Using Topological Data Analysis to Interpret Cancer Images

Jagdeep Dhesi

Wolfson College

University of Oxford

A thesis submitted for the degree of

M.Sc. in Mathematical Modelling and Scientific Computing

Trinity Term 2022

Abstract

In the immune response to a tumour, specialised white blood cells called macrophages are released by blood vessels to attack and destroy the tumour. Environmental cues in the tumour microenvironment may alter the behaviour of macrophages causing them to enact pro-tumour tendencies, promoting the growth and spreading of the tumour. This heterogeneity in the macrophage population gives rise to complex spatial patterns. We employ methods from topological data analysis (TDA) to determine the concentration of pro-tumour macrophages, given only dynamic spatial data of the involved cells from an agent-based model. Moreover, we use machine learning methods to predict the future concentration of pro-tumour macrophages given spatial data at an earlier time.

Contents

List of Figures

Chapter 1

Introduction

The tumour microenvironment is a complex landscape consisting of many distinct agents which each play an important role in the response of the immune system [4]. These agents include tumour cells, stromal cells, necrotic cells, macrophages and blood vessels. Macrophages are the immune cells which are released in response to tumours, however their behaviour in the proximity of the tumour is heavily dictated by microenvironmental cues relating to concentrations of different diffuse cytokines and chemicals [4, 19]. Given the proximal conditions, macrophages either exhibit anti- or pro-tumour tendencies [19], where the presence of a large number of pro-tumour macrophages is heavily associated with unfavourable outcomes [19, 23]. Anti-tumour macrophages typically migrate towards and fight the tumour, whereas pro-tumour macrophages tend to mediate the transport of tumour cells away from the primary tumour and promote tumour growth and metastasis [4]. Such complex behaviours give rise to a range of spatial patterns in the tumour microenvironment, and these may be indicative of the nature of the interaction. A key deterministic metric for the outcome of the interaction is the ratio of macrophages which have these pro-tumour tendencies. As an example, in [23] they show an association between a higher concentration of pro-tumour macrophages and decreased survival rate in patients with colorectal cancer. We use the spatial features of the different cell types in the tumour microenvironment to classify if there is a majority of pro-tumour macrophages in a given image, moreover we build a predictive model based on these features to determine future pro-tumour macrophage concentrations. We obtain topological features by employing methods from the emerging field of topological data analysis. We compare our classification accuracies using several different constructions of our topological features, involving a well established method in the Rips filtration, a method which has not been used for image analysis in the Dowker filtration, and finally a novel construction which takes

into account the time variable in these dynamic tumour microenvironments.

This report is structured in the following way. We start in Chapter 2 with a review of the relevant notions from algebraic topology. Following from this, in Chapter 3, we introduce our biological problem formally, and in particular we introduce aspects of the agent-based model which generates the spatial data that we use for our topological data analysis. In Chapter 4 we introduce the specific methods we use to determine the topological features of our data set, and the machine learning methods we use for the subsequent data analysis. In Chapter 5 we present and interpret the results of applying the methods laid out in Chapter 4 to the obtained data from the agent-based model described in Chapter 3. We conclude, in Chapter 6 by discussing the implications of our findings, and possible directions for future analysis.

Chapter 2

Topology Background

In this Chapter we review notions from topology that form the mathematical basis of this report. We start by introducing simplicial complexes and noting some important features of these high dimensional structures. We then consider the quantification of 'holes' in a given simplicial complex through notions from simplicial homology. Following from this, we further quantify the change in homology over a sequence of simplicial complexes. We then consider the stability of persistent homology before finally introducing the main, vectorised, topological representations we use in our analysis. This section is entirely a review pertaining to a vast amount of literature on algebraic topology. We consider standard definitions for simplicial complexes and simplicial homology from [21], and review results regarding persistent homology from [13, 14, 27]. We note that this report heavily makes use of the computational aspects of persistent homology, and these are discussed in great detail in [10, 27].

2.1 Simplicial Complexes

Before we discuss simplicial complexes, it is sensible to recall the definition of a simplex. In a combinatorics setting a k-simplex is defined as a finite set $\{v_0, \dots v_k\}$ of elements called vertices. The geometric counterpart follows intuitively from this definition if we think of the points as being embedded in $\mathbb{R}^k$, formally we have the following definition.

Definition 2.1 (Simplex) *A k-simplex is the convex hull formed by $k+1$ affinely independent points $v_0, \cdots v_k \in \mathbb{R}^k$.*

By this definition, 0-simplices are vertices, 1- simplices are edges, 2-simplices are

triangles and so on[1]. The requirement of affine independence ensures that the set of $k+1$ vertices in the k-dimensional space do not lie on the same $k-1$ dimensional hyperplane, for example the three vertices of a triangle in two dimensional space cannot lie on the same edge. We can also define a face of a simplex:

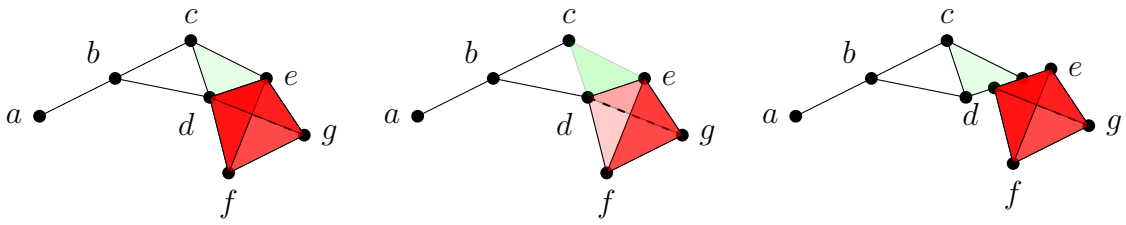
Definition 2.2 (Face) *A face τ of a k-simplex σ is a d-simplex formed by a subset of the vertices of σ, where $d < k$.*

A k-dimensional simplicial complex, in essence, is a collection of simplices of dimension at most k, where we include the faces of any included simplex. To interpret this geometrically we impose an extra condition on the intersection of simplices. The formal definition is the following.

Definition 2.3 (Simplicial Complex) *A k-dimensional simplicial complex $\mathcal{X}$ is a collection of simplices of dimension at most k that satisfies the following properties:*

1. *If $\sigma \in \mathcal{X}$ and $\tau \subset \sigma$, then $\tau \in \mathcal{X}$*
2. *If $\sigma_1, \sigma_2 \in \mathcal{X}$, and $\sigma_1 \cap \sigma_2 = \tau \neq \emptyset$, then $\tau \subset \sigma_1$ and $\tau \subset \sigma_2$.*

Here, the first condition imposes the closure under subsets (faces) requirement, and the second condition ensures that the intersection of two simplices in the simplicial complex is a common face of both. Figure 2.1 shows an example of a simplicial complex along with non-examples that showcase the necessity for our conditions.



simplicial complex without property (1) without property (2)

Figure 2.1: An example of a 3-dimensional simplicial complex (left), and examples of possible structures that disregard property (1) (middle) and property (2) (right).

[1]For computation purposes we generally do not consider tetrahedra or higher dimensional simplices.

2.2 Simplicial Homology

For our purposes, it is sufficient to note that by simplicial homology, we refer to the space of i-dimensional holes in a simplicial complex. A 0-dimensional hole describes connected component, a 1-dimensional hole describes a cycle, a 2-dimensional hole describes a void, and so on. As an example, the simplicial complex shown in figure 2.1 has a single connected component, a single cycle, and zero higher dimensional holes. Equivalently, we could say that the ranks of the kth homology groups are $\operatorname{rank} H_k = 1$ for $k = 0, 1$ and $\operatorname{rank} H_k = 0$ for all $k \in \mathbb{N}_{>1}$. The goal of this section is to provide a brief but formal introduction to this concept.

It shall be helpful in our discussions to start by introducing the notion of an oriented simplex, which we can do by assigning the constituent vertices of a given simplex with positive integers. Namely, when we refer to an oriented k-simplex defined by the vertices $v_0, \ldots, v_k$, we traverse the simplex in the order of the indices. If two simplices are differently oriented we denote this with a minus sign. To start considering the space of k-dimensional holes in a given simplicial complex $\mathcal{X}$, we need to consider the isolated space of simplices in $\mathcal{X}$ whose dimension is exactly k. This information is defined by the kth chain group. Let $|\sigma|$ denote the number of vertices that constitute the simplex σ, then we have the following definition.

Definition 2.4 (Chain Group and k-chain) *The kth chain group of a simplicial complex $\mathcal{K}$, $C_k(\mathcal{K})$ is the free abelian group generated by the basis including all distinct oriented k-simplices of $\mathcal{K}$. Namely, we have*

$$C_k(\mathcal{K}) = \left\{ \sum_i a_i \sigma_i \quad \middle| \quad |\sigma_i| = k+1, \sigma_i \in \mathcal{K}, a_i \in \mathbb{Z}_2 \right\}. \tag{2.1}$$

We call an element of $C_k(\mathcal{K})$ a k-chain of $\mathcal{K}$.

Here, $\mathbb{Z}_2$ denotes the quotient group $\mathbb{Z}/2\mathbb{Z}$. In the literature it is very common to use general ring coefficients, however the upcoming derivations become far more complicated. For our purposes[2], taking the ground ring to be $\mathbb{Z}_2$ is sufficient. Then the notion of addition between simplices with these coefficients is simply analogous to $\mathbb{Z}_2$ addition, and the fact that we have binary coefficients allows for an easy interpretation to a k-chain; if the coefficient $a_i = 1$ then the simplex σ_i belongs to the chain, and

[2]It is also sufficient in most practical, computational, aspects of simplicial homology to take $\mathbb{Z}_2$ coefficients.

if $a_i = 0$, then σ_i does not belong to the chain. Also, since $\mathbb{Z}_2$ is a field, our chain groups are actually vector spaces. Now that we can classify between the different dimensions of our simplicial complex, we can define a map between them.

Definition 2.5 (Boundary Operator) *Given some simplicial complex $\mathcal{K}$, the boundary operator $\partial_k : C_k(\mathcal{K}) \mapsto C_{k-1}(\mathcal{K})$ is the linear map[3] which is defined, on an oreiented k-simplex $\sigma \in \mathcal{K}$, by the formal sum*

$$\partial_k \sigma = \sum_{i=0}^{k} (-1)^i [v_0, \cdots, \hat{v_i}, \cdots, v_k], \tag{2.2}$$

and on an arbitrary k-chain c by the formal sum

$$\partial_k c = \sum_{i=1} a_i \partial_k \sigma_i, \tag{2.3}$$

where $a_i \in \mathbb{Z}_2$ are the corresponding coefficients for k-simplices $\sigma_i \in \mathcal{K}$.

The notation $[v_0, \cdots, \hat{v_i}, \cdots, v_k]$ means that the vertex v_i is not included in the resulting $(k-1)$-simplex. The defining property of the boundary operator is the result of the composition

$$\partial_k \circ \partial_{k+1} \sigma = 0 \tag{2.4}$$

for any k-simplex σ. This tells us that the boundary of a boundary is empty. Sequences of vector spaces (2.1) and corresponding linear maps (2.2) that satisfy (2.4) are fundamental objects in the theory of simplicial homology. We summarise these in the following definition.

Definition 2.6 (Simplicial chain complex) *The simplicial chain complex of the simplicical complex $\mathcal{K}$, (C_*, ∂_*), is the sequence of chain groups C_k for $k \geq 0$ along with the linear maps ∂_k.*

The relationship between chain groups and boundary operators are perhaps more easily seen in the following diagram.

$$\cdots \xrightarrow{\partial_{k+2}} C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{\partial_{k-1}} \cdots$$

Now that we have a framework to handle simplicial complexes, and in particular, a tool to traverse their hierarchical structure (the boundary operator), we can finally consider homology groups of simplicial complexes. Cycle and boundary groups allow us to do this.

[3] In the case of general ring coefficients, the map is a homomorphism.

Definition 2.7 (Cycle Group) *The kth cycle group of $\mathcal{K}$ is defined as $Z_k = \ker \partial_k$*

Definition 2.8 (Boundary group) *The kth boundary group of $\mathcal{K}$ is defined as $B_k = \operatorname{im} \partial_{k+1}$*

Again, since we are working with $\mathbb{Z}_2$ coefficients, both of these structures are vector spaces. The kth boundary group is the set of boundaries of $(k+1)$ chains, and the kth cycle group is the subset of k-chains that form closed loops. In particular, every boundary of a $(k+1)$-chain forms a closed loop in C_k, that is $B_k \subseteq Z_k \subseteq C_k$. Moreover, we notice that the space of cycles in the absence of those that are boundaries of $(k+1)-$chains are simply those cycles that enclose voids, thus we have found our kth-homology group.

Definition 2.9 (Homology group) *The kth homology group of $\mathcal{K}$, H_k is given by the quotient group*

$$H_k = Z_k / B_k. \tag{2.5}$$

This vector space encompasses the class of k-dimensional voids in the simplicial complex, and therefore the number of such voids is given by its dimension[4], and we formally encode this in the k-th Betti number.

Definition 2.10 (Betti number) *The kth Betti number of $\mathcal{K}$ is $\beta_k = \dim H_k$.*

We therefore have found, formally, the number of holes in a given simplicial complex. What we are especially interested with in topological data analysis is how the homology changes over a sequence of simplicial complexes. This concept belongs to an area of study called persistent homology, which we consider next.

2.3 Persistent Homology

Consider a simplicial $\mathcal{K}$. We want to build up a sequence of simplicial complexes, with each prior simplicial complex being contained in the next, and where the sequence eventually ends up at a $\mathcal{K}$. This construction is called a filtration, and we define it formally after a series of complementary definitions.

Definition 2.11 (Subcomplex) *A subcomplex $\mathcal{L}$ of a simplicial complex $\mathcal{K}$ is a subset $\mathcal{L} \subseteq \mathcal{K}$ that also satisfies the properties of a simplicial complex.*

[4]Since we are dealing with vector spaces, we can take the number of voids as the dimension of H_k, however in the general case H_k would define a free abelian group, and thus we would be required to take the dimension of its smallest generating set; rank H_k.

Definition 2.12 (Simplicial map) *A simplcial map is a map $f : \mathcal{K}_1 \mapsto \mathcal{K}_2$ that assigns every vertex in $\mathcal{K}_1$ to a vertex in $\mathcal{K}_2$, with the additional property that simplices get mapped to simplices. That is if $\sigma = \{v_0, \cdots, v_k\} \in \mathcal{K}_1$, then $f(\sigma) = \{f(v_1), \cdots, f(v_2)\} \in \mathcal{K}_2$.*

Definition 2.13 (Inclusion map) *An inclusion map $g : \mathcal{L} \hookrightarrow \mathcal{K}$ is a simplicial map between a simplicial complex $\mathcal{K}$ and its subcomplex $\mathcal{L} \subseteq \mathcal{K}$, with the additional property that any simplex in $\mathcal{L}$ is mapped to the same simplex in $\mathcal{K}$.*

Definition 2.14 (Identity map) *The identity map of a simplicial complex $\mathcal{K}$ is the inclusion map $1_{\mathcal{K}} : \mathcal{K} \hookrightarrow \mathcal{K}$.*

Definition 2.15 (Filtration) *A filtration of a simplicial complex $\mathcal{K}$ is a nested sequence of subcomplexes $\emptyset \subseteq \mathcal{K}_0 \subseteq \mathcal{K}_1 \subseteq \cdots \subseteq \mathcal{K}_m = \mathcal{K}$. In this case we refer to $\mathcal{K}$ as a filtered simplicial complex.*

Given an appropriate sequence of simplicial maps between elements of the filtration, one could alternatively write a filtration as the following list of inclusions.

$$\emptyset \xhookrightarrow{g} \mathcal{K}_0 \xhookrightarrow{g_0} \mathcal{K}_1 \xhookrightarrow{g_1} \cdots \xhookrightarrow{g_{m-1}} \mathcal{K}_m \xhookrightarrow{1_{\mathcal{K}}} \mathcal{K}$$

When we considered the homology of a singular simplicial complex, we made use of the simplicial chain complex, namely a chain group together with its corresponding boundary operator. We simply extend this to the case of a filtration by introducing a map between simplicial chain complexes.

