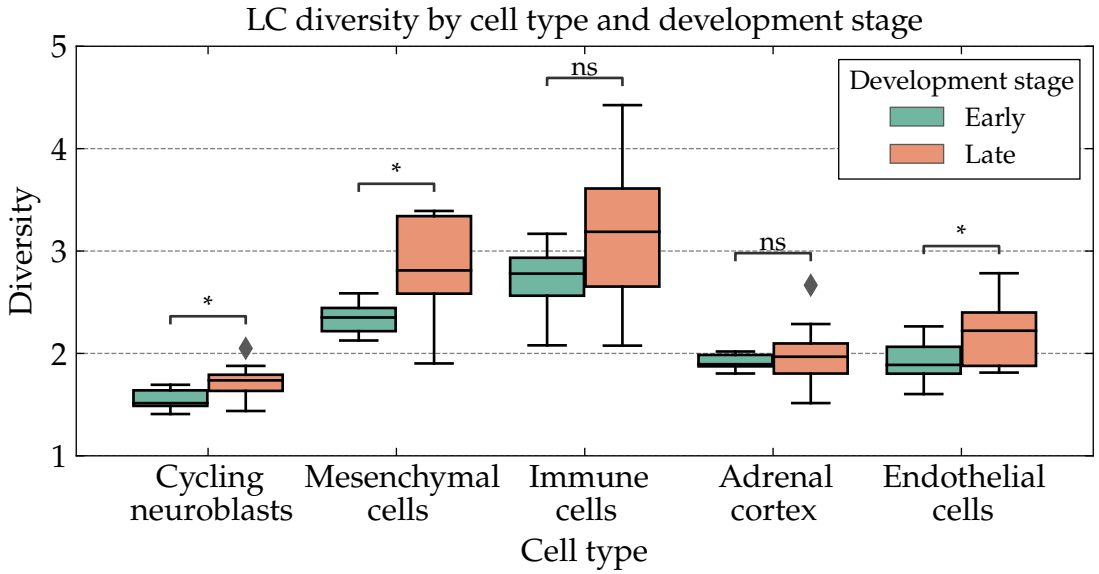


Figure 5.3. Transcription-based LC diversity of individual cell types in the human adrenal gland by early (green) and late (orange) development stages. Centre lines show median; box limits show interquartile range; whiskers show range; diamonds are outliers. Statistical tests are two-sided Welch's t-tests.

Next, we aimed to make rigorous the observation in the original study that the adrenal medulla becomes less transcriptionally diverse as the tissue develops. Jansky *et al.* arrived at this conclusion by studying the relative abundances of medullary cells and noticing that neuroblasts dominate in later development stages [Jan+21, Figure 1f]. To test this claim using our diversity scores, we computed Hill numbers and LC diversity as before, but restricted to medullary cells. Because medullary cells make up a relatively small proportion of cells in the adrenal gland (~10 %), we computed diversity of bulked data at each time point to ensure enough cells for a robust computation.

Figure 5.4 shows each of these scores at each development stage. All three measures show a decrease in diversity in the late 14pcw and 17pcw stages, aligning with the increased abundances of neuroblasts and confirming the observation of Jansky *et al.*. However, a discrepancy emerges between the Hill number and the LC diversities in the early development: while the Hill number is at its highest in early development (CS18–CS20), the LC diversity scores increase up to peaks at the 11pcw stage. This suggests that there is an increase in transcriptional diversity in the early development that is not reflected in a shift in cell-type compositions.

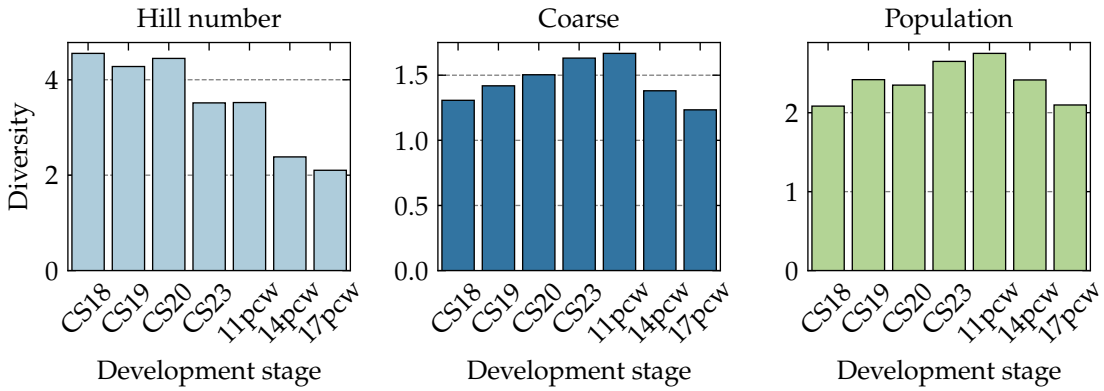


Figure 5.4. Transcription-based diversity scores of medullary cells in the human adrenal gland by development stage. Left: Hill number of order 2. Middle: LC diversity of order 2 with cell-type partition. Right: LC diversity of order 2 without cell-type partition.

5.2.3 Application: stem-cell-based mouse embryo development

For our second application we considered single-cell RNA sequencing (scRNA-seq) data of a murine SEM [Vil+25]. In this model, mouse embryonic stem cell aggregates called gastruloids develop for 48 h before being introduced to agonist CHIR99021 (CHIR), prompting cellular organisation and subsequent development into structures resembling the embryonic trunk. This development is prone to failure, with only 40 % of gastruloids developing successfully at 120 h. In [Vil+25] the authors gathered scRNA-seq and morphological data from gastruloids at time points $t = 48\,\text{h}$, 72 h, 96 h and 120 h, with the aim of determining which early factors are indicative of successful development.

Notably, the authors make several claims throughout the paper on the cellular variability of the developing gastruloids. Lacking a sample-level measure of diversity, they use the squared coefficient of variation CV^2 computed on individual cell types at each time point as a proxy. However, as we saw in the previous application, the diversity of individual cell types can increase even as the diversity of the embryo as a whole decreases. The immediate goal in our analysis is therefore to more rigorously verify the following two claims made by the authors of [Vil+25]:

1. There is a decrease in tissue variability between the 48 h and 72 h time points, prompted by the introduction of the CHIR agonist.

2. There is a steady increase in tissue variability between the 72 h and 120 h time points as the embryo develops and cell types become more mature.

Preprocessing was performed identically to the original study. In brief, we filtered out cells with fewer than 2,000 (48 h and 120 h), fewer than 500 (72 h and 96 h), or more than 80,000 counts, and any cells with more than 2.5 % reads mapping to mitochondrial genes. This left 21,026 cells for further processing, split across 166 samples. For each of the four time points we used an independently selected set of 2,000 differentially expressed genes as defined previously.

Given the relatively small sizes of the early-development gastruloids (~40 cells at 48 h compared with ~500 cells at 120 h), it is impractical to compute diversity measures using cell-type distributions. This setting therefore provides an ideal testing ground for the cell-type-agnostic approach. To this end, we computed the order-2 LC diversity for each sample directly on the cell population, forgoing cell types.

Figure 5.5 shows the distributions of the diversity statistics separated by time point. There is a decrease (two-sided Welch's t-test; $p = 3.2 \times 10^{-12}$) in diversity from 48 h ($n = 71$) to 72 h ($n = 47$). This corresponds to the expectation based on the effect of the CHIR agonist. Furthermore, we then see a steady increase in diversity from 72 h to 120 h, corresponding to tissue development and maturation of cell types. The LC diversity is therefore able to precisely quantify, with significance levels, statements on cell-type variability which previously lacked direct evidence.

We noted earlier that 60 % of gastruloid structures develop unsuccessfully at the 120 h mark [Vil+25, p. 761]. A natural question is whether there is a difference in the transcriptional heterogeneity of the successful (*TLS*) vs unsuccessful (*Neural bias*) morphotypes of developed gastruloids. As can be seen in Figure 5.6, we found no evidence for such a difference using LC diversity of order 2. This suggests that abnormally developing gastruloids still mature into diverse cell-type populations.

It is also interesting to compare the y-axes of the plots in Figures 5.2 and 5.5. In particular, the LC diversities in the human adrenal experiment lie in the range $[2, 5]$ while those in this gastruloid experiment are much closer to 1. This behaviour reflects the fact that the gastruloid data are at much earlier time points (days rather than weeks), and therefore capture tissue in very early development before mature cell types have been able to develop. In other words, the diversity score is also capturing some information about the maturity of cell differentiation in the tissue.

