PERSISTENT HOMOLOGY BASED CHARACTERIZATION