Definition 2.16 *Given two chain complexes (C^1_*, ∂^1_*) and (C^2_*, ∂^2_*), the chain map F_* is the sequence of linear maps indexed by k, $F_k : C^1_k \mapsto C^2_k$ such that $\partial^1_k \circ F_k = F_{k-1} \circ \partial^1_k$ for all $k \geq 0$.*

This definition is equivalent to requiring the following diagram to commute:

$$\begin{array}{ccccccccc}
\cdots & \xrightarrow{\partial^1_{k+2}} & C^1_{k+1} & \xrightarrow{\partial^2_{k+1}} & C^1_k & \xrightarrow{\partial^1_k} & C^1_{k-1} & \xrightarrow{\partial^1_{k-1}} & \cdots \\
 & & \downarrow{\scriptstyle F_{k+1}} & & \downarrow{\scriptstyle F_k} & & \downarrow{\scriptstyle F_{k-1}} & & \\
\cdots & \xrightarrow{\partial^2_{k+2}} & C^2_{k+1} & \xrightarrow{\partial^2_{k+1}} & C^2_k & \xrightarrow{\partial^2_k} & C^2_{k-1} & \xrightarrow{\partial^2_{k-1}} & \cdots
\end{array}$$

Maps between chain complexes also induce linear maps between their corresponding kth Homology groups. The collection of these is known as a persistence module.

Definition 2.17 (Persistence Module) *A persistence module over $\mathbb{Z}_2$ is a sequence $\mathcal{M} = (M_*, a_*)$ of vector spaces over $\mathbb{Z}_2$ and linear maps a_k such that $a_k : M_k \mapsto M_{k+1}$.*

We typically consider the persistence module induced by a filtered simplicial complex. Namely, the vector spaces are the kth homology groups and the linear maps are the induced maps between kth homology groups. Another important type of persistence module we make use of are interval modules.

Definition 2.18 (Interval Module) *An interval module described by indices i, j with $i \leq j \leq \infty$ with $i \neq \infty$ is the persistence module $(I_*^{i,j}, \phi_*^{i,j})$ of vector spaces*

$$I_p^{i,j} = \begin{cases} \mathbb{Z}_2 & i \leq p \leq j, \\ 0 & otherwise \end{cases} \tag{2.6}$$

and linear maps

$$\phi_p^{i,j} = \begin{cases} id_{\mathbb{Z}_2} & i \leq p < j \\ 0 & otherwise. \end{cases} \tag{2.7}$$

where $id_{\mathbb{Z}_2}$ denotes the identity map on $\mathbb{Z}_2$.

Consider the pesistence module of homology groups of our filtered simplicial complex $\mathcal{K}$, $\mathcal{M} = (H_*, a_*)$. What this says exactly is that for each simplicial complex $\mathcal{K}_m$ in our filtration, we have our kth Homology group, $H_k(\mathcal{K}_m)$, and a map a_k which takes us to the next kth homology group in the filtration, $H_k(\mathcal{K}_{m+1})$. The goal of persistent homology is to know how 'persistent' topological features (k-dimensional voids) are over the course of a filtration of $\mathcal{K}$. This is uniquely described by the indices of the simplicial complexes in the filtration at which the void is 'born' and at which it 'dies'. The structure theorem is a fundamental theorem in the study of persistent homology, as it says that, given all the pairs of indices which describe the birth and death times of k-dimensioanl voids in the filtration, the persistence module of the filtration can be decomposed into a direct sum of interval modules defined by these indices [13, 27]. Let $\mathcal{B}$ denote the set of pairs of birth and death indices for an arbitrary dimension.

Theorem 2.1 *[Structure Theorem] Given the persistence module $\mathcal{M} = (H_*, a_*)$ of homology groups of a filtered simplicial complex $\mathcal{K}$. There exists a set $\mathcal{B}$ of index pairs and a function $\mu : \mathcal{B} \mapsto \mathbb{N}$ such that we have the following decomposition:*

$$\mathcal{M} \cong \bigoplus_{[i,j] \in \mathcal{B}} (I_*^{i,j}, \phi_*^{i,j})^{\mu(i,j)}. \tag{2.8}$$

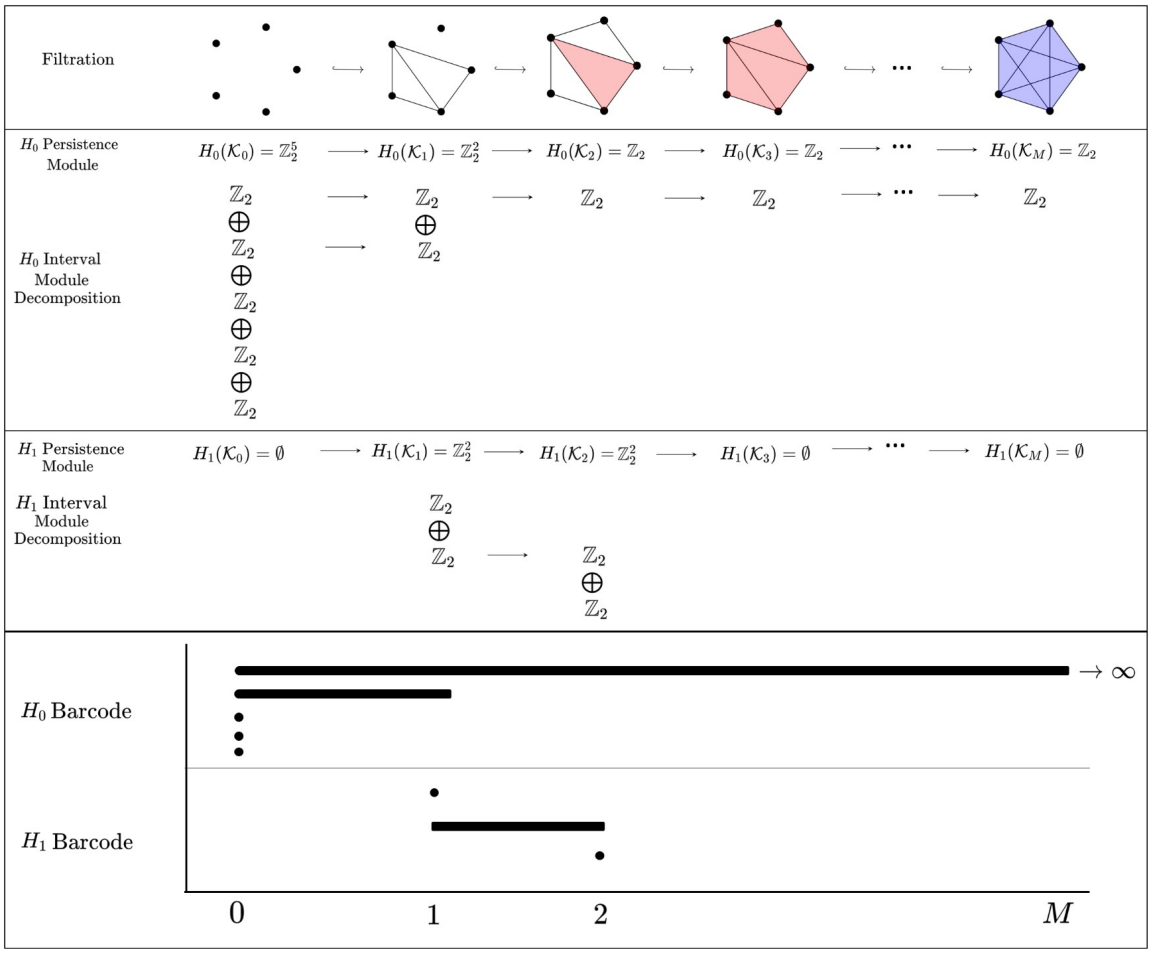


Figure 2.2: Example of a filtration of a simplicial complex, along with H_0 and H_1 persistence modules, and their corresponding interval module decompositions. The interval modules are $(0,\infty) \oplus (0,1) \oplus (0,0)^3$ for H_0 and $(1,1) \oplus (1,2) \oplus (2,2)$ for H_1. We also see the corresponding barcode for the filtration.

The elements of $\mathcal{B}$ are called bars, and the visualisation of all bars of a given filtration is called a barcode. The relationship between bars and the interval module decomposition can clearly be seen in figure 2.2. It is important to note that for our computational purposes, we do not consider homology past $k = 1$. We now have a well defined, quantitative method to determine the topological features of a given filtered simplicial complex. The existence of a barcode representation of our topological features forms the basis of our topological data analysis. In the next section we introduce an equivalent representation of barcodes that gives rise to a stability theorem, and in the final section of this Chapter we intrdouce a vectorised reprsentation which shall be used in our analysis.

2.4 Stability of Persistent Homology

Since the barcode of a given filtered simplicial complex is defined only by its interval module decomposition, and thus by a multiset of index pairs defining the birth and death points, an alternate representation would be to consider the plot of birth times against death times. Such a representation is called a persistence diagram, and will be very helpful when we discuss vectorised representations of topological features. For a filtered simplicial complex $\mathcal{K}$, the persistence diagram representation, $D(\mathcal{K}) \subset \mathbb{R}^2$ simply considers the multiset of elements of $\mathcal{B}$ as points embedded in the plane.

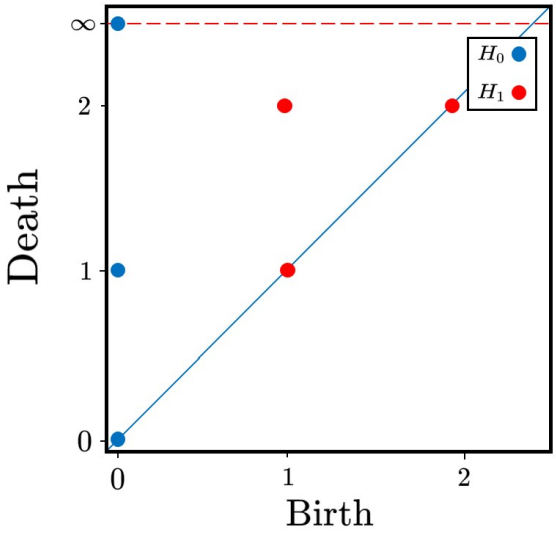


Figure 2.3: Persistence diagram representation of the filtration shown in figure 2.2.

Figure 2.3 shows an example of a persistence diagram for the filtration shown in figure 2.2. Points higher above the diagonal on this diagram can be interpreted as being more persistent over the filtration. Such a representation is useful since the space of persistence diagrams can be equipped with a metric, with typical examples being the Wasserstein or Bottleneck distances [1, 11, 12]. We formally define the Bottleneck distance between multisets of points.

Definition 2.19 (Bottleneck Distance) *Let X and Y denote two multisets. Let $\gamma : X \mapsto Y$ be an arbitrary bijection between our multisets. The bottleneck between distance between X and Y is*

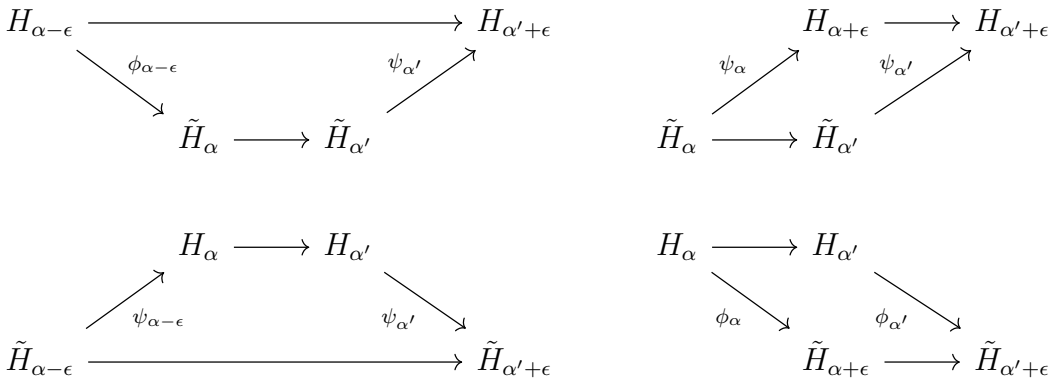
$$d(X, Y) = \inf_{\gamma} \sup_{x} ||x - \gamma(x)||_{\infty} \tag{2.9}$$

where we take the supremum over all $x \in X$ and the infimum over all possible bijections between X and Y.

And we say the botteneck distance between two persistence diagrams is the botteneck distance between their representative multisets of elements. Having such a metric is useful as it gives us a quantitative way to compute the dissimilarity between two persistence diagrams. A high bottleneck distance between two persistence diagrams is indicative of a significant difference between the persistent homology of their respective filtered simplicial complexes. Equipping the space of persistence diagrams with the bottleneck metric gives rise to a stability theorem for small perturbations in the original data [1, 8]. We state this theorem after defining a notion of closeness between persistence modules.

Recall that a filtration will induce a persistence module of kth homology groups and linear maps between them. Consider the two filtrations we would like to compare, and in particular let the persistence modules of their kth Homology groups be $\mathcal{M}_1 = (H, f)$ and $\mathcal{M}_2 = (\tilde{H}, g)$.

Definition 2.20 (ϵ-interleaving) *$\mathcal{M}_1$ and $\mathcal{M}_2$ are ϵ-interleaved if there exists two families of linear maps $\{\phi_\alpha : H_\alpha \mapsto \tilde{H}_{\alpha+\epsilon}\}_{\alpha\in\mathbb{R}}$ and $\{\psi_\alpha : \tilde{H}_\alpha \mapsto H_{\alpha+\epsilon}\}_{\alpha\in\mathbb{R}}$. Such that for all $\alpha \leq \alpha'$, the following diagrams commute.*



The intuition behind the required commutative relationships is that for the two persistence modules to be ϵ-interleaved, we should be able to find a linear mapping between the kth homology group indexed by α in one filtration, with the kth homology group of the other filtration indexed by $\alpha+\epsilon$, for arbitrary $\alpha \in \mathbb{R}$. In terms of our topological features, an ϵ-interleaving between two persistence modules induced by the filtrations says that if a feature in one filtration is born at index α, then it will be born in the other filtration by index $\alpha + \epsilon$. This allows us to define a notion of similarity between two filtrations, in the case where ϵ is sufficiently small. We can further quantify the similarity between the topological features formed in these filtrations with the following stability theorem regarding the resulting persistence diagrams of ϵ-interleaved persistence modules.

Theorem 2.2 (Stability of ϵ-interleaved Persistence Modules) *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two ϵ-interleaved persistence modules induced by filtered simplicial complexes $\mathcal{K}_1$ and $\mathcal{K}_2$ respectively. Then*

$$d(D(\mathcal{K}_1), D(\mathcal{K}_2)) \leq \epsilon. \tag{2.10}$$

where we have used the Bottleneck distance between persistence diagrams.

This theorem says that similar filtrations give rise to similar persistence diagrams, and the level of similarity is determined by ϵ. The existence of such a stability theorem for persistent homology is fundamental for its applicability in data analysis.
Persistence diagrams are representations of persistent homology that are both stable and easy to interpret, however, their shortcoming is with their lack of suitability with standard machine learning algorithms, which typically require a vectorised input. As such we instead employ the vectorised representation of persistence diagrams; persistence images, which we introduce next.