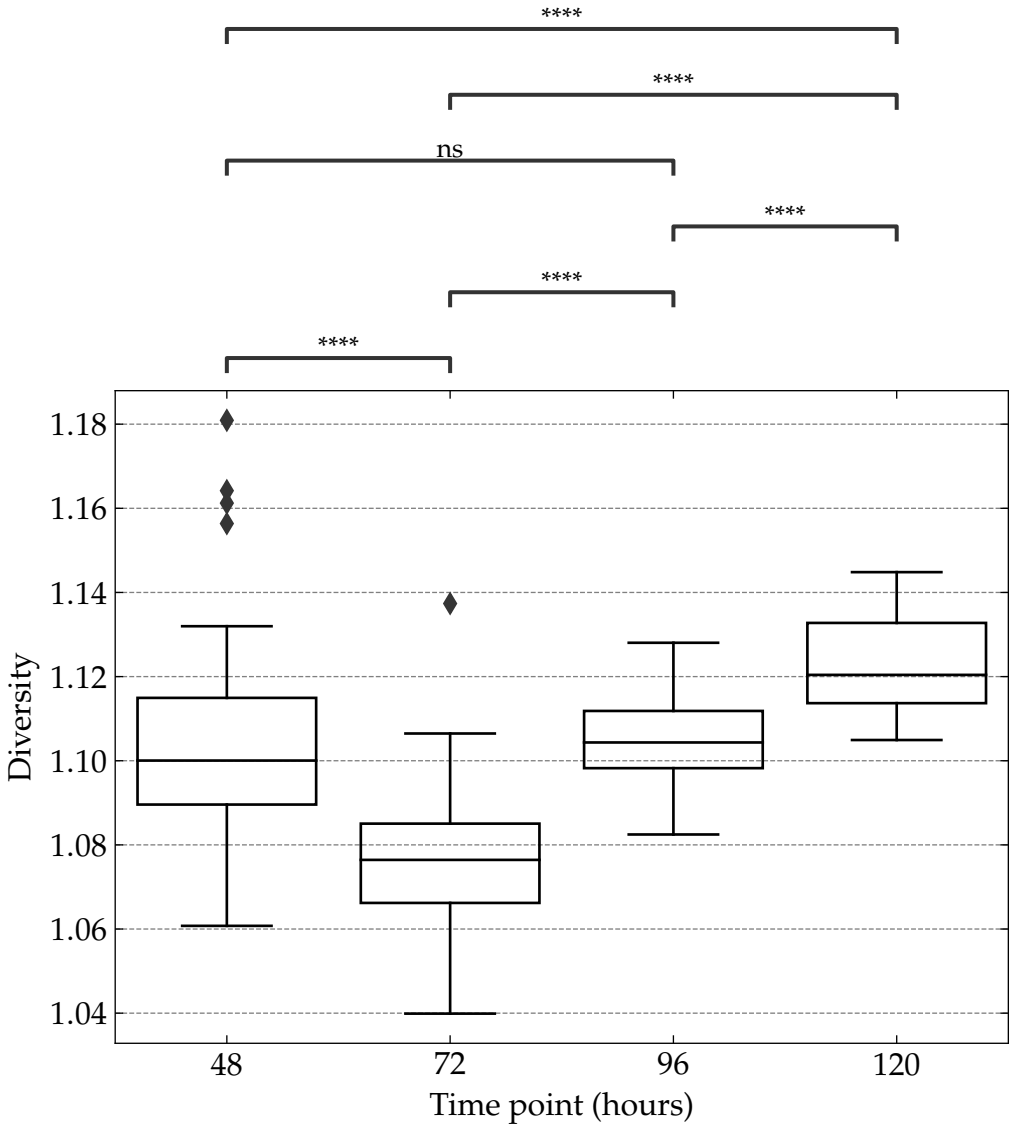


Figure 5.5. LC diversity (order 2) across four time points of gastruloid development: 48 h ($n = 71$), 72 h ($n = 47$), 96 h ($n = 24$), 120 h ($n = 24$). Centre lines show median; box limits show interquartile range; whiskers show range; diamonds are outliers. Statistical tests are two-sided Welch's t-tests.

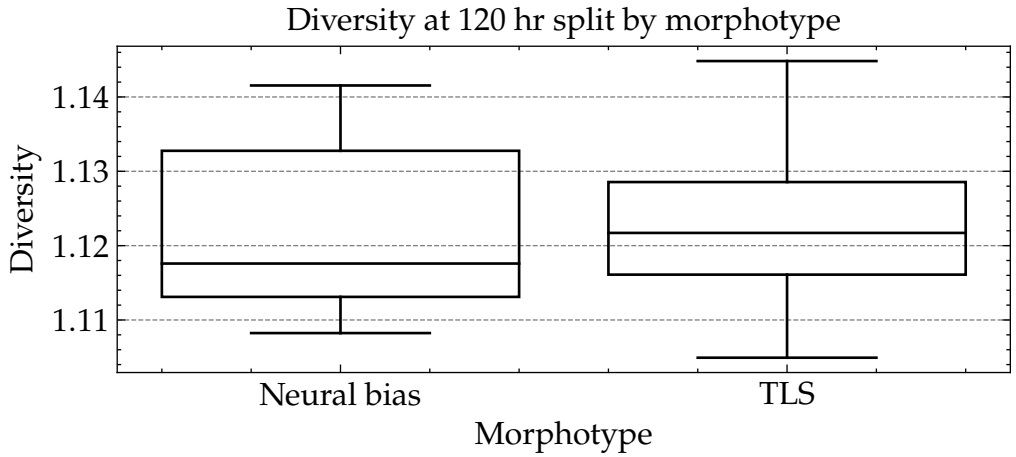


Figure 5.6. LC diversity (order 2) of mouse embryos across abnormal (Neural bias) and normal (TLS) morphotypes at 120 h. There is no evidence that the LC diversity measure can distinguish between the two morphotypes. Centre lines show median; box limits show interquartile range; whiskers show range.

5.2.4 Application: spatial transcriptomics

As a final application, we show how diversity measures can be used to profile *in situ* heterogeneity in spatial transcriptomics data. Here we use a slide from the mouse hippocampus profiled using Slide-seqV2 [Sti+21]. This data set features 41,786 beads at 10 µm resolution. Data are provided with 4,000 HVGs and a pre-computed 14-cell-type assignment which we present in Figure 5.7(a) for visualisation.

For spatial data, some extra preprocessing was required. First, because Slide-seqV2 has very high resolution, we binned neighbouring beads into 4,251 40 µm pseudocells each containing roughly 10 beads. To further protect against gene dropout, we performed a smoothing procedure as follows. We used PCA to project beads into 50-component PC space and then averaged each bead's gene expression with its $k = 15$ nearest neighbours in PC space. Finally, we partitioned the slide into 133 hexagonal regions of radius 225 µm to use as the units on which to compute diversity.

For each region, we computed two different kinds of diversity, in the style of Reeve *et al.* [Ree+16]. The first, (normalised) α-diversity, is the standard diversity measure, as used elsewhere in this thesis, computed independently on each region. One can think of α-diversity as measuring the local heterogeneity within each region. The second measure, γ-diversity, is computed by replacing the similarity matrix within each diversity computation with the whole-slide similarity matrix. The γ-diversity therefore computes the contribution of the region to the diversity of

the slide as a whole. In both cases we used the population-level diversity, and in particular did not use the provided cell-type annotations.

Figure 5.7 panels (b) and (c) show the region-wise α- and γ-diversity of the slide respectively. We see increased α-diversity in the hippocampal interior where dendrites, interneurons, and astrocytes intermix. There is especially raised α-diversity at the borders of these regions indicating increased heterogeneity as the tissue makeup transitions. The γ-diversity shows a strong peak over the choroid plexus in the upper left, indicating that this bundle of epithelial and ependymal cells is particularly transcriptionally unique across the whole slide.

5.3 Discussion

In this chapter we have shown how similarity-sensitive diversity measures as developed by ecologists can be applied to single-cell and spatial transcriptomics. We argue that the cell-type-agnostic behaviour of such measures makes them ideal tools for data analysis in a field where the unreliability of discrete cell-type partitions has become a point of central concern. We have demonstrated the efficacy of LC diversity on two data sets from embryonic development, where the proposed method is able to rigorously justify claims which were previously only made on an intuitive basis. Moreover, by leveraging LC diversity at the single-cell level, we revealed new insights into the composition of cell types that cannot be seen by merely inspecting the relative abundances of each type. In spatial transcriptomics data from Slide-seqV2 we showed how partitioned diversity reveals transcriptomic structure in the mouse hippocampus, including pinpointing the choroid plexus.

One aspect of LC diversity that we have neglected in this thesis is its connections with the categorical notion of the *magnitude* of a finite metric space [Lei13]. We avoid details here, but the general idea is that the magnitude of a finite metric space is a measure of its size analogous to the Euler characteristic of a CW complex. The connection between magnitude and diversity is the following [Lei11]. Let (A, d) be a finite metric space and define $Z_{ij} = \exp(-d(a_i, a_j))$. Then it turns out that there is a single choice of distribution p^*, independent of q, for which the LC diversity $D_q(p^*; Z)$ is maximised for all q. Moreover, this maximum diversity $D_q(p^*; Z)$ is precisely the magnitude of the metric space A, and is therefore also independent of q. It is possible that this connection could be exploited to develop more powerful tools for data analysis which draw on ideas from applied category theory.

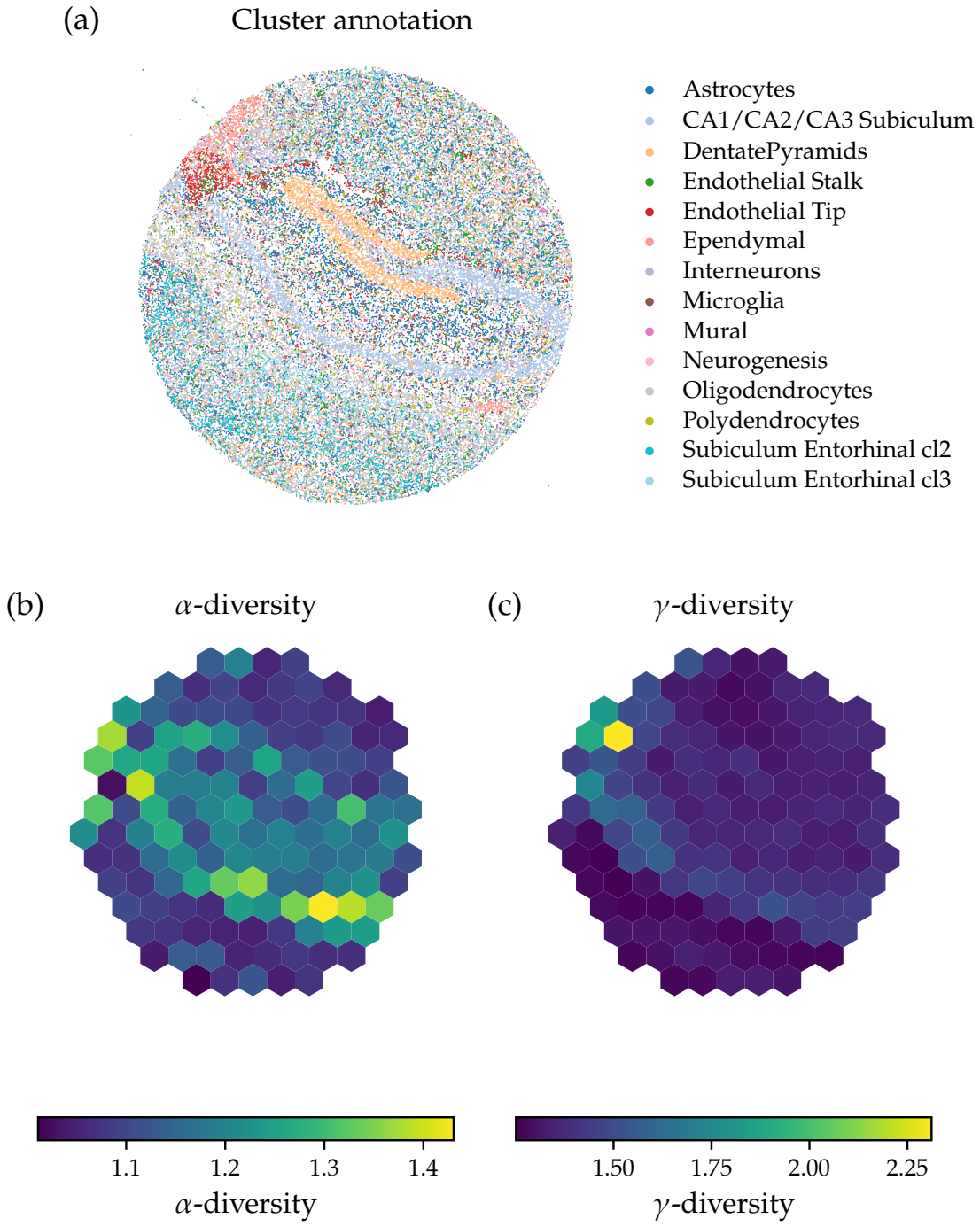


Figure 5.7. Diversity analysis of Slide-seqV2 mouse hippocampus data. (a) Cluster annotations. (b–c): Per-region diversity scores. (b) α-diversity measures heterogeneity of each region as its own unit. (c) γ-diversity measures contribution of each region to the diversity of the slide as a whole.

It is also possible that the general methodology underpinning LC diversity could be extended to solve problems more specific to single-cell transcriptomics. A simple description of LC diversity is that it is a computation of the average uniqueness of a cell with respect to all the other cells in its sample. A very common idea in the single-cell setting, roughly speaking, is to quantify the 'difference' between two distinct single-cell samples. At present the standard way to achieve this is to embed both samples in a shared latent space (for example by computing a low-dimensional embedding of their bulked data) and then compare the two samples in the shared embedding. Such approaches are in general quite unsound [CP23]. As an alternative approach, one could imagine computing the average uniqueness of a cell with respect to all the cells in the *other* sample as a way of measuring such a difference. In this sense the 'difference' between a sample and itself would be the sample's diversity, and in general non-zero, so the resulting measure would not be a metric. Exploring this possibility will be the focus of future work.

Another direction for future research is uncertainty quantification. In this thesis we have presented confidence intervals for between-sample uncertainty, and provided robustness to cell-type-assignment uncertainty. One source of uncertainty which we have not considered here is at the level of transcription within cells. It would be interesting to see how robust the LC diversity is, in both the cell-type and population-level regimes, with respect to different stochastic models of gene expression.

Chapter 6

Discussion

In this thesis we have presented three new methods sitting at the interface of mathematics and biology, addressing distinct problems in topological data analysis, spatial transcriptomics, and single-cell transcriptomics. In Chapter 2 we developed the sparse representation of a persistence module, which allows for fast computation of the rank invariant in 3-parameter persistent homology. In Chapter 4 we introduced TopACT, a state-of-the-art method for cell-type annotation in next-generation subcellular spatial transcriptomics data. In Chapter 5 we proposed the use of similarity-sensitive diversity measures for quantifying cellular diversity in single-cell and spatial transcriptomics data. Now, in this chapter, we will revisit these contributions and consider future research directions.

Contributions Chapter 2 pushes forward the possibilities for applications in topological data analysis. Building on work of Bender *et al.* [BGLa], we present the *sparse representation* of a multiparameter persistence module. The sparse representation allows for the rank invariant, and consequently the persistence landscape, of a multiparameter persistence module to be computed extremely quickly from an appropriate minimal presentation. In particular, this work and its implementation in the Muphasa software package [BGLb] now allow for practitioners of topological data analysis to compute persistence landscapes of filtrations in three or more parameters for the first time. As a proof of concept, we showcase the sparse representation on simulated swarm data from work of Giusti & Lee [GL23], showing how 3-parameter persistence landscapes can be used as feature vectors in regression tasks with time-varying point clouds.

Chapter 4 addresses a major open question in spatial transcriptomics: how do we resolve single-cell information from next-generation data sets with subcellular resol-

ution? We introduce TopACT to produce fine-grained spatial cell-type annotations from spatial transcriptomics data sets, without requiring a separate segmentation step. TopACT enables practitioners to pinpoint individual sparsely dispersed cells, enabling characterisation of a range of phenotypes including autoimmune disease. On imaging-based data, TopACT provides more biologically plausible cell-type segmentations than existing industry approaches that require a separate segmentation step.

Chapter 5 concerns itself with a fundamental problem in single-cell transcriptomics: how can we ensure our methods are robust with respect to cell-type assignments, when those assignments can vary wildly from practitioner-to-practitioner and software-to-software? In this direction, the chapter takes inspiration from recent work in ecological diversity on similarity-sensitive diversity measures [LC12]. We show that the robustness of these diversity measures to species relabelling translates precisely to robustness to cell-type reassignment in single-cell transcriptomics. This method therefore provides a new tool with which biologists can analyse their single-cell data while remaining assured that upstream choices on cell-type assignment will not affect the reproducibility of their results. We demonstrate the applicability of these ideas to two scRNA-seq studies in developmental biology, as well as a spatial transcriptomics data set of the mouse hippocampus.

Limitations The spatiotemporal persistent homology pipeline presented here has two major limitations. Firstly, while computing the underlying minimal presentation is now possible [BGLa], it remains a computationally expensive and memory-hungry task. It is therefore impractical to compute persistence of Interlevel-Rips-DMS filtrations of more than a hundred agents or time points. Furthermore, interpretation of 3PH landscapes remains a hard problem.

TopACT is also limited by its performance. In contrast to the fixed-window approach at Bin 20, it requires classification of over a thousand times more input vectors. Furthermore, the method is unable to see the boundaries between neighbouring cells of the same cell type. Finally, while we offered verification here on synthetic data, the lack of ground truth labels in real-world data limits how sure we can be of the method's effectiveness in those settings.

Our transcriptomic diversity pipeline needs further work in order to quantify its uncertainty with respect to the underlying transcription process. Furthermore, it still requires preprocessing choices in terms of feature selection and determining a suitable similarity matrix. The choice of similarity measure in particular can change the magnitude of the measured diversity significantly.

Future directions It is an exciting time for biological data analysis, and this thesis provides a range of opportunities for future work in the area.

In multiparameter persistent homology, we have now provided an end-to-end pipeline for computation of vector summaries of 3-parameter persistence modules. The future of research in this direction is now to use this pipeline to study as many different data analysis problems as possible. The spatiotemporal analysis we applied to swarm data in Section 2.4 already generalises to many different data types, including embryonic development data sets such as the Zebrahub atlas [Lan+24] and molecular dynamics simulations. In 2-parameter persistence, authors have been forced to choose between two different options for the second persistence parameter, and the availability of 3-parameter persistence will now enable them to avoid these choices. In general, the computational barrier has been lifted, and now the hard work of applying these tools to the real world begins.

In transcriptomics, it remains an open problem to put the concept of cell types on firm mathematical footing: 'the concept of "cell type" is poorly defined and incredibly useful' [Kle17]. While this thesis has offered a flexible method for cell-type assignment in spatial transcriptomics, and a method in single-cell transcriptomics which sidesteps the cell-type issue entirely, future research should tackle the mathematical problem of cell types head-on. How should we think about discrete cell types in a continuous gene expression space? Should our notion of cell types be equipped with a richer mathematical structure than simple sets? How do we handle platform effects, and does the idea of a 'platonic' cell type divorced from any specific data set even make sense? What is the geometric structure of gene expression space; is the manifold assumption valid? Going forward, these are the questions that need to be at the forefront of mathematical thinking in transcriptomics.

Conclusion While the three contributions in this thesis all use seemingly distinct methods, we finish by noting that they do share a common thread. Biological data analysis is filled with assumptions: choose a scale, pick a segmentation, fix a cell-type assignment. While these assumptions are often helpful, they undermine reproducibility, and making the wrong choice can lead us astray. Our approach in this thesis has been to sidestep these assumptions: look at all scales, don't segment your data, don't fix a cell-type assignment. In doing so, we have created flexible mathematical approaches which are not just more robust, but can also tell us more about our data.

Bibliography

[10x19] 10x Genomics. *Visium Spatial v1 3' Gene Expression*. 2019. URL: https://www.10xgenomics.com/products/visium-v1-gene-expression (visited on 26th Dec. 2025).

[10x22] 10x Genomics. *Visium Spatial v2 WT Panel Gene Expression*. 2022. URL: https://www.10xgenomics.com/products/visium-v2-gene-expression (visited on 26th Dec. 2025).

[Ali+23] Dashti Ali et al. 'A Survey of Vectorization Methods in Topological Data Analysis'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.12 (Dec. 2023), pp. 14069–14080.

[Alo19] Uri Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. 2nd ed. New York: CRC Press, Aug. 2019.

[And+20] Alma Andersson et al. 'Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography'. In: *Communications Biology* 3.1 (Oct. 2020).

[AOM19] Maria I. Arnone, Paola Oliveri and Pedro Martinez. 'A conceptual history of the "regulatory genome": From Theodor Boveri to Eric Davidson'. In: *Marine Genomics* 44 (Apr. 2019), pp. 24–31.

[Axe+18] [Software] Shannon Axelrod et al., *Starfish: Open Source Image Based Transcriptomics and Proteomics Tools* 2018. LIC: MIT. VCS: http://github.com/spacetx/starfish. Reference article: Shannon Axelrod et al. 'starfish: scalable pipelines for image-based transcriptomics'. In: *Journal of Open Source Software* 6.61 (May 2021), p. 2440.

[Bae+22] Sungwoo Bae et al. 'CellDART: cell type inference by domain adaptation of single-cell and spatial transcriptomic data'. In: *Nucleic Acids Research* 50.10 (June 2022), e57.

[BBE22] Leo Betthauser, Peter Bubenik and Parker B. Edwards. 'Graded Persistence Diagrams and Persistence Landscapes'. In: *Discrete & Computational Geometry* 67.1 (Jan. 2022), pp. 203–230.

[BC26] Perry Beamer and Zixuan Cang. 'Multiscale domain identification for spatial transcriptomics via persistent homology'. In: *Cell Reports Methods* 6.5 (May 2026), p. 101376.

[Bec+19] Etienne Becht et al. 'Dimensionality reduction for visualizing single-cell data using UMAP'. In: *Nature Biotechnology* 37.1 (Jan. 2019), pp. 38–44.

[Bee+23] David Beers et al. 'Barcodes distinguishing morphology of neuronal tauopathy'. In: *Physical Review Research* 5.4 (Oct. 2023), p. 043006.

[Ben+13] Paul Bendich et al. 'Homology and robustness of level and interlevel sets'. In: *Homology, Homotopy and Applications* 15.1 (Mar. 2013), pp. 51–72.

[Ben+23] Katherine Benjamin et al. 'Homology of homologous knotted proteins'. In: *Journal of the Royal Society Interface* 20.201 (Apr. 2023), p. 20220727.

[Ben+24] Katherine Benjamin et al. 'Multiscale topology classifies cells in subcellular spatial transcriptomics'. In: *Nature* 630.8018 (June 2024), pp. 943–949.

[BGLa] Matías R. Bender, Oliver Gäfvert and Michael Lesnick. 'Computing minimal presentations of multiparameter persistent homology modules'. To appear.

[BGLb] [Software] Matías R. Bender, Oliver Gäfvert and Michael Lesnick, *Mup-hasa*. LIC: MIT. VCS: `https://github.com/olivergafvert/muphasa`. Reference article: 'Computing minimal presentations of multiparameter persistent homology modules'. To appear.

[BGV92] Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik. 'A training algorithm for optimal margin classifiers'. In: *Proceedings of the fifth annual workshop on Computational learning theory*. Pittsburgh Pennsylvania USA: ACM, July 1992, pp. 144–152.

[Bha+19] Dhananjay Bhaskar et al. 'Analyzing collective motion with machine learning and topology'. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 29.12 (Dec. 2019), p. 123125.

[Bia+21] Tommaso Biancalani et al. 'Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram'. In: *Nature Methods* 18.11 (Nov. 2021), pp. 1352–1362.