2.5 Vectorisation of Topological Summary via Persistence Images

Persistence images offer a vectorised representation of persistence diagrams. We next briefly review how these are computed, for more details see [1]. We note that for the convenience of this method, we apply the mapping $(b, d) \mapsto (b, d - b)$ to our persistence diagrams. In applying this mapping we are considering the plot of birth time, b, against $d-b$. We define $p = d-b$ as the persistence of our topological feature. Let B denote our persistence diagram, and consider a space $\mathbb{R}^2$ where our persistence image will live. Let $f : \mathbb{R}^2 \mapsto \mathbb{R}$ denote an arbitrary weighting function that is continuous, piecewise differentiable and satisfies $f(x, y) \geq 0$ and $f(x, 0) = 0$ for all $(x, y) \in \mathbb{R}^2$. Then, for each point $(x, y) \in \mathbb{R}^2$, apply the mapping

$$\rho_B(x, y) = \sum_{(b,p)\in B} f(b, p)\, \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x-b)^2 + (y-p)^2}{2\sigma^2}\right) \tag{2.11}$$

where the sum is taken over all points in our persistence diagram, and σ^2 is an arbitrary variance of the Gaussian part. This simply weights each point with a Gaussian radial basis function, so that points that appear in the persistence diagram have high intensities, which decrease radially further from the point. For our analysis, we always consider the weighting function f to be 0 on the x-axis, and linearly increasing to the highest point on our transformed persistence diagram [28]. Namely, to the point

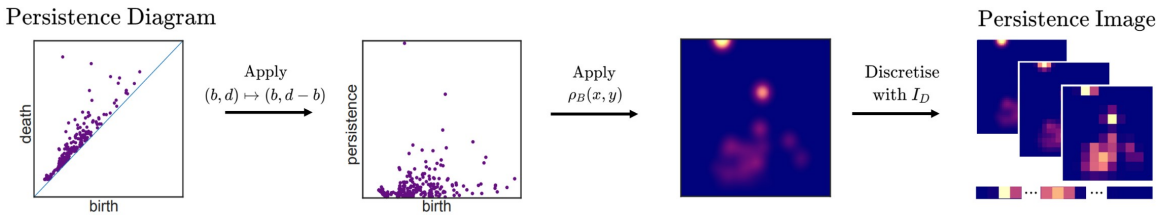


Figure 2.4: Pipeline to compute a persistence image from a persistence diagram. We see the visualisation at each step of the computation, along with three persistence images representing different sizes of discrete domains. Figure adapted from [1].

which is most persistent. The persistence image is computed from this mapping by considering an arbitrary discrete domain of n^2 boxes $D_i \subset \mathbb{R}^2$ for $i = 1, \ldots, n^2$ and assigning each of the boxes with an intensity given by

$$I_D = \iint_D \rho_B \, \mathrm{d}y\mathrm{d}x \tag{2.12}$$

We then have a $n \times n$ grid with an intensity assigned to each 'pixel', and thus a vectorised form of our persistence diagram, since we can reshape the $n \times n$ matrix into a size n^2 vector. The distance between persistence images is then simply the Euclidean norm of the difference between these vectors. Figure 2.4 shows the pipeline used to compute a persistence image from a persistence diagram. We frequently make use of this pipeline in our analysis, and refer to it repeatedly in Chapter 5.

Chapter 3

Biological Background

In this Chapter we first motivate the biological background of the interaction we study using topological data analysis. We then briefly describe aspects of the agent-based model used to generate the corresponding data, taking a particular focus on the different spatial patterns that arise in the synthetic data.

3.1 Motivation

We consider a model of the immune response to tumours, where macrophages exhibit anti- and pro-tumour tendencies dependent on a continuous phenotype. In this section we motivate the roles of different cells in the tumour microenvironment over the course of the interaction. The heterogeneous landscape gives rise to complex behaviours, for example those between macrophages and tumour cells [4, 6, 18]. Several diffusible species define the environmental cues that determine the phenotype of macrophages. Colony stimulating factor-1 (CSF-1) is produced by tumour cells and acts as a chemoattractant for macrophages, causing their migration towards the tumour and an increased sensitivity of macrophages towards CSF-1 promotes favourable, anti-tumour tendencies. In contrast, tumour cell induced transforming growth factor beta (TGF-β) increases the sensitivity of macrophages to C-X-C motif chemokine 12 (CXCL12), which results in unfavourable, pro-tumour tendencies in the process which we describe next. Cancer associated fibroblasts near blood vessels produce CXCL12 [2, 4]. Exposure of macrophages to this chemotactic cytokine stimulates the macrophages to migrate towards higher concentrations of CXCL12. It also increases expression by macrophages of epidermal growth factor (EGF), which acts as a chemoattractant for the tumour cells [4]. We observe migration of tumour

cells in the vicinity of these now pro-tumour macrophages along a common trajectory towards blood vessels, increasing the likelihood that tumour cells enter into the blood stream and, ultimately, metastasise [2, 4]. The interplay between macrophage phenotype and behaviour gives rise to complex spatial patterns which may contain information about the interactions. Time series imaging data of such interactions is not easy to collect. We instead exploit an agent-based model (ABM) [4] that describes these interactions, and use synthetic data generated from the ABM in our analysis. In the next section we describe aspects of the agent-based model that are relevant to the interactions we have described.

3.2 Agent-based Model

In this section we introduce the ABM and the generated data which we shall analyse. We provide examples of spatial patterns which we shall refer to in Chapter 5, where we discuss our results. For more details regarding the nuances of the ABM, see [4].

The agent based model considers an in vivo tumour-immune interaction on a domain $\Omega = [0, 50] \times [0, 50]$. It considers four agents; macrophages, tumour cells, necrotic cells and stromal cells. Blood vessels exist as point sources for macrophages and oxygen, however they do not occupy area or interact with other agents directly. Mechanical forces exist between sufficiently close agents, and these govern the basic movement of agents. Macrophages and tumour cells also experience chemotactic forces, which depend on the diffusible species CSF-1, TGF-β, CXCL12 and EGF. Put simply, the chemotactic forces on macrophages bias their movement either towards tumour cells, or towards blood vessels (where we assume CXCL12 to be produced), and this depends on concentrations of CSF-1 and CXCL12, the macrophages' sensitivity towards CSF-1 and CXCL12 (recall that this is influenced by TGF-β) and finally the phenotype of macrophages. The chemotactic forces experienced by tumour cells are only dependent on the local concentrations of EGF, which we recall is expressed by pro-tumour macrophages and acts as a chemoattractant for tumour cells. High concentrations of EGF cause tumour cells to migrate towards blood vessels with macrophages, promoting metastasis. The interplay between the diffusible species, macrophage phenotype and the movement of macrophages and tumour cells give rise to complex spatial patterns which we shall analyse. We show some examples of these next.

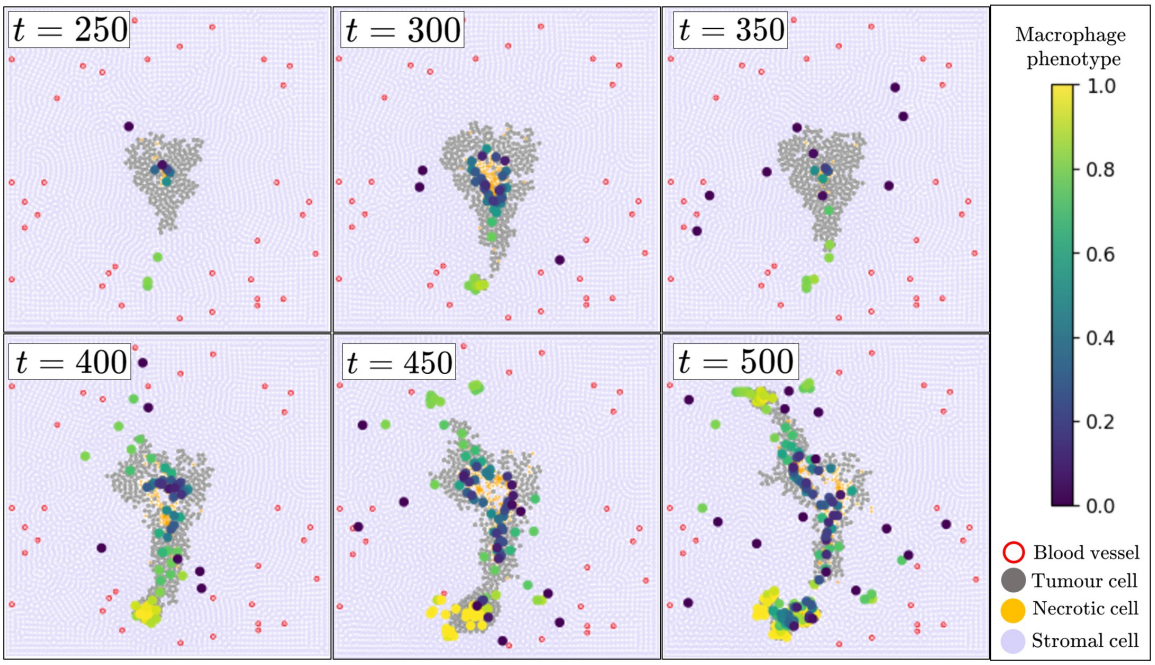


Figure 3.1: Example simulation from the agent-based model. We see how the spatial patterns of the different cell types change over times 250 - 500 hours.

In our analysis, we consider data generated generated from this ABM. We obtain spatial data for the different parameters that govern the equations at times 250, 300, 350, 400, 450 and 500 hours. For each time step we obtain the Cartesian coordinates of our different agents (and blood vessels), and also the phenotype of macrophages. We make use of such data generated from a 2- and a 6-parameter sweep[1] of the governing parameters[2]. In our analysis, we do not consider the exact parameters used to form these simulations, but rather attempt to classify based on spatial patterns alone. Figure 3.1 shows an example of a simulation based on a specific set of parameters. In this example we see a range of different phenotypes of macrophages, where the lower phenotype, anti-tumour macrophages are killing the tumour, and the higher phenotype, pro-tumour macrophages are attracting the tumour cells towards blood vessels, promoting metastasis. It is important to note that in our analysis, we threshold the phenotype; macrophages with phenotype ≤ 0.5 are considered anti-tumour, and those with phenotype > 0.5 are pro-tumour. Figure 3.2 shows some key spatial patterns that arise in both macrophages and tumour cells. We see some varying behaviours,

[1]Meaning that in the former we vary 2 parameters and in the latter we vary 6-parameters, and for each new set of parameters we obtain our dyanmic spatial data.

[2]Examples of these parameters include sensitivities of macrophages and tumour cells to the diffusible species, and also other metrics such as the number of blood vessels.

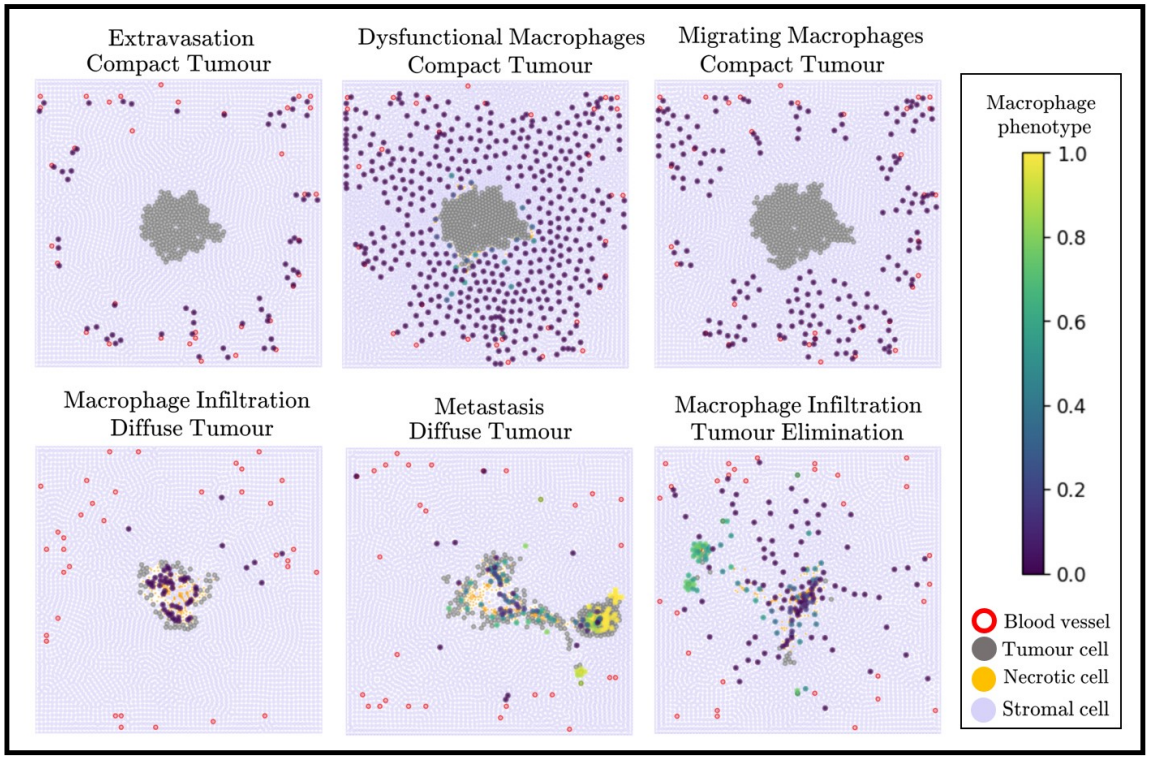


Figure 3.2: Examples of different spatial patterns that arise in the ABM simulations. We see different spatial patterns of macrophages and tumour cells, which are highly dependent on macrophage phenotype.

and we shall constantly refer back to these when analysing our results in Chapter 5. In particular we see simulations in which the tumour is compact, and those in which the tumour becomes more diffuse or even eliminated. We see early time simulations where we have extravasation of anti-tumour macrophages from blood vessels, and intermediete times where anti-tumour macrophages migrate towards increased CSF-1 concentrations induced by tumour cells. We also see the special case of dysfunctional macrophages, where anti-tumour macrophages surround the tumour but do not infiltrate. In the next Chapter we shall introduce the methods we use to analyse the point clouds formed by the generated Cartesian coordinates of different agents.

Chapter 4

Methods

In this Chapter we introduce existing and novel methods that we developed to analyse ABM point cloud data described in Chapter 3. We present three different filtrations that we build on given point clouds and an interpretation of the resulting barcodes. The first method is the Vietoris-Rips filtration, which is a popular tool in topological data analysis [13, 27]. We then introduce the Dowker filtration , which captures the homology of relations between distinct types of points in a given point cloud data set [9, 7, 26]. The Dowker filtration has been introduced to the research community relatively recently [7, 26], and this thesis provides its first application in image analysis. The Vietoris-Rips and Dowker and filtrations can capture topological features of a single time frame in our data set. To this end, we also discuss a construction that incorporates the time variable and one that is a novel methodological contribution of this dissertation. For all of these methods we describe how to obtain their corresponding persistence modules, and thus their persistence images. We conclude this Chapter by explaining how we use machine learning techniques to analyse the vectorised topological data.

4.1 Method 1: Vietoris-Rips Filtration

The Vietoris-Rips filtration is widely used to analyse topological features of point cloud data [13, 27]. Suppose we are given a point cloud $\mathcal{P} \in \mathbb{R}^2$. For every vertex $p \in \mathcal{P}$, we consider the ϵ-neighbourhood, $\mathcal{N}_\epsilon(p)$ centred at that vertex for varying parameter $\epsilon \geq 0$. As we increase ϵ, more and more of the neighbourhoods intersect. The Vietoris-Rips filtration at parameter ϵ is built by including a k-simplex $\{p_{i_0}, \ldots, p_{i_k}\}$ wherever there are $k+1$ vertices $p_{i_0}, \ldots, p_{i_k}$ whose respective neighbourhoods have

non-empty pairwise intersections. For example when the ϵ-neighbourhoods of two vertices intersect, a 1-simplex (an edge) is included between those points. Similarly, when the ϵ-neighbourhoods of three vertices form three non-empty pairwise intersections, we include not only the edges between them, but also the 2-simplex (triangle) which would occupy the void. Such a construction clearly includes all faces of simplices in the simplicial complex, as required by the definition. As will be the case for all the filtrations we discuss, the interpretation is perhaps best gained from a diagram.