[BL23] Magnus Bakke Botnan and Michael Lesnick. 'An introduction to multiparameter persistence'. In: *EMS Series of Congress Reports*. Ed. by Aslak Bakke Buan, Henning Krause and Øyvind Solberg. 1st ed. Vol. 19. EMS Press, Nov. 2023, pp. 77–150.

[Bla+25] Quentin Blampey et al. 'Novae: a graph-based foundation model for spatial transcriptomics data'. In: *Nature Methods* 22.12 (Dec. 2025), pp. 2539–2550.

[Bub15] Peter Bubenik. 'Statistical topological data analysis using persistence landscapes'. In: *Journal of Machine Learning Research* 16 (Jan. 2015), pp. 77–102.

[BW20] David Bramer and Guo-Wei Wei. 'Atom-specific persistent homology and its application to protein flexibility analysis'. In: *Computational and Mathematical Biophysics* 8.1 (Jan. 2020), pp. 1–35.

[Cab+22] Dylan M. Cable et al. 'Robust decomposition of cell type mixtures in spatial transcriptomics'. In: *Nature Biotechnology* 40.4 (Feb. 2022).

[CB20] Mathieu Carrière and Andrew Blumberg. 'Multiparameter Persistence Images for Topological Machine Learning'. In: *Advances in Neural Information Processing Systems*. Vol. 34. Dec. 2020.

[CBS23] Daniel Charytonowicz, Rachel Brody and Robert Sebra. 'Interpretable and context-free deconvolution of multi-scale whole transcriptomic data with UniCell deconvolve'. In: *Nature Communications* 14.1350 (Mar. 2023).

[Cd10] Gunnar Carlsson and Vin de Silva. 'Zigzag Persistence'. In: *Foundations of Computational Mathematics* 10 (Apr. 2010), pp. 367–405.

[CEH05] David Cohen-Steiner, Herbert Edelsbrunner and John Harer. 'Stability of Persistence Diagrams'. In: *Proceedings of the Twenty-first Annual Symposium on Computational Geometry*. SCG '05. June 2005.

[CEM06] David Cohen-Steiner, Herbert Edelsbrunner and Dmitriy Morozov. 'Vines and vineyards by updating persistence in linear time'. In: *Proceedings of the twenty-second annual symposium on Computational geometry*. Sedona Arizona USA: ACM, June 2006, pp. 119–126.

[Che+15] Kok Hao Chen et al. 'Spatially resolved, highly multiplexed RNA profiling in single cells'. In: *Science* 348.6233 (Apr. 2015), aaa6090.

[Che+22] Ao Chen et al. 'Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays'. In: *Cell* 185.10 (May 2022), 1777–1792.e21.

[Che+23] Jiawen Chen et al. 'Cell composition inference and identification of layer-specific spatial transcriptional profiles with POLARIS'. In: *Science Advances* 9.9 (Mar. 2023), eadd9818.

[Cho+21] Chun Seok Cho et al. 'Microscopic examination of spatial transcriptome using Seq-Scope'. In: *Cell* 184.13 (June 2021), 3559–3572.e22.

[Chu+07] Yao-li Chuang et al. 'State Transitions and the Continuum Limit for a 2D Interacting, Self-Propelled Particle System'. In: *Physica D: Nonlinear Phenomena* 232.1 (2007), pp. 33–47.

[Chu+23] Salvador Chulián et al. 'The shape of cancer relapse: Topological data analysis predicts recurrence in paediatric acute lymphoblastic leukaemia'. In: *PLOS Computational Biology* 19.8 (Aug. 2023), e1011329.

[CJ17] Padraig Corcoran and Christopher B. Jones. 'Modelling Topological Features of Swarm Behaviour in Space and Time With Persistence Landscapes'. In: *IEEE Access* 5 (Sept. 2017), pp. 18534–18544.

[Coo+23] David P. Cook et al. *A Comparative Analysis of Imaging-Based Spatial Transcriptomics Platforms*. Dec. 2023. bioRxiv: 2023.12.13.571385.

[CP23] Tara Chari and Lior Pachter. 'The specious art of single-cell genomics'. In: *PLOS Computational Biology* 19.8 (Aug. 2023), e1011288.

[Cro+23] Helena L. Crowell et al. 'The shaky foundations of simulating single-cell RNA sequencing data'. In: *Genome Biology* 24.1 (Mar. 2023), p. 62.

[CRW24] Yunlu Chen, Feng Ruan and Ji-Ping Wang. 'NLSDeconv: an efficient cell-type deconvolution method for spatial transcriptomics data'. In: *Bioinformatics* 41.1 (Dec. 2024). Ed. by Christina Kendziorski, btae747.

[CS17] 'What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism?' In: *Cell Systems* 4.3 (Mar. 2017), pp. 255–259.

[CSG21] Yuzhou Chen, Ignacio Segovia-Dominguez and Yulia R. Gel. 'Z-GCNETs: Time Zigzags at Graph Convolutional Networks for Time Series Forecasting'. In: *Proceedings of the 38th International Conference on Machine Learning, PMLR*. May 2021.

[CSV17] W. Chachólski, M. Scolamiero and F. Vaccarino. 'Combinatorial presentation of multidimensional persistent homology'. In: *Journal of Pure and Applied Algebra* 221.5 (May 2017), pp. 1055–1075.

[CV95] Corinna Cortes and Vladimir Vapnik. 'Support-Vector Networks'. In: *Machine Learning* 20.3 (Sept. 1995), pp. 273–297.

[CW25] Sean Cottrell and Guo-Wei Wei. 'Multiscale Cell–Cell Interactive Spatial Transcriptomics Analysis'. In: *Advanced Science* (Sept. 2025), e08358.

[CZ09] Gunnar Carlsson and Afra Zomorodian. 'The theory of multidimensional persistence'. In: *Discrete and Computational Geometry* 42 (Apr. 2009), pp. 71–93.

[Dan+22] Patrick Danaher et al. 'Advances in mixed cell deconvolution enable quantification of cell types in spatial transcriptomic data'. In: *Nature Communications* 13.385 (Jan. 2022).

[DH21] Tamal K. Dey and Tao Hou. 'Computing Zigzag Persistence on Graphs in Near-Linear Time'. In: *37th International Symposium on Computational Geometry (SoCG 2021)*. Ed. by Kevin Buchin and Éric Colin de Verdière. Vol. 189. Leibniz International Proceedings in Informatics (LIPIcs). June 2021, 30:1–30:15.

[DH22] Tamal K. Dey and Tao Hou. 'Fast Computation of Zigzag Persistence'. In: *30th Annual European Symposium on Algorithms (ESA 2022)*. Ed. by Shiri Chechik et al. Vol. 244. Leibniz International Proceedings in Informatics (LIPIcs). Sept. 2022, 43:1–43:15.

[DOr+06] M. R. D'Orsogna et al. 'Self-Propelled Particles with Soft-Core Interactions: Patterns, Stability, and Collapse'. In: *Physical Review Letters* 96.10 (Mar. 2006), p. 104302.

[DS25] Tamal K. Dey and Shreyas N. Samaga. *Quasi Zigzag Persistence: A Topological Framework for Analyzing Time-Varying Data*. Feb. 2025. arXiv: 2502.16049 [cs.LG].

[DW07] Tamal K. Dey and Rephael Wenger. 'Stability of Critical Points with Interval Persistence'. In: *Discrete & Computational Geometry* 38.3 (Oct. 2007), pp. 479–512.

[DY21] Rui Dong and Guo-Cheng Yuan. 'SpatialDWLS: accurate deconvolution of spatial transcriptomic data'. In: *Genome Biology* 22.145 (Dec. 2021).

[Eis95] David Eisenbud. *Commutative Algebra: with a View Toward Algebraic Geometry*. 1st ed. Vol. 150. Graduate Texts in Mathematics. New York, NY: Springer, 1995.

[Elo+21] Marc Elosua-Bayes et al. 'SPOTlight: Seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes'. In: *Nucleic Acids Research* 49.9 (May 2021).

[Elo17] Michael Elowitz. 'Cellular Demographies, Recorded'. In: *Cell Systems* 4.3 (Mar. 2017). Part of *What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism?*, p. 225.

[Eng+19] Chee-Huat Linus Eng et al. 'Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+'. In: *Nature* 568.7751 (Apr. 2019), pp. 235–239.

[Erg+24] Can Ergen et al. 'Consensus prediction of cell type labels in single-cell data with popV'. In: *Nature Genetics* 56.12 (Dec. 2024), pp. 2731–2738.

[Fem+98] Andrea M. Femino et al. 'Visualization of Single RNA Transcripts in Situ'. In: *Science* 280.5363 (Apr. 1998), pp. 585–590.

[FH24] Martina Flammer and Knut Hüper. *Spatiotemporal Persistence Landscapes*. Dec. 2024. arXiv: 2412.11925 [math.AT].

[Gam+14] Marcio Gameiro et al. 'A topological measurement of protein compressibility'. In: *Japan Journal of Industrial and Applied Mathematics* 32 (Oct. 2014), pp. 1–17.

[Gar+22] Richard J. Gardner et al. 'Toroidal topology of population activity in grid cells'. In: *Nature* 602.7895 (Feb. 2022), pp. 123–128.

[Gas+25] Lucie C. Gaspard-Boulinc et al. 'Cell-type deconvolution methods for spatial transcriptomics'. In: *Nature Reviews Genetics* 26.12 (Dec. 2025), pp. 828–846.

[GGB16] Chad Giusti, Robert Ghrist and Danielle S. Bassett. 'Two's company, three (or more) is a simplex: Algebraic-topological tools for understanding higher-order structure in neural data'. In: *Journal of Computational Neuroscience* 41 (Aug. 2016), pp. 1–14.

[GL23] Chad Giusti and Darrick Lee. 'Signatures, Lipschitz-Free Spaces, and Paths of Persistence Diagrams'. In: *SIAM Journal on Applied Algebra and Geometry* 7.4 (Dec. 2023), pp. 828–866.

[Goo+24] Christian Goodbrake et al. 'Brain chains as topological signatures for Alzheimer's disease'. In: *Journal of Applied and Computational Topology* 8.5 (Oct. 2024), pp. 1257–1298.

[GW09] Nicholas Geard and Kai Willadsen. 'Dynamical approaches to modeling developmental gene regulatory networks'. In: *Birth Defects Research Part C: Embryo Today: Reviews* 87.2 (June 2009), pp. 131–142.

[GYC19] Kiya W. Govek, Venkata S. Yamajala and Pablo G. Camara. 'Clustering-independent analysis of genomic data using spectral simplicial theory'. In: *PLOS Computational Biology* 15.11 (Nov. 2019), e1007509.

[Hao+21] Yuhan Hao et al. 'Integrated analysis of multimodal single-cell data'. In: *Cell* 184.13 (June 2021), 3573–3587.e29.

[Hat02] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.

[HC24] Tram Huynh and Zixuan Cang. 'Topological and geometric analysis of cell states in single-cell transcriptomic data'. In: *Briefings in Bioinformatics* 25.3 (Mar. 2024), bbae176.

[Hil73] M. O. Hill. 'Diversity and Evenness: A Unifying Notation and Its Consequences'. In: *Ecology* 54.2 (Mar. 1973), pp. 427–432.

[Hoe+22] Renee S. Hoekzema et al. 'Multiscale Methods for Signal Selection in Single-Cell Data'. In: *Entropy* 24.8 (Aug. 2022), p. 1116.

[Hua12] Sui Huang. 'The molecular and mathematical basis of Waddington's epigenetic landscape: A framework for post-Darwinian biology?' In: *BioEssays* 34.2 (Feb. 2012), pp. 149–157.

[IHGSC04] International Human Genome Sequencing Consortium. 'Finishing the euchromatic sequence of the human genome'. In: *Nature* 431.7011 (Oct. 2004), pp. 931–945.

[Jan+21] Selina Jansky et al. 'Single-cell transcriptomic analyses provide insights into the developmental origins of neuroblastoma'. In: *Nature Genetics* 53.5 (May 2021), pp. 683–693.

[Jan+23] Amanda Janesick et al. 'High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis'. In: *Nature Communications* 14.8353 (Dec. 2023).

[Kan+18] Lida Kanari et al. 'A Topological Representation of Branching Neuronal Morphologies'. In: *Neuroinformatics* 16.1 (Jan. 2018), pp. 3–13.

[Kim20] Woojin Kim. 'The Persistent Topology of Dynamic Data'. PhD thesis. The Ohio State University, 2020.

[KL21] Dmitry Kobak and George C. Linderman. 'Initialization is critical for preserving global data structure in both t-SNE and UMAP'. In: *Nature Biotechnology* 39.2 (Feb. 2021), pp. 156–157.

[Kle+22] Vitalii Kleshchevnikov et al. 'Cell2location maps fine-grained cell types in spatial transcriptomics'. In: *Nature Biotechnology* 40.5 (Jan. 2022), pp. 661–671.

[Kle17] Allon Klein. 'Farewell, "Cell Type."' In: *Cell Systems* 4.3 (Mar. 2017). Part of *What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism?*, p. 225.

[KM21] Woojin Kim and Facundo Mémoli. 'Generalized persistence diagrams for persistence modules over posets'. In: *Journal of Applied and Computational Topology* 5.4 (Dec. 2021), pp. 533–581.

[KMS22] Tanya T Karagiannis, Stefano Monti and Paola Sebastiani. 'Cell Type Diversity Statistic: An Entropy-Based Metric to Compare Overall Cell Type Composition Across Samples'. In: *Frontiers in Genetics* 13 (Apr. 2022), p. 855076.

[Kov+16] Violeta Kovacev-Nikolic et al. 'Using persistent homology and dynamical distances to analyze protein binding'. In: *Statistical Applications in Genetics and Molecular Biology* 15 (Jan. 2016).

[Läh+20] David Lähnemann et al. 'Eleven grand challenges in single-cell data science'. In: *Genome Biology* 21 (Feb. 2020), p. 31.

[Lan+24] Merlin Lange et al. 'A multimodal zebrafish developmental atlas reveals the state-transition dynamics of late-vertebrate pluripotent axial progenitors'. In: *Cell* 187.23 (Nov. 2024), 6742–6759.e17.

[Law+19] Peter Lawson et al. 'Persistent Homology for the Quantitative Evaluation of Architectural Features in Prostate Cancer Histology'. In: *Scientific Reports* 9 (Feb. 2019), p. 1139.

[LBL17] Ed Lein, Lars E. Borm and Sten Linnarsson. 'The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing'. In: *Science* 358.6359 (Oct. 2017), pp. 64–69.

[LC12] Tom Leinster and Christina A. Cobbold. 'Measuring diversity: the importance of species similarity'. In: *Ecology* 93.3 (Mar. 2012), pp. 477–489.

[Lei11] Tom Leinster. *A maximum entropy theorem with applications to the measurement of biodiversity*. Jan. 2011. arXiv: `0910.0906v4 [cs.IT]`.

[Lei13] Tom Leinster. 'The Magnitude of Metric Spaces'. In: *Documenta Mathematica* 18 (2013), pp. 857–905.

[Lei21] Tom Leinster. *Entropy and Diversity: the Axiomatic Approach*. Cambridge, United Kingdom: Cambridge University Press, 2021.

[Li+22] Haoyang Li et al. 'SD2: spatially resolved transcriptomics deconvolution through integration of dropout and spatial information'. In: *Bioinformatics* 38.21 (Oct. 2022), pp. 4878–4884.

[Liu+20] Yang Liu et al. 'High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue'. In: *Cell* 183.6 (Dec. 2020), 1665–1681.e18.

[Liu+23] Zhiyuan Liu et al. 'SONAR enables cell type deconvolution with spatially weighted Poisson-Gamma model for spatial transcriptomics'. In: *Nature Communications* 14.4727 (Aug. 2023).

[LL24] Yawei Li and Yuan Luo. 'STdGCN: spatial transcriptomic cell-type deconvolution using graph convolutional networks'. In: *Genome Biology* 25.206 (Aug. 2024).

[Lon+23] Yahui Long et al. 'Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST'. In: *Nature Communications* 14.1155 (Mar. 2023).

[Lop+22] Romain Lopez et al. 'DestVI identifies continuums of cell types in spatial transcriptomics data'. In: *Nature Biotechnology* 40.9 (Apr. 2022), pp. 1360–1369.

[Low04] David G. Lowe. 'Distinctive image features from scale-invariant keypoints'. In: *International Journal of Computer Vision* 60.2 (2004).

[LS24] [Software] David Loiseaux and Hannah Schreiber, *Multipers* 2024. LIC: MIT. VCS: `https://github.com/DavidLapous/multipers`. Reference article: 'Multipers: Multiparameter Persistence for Machine Learning'. In: *Journal of Open Source Software* 9.103 (Nov. 2024), p. 6773.

[Mad+25a] Hiren Madhu et al. *HEIST: A Graph Foundation Model for Spatial Transcriptomics and Proteomics Data*. June 2025. arXiv: `2506.11152 [q-bio.GN]`. URL: `https://arxiv.org/abs/2506.11152`.

[Mad+25b] Christian D. Madsen et al. 'The topological properties of the protein universe'. In: *Nature Communications* 16.7503 (Aug. 2025).

[Mañ+24] Diego Mañanes et al. 'SpatialDDLS: an R package to deconvolute spatial transcriptomics data using neural networks'. In: *Bioinformatics* 40.2 (Feb. 2024), btae072.

[Mar+25] Sergio Marco Salas et al. 'Optimizing Xenium In Situ data utility by quality assessment and best-practice analysis workflows'. In: *Nature Methods* 22.4 (Apr. 2025), pp. 813–823.

[Mar21] Vivien Marx. 'Method of the Year: spatially resolved transcriptomics'. In: *Nature Methods* 18 (Jan. 2021), pp. 9–14.

[McD+23] Robert A. McDonald et al. 'Zigzag persistence for coral reef resilience using a stochastic spatial model'. In: *Journal of the Royal Society Interface* 20 (Aug. 2023).

[MCN21] Floyd Maseda, Zixuan Cang and Qing Nie. 'DEEPsc: A Deep Learning-Based Map Connecting Single-Cell Transcriptomics and Spatial Imaging Data'. In: *Frontiers in Genetics* 12 (Mar. 2021), p. 636743.

[MH08] Laurens van der Maaten and Geoffrey Hinton. 'Visualizing Data using t-SNE'. In: *Journal of Machine Learning Research* 9.86 (Nov. 2008), pp. 2579–2605.

[MHM18] Leland McInnes, John Healy and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. Feb. 2018. arXiv: 1802.03426 [stat.ML].

[MP22] Lambda Moses and Lior Pachter. 'Museum of spatial transcriptomics'. In: *Nature Methods* 19 (Mar. 2022), pp. 534–546.

[Mül+25] Niklas Müller-Bötticher et al. 'Sainsc: A Computational Tool for Segmentation-Free Analysis of In Situ Capture Data'. In: *Small Methods* 9.5 (May 2025), p. 2401123.

[Mye+23] Audun Myers et al. 'Temporal network analysis using zigzag persistence'. In: *EPJ Data Science* 12.1 (Mar. 2023), p. 6.

[MZ22] Ying Ma and Xiang Zhou. 'Spatially informed cell-type deconvolution for spatial transcriptomics'. en. In: *Nature Biotechnology* 40.9 (Sept. 2022), pp. 1349–1359.

[NLC11] Monica Nicolau, Arnold J. Levine and Gunnar Carlsson. 'Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival'. In: *Proceedings of the National Academy of Sciences* 108.17 (Apr. 2011), pp. 7265–7270.

[Oli+25] Michelli Faria De Oliveira et al. 'High-definition spatial transcriptomic profiling of immune cell populations in colorectal cancer'. In: *Nature Genetics* 57.6 (June 2025), pp. 1512–1523.

[Ott+17] Nina Otter et al. 'A roadmap for the computation of persistent homology'. In: *EPJ Data Science* 6.1 (Dec. 2017), pp. 1–38.

[Pas+21] Giovanni Pasquini et al. 'Automated methods for cell type annotation on scRNA-seq data'. In: *Computational and Structural Biotechnology Journal* 19 (Jan. 2021), pp. 961–969.

[Pee11] Irena Peeva. *Graded Syzygies*. 1st ed. Algebra and Applications. Springer London, 2011.

[Pla+99] John Platt et al. 'Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods'. In: *Advances in large margin classifiers* 10.3 (1999).

[Poo+24] Suresh Poovathingal et al. 'Nova-ST: Nano-patterned ultra-dense platform for spatial transcriptomics'. In: *Cell Reports Methods* 4.8 (Aug. 2024), p. 100831.

[Pre+25] Stephan Preibisch et al. 'Scalable image-based visualization and alignment of spatial transcriptomics datasets'. In: *Cell Systems* 16.5 (May 2025), p. 101264.

[Raj+08] Arjun Raj et al. 'Imaging individual mRNA molecules using multiple singly labeled probes'. In: *Nature Methods* 5.10 (Oct. 2008), pp. 877–879.

[Ree+16] Richard Reeve et al. *How to partition diversity*. 2016. arXiv: 1404.6520 [q-bio.QM].

[Rén61] Alféd Rényi. 'On measure of entropy and information'. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Jan. 1961, pp. 547–561.

[Ric+26] Joseph M. Rich et al. 'The impact of package selection and versioning on single-cell RNA-seq analysis'. In: *Cell Systems* 17.4 (Apr. 2026), p. 101560.