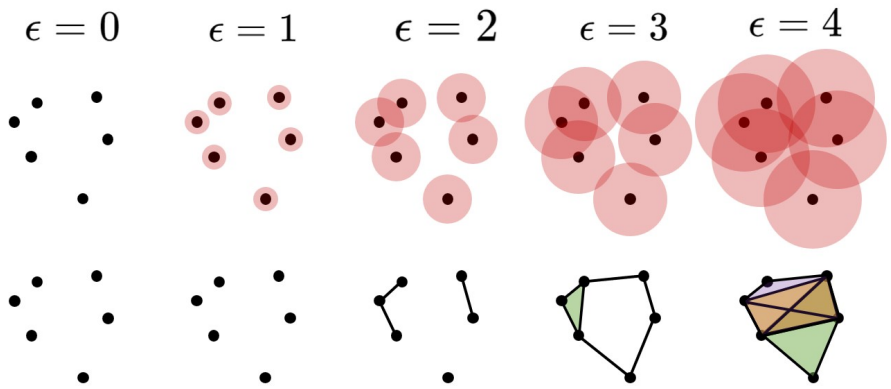


Figure 4.1: Schematic of a Vietoris-Rips filtration built on a point cloud. We see the neighbourhoods for increasing ϵ (top), and the corresponding simplicial complex in the filtration (bottom).

Fix a point cloud $\mathcal{P}$ with elements in $\mathbb{R}^2$. Let the simplicial complex formed at parameter ϵ be denoted $\mathcal{R}_\epsilon$. Formally, we have

$$\emptyset \hookrightarrow \mathcal{R}_{\epsilon_1} \hookrightarrow \cdots \hookrightarrow \mathcal{R}_{\epsilon_N} \qquad \text{(Vietoris-Rips filtration)}$$

$$\overset{\text{apply homology}}{\Longrightarrow} \quad H_k[\emptyset] \longrightarrow H_k[\mathcal{R}_{\epsilon_1}] \longrightarrow \cdots \longrightarrow H_k[\mathcal{R}_{\epsilon_N}] \qquad \text{(Persistence module)}$$

where $\epsilon_1 < \ldots < \epsilon_N$. Given our persistence module of vector spaces and linear maps, we can apply the Structure Theorem (Theorem 2.1) to obtain our interval module decomposition and the corresponding persistence diagram representation. We then apply the pipeline as described in Chapter 2 to obtain our persistence image. In practical computations we terminate at some sufficiently large ϵ_N, where by sufficient we mean that no more topological features of interest will be generated by further increasing the parameter. This can usually be determined *a priori* since we will know the positions of the vertices of the initial point cloud. In our analysis we use the `Julia` based packages `Eirene` [16] and `Ripser` [28] to compute the Vietoris-Rips filtration.

4.2 Method 2: Dowker Filtration

In this section we describe the Dowker filtration, and motivate its interpretation and relevance for our data. Unlike the Vietoris-Rips filtration (which is defined on a single point cloud), the Dowker filtration is constructed upon two point clouds $\mathcal{P}_1$ and $\mathcal{P}_2$ whose relational homology is of interest.

4.2.1 Relations

We define a relationship [14, 9] $\mathcal{R}$ to be a subset of the space of tuples of vertices in these point clouds, that is $\mathcal{R} \subset \mathcal{P}_1 \times \mathcal{P}_2$. We say an element $p_i \in \mathcal{P}_1$ is related to an element $q_j \in \mathcal{P}_2$ if and only if they are within a distance ϵ of each other [14]. Since we are working with point clouds in $\mathbb{R}^2$ we use the Euclidean distance. Here ϵ is once again the parameter which will eventually determine elements in our filtration. For now, however, we focus on the construction of a single simplicial complex. The first step to building the simplicial complex is to construct a binary matrix which represents our relations.

Suppose $\mathcal{P}_1$ has m elements and $\mathcal{P}_2$ has n elements. The binary matrix that represents the relation $\mathcal{R}$ is given by the $m \times n$ matrix

$$R_{ij} = \begin{cases} 1 & d(p_i, q_j) \leq \epsilon \\ 0 & d(p_i, q_j) > \epsilon. \end{cases}$$

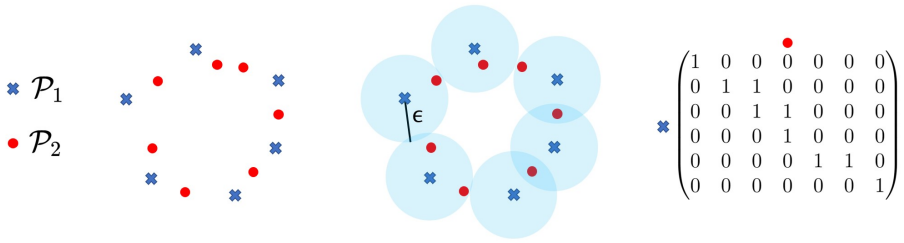


Figure 4.2: Example of two point clouds and the corresponding relation matrix formed by thresholding the distance matrix the some parameter ϵ.

In practice this matrix can be computed by constructing a distance matrix between the two point clouds and thresholding by the value ϵ [14, 26]. Thus if the distance between two points is less than or equal to this threshold we set the entry to 1,

otherwise we set it to 0. We also note that the elements of the point cloud from which we compute distances (corresponding to the rows of the matrix) are called 'landmarks', and the elements in the other point cloud are called 'witnesses'.

4.2.2 Dowker Complex

The Dowker simplicial complex formed at threshold ϵ, $\mathcal{D}_\epsilon(\mathcal{P}_1, \mathcal{P}_2)$, is constructed from the binary matrix we obtain from our relation. However, there are two options for its construction. The first is to consider the row major ordering of the matrix. We let each column in the matrix represent a vertex in our simplicial complex. We then consider the entries in each row to determine the simplices included between these vertices. For example if a row had $k+1$ non zero entries in columns representing vertices $v_0 \dots v_k$, then we construct the corresponding k-simplex between these vertices in the simplicial complex, and also every corresponding face which is defined by subsets of non zero entries[1]. The second method considers the column major ordering of the matrix and follows similarly by transposing the matrix and computing the matrix as for the row major method. Dowker's theorem says that these methods yield equivalent homology groups [7, 9], and can be stated as follows.

Theorem 4.1 (Dowker's Theorem) *Let D_ϵ^{row} and D_ϵ^{col} denote the two Dowker complexes obtained from a relation. The following statement holds for all $k \geq 0$.*

$$H_k\left[D_\epsilon^{row}\right] \cong H_k\left[D_\epsilon^{col}\right] \tag{4.1}$$

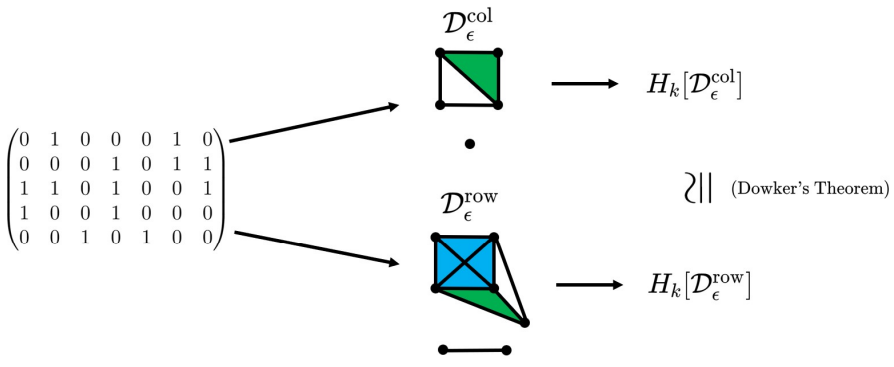


Figure 4.3: An example of the two Dowker simplicial complexes generated by some binary distance matrix. We see that constructing the complex in either way yields equivalent homology groups.

[1]Thus satisfying the closure under subsets property of simplicial complexes.

A consequence of this theorem is that in our computations, the choice of landmarks and witnesses between $\mathcal{P}_1$ and $\mathcal{P}_2$ is, in theory, arbitrary. In a practical setting, we take the smaller point cloud to represent the rows of our relation matrix, as this leads to faster computations.

4.2.3 Dowker Filtration

The Dowker filtration is found by constructing Dowker complexes for an increasing sequence of parameters $\epsilon_1 < \epsilon_2 < \ldots < \epsilon_N$. Formally we have

$$\emptyset \hookrightarrow \mathcal{D}_{\epsilon_1}(\mathcal{P}_1, \mathcal{P}_2) \hookrightarrow \cdots \hookrightarrow \mathcal{D}_{\epsilon_N}(\mathcal{P}_1, \mathcal{P}_2) \qquad \text{(Dowker filtration)}$$

$$\overset{\text{apply homology}}{\Longrightarrow} \quad H_k[\emptyset] \to H_k[\mathcal{D}_{\epsilon_1}(\mathcal{P}_1, \mathcal{P}_2)] \to \cdots \to H_k[\mathcal{D}_{\epsilon_N}(\mathcal{P}_1, \mathcal{P}_2)] \qquad \text{(Persistence module)}$$

Given our persistence module of vector spaces and linear maps, we apply the Structure Theorem (Theorem 2.1) to obtain our interval module decomposition and, thus, our corresponding Dowker persistence diagram. As before we can then compute the corresponding persistence image. This method captures those topological features of the two respective point clouds that are similar. For example, if the two point clouds of interest had persistent loops, we would expect our Dowker filtration also to have a persistent loop. However, this would not be the case if only one of the point clouds had a persistent loop. This type of filtration is relevant to our ABM data set since the behaviour of the agents dependent on each other. Therefore studying the relative topological features of their point clouds may yield increased insight into the interaction and, thus, improve classification accuracy. In our analysis we use the `Julia` based code developed by Yoon, Ghrist and Giusti [26] to compute the Dowker filtration.

4.3 Method 3: Dowker-time Filtration

We present a novel contribution of this thesis, which is an extension of the Dowker filtration to encode dynamic spatial distributions. Let $\mathcal{P}_t$ denote a given point cloud at time t. We consider the distance relationship formed by $\mathcal{P}_{t_i}$ and $\mathcal{P}_{t_j}$ with $j \geq i$. Figure 4.4 shows the choices of landmarks and witnesses in this construction. Then,

we obtain the following filtration and persistence module.

$$\emptyset \hookrightarrow \mathcal{D}_{\epsilon_1}\left(\mathcal{P}_{t_i}, \mathcal{P}_{t_j}\right) \hookrightarrow \cdots \hookrightarrow \mathcal{D}_{\epsilon_N}\left(\mathcal{P}_{t_i}, \mathcal{P}_{t_j}\right) \qquad \text{(Dowker-time filtration)}$$

$$\overset{\text{apply homology}}{\Longrightarrow} \quad H_k\left[\emptyset\right] \to H_k\left[\mathcal{D}_{\epsilon_1}\left(\mathcal{P}_{t_i}, \mathcal{P}_{t_j}\right)\right] \to \cdots \to H_k\left[\mathcal{D}_{\epsilon_N}\left(\mathcal{P}_{t_i}, \mathcal{P}_{t_j}\right)\right] \qquad \text{(Persistence module)}$$

We apply the Structure Theorem (Theorem 2.1) to obtain the interval module decomposition and corresponding Dowker-time persistence diagram. Applying the pipeline from Chapter 2 then yields the persistence image. Since Dowker filtrations capture the persistent topological features that are common to the two point clouds, this filtration quantifies the relative change in persistent features over time.

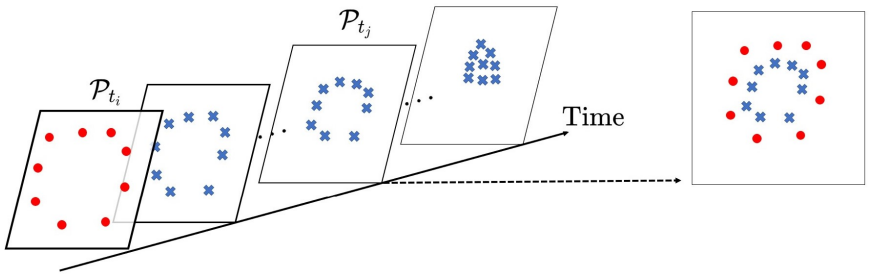


Figure 4.4: Example showing how we select the landmarks and witnesses in the Dowker-time filtration. We see dynamic spatial data, along with the projection of these into a single plane. In the Dowker-time filtration we consider the persistent homology of the Dowker filtration formed by these point clouds.

This filtration is applicable for our data set since we consider the dynamics of the agents involved in the tumour microenvironment. As described in Chapter 3, the behaviour of agents, specifically macrophages, are highly dependent on concentrations of diffusible species in the microenvironment. Recall that concentrations of CSF-1 and CXCL12 bias the movement of the macrophages towards and away from the tumour respectively. As such, knowing how topological features change over two timepoints may provide further insight into the concentrations of anti- and pro-tumour macrophages. In our analysis we adapt the code used for standard Dowker filtrations [26] to incorporate our time-dependent filtration.

4.4 Machine Learning Methods

Since we can represent the persistent homology of all of our point cloud filtrations as vectorised persistence images, we have vectors that describe the topological features

of any given point cloud. When we consider point cloud data from the ABM, we shall compute persistence images for different filtrations and different cell types. Given an ensemble of persistence images, we can perform clustering and classification. Here we briefly describe the methods that we use and the rationale behind them.

4.4.1 Multidimensional Scaling

Multidimensional scaling (MDS) is a technique used to reduce the dimensionality of data whilst preserving distances between data points [3, 15]. In our case, the data points are persistence images, and the technique projects the high dimensional vectorised representations onto two dimensions, where pairwise Euclidean distances between vectors are preserved. This method allows us to visualise the dissimilarity between persistence images, and, thus, the dissimilarity between the topological features of different ABM simulations. We stress that this technique is used solely to visualise the topological features. For classification, we use the entire feature vectors, not their two dimensional projections. To perform MDS, we make use of the `Julia` based library `MultivariateStats`, see [20].

4.4.2 Support-Vector Machine

Our first classification task will be to distinguish simulations with a majority of pro-tumour macrophages from those with a majority of anti-tumour macrophages. This is a binary classification problem and, thus, well suited to the application of support-vector machines (SVMs). An SVM finds the separating hyperplane which maximises the distance between both classes and itself [17, 5]. We construct the SVM from a randomly selected sample of our (high dimensional) topological feature vectors for training and then compute its accuracy by testing on the remainder of the data. We shall apply this procedure to each filtration and compare their accuracies. This will allow us to determine which cell types and which filtrations yield the best classification of pro-tumour rich environments given only the spatial distributions of the involved cells. To compute our SVMs, we make use of the `Julia` API of the `C++` library `LIBSVM`, see [5].

4.4.3 Multilayer Perceptron

We also consider a regression problem, where we predict the concentration of pro-tumour macrophages from earlier times. A Feedforward artificial neural network is

an appropriate tool when trying to predict an output (final pro-tumour macrophage concentration) given an input (topological features at an earlier time point) when there is no clear analytic relationship between them [15]. As such we make use of these, and in particular we use a Multilayer perceptron (MLP). MLPs (and feedforward neural networks in general) make use of the backpropogation algorithm. We omit details regarding this algorithm, however there is a vast amount of literature regarding this. We refer the reader to [15] for more details. We also perform hyperparameter optimisation using a randomised grid search which includes searching for the optimal activation functions, hidden layer sizes and solvers, amongst other standard parameters. To construct MLPs, we use the `Julia` API of the `Python` based library `ScikitLearn`, see [22].

Chapter 5

Results

In this Chapter we apply the methods from Chapter 4 to our synthetic data set. We obtain dynamic 2-dimensional data from the ABM at times $t = 250$, 300, 350, 400, 450, and 500 hours for each set of parameters, as described in Chapter 3. In our discussions we only consider what the spatial patterns arising in different cell types and combinations of cell types can infer about the concentration of pro-tumour macrophages in the simulation, and thus we do not take into account the specific parameters[1] giving rise to the patterns. We first consider, in sections 5.1 and 5.2 the classification task of determining whether or not a given simulation has a majority of pro-tumour macrophages. We then consider, in section 5.3 the regression task of predicting the final pro-tumour macrophage concentration from earlier time points. We note that in the binary classification task we do not consider the time variable, but rather view all simulations across all time points as an ensemble and perform clustering on this. Let $\mathcal{T}$, $\mathcal{M}$, $\mathcal{N}$ and $\mathcal{V}$ denote the point clouds representing tumour cells, macrophages, necrotic cells, and vessels of an arbitrary simulation respectively. Throughout this section we shall apply our methods on these point clouds to generate persistence images. Figure 5.1 shows the output from a typical simulation together with corresponding point clouds which we shall analyse.