[RIV20] [Software] The RIVET Developers, *RIVET* version 1.1.0, 2020. LIC: GPLv3. VCS: https://github.com/rivetTDA/rivet/. Reference article: Michael Lesnick and Matthew Wright. *Interactive Visualization of 2-D Persistence Modules*. Dec. 2015. arXiv: 1512.00180 [math.AT].

[Riz+17] Abbas H. Rizvi et al. 'Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development'. In: *Nature Biotechnology* 35.6 (June 2017), pp. 551–560.

[Rod+19] Samuel G. Rodriques et al. 'Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution'. In: *Science* 363.6434 (Mar. 2019), pp. 1463–1467.

[Rus+24] Andrew J. C. Russell et al. 'Slide-tags enables single-nucleus barcoding for multimodal spatial genomics'. In: *Nature* 625.7993 (Jan. 2024), pp. 101–109.

[Sal+21] Andrew Salch et al. 'From mathematics to medicine: A practical primer on topological data analysis (TDA) and the development of related analytic tools for the functional discovery of latent structure in fMRI data'. In: *PLOS ONE* 16.8 (Aug. 2021), e0255859.

[Sat+21] [Software] Satija Lab and Collaborators, *Seurat* version 4, 2021. LIC: MIT. URL: https://satijalab.org/seurat/. Reference article: Yuhan Hao et al. 'Integrated analysis of multimodal single-cell data'. In: *Cell* 184.13 (June 2021), 3573–3587.e29.

[Sat+23] Satija Lab and Collaborators. *Seurat - Guided Clustering Tutorial*. 2023. URL: https://satijalab.org/seurat/articles/pbmc3k_tutorial (visited on 13th Mar. 2025).

[Sch+24] Marie Schott et al. 'Open-ST: High-resolution spatial transcriptomics in 3D'. In: *Cell* 187.15 (July 2024), 3953–3972.e26.

[SMC07] Gurjeet Singh, Facundo Memoli and Gunnar Carlsson. 'Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition'. In: *Eurographics Symposium on Point-Based Graphics*. Ed. by M. Botsch et al. The Eurographics Association, 2007.

[SS21] Qianqian Song and Jing Su. 'DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence'. In: *Briefings in Bioinformatics* 22.5 (Sept. 2021), bbaa414.

[Stå+16] Patrik L. Ståhl et al. 'Visualization and analysis of gene expression in tissue sections by spatial transcriptomics'. In: *Science* 353.6294 (July 2016), pp. 78–82.

[Sti+21] Robert R. Stickels et al. 'Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2'. In: *Nature Biotechnology* 39 (Dec. 2021), pp. 313–319.

[Sto+21] Bernadette J. Stolz et al. 'Topological data analysis of task-based fMRI data from experiments on schizophrenia'. In: *Journal of Physics: Complexity* 2.3 (May 2021), p. 035006.

[Sto+22] Bernadette J. Stolz et al. 'Multiscale topology characterizes dynamic tumor vascular networks'. In: *Science Advances* 8.23 (June 2022), eabm2456.

[Sto+24] Bernadette J. Stolz et al. 'Relational Persistent Homology for Multispecies Data with Application to the Tumor Microenvironment'. In: *Bulletin of Mathematical Biology* 86.11 (Nov. 2024), p. 128.

[Sui+25] Xin Sui et al. 'Scalable spatial single-cell transcriptomics and translatomics in 3D thick tissue blocks'. In: *Nature Methods* 22.12 (Dec. 2025), pp. 2574–2584.

[SWH21] Duluxan Sritharan, Shu Wang and Sahand Hormoz. 'Computing the Riemannian curvature of image patch and single-cell RNA sequencing data manifolds using extrinsic differential geometry'. In: *Proceedings of the National Academy of Sciences* 118.29 (July 2021), e2100473118.

[TMK20] Sarah Tymochko, Elizabeth Munch and Firas A. Khasawneh. 'Using Zigzag Persistent Homology to Detect Hopf Bifurcations in Dynamical Systems'. In: *Algorithms* 13.11 (Oct. 2020), p. 278.

[Tor+25] Maria Torras-Pérez et al. 'Topology across scales on heterogeneous cell data'. In: *PLOS Computational Biology* 21.10 (Oct. 2025). Ed. by Calina Copos, e1013460.

[TSB25] [Software] Christopher Tralie, Nathaniel Saul and Rann Bar-On, *Ripser.py* version 0.6.14, 2025. LIC: MIT. VCS: `https://github.com/scikit-tda/ripser.py/tree/master`. Reference article: 'Ripser.py: A Lean Persistent Homology Library for Python'. In: *The Journal of Open Source Software* 3.29 (Sept. 2018), p. 925.

[TZH15] Chad M. Topaz, Lori Ziegelmeier and Tom Halverson. 'Topological Data Analysis of Biological Aggregation Models'. In: *PLOS ONE* 10.5 (May 2015), e0126383.

[UKB25] The UK Biobank Whole-Genome Sequencing Consortium. 'Whole-genome sequencing of 490,640 UK Biobank participants'. In: *Nature* 645.8081 (Sept. 2025), pp. 692–701.

[vdW+14] [Software] Stéfan van der Walt et al., *scikit-image*. LIC: BSD 3-Clause. VCS: `https://github.com/scikit-image/scikit-image`. Reference article: 'scikit-image: image processing in Python'. In: *PeerJ* 2 (June 2014), e453.

[Vic+19] Sanja Vickovic et al. 'High-definition spatial transcriptomics for in situ tissue profiling'. In: *Nature Methods* 16 (Sept. 2019), pp. 987–990.

[Vil+25] Alba Villaronga-Luque et al. 'Integrated molecular-phenotypic profiling reveals metabolic control of morphological variation in a stem-cell-based embryo model'. In: *Cell Stem Cell* 32.5 (May 2025), 759–777.e13.

[Vip+21] Oliver Vipond et al. 'Multiparameter persistent homology landscapes identify immune cell spatial patterns in tumors'. In: *Proceedings of the National Academy of Sciences* 118.41 (Oct. 2021).

[Vip20] Oliver Vipond. 'Multiparameter Persistence Landscapes'. In: *Journal of Machine Learning Research* 21.61 (Mar. 2020), pp. 1–38.

[Wad57] Conrad Hal Waddington. *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology*. London: George Allen & Unwin, 1957.

[Wet23] Kris A. Wetterstrand. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program*. 2023. URL: `https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data` (visited on 2nd Jan. 2026).

[WNW20] Rui Wang, Duc Duy Nguyen and Guo-Wei Wei. 'Persistent spectral graph'. In: *International Journal for Numerical Methods in Biomedical Engineering* 36.9 (Sept. 2020), e3376.

[Xia+19] Chenglong Xia et al. 'Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression'. In: *Proceedings of the National Academy of Sciences* 116.39 (Sept. 2019), pp. 19490–19499.

[Xin+23] Cheng Xin et al. 'GRIL: A 2-parameter Persistence Based Vectorization for Machine Learning'. In: *Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML)*. Vol. 221. July 2023, pp. 313–333.

[Xu+23] Hao Xu et al. 'SPACEL: deep learning-based characterization of spatial transcriptome architectures'. In: *Nature Communications* 14.7603 (Nov. 2023).

[Yan+25] Jingjie Yang et al. 'Topological classification of tumour-immune interactions and dynamics'. In: *Journal of Mathematical Biology* 91.3 (Sept. 2025), p. 25.

[Yok+14] Maki Yokogawa et al. 'Epicutaneous Application of Toll-like Receptor 7 Agonists Leads to Systemic Autoimmunity in Wild-Type Mice: A New Model of Systemic Lupus Erythematosus'. In: *Arthritis and Rheumatology* 66.3 (Feb. 2014).

[Yoo+16] Jaejun Yoo et al. 'Topological persistence vineyard for dynamic functional brain connectivity during resting and gaming stages'. In: *Journal of Neuroscience Methods* 267 (July 2016), pp. 1–13.

[You+24] Yue You et al. 'Systematic comparison of sequencing-based spatial transcriptomic methods'. In: *Nature Methods* 21 (July 2024), pp. 1743–1754.

[YWZ24] Wang Yin, You Wan and Yuan Zhou. 'SpatialcoGCN: deconvolution and spatial information–aware simulation of spatial transcriptomics data via deep graph co-embedding'. In: *Briefings in Bioinformatics* 25.3 (Mar. 2024), bbae130.

[ZC05] Afra Zomorodian and Gunnar Carlsson. 'Computing persistent homology'. In: *Discrete and Computational Geometry* 33 (2005), pp. 249–274.

[Zei+18] Amit Zeisel et al. 'Molecular Architecture of the Mouse Nervous System'. In: *Cell* 174.4 (Aug. 2018), 999–1014.e22.

[Zen22] Hongkui Zeng. 'What is a cell type and how to define it?' In: *Cell* 185.15 (July 2022), pp. 2739–2755.

[Zha+25] Yu Zhao et al. 'Stereo-seq V2: Spatial mapping of total RNA on FFPE sections with high resolution'. In: *Cell* 188.23 (Nov. 2025), 6554–6571.e21.

[Zhe+21] Fan Zheng et al. 'HiDeF: identifying persistent structures in multiscale 'omics data'. In: *Genome Biology* 22.1 (Dec. 2021), p. 21.