[1] See [4] for details regarding the specific parameters.

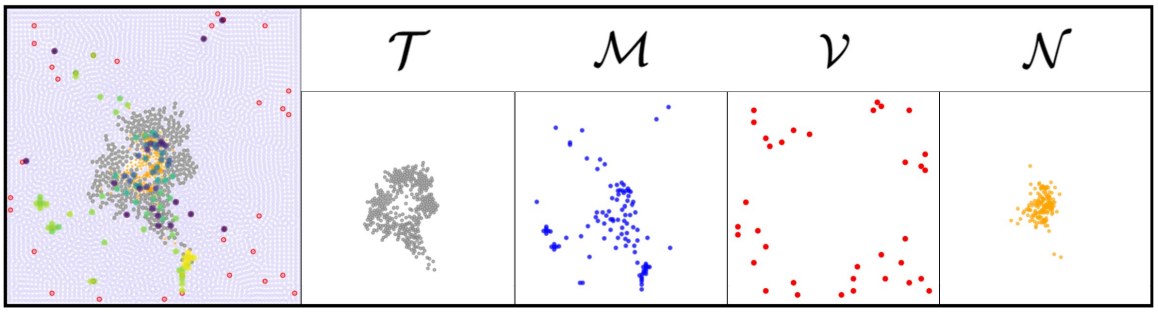


Figure 5.1: An example simulation, along with its constituent tumour, macrophage, vessel and necrotic cell point clouds. In our analysis we consider only the spatial patterns of macrophages and so in our macrophage point cloud we do not include phenotypes.

5.1 Binary Classification

Our first task is to classify, from an ensemble of simulations such as the one shown in figure 5.1, which simulations have a majority or pro tumour macrophages. Consider an arbitrary simulation S. Let M_1 denote the number of macrophages with an anti-tumour phenotype, that is $0 \leq p < 0.5$ and let M_2 denote the number of macrophages with a pro-tumour phenotype, $0.5 \leq p \leq 1$. We apply binary labels

$$\text{Label}(S) = \begin{cases} 1 & \frac{M_2}{M_1+M_2} \geq \frac{1}{2} \\ 0 & \frac{M_2}{M_1+M_2} < \frac{1}{2} \end{cases} \tag{5.1}$$

to our simulations and then classify these binary outcomes based on the topological features derived by our various methods. For each simulation, we compute a filtration on the different point clouds. This yields a barcode, and we then apply the pipeline highlighted in Chapter 2 to obtain persistence image representations. We use persistence images of size[2] 20×20, with variance $\sigma^2 = 1$ and the standard weighting function, as described in Chapter 2. We consider only H_0 and H_1 in our analysis, namely we consider the persistence of connected components and loops in our filtrations on the point cloud data.

We structure this section in the following way. For each method, we recall their key properties and describe how they are applied to our simulations. We then show the multi-dimensional scaling (MDS) results; the 2-dimensional visualisation of the persistence images of the ensemble of simulations such that distances between persistence images are preserved. The MDS results are coloured by 'red' simulations

[2]We consider 20×20 persistence images in $\mathbb{R}^2$, and thus dimension 400 vectors in $\mathbb{R}$.

which are labelled '1' and 'blue' simulations which are labelled '0'. We then consider the strengths and weaknesses of the clustering results, and finally elaborate further on the clustering by showing results for individual simulations. We compare these simulations to understand and interpret results. We note here that in the next section we shall compare the classification accuracies for each method, with a benchmark set by classifying the data using simpler features.

5.1.1 Vietoris-Rips Filtration

We first apply the Vietoris-Rips filtration to our simulations, focusing on macrophage and tumour cell point clouds. Recall that the Rips filtration is computed by forming simplicial complexes based on the pairwise intersections of ϵ-neighbourhoods of points in the point cloud. In the Rips filtration, connected components are always born at $\epsilon = 0$, and, thus, the H_0 persistence images only have non-zero intensities in the first column. We therefore show these as 20×1 vectors.

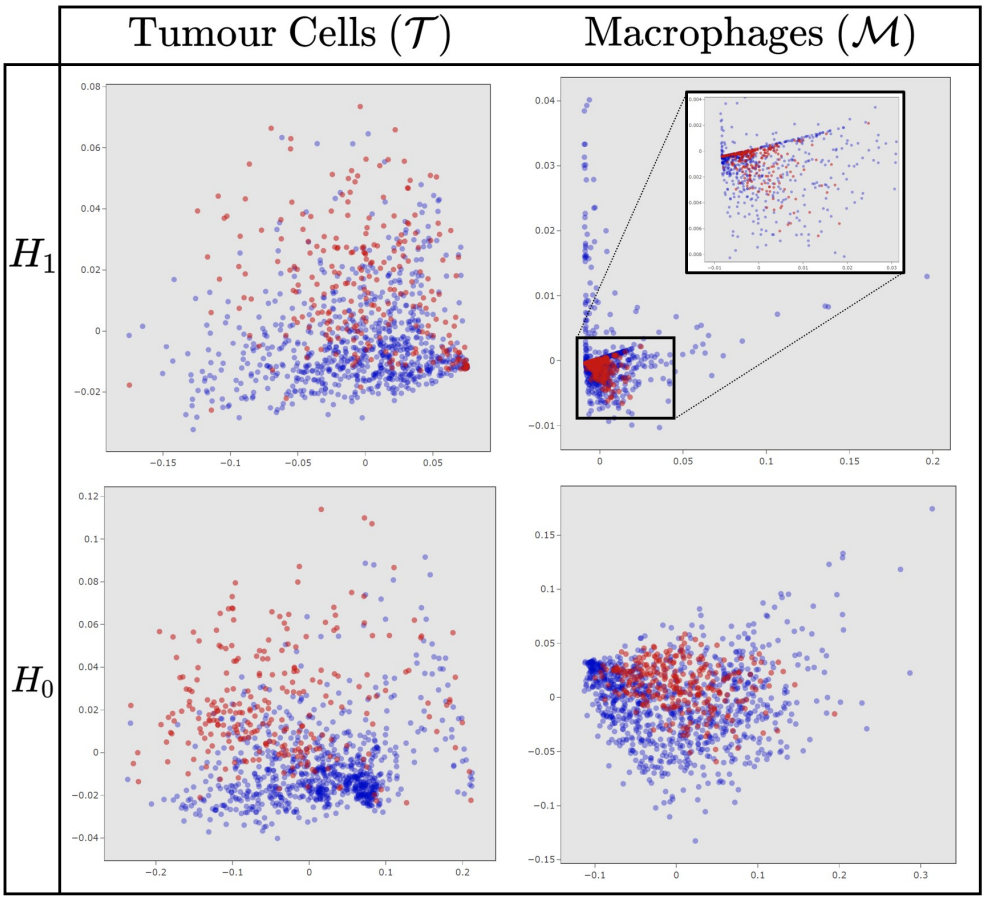


Figure 5.2: Two-dimensional projection of our Vietoris-Rips feature vectors. Simulations where pro-tumour macrophages are in excess are coloured red and simulations where anti-tumour macrophages are in excess are coloured blue. We see clustering of H_1 (top row) and H_0 (bottom row) persistence images, for both tumour cell point clouds (left column) and macrophage point clouds (right column). The axes are the principal coordinates obtained from the multidimensional scaling.

Figure 5.2 shows the projections of our persistence images after applying MDS. We see the 2-dimensional projection of our 400-dimensional persistence images, with distances between persistence images preserved.

5.1.1.1 Vietoris-Rips Filtration on Tumour Cells

We first consider the clustering of the persistence of H_0 features for tumour cells. The projection shows a slight separation between the two classes. For tumour cells, there are two main outcomes for which we topological features differ significantly. Firstly in the case of a compact tumour, the tumour cells are densely packed in the domain, and,

thus, when growing ϵ-neighbourhoods around them, we expect intersections between the neighbourhoods to appear at low values of ϵ. Therefore, within a very small range of the parameter, we shall see the birth of all connected components and their subsequent death due to components quickly joining together. In contrast, we may have a diffuse tumour, in which larger values of ϵ are needed to form a significant number of intersections and, thus, to cause the death of connected components.

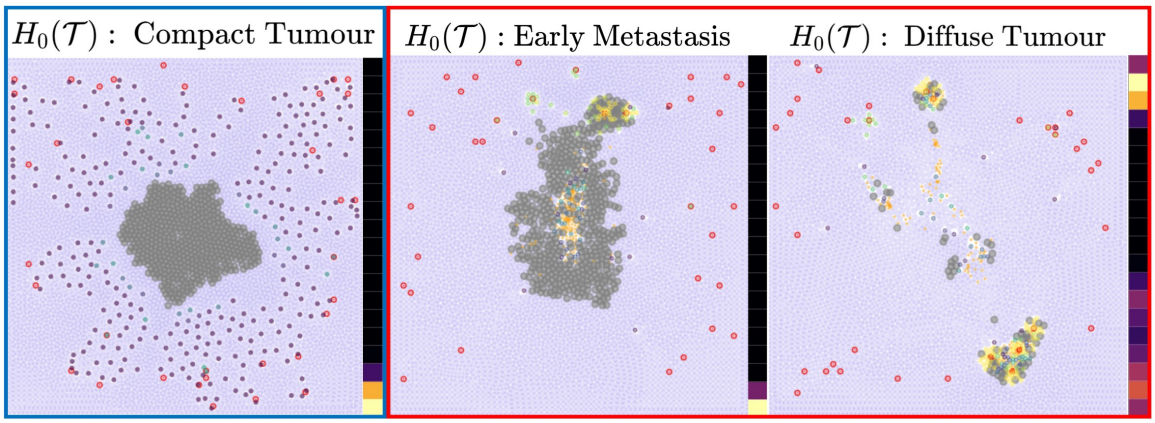


Figure 5.3: Examples of simulations along with their persistence images representing the H_0 features of the tumour point clouds. We see an example of a simulation labelled 0 (left) and two simulations labelled 1 (right). Since H_0 persistence images of the Rips filtration only have non-zero intensities at $\epsilon = 0$, we show only the first column of the persistence image.

The resulting persistence images represent compact tumours by high intensities at low persistence (i.e. lower on the y-axis), and very sparse/diffuse tumours with high intensities at high persistence. Further, compact tumours almost always occur in the presence of a significant number of anti-tumour macrophages. This could be due to the simulation showing the intermediate stages between extravasation, and the subsequent migration of low phenotype macrophages with the CSF-1 gradient towards the tumour cells, and thus we have an in-tact dense tumour. It may also be due to simulations involving dysfunctional macrophages, in which anti-tumour macrophages surround the tumour with very little infiltration. In the case of an environment rich in pro-tumour macrophages, we almost always see dispersion of tumour cells in response to the high levels of EGF expressed by the macrophages which migrate along the CXCL12 gradient towards blood vessels. With these different spatial patterns arising from simulations that are pro- and anti-tumour rich environments, one might expect exceptional results from the clustering of H_0 persistence images of tumour point

clouds. However, these only apply in idealised simulations. During the migration of the tumour cells, we may see equivalent tumour spatial patterns to those that arise when there is a majority of anti-tumour macrophages[3]. Furthermore, even in the case of metastasis, the tumour may remain compact at earlier times, resulting in a similar persistence image to that of a fully compact tumour. Such behaviours can lead to some erroneous results in our classification, which we shall see in the next section.

We next consider the persistence of H_1 features for tumour cells. This ultimately leads to a similar classification to the case of H_0 (which we shall see in the next section) but for a different reason. In a considerable number of cases, there are very few significant loops formed by the point cloud of tumour cells. We therefore have persistence images containing many noisy loops, in particular high intensities at early birth and low persistence. A consolation to this is the fact that significant loops are formed almost exclusively in cases where macrophages infiltrate the tumour, causing inner tumour cells to become necrotic. These simulations are rich in anti-tumour macrophages.

[3]For example, some instances of tumour infiltration yield similar spatial patterns of the tumour, and, thus, similar persistence images.

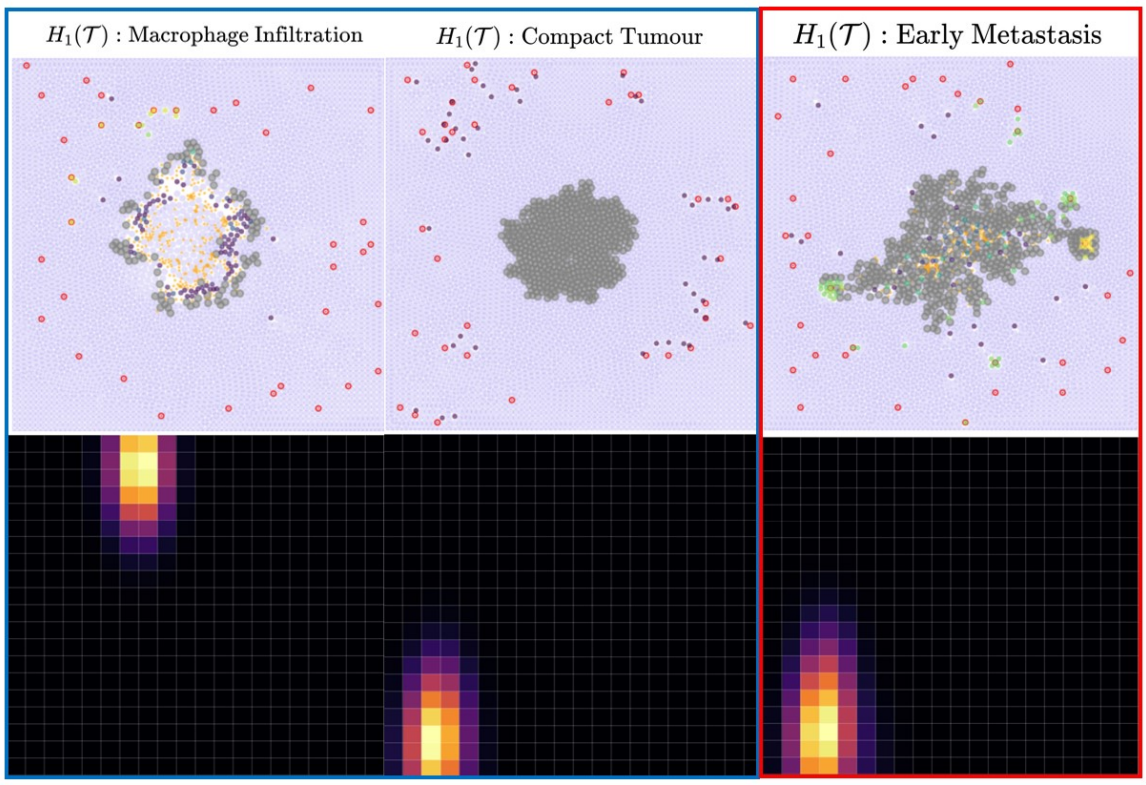


Figure 5.4: Examples of simulations and their persistence images representing the H_1 features of the tumour point clouds. We see examples of two simulations labelled 0 (left and middle), recall that these are rich in anti-tumour macrophages. We also see one simulation labelled 1 (right), recall that this is rich in pro-tumour macrophages.

Figure 5.4 shows typical H_1 persistence images corresponding to different spatial distributions of tumour cells. In the case of macrophage infiltration and tumour elimination, a significant loop forms at low ϵ. We also see that in cases where the tumour is compact (rich in anti-tumour macrophages) and during early metastasis (rich in pro-tumour macrophages), persistence images are similar, leading to some missclassification. We note that there are still slight differences in these cases (for example the height difference in the latter two simulations). The early metastasis persistence image has higher persistence noisy loops, and since we cluster over a large number of simulations, even such minute differences are magnified, thus leading to an improved separation of the classes.