Extended Figures

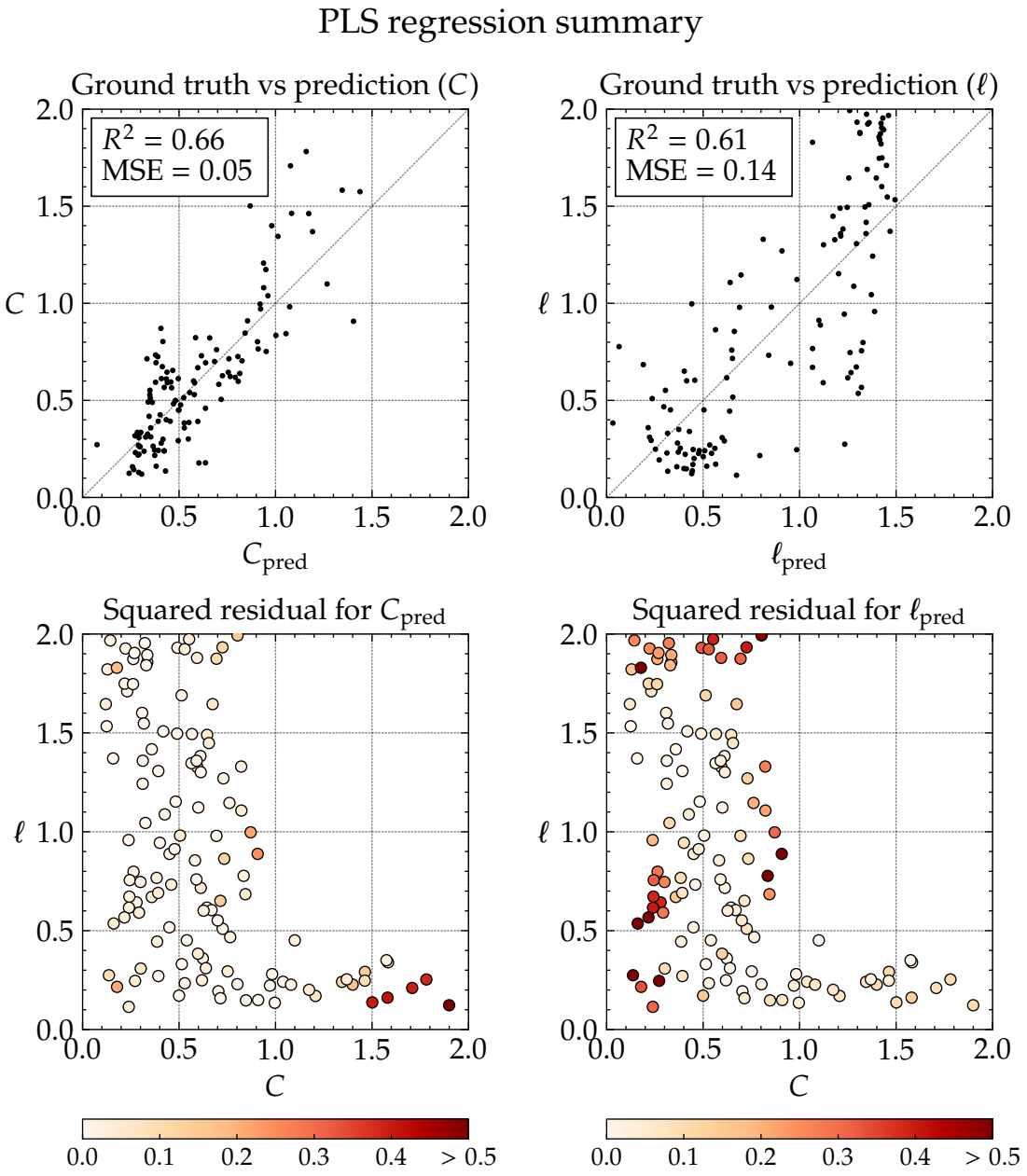


Figure E1. PLS regression of D'Orsogna swarm parameters. Top row: ground truth parameter values vs predicted values for the C (left) and ℓ (right) parameters. Bottom row: Test parameter pairs coloured by squared residual of predictions for C (left) and ℓ (right).

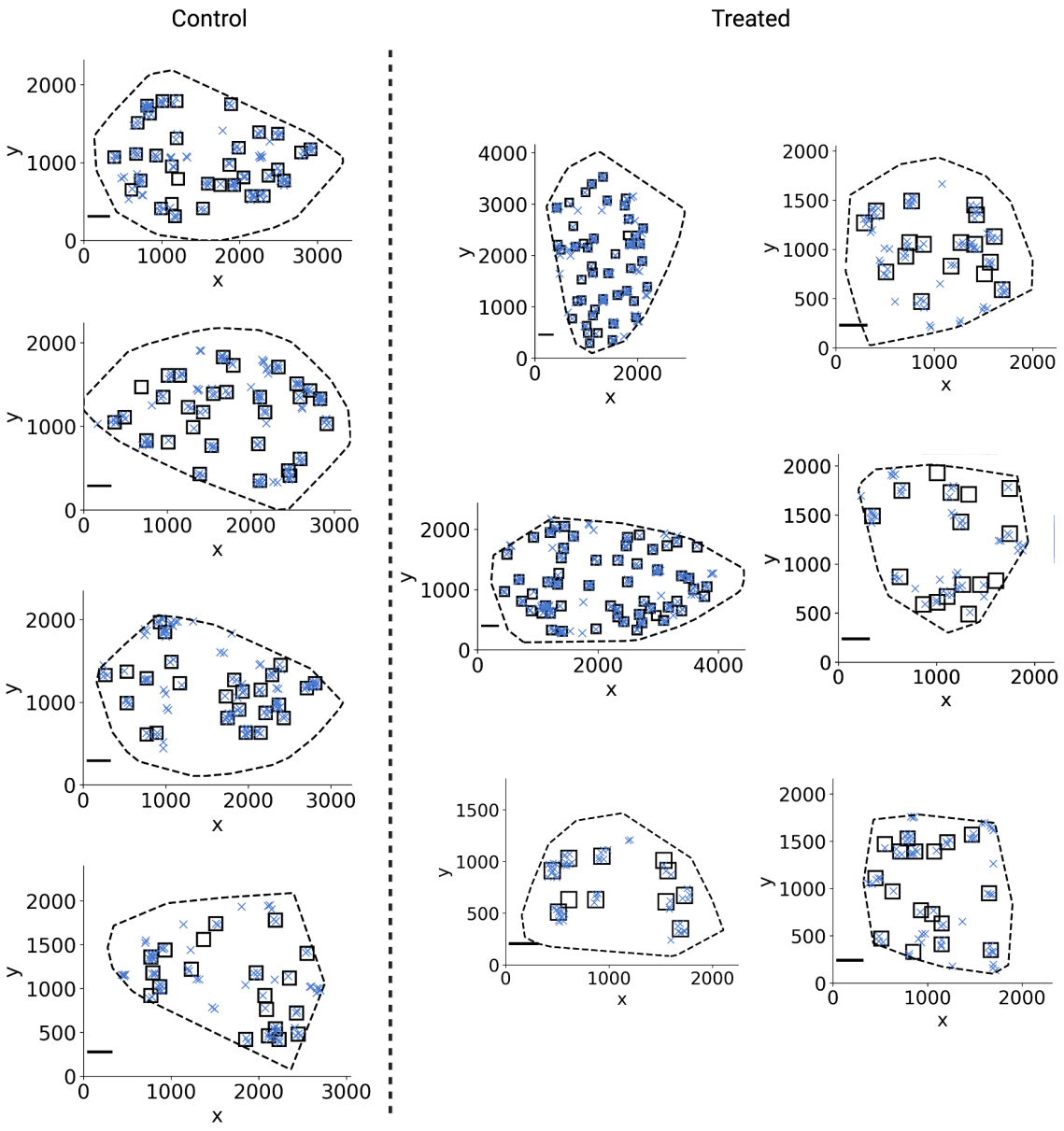


Figure E2. TopACT predicted podocyte cells (blue cross) and ground truth glomeruli (black square) for each mouse kidney Stereo-seq sample. Note that predicted podocytes colocalise with glomeruli, as expected, validating the use of TopACT on mouse kidney data. Dashed black lines show sample boundaries. Scale bars: 0.2 mm.

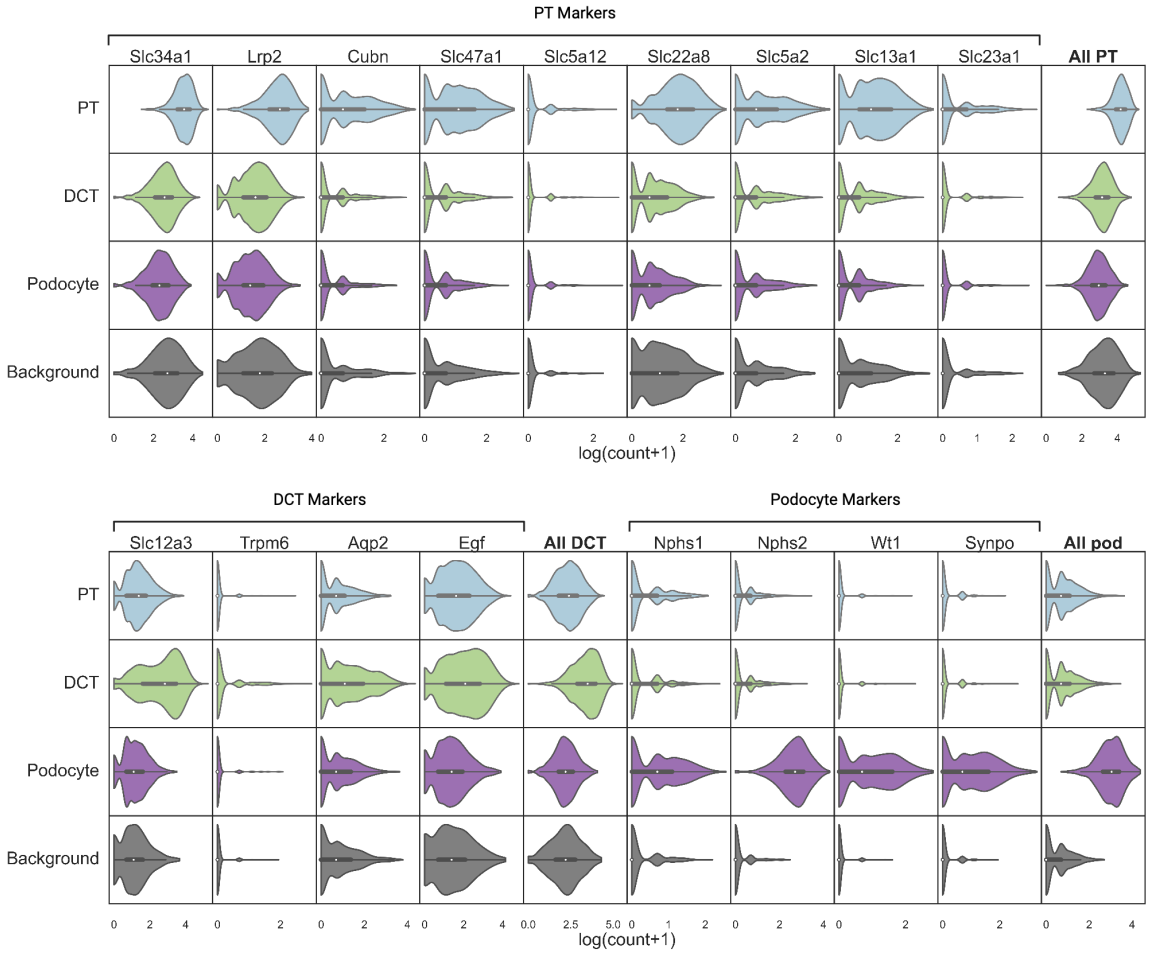


Figure E3. Violin plots showing expression of common markers of proximal tubule (PT), distal convoluted tubule (DCT), and podocyte cells, for TopACT predicted PT cells (blue), TopACT predicted DCT cells (green), TopACT predicted podocyte cells (purple), and randomly sampled background cells (grey), across all mouse kidney samples. Each plot corresponds to the expression counts of a single given marker gene in cells labelled with the given cell type across all samples. Top rows: PT markers. Bottom row, first half: DCT markers. Bottom row, second half: podocyte markers. Log scale.