5.1.1.2 Vietoris-Rips Filtration on Macrophages

We now consider the topological features formed by the Rips filtration on our macrophage point clouds. At early stages in the simulations, where we see extravasation of macro-phages from blood vessels, we expect relatively sparse distributions of macrophages, resulting in a persistence image showing very persistent (dis)connected components. In contrast, high pro-tumour macrophage concentrations usually correspond to more dense distributions of macrophages. This is due to the clustering around vessels for simulations in the later stages of metastasis. Greater densities are also seen in the earlier stages of metastasis. This is because when macrophages change phenotype in response to TGF-β, they migrate towards the higher concentrations of CXCL12 near blood vessels. We therefore have some macrophages migrating towards the tumour and some towards the blood vessels, thus, the distribution of macrophages becomes more dense. These arrangements of macrophages over the course of metastasis lead to very similar persistence images, with high intensities at low persistence. The differences in persistence images in the cases of metastasis and the migration of anti-tumour macrophages towards the tumour result in good classification. However, in some extreme cases of tumour elimination, a significant number of macrophages are scattered across the domain, with particularly high concentrations at locations previously occupied by the tumour. Such simulations also give rise to connected components with very low persistence due to the large number of macrophages, as seen in figure 5.5.

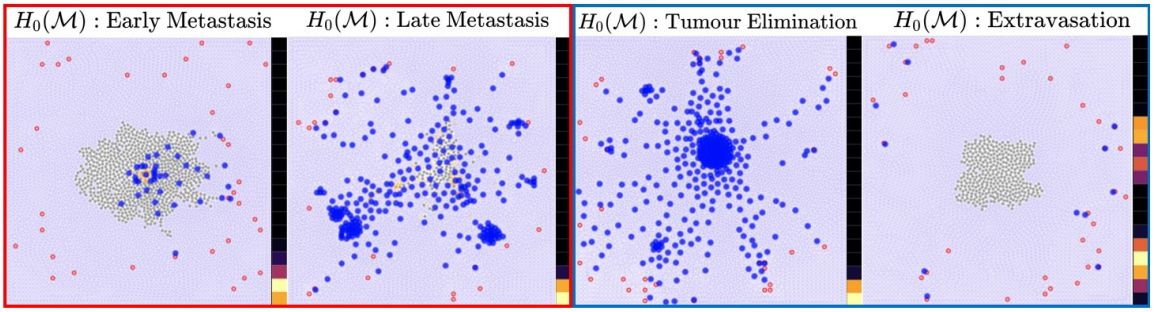


Figure 5.5: Examples of simulations and their persistence images summarising the H_0 features of the macrophage point clouds. We see examples of two simulations which are rich in pro-tumour macrophages, and thus labelled 1 (left) and two simulations which are rich in anti-tumour macrophages, labelled 0 (right).

The clustering of the H_1 feature vectors for the macrophage point clouds give interesting results. Firstly, this method is successful in distinguishing, in most cases,

between environments which are anti- and pro-tumour macrophage rich, due to the significant difference in direction of movement between macrophages of differing phenotype.

As discussed in Chapter 3, the forces on macrophages are highly dependent on their sensitivities to CSF1 and CXCL12, which are both dependent on phenotype. Since, at early times, macrophages have anti-tumour phenotypes, we see macrophages migrating radially inwards towards the tumour. Such simulations give rise to persistent loops. In contrast, when macrophages start exhibiting pro-tumour tendencies due to the high concentrations of TGF-β near tumour cells, we see much more disorderly movement and, thus, the loops formed become less and less significant. Such behaviours give rise to persistence images with high persistence for environments rich in anti-tumour macrophages, and lower persistence for environments rich in pro-tumour macrophages. This method's ability to distinguish between these behaviours is beneficial for our classification task (which we shall see in the next section), however the clustering also reveals more information about simulations.

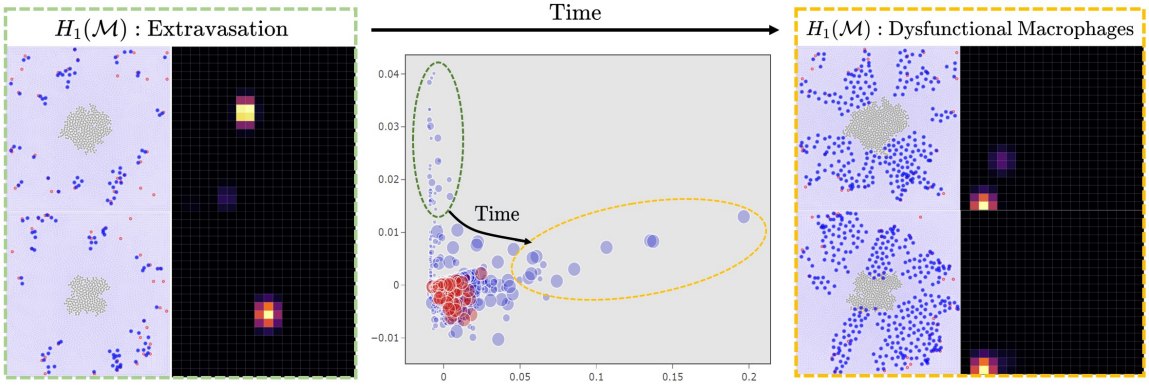


Figure 5.6: We see our MDS visualisation of our feature vectors formed by the H_1 persistence images of macrophages. We see marker sizes representing the time steps of the individual simulations, with smaller markers representing earlier times. We also see examples of simulations corresponding to each arm, with the left simulations showing early time behaviour and the right simulations showing late time behaviour.

In our multidimensional scaling we see two 'arms' in which simulations are particularly different to other simulations[4]. We can understand why these arms are formed by including our time variable. In figure 5.6 we present the MDS results again, but with marker sizes indicating the simulation time at which each synthetic image was

[4]In the context of the MDS, simulations are particularly different when they are further away from each other, since this dimensionality reduction preserves distances between persistence images.

generated. In most cases, simulations in the upper arm are from the earlier time points of those in the right arm. We also note that in all cases for which we can distinguish between early and late times between the arms, we see the final time simulations representing dysfunctional macrophages[5]. Having such a clear distinction between early and late times between these simulations suggests that we could predict such outcomes from earlier times.

5.1.2 Dowker Filtration

We next apply the Dowker filtrations to our simulated data. We consider combinations of macrophages, tumour cells, blood vessels and necrotic cells. Recall that the Dowker filtration considers one point cloud as 'landmarks' and the other as 'witnesses'. It then constructs a distance matrix between landmarks and witnesses and forms a filtration by thresholding the matrix at different values and creating simplices based on the binary entries. We consider the persistence of both H_0 (connected components) and H_1 (loops) features of the resulting Dowker filtrations.

[5]Dysfunctional macrophages are anti-tumour macrophages which surround the tumour however do not infiltrate or destroy it

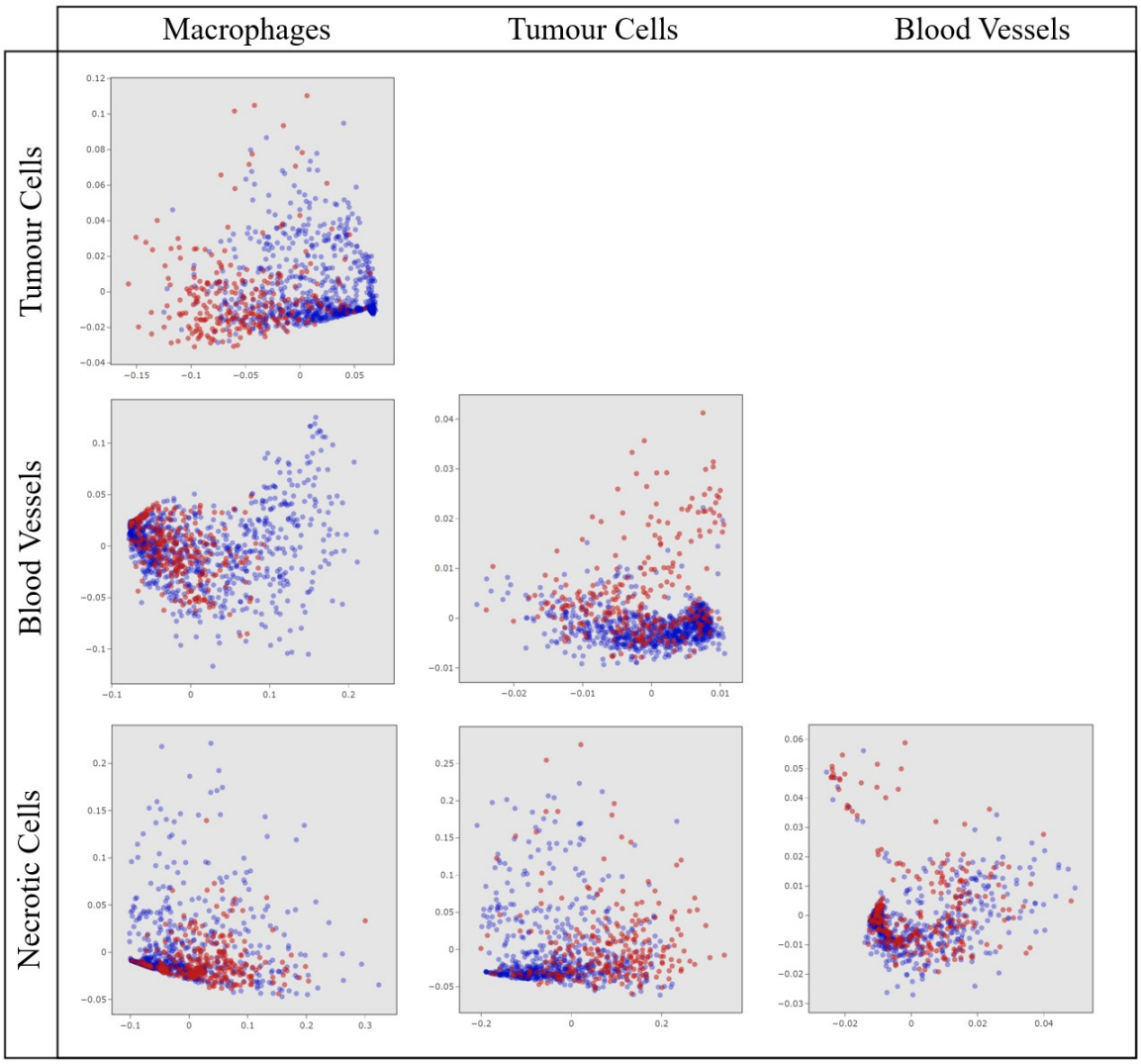


Figure 5.7: Two-dimensional projection of our Dowker feature vectors, coloured by simulations rich in pro-tumour macrophages (red) and those rich in anti-tumour macrophages (blue). We see the clustering of H_0 persistence images of the Dowker filtration of combinations of macrophages, tumour cells, necrotic cells, and blood vessels relative to each other.

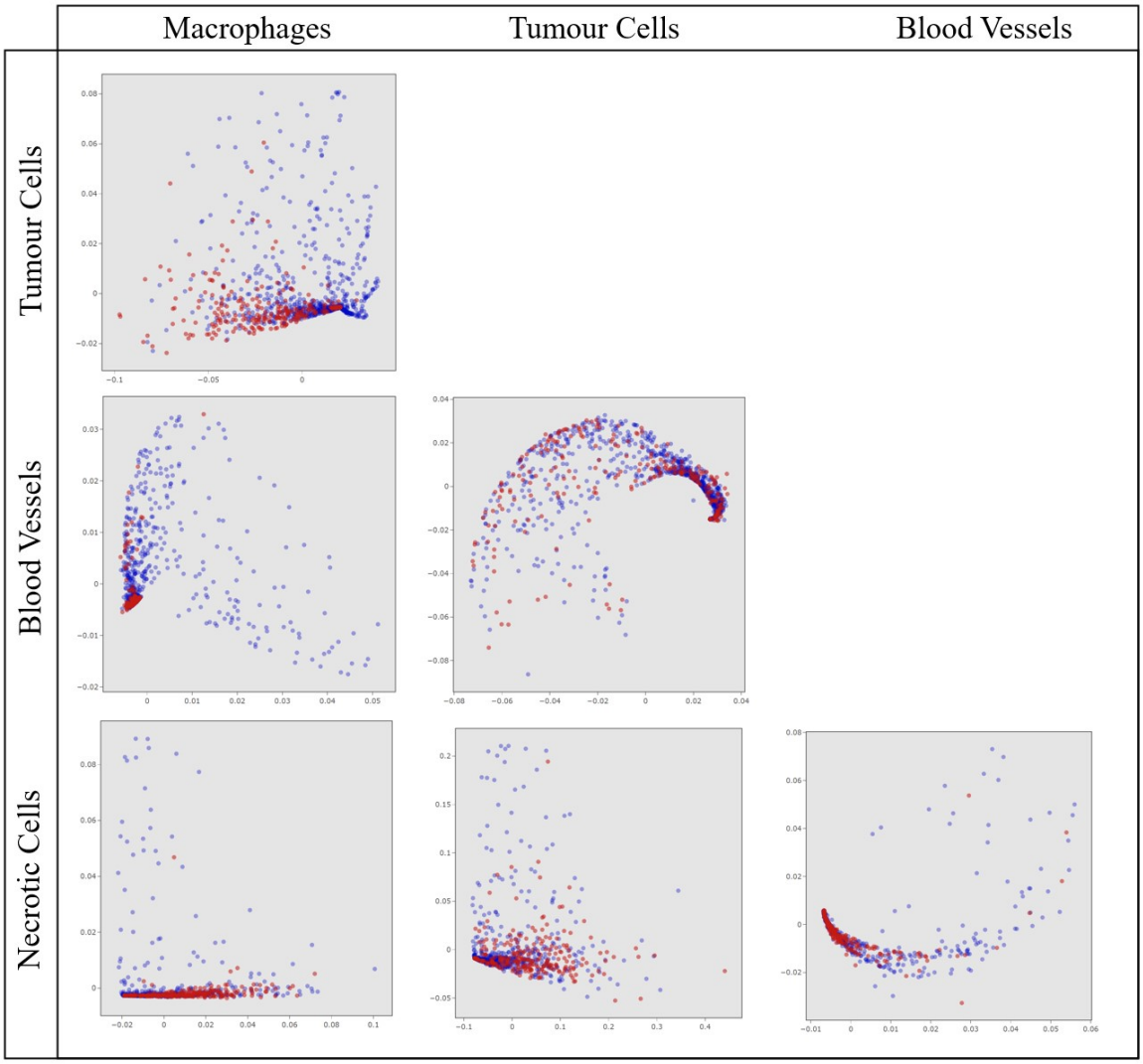


Figure 5.8: Two-dimensional projection of our Dowker feature vectors, coloured by simulations which are rich in pro-tumour macrophages (red) and those that are rich in anti-tumour macrophages (blue). We see the clustering of H_1 persistence images of the Dowker filtration of combinations of macrophages, tumour cells, necrotic cells, and blood vessels relative to each other.

Figures 5.7 and 5.8 show the clustering results of our H_0 and H_1 Dowker filtrations respectively. We next consider some specific cases where we see good separation between the classes and analyse them further.