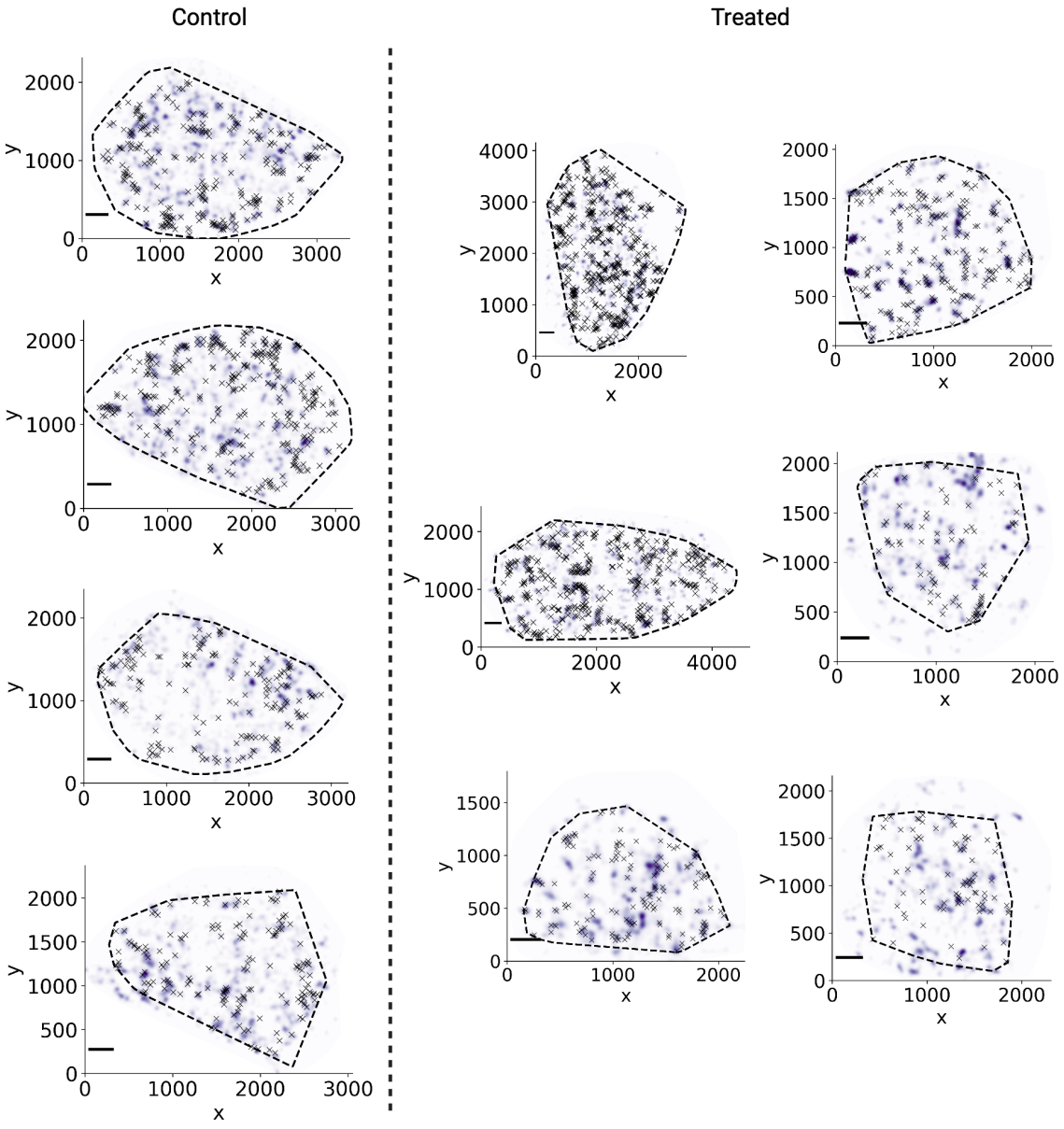


Figure E4. TopACT-predicted DCT cells (black cross) overlaid on map of combined density of DCT marker genes (blue background). Markers are *Slc12a3*, *Trpm6*, *Egf*, and *Aqp2*. Note that predicted DCT cells are found in areas of high marker gene expression as expected. Dashed black lines show sample boundaries. Scale bars: 0.2 mm.

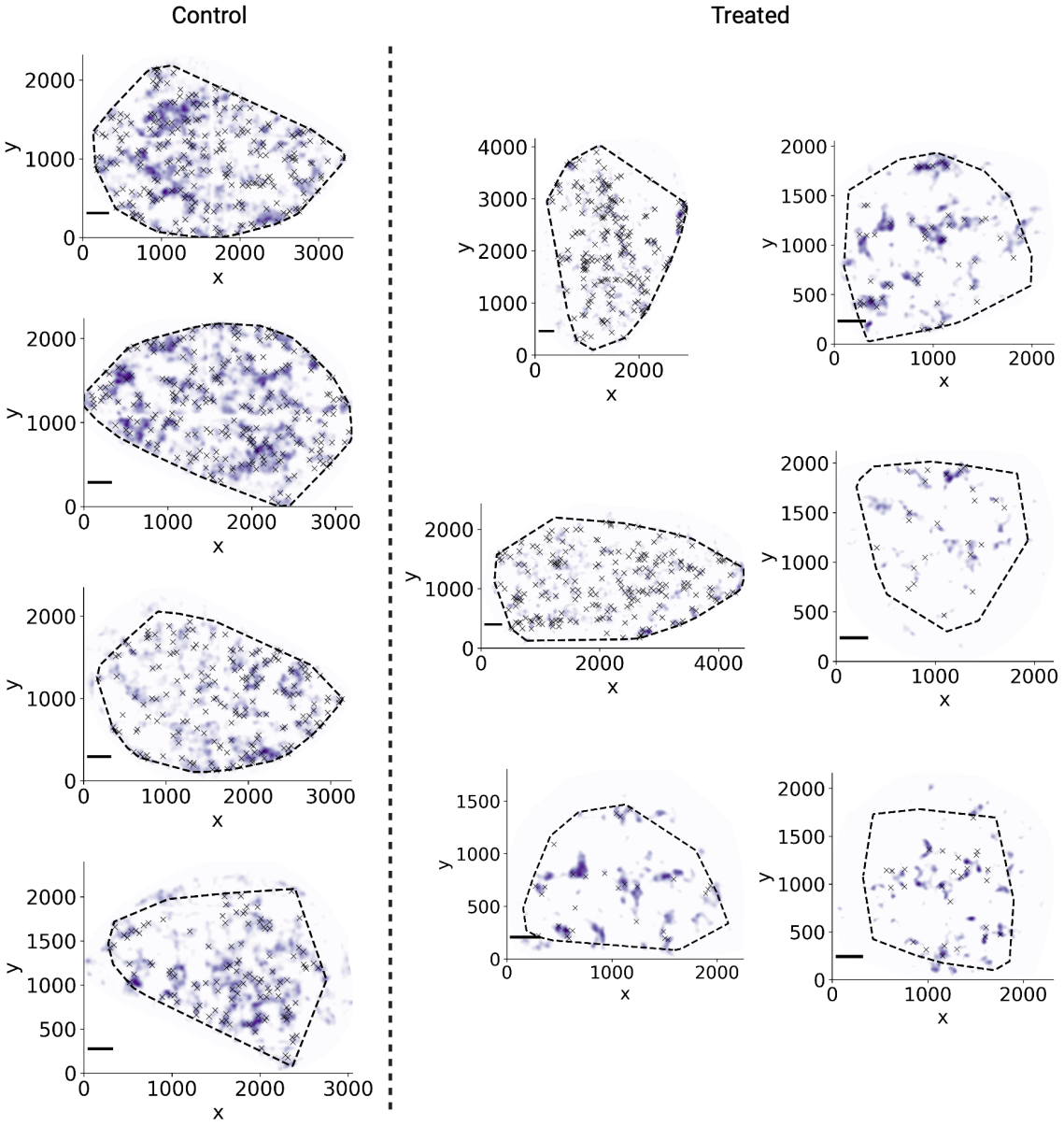


Figure E5. TopACT-predicted PT cells (black cross) overlaid on map of combined density of PT marker genes (blue background). Markers are *Slc34a1*, *Lrp2*, *Cubn*, *Slc47a1*, *Slc5a12*, *Slc22a8*, *Slc5a2*, *Slc13a1*, and *Slc23a1*. Note that predicted PT cells are found in areas of high marker gene expression as expected. Dashed black lines show sample boundaries. Scale bars: 0.2 mm.

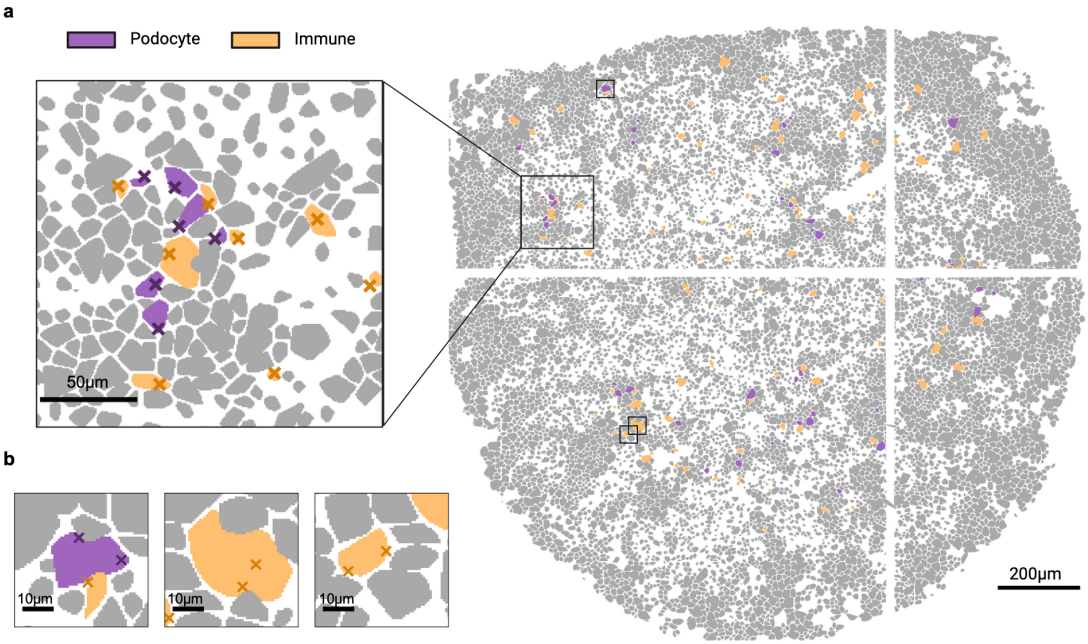


Figure E6. (a) ssDNA-based cell segmentation annotated with TopACT-predicted immune (orange) and podocyte (purple) cells. Grey regions indicate ssDNA cell bins without a corresponding TopACT annotation. Right: the whole sample. Left: a magnified representative patch. Crosses denote TopACT-predicted cell centres. Note that each cell bin in the representative patch contains at most one TopACT prediction. (b) Magnification of the three ssDNA bins with more than one assigned TopACT cell. In each case the bin is assigned two TopACT cells of the same type.