5.1.2.1 Dowker Filtration: Macrophages and Tumour cells

We first consider the Dowker filtration of the macrophage point clouds relative to the tumour cell point clouds. These are amongst the best methods for our classification

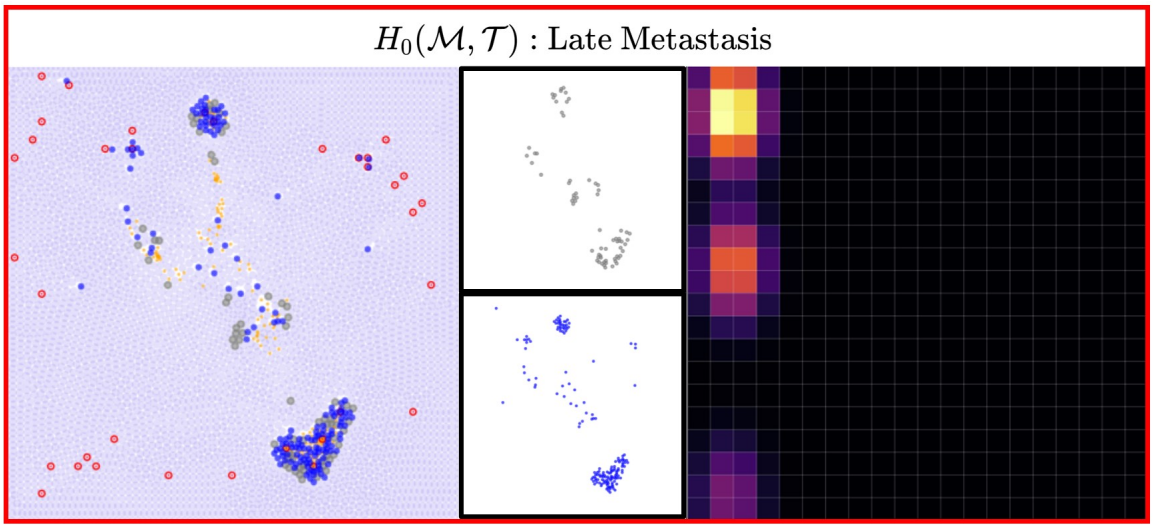


Figure 5.9: Example simulation along with its H_0 persistence image formed by the Dowker filtration of macrophages relative to tumour cells. We also see each individual point cloud. This simulation is labelled 1.

task because macrophages and tumour cells rarely share any common topological features unless there is a significant number of pro-tumour macrophages.

This is because of the increased expression of EGF by macrophages with higher phenotype, which we recall acts as a chemoattractant for tumour cells.

If tumour cells start to migrate towards blood vessels with macrophages, we begin to see features in our Dowker filtration. This is made particularly prominent with our H_0 features. Figure 5.9 shows the similarities between the two point clouds for a simulation which has a majority of pro-tumour macrophages. The H_0 persistence image shows clearly the prominent connected components shared by both point clouds.

5.1.2.2 Dowker Filtration: Tumour cells and Blood Vessels

We next consider the Dowker filtrations formed by our tumour cell and blood vessel point clouds. The H_1 features of the Dowker filtration give very little insight since any significant loops only form when the tumour cells migrate towards the blood vessels. However, in these simulations we rarely see tumour cells clustering at more than two blood vessels. In contrast, the H_0 features of our Dowker filtrations give excellent clustering results (which we shall see, in the next section, leads to a good classification accuracy). This is because tumour cells typically form a compact mass, and do not come into contact with blood vessels when the pro-tumour macrophage

concentration is small. This is highlighted clearly in the persistence images shown in figure 5.10.

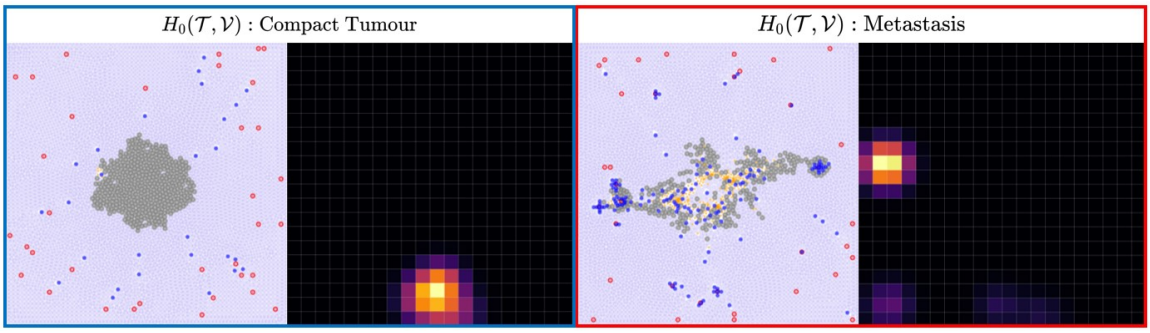


Figure 5.10: Examples of simulations along with their persistence images representing the H_0 features of the tumour cells relative to blood vessels. We see examples of one simulation labelled 0 (left) and one simulations labelled 1 (right).

5.1.2.3 Dowker Filtration: Macrophages and Blood Vessels

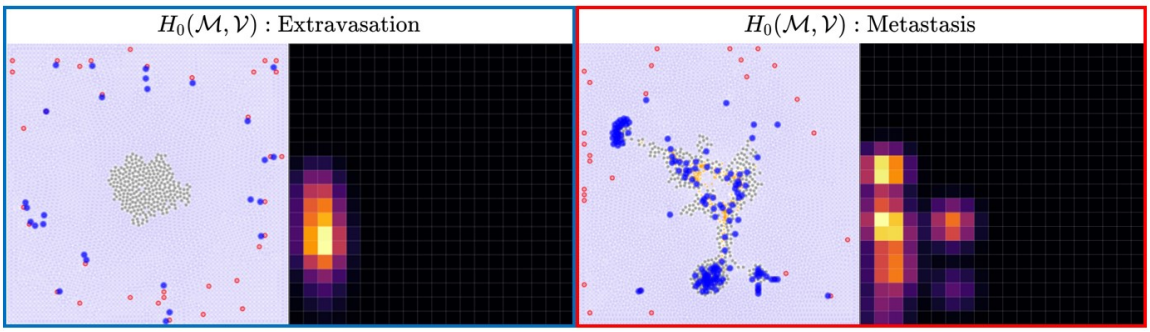


Figure 5.11: Examples of simulations along with their persistence images representing the H_0 features of the macrophages relative to blood vessels. We see examples of one simulations labelled 0 (left) and one simulation labelled 1 (right).

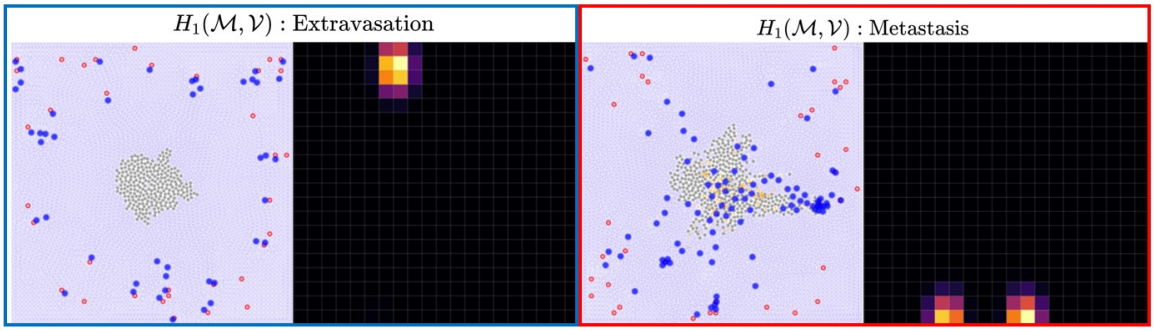


Figure 5.12: Examples of simulations along with their persistence images representing the H_1 features of the macrophages relative to blood vessels. We see examples of one simulations labelled 0 (left) and one simulation labelled 1 (right).

The topological features of the Dowker filtrations of macrophages relative to blood vessels give excellent results. The H_1 features exemplify the differences in the H_1 features of macrophages for our Rips filtrations in the cases of simulations rich in anti- or pro-tumour macrophages. Recall that H_1 features of the Rips filtration of the macrophage point clouds gave a good separation in our clustering, since significant loops only formed in rich anti-tumour macrophage simulations. Considering the relative loops formed by macrophages and blood vessels further exemplifies this since anti-tumour macrophages migrate inwards from the blood vessels from which they were released. We therefore have shared significant loops by blood vessels and macrophages in simulations which are rich in anti-tumour macrophages, and these become more noisy for simulations rich in pro-tumour macrophages. The H_0 features of macrophages relative to blood vessels give a significant separation of the classes (and, as will be shown in the next section, give rise to the best classification results).

It is perhaps surprising, given the persistence images shown in figures 5.11 and 5.12, that these features give better results for our classification than their H_1 counterparts. This is because the persistence image of metastasis shown in figure 5.11 has intensities distributed similarly to the persistence image showing extravasation, whereas these show significantly different persistence images in figure 5.12.

Recall that we expect to see persistence images which represent common H_0 behaviours of both individual point clouds. Note also that blood vessels remain static throughout all simulations. We propose that H_0 features better capture the concentration of pro-tumour macrophages than H_1 features due to their improved ability to detect differences in spatial patterns at edge cases, where $M_2/(M_1 + M_2) \approx 0.5$. This is because at the interface between a majority anti- and majority pro-tumour

macrophage concentration (recall that we will always start with a majority anti-tumour macrophage concentration), we expect a roughly 50% split of macrophages moving inwards towards the tumour due to their increased sensitivity to CSF1 and macrophages moving outwards due to their increased sensitivity to CXCL12. Since H_1 features capture loops that are common between blood vessels and macrophages, at the edge case a significant loop will likely still exist, and thus, we may see missclassification. When considering H_0 features, however, we assess how sparse macrophages are relative to blood vessels. Since the blood vessels are stationary, this is very different at the interface where $M_2/(M_1 + M_2) \approx 0.5$ compared to when $M_2/(M_1 + M_2) < 0.5$, and we therefore expect a better separation of classes in this case.

As a final remark, we note that the homology of necrotic cells relative to the other cells gives little insight into the concentration of pro-tumour macrophages, as seen by the clustering (and, later, classification) results of our filtrations.

5.1.3 Dowker-time Filtration

We finally consider the results of our time dependent filtration for tumour cells and macrophages. Our feature vectors are found by computing

$$\begin{cases} \mathcal{D}\left(\mathcal{P}(t), \mathcal{P}(t)\right) & t = 250 \\ \mathcal{D}\left(\mathcal{P}(t-50), \mathcal{P}(t)\right) & t = 300,\ 350,\ 400,\ 450,\ 500 \end{cases} \tag{5.2}$$

for $\mathcal{P} = \mathcal{M}, \mathcal{T}$. Namely, we consider the Dowker-time filtration between adjacent time-steps in our ABM data. Upon computing our feature vectors for this filtation and performing MDS, we instantly notice a striking resemblance between the resulting two-dimensional projection (figure 5.13) and that of the standard Rips filtration for H_1 features (figure 5.2).

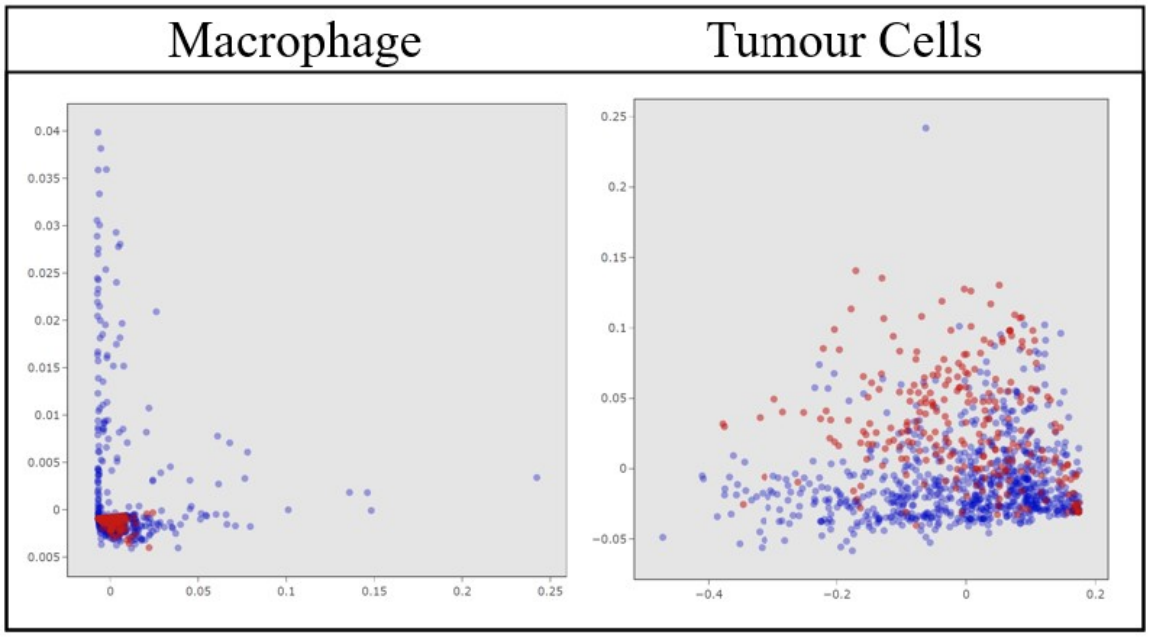


Figure 5.13: Two-dimensional projection of the Dowker-time feature vectors coloured by simulations which are rich in pro-tumour macrophages (red) and those rich in anti-tumour macrophages (blue). We see the clustering of H_1 persistence images for macrophages and tumour cells.

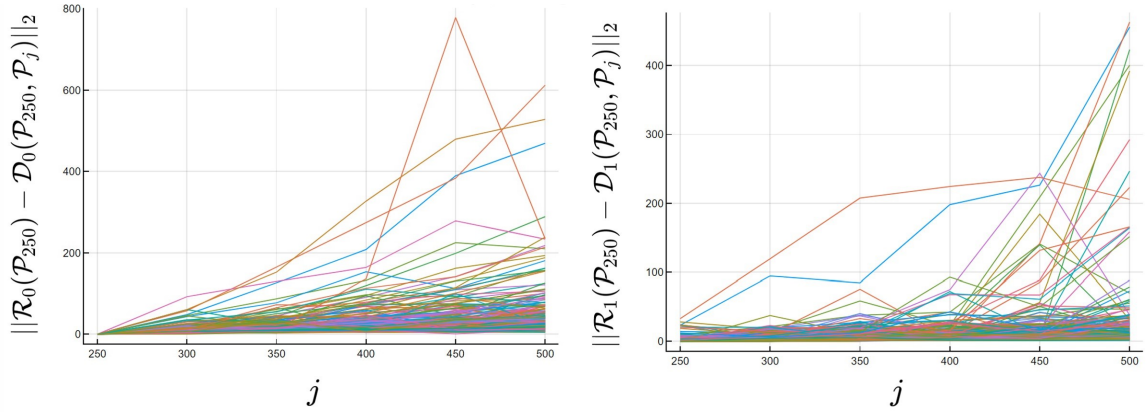
Although the MDS visualisation is similar, this method appears to separate the classes slightly better than the corresponding Rips filtrations (and this shall be exemplified when we consider the classification accuracies). This result suggests that changes in topological features over short time periods may provide more insight than in a single time step. A possible reason for this is the inherent ability of this method to determine significant changes in the behaviour of the tumour cells and macrophages, which become significant when we see more pro-tumour macrophages. This is because the forces on the macrophages change for these different phenotypes.

We next explain the resemblance between the Dowker-time filtration between adjacent time points, and the Rips filtration at a single time point, which is shown in figures 5.13 and 5.2.

5.1.3.1 Comparison of Dowker-time and Vietoris-Rips Filtrations

Recall that the Dowker filtration is constructed by forming binary distance matrices between the two point clouds. When two point clouds are sufficiently close, we require small ϵ for any given landmark to be within ϵ of its nearby witness. In the extreme case where we consider identical, overlapping point clouds, the Dowker filtration

built on the two point clouds (namely the Dowker-time filtration $\mathcal{D}(\mathcal{P}(t), \mathcal{P}(t))$) is very similar to the Rips filtration on either one of the point clouds. This is because at $\epsilon = 0$, all connected components are born (similarly to the Rips filtration), and we include k-simplices in our Dowker complex when $k+1$ witnesses are within ϵ of a given landmark. This condition is similar (though not identical) to the Rips filtration, where we include k-simplices if $k+1$ nodes have non-empty pairwise intersections. We propose that there exists an ϵ-interleaving between the Dowker-time filtration $\mathcal{D}(\mathcal{P}(t), \mathcal{P}(t))$, and the Rips filtration $\mathcal{R}(\mathcal{P}(t))$. Recall that an ϵ-interleaving between two filtrations formalises their 'similarity'. In particular it implies that features born in one filtration at parameter α will always be born in the other by parameter $\alpha + \epsilon$, for some ϵ. We test our hypothesis numerically, and note that a future direction would be to prove, or disprove this statement. Let $\mathcal{R}(\mathcal{P}_i)$ denote the persistence image formed by the Rips filtration of the macrophage point clouds at time i, and let $\mathcal{D}(\mathcal{P}_i, \mathcal{P}_j)$ be the persistence image of the Dowker-time filtration between times i and j. We consider the change in $||\mathcal{R}(\mathcal{P}_i) - \mathcal{D}(\mathcal{P}_i, \mathcal{P}_j)||_2$ for fixed $i = 250$ and $j = 250, 300, 350, 400, 450, 500$ hours in our data set.



(a) H_0 (b) H_1

Figure 5.14: We see the change in Euclidean norm of the difference between persistence images of our Dowker-time filtrations between time 250 hours and times 250, 300, 350, 400, 450 and 500 hours, and the Rips persistence images at time 250 hours. We see results for both H_0 and H_1, with persistence images of size 100×100. At time 250 where the Dowker-time filtration considers identical point clouds, we see very similar persistence images compared to the Rips persistence image. Moreover, the farther j is from time 250, the more different the Dowker-time filtration is from the Rips filtration at time 250.

Figure 5.14 shows the change in difference between our Dowker-time filtration over time and the static Rips filtation at time 250 hours. As proposed, we see similarity[6] between the Rips filtration at time 250, and the Dowker-time filtration $\mathcal{D}(\mathcal{P}_{250}, \mathcal{P}_{250})$. Moreover, we see that as time increases, the difference between our Dowker-time filtration and the Rips filtration increases.

In our analysis of the ABM data using the Dowker-time filtration we considered adjacent time steps, as shown by (5.2). Since we have now shown that we expect the Dowker-time filtration to be in good agreement with the Rips filtration over time steps in which the amount of cell movement is small, the MDS projections between the two being similar are expected results.

[6]And in fact, identical persistence images for H_0

5.2 Comparison of Classification Performance

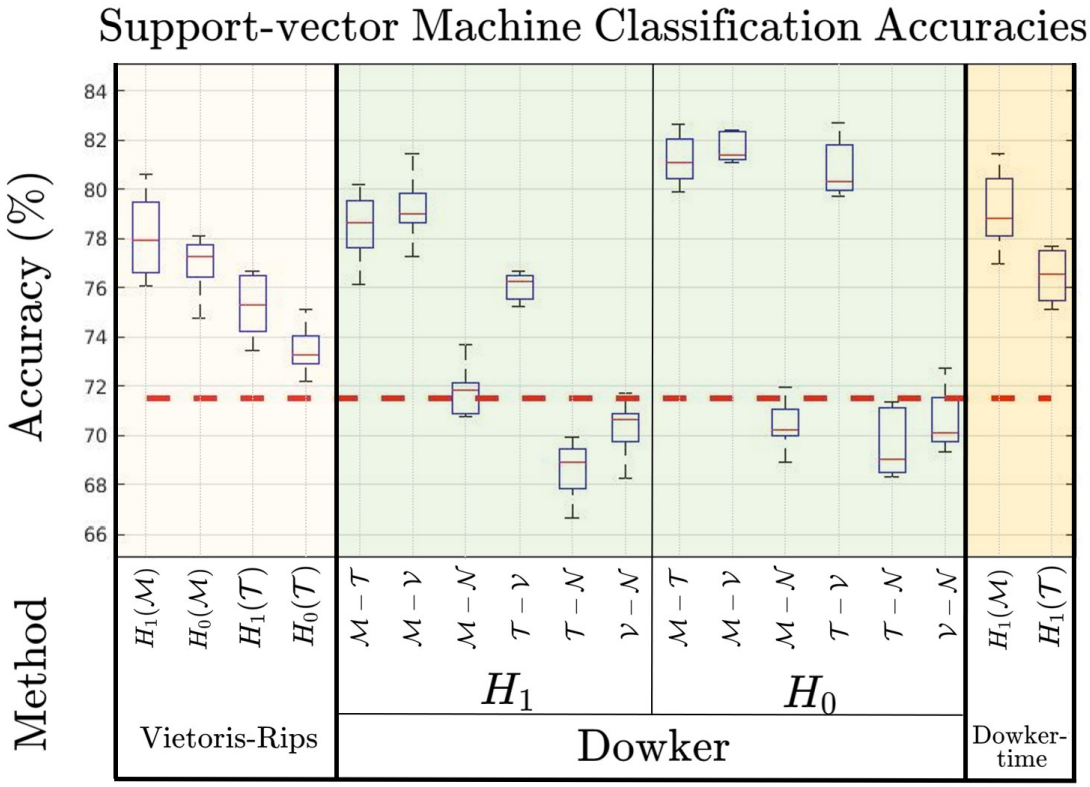


Figure 5.15: Support-vector machine classification accuracies. We see the accuracies for each of the methods described compared to a benchmark value set using simple statistics only involving the numbers of each cell type and distances between macrophages and blood vessels.

We finally report the SVM classification accuracies for the different methods and celltypes. Figure 5.15 shows our overall classification accuracies, with error bars reflecting the range of values obtained by repeating our classification problem on randomly selected simulations. We also see a benchmark accuracy. This benchmark is the classification accuracy obtained by using the amounts of each cell type in an ABM simulation, and also the average distance between blood vessels and their closest macrophage as feature vectors. These are standard statistics one may consider when trying to determine the pro-tumour macrophage concentrations given only information about cell locations, and thus serve as a viable benchmark for our more involved topological statistics.

The best three filtrations are all Dowker based filtrations, and all consider H_0

features. In particular, H_0 features of macrophages relative to blood vessels give the highest average accuracy, followed closely by H_0 features of macrophages relative to tumour cells, and the third best performer is tumour cells relative to blood vessels. We note that, two of the top three performing methods include information about blood vessels. This result makes sense as blood vessels play a significant role for both anti- and pro-tumour macrophages. Of particular importance is the fact that anti-tumour macrophages are released from the blood vessels that surround the domain, and thus form common loops in their migration towards the higher CSF-1 concentration near tumour cells. Moreover, pro-tumour macrophages migrate outwards from the tumour, towards blood vessels, in response to higher concentrations of CXCL12. This interplay between macrophages and blood vessels results in spatial patterns of both tumour cells and macrophages that are indicative of the pro-tumour macrophage concentration. It is an interesting result that the H_0 features for macrophages relative to blood vessels outperform that of tumour cells relative to blood vessels. The successful results of the latter are perhaps more easy to interpret in terms of how we believe the interaction to occur, however the results of the former exemplify the importance of the spatial positions of macrophages relative to blood vessels in knowing the concentration of pro-tumour macrophages. We note that there is a significant overlap in errors between methods. Thus, it is difficult to state which combinations of celltypes leads to the best classification. A common feature of the worst performing methods is that they include information about necrotic cells. Thus, it appears that for the methods considered, the topological features of necrotic cells do not assist with our classification[7]. We note also that the Dowker-time filtration which considers H_1 features of macrophages over time outperforms many of the other methods, including the Rips filtration with which we previously compared it. This exemplifies the importance of the change in topological features over time of macrophages and tumour cells.

[7]This is not to say that information about necrotic cells is not useful, as this is likely dependent on the classification task and methods used.

5.3 Prediction of Outcome from Early Snapshots

Figure 5.16: Multilayer perceptron error values over predicting the final pro-tumour macrophage concentration from earlier time steps. We see the errors for our three top performing methods in the previous section.

Thus far we have shown that we can use topological methods to classify simulations based on whether or not they have a majority of pro- or anti-tumour macrophages, with improved accuracy compared to simpler features. We next consider a scenario in which we may only have spatial data from a single time point, and our goal is to predict the concentration of pro-tumour macrophages at the final time ($t = 500$ hours). We apply a feed-forward neural network, in particular a multilayer perceptron (MLP), using the feature vectors of simulations at single times 250, 300, 350, 400 and 450 hours as inputs, and concentrations of pro-tumour macrophages at time 500 hours as our target variables. We perform hyper-parameter optimisation by using a random grid search over a series of input variables and consider the accuracy of the output based on the average percentage difference between our test set labels and our predictor estimates. We apply this pipeline to the top three performers of our previous classification task and compare results. Recall that these top performing methods all considered H_0 features of Dowker filtrations; macrophages relative

to blood vessels, tumour cells relative to blood vessels, and macrophages relative to tumour cells. Figure 5.16 shows how the average prediction error changes over time for these three methods. An interesting comparison between the SVM results and the MLP results are that we do not always see a correspondence between the accuracy with which the methods predict if there is a majority of pro-tumour macrophages at a particular time, and predicting the concentration of pro-tumour macrophages at time 500 hours. For our MLP, we see particularly good results for H_0 features of macrophages relative to tumour cells at times $300 - 450$ hours, whereas we see much worse results at the outset. This indicates that at stages beyond the initial extravasation, the relative topological features of macrophages and tumour cells are indicative of late time pro-tumour macrophage concentrations.

In summary of the results of this Chapter, we have shown that by employing the different methods from Chapter 4 to the ABM data, we can significantly increase the benchmark accuracy in determining whether or not a given simulation has a majority of pro-tumour macrophages. We have also shown that the topological features from earlier time points can accurately predict the final concentration of pro-tumour macrophages.

Chapter 6

Discussion and Future Directions

In this report we have implemented a series of topological methods to analyse point cloud data representing the heterogeneous immune landscape in a tumour microenvironment. In particular we have shown that the standard Rips filtration outperforms methods which take into account simple features for the classification of the heterogenity in the macrophage population. Moreover we have implemented the Dowker filtration to further enhance our classification accuracies, by considering the homology of cell types relative to each other. In particular we have shown that the relative topological features of macrophages relative to tumour cells, macrophages relative to blood vessels, and tumour cells relative to blood vessels significantly improve the benchmark accuracy for classification. Furthermore we have implemented a novel time dependent Dowker filtration, and shown its success in the classification task, improving on the closely related Rips filtrations. It should be noted that although we applied our techniques to the tumour microenvironment, such methodologies have broader potential for dynamic data sets including populations which may adapt to their environment. Possible future directions of this work include both further topological and biological considerations.

Multiparameter persistent homology (MPH) is an extension of persistent homology which considers how the homology of a point cloud changes over the course of a filtration which depends on two parameters rather than one [24] as used by the methods in this report. These multi-parameter dependent filtrations are called bi-filtrations. Such techniques have been shown to be effective in similar classification tasks [24, 25], and thus such methods warrant applications in our classification tasks. One possible bi-filtration we could consider involves varying the ϵ-neighbourhood of our macrophages and constructing the standard Rips filtration, with the second parameter being macrophage phenotype. In this way, we could investigate whether

the Rips filtration at different thresholds of macrophage phenotype may give further insights or classification accuracy compared to the methods used in this report. Another possible bi-filtration would be to construct the Dowker filtration of tumour cells or macrophages relative to blood vessels, where the second parameter encompasses the density of the corresponding cell. Such a bi-filtration may improve classification accuracy by removing noisy, dense, areas of the cells over the filtration.

In this report, we have focused on the spatial patterns arising in the interactions between tumour cells and macrophages, and what we can learn about the proportion of pro-tumour macrophages from these. Further biological considerations would be to find a correspondence between our findings and the relevant parameters in the agent-based model used to generate the data. Such a relationship may provide further insights into the mechanisms that give rise to unfavourable behaviours, and thus may provide further guidance on possible treatments in a clinical setting.

A final extension to our results is testing these methods on clinical data. While we have presented results for idealised synthetic data, the methods used may be generally applicable. Thus, our research provides motivation for their application to real data sets. In particular, our predictive models may prove insightful in real life applications when it is difficult to obtain time series imaging data.

References

[1] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. Journal of Machine Learning Research, 18(8):1–35, 2017.

[2] Dinesh K. Ahirwar, Mohd W. Nasser, Madhu M. Ouseph, Mohamad Elbaz, Maria C. Cuitiño, Raleigh D. Kladney, Sanjay Varikuti, Kirti Kaul, Abhay R. Satoskar, Bhuvaneswari Ramaswamy, Xiaoli Zhang, Michael C. Ostrowski, Gustavo Leone, and Ramesh K. Ganju. Fibroblast-derived cxcl12 promotes breast cancer metastasis by facilitating tumor cell intravasation. Oncogene, 37(32):4428–4442, Aug 2018.

[3] Ingwer Borg and Patrick JF Groenen. Modern multidimensional scaling: Theory and applications. Springer Science & Business Media, 2005.

[4] Joshua A. Bull and Helen M. Byrne. Macrophage sensitivity to microenvironmental cues influences spatial heterogeneity of tumours. bioRxiv, 2022.

[5] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3):1–27, 2011.

[6] Theerawut Chanmee, Pawared Ontong, Kenjiro Konno, and Naoki Itano. Tumor-associated macrophages as major players in the tumor microenvironment. Cancers, 6(3):1670–1690, 2014.

[7] Samir Chowdhury and Facundo Mémoli. A functorial dowker theorem and persistent homology of asymmetric networks, 2016.

[8] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. Discrete & Computational Geometry, 37(1):103–120, Jan 2007.

[9] C. H. Dowker. Homology groups of relations. Annals of Mathematics, 56(1):84–95, 1952.

[10] Jean-Guillaume Dumas, Frank Heckenbach, B. David Saunders, and Volkmar Welker. Computing simplicial homology based on efficient smith normal form algorithms. In Algebra, Geometry, and Software Systems, 2003.

[11] Alon Efrat, Alon Itai, and Matthew J. Katz. Geometry helps in bottleneck matching and related problems, 1999.

[12] Barbara Di Fabio and Massimo Ferri. Comparing persistence diagrams through complex vectors. In ICIAP, 2015.

[13] Robert Ghrist. Barcodes: The persistent topology of data. BULLETIN (New Series) OF THE AMERICAN MATHEMATICAL SOCIETY, 45, 02 2008.

[14] Robert W Ghrist. Elementary applied topology, volume 1. Createspace Seattle, 2014.

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

[16] G. Henselman and R. Ghrist. Matroid Filtrations and Computational Persistent Homology. ArXiv e-prints, June 2016.

[17] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An introduction to statistical learning, volume 112. Springer, 2013.

[18] Yoshihiro Komohara and Motohiro Takeya. Cafs and tams: maestros of the tumour microenvironment. The Journal of Pathology, 241(3):313–315, 2017.

[19] Zhuo Li, Daichi Maeda, Makoto Yoshida, Michinobu Umakoshi, Hiroshi Nanjo, Kouya Shiraishi, Motonobu Saito, Takashi Kohno, Hayato Konno, Hajime Saito, Yoshihiro Minamiya, and Akiteru Goto. The intratumoral distribution influences the prognostic impact of CD68- and CD204-positive macrophages in non-small cell lung cancer. Lung Cancer, 123:127–135, July 2018.

[20] Dahua Lin. Multivariatestats.jl. https://github.com/JuliaStats/MultivariateStats.jl, 2014.

[21] James R Munkres. Elements of algebraic topology, 1984.

[22] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. Journal of machine learning research, 12(Oct):2825–2830, 2011.

[23] Marta L Pinto, Elisabete Rios, Cecília Durães, Ricardo Ribeiro, José C Machado, Alberto Mantovani, Mário A Barbosa, Fatima Carneiro, and Maria J Oliveira. The two faces of Tumor-Associated macrophages and their clinical significance in colorectal cancer. Front Immunol, 10:1875, August 2019.

[24] Oliver Vipond. Multiparameter persistence landscapes. Journal of Machine Learning Research, 21(61):1–38, 2020.

[25] Oliver Vipond, Joshua A. Bull, Philip S. Macklin, Ulrike Tillmann, Christopher W. Pugh, Helen M. Byrne, and Heather A. Harrington. Multiparameter persistent homology landscapes identify immune cell spatial patterns in tumors. Proceedings of the National Academy of Sciences, 118(41):e2102166118, 2021.

[26] Hee Rhang Yoon, Robert Ghrist, and Chad Giusti. Persistent extension and analogous bars: Data-induced relations between persistence barcodes, 2022.

[27] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. Discrete & Computational Geometry, 33(2):249–274, Feb 2005.

[28] Matija Čufar. Ripserer.jl: flexible and efficient persistent homology computation in julia. Journal of Open Source Software, 5(54):2614, 2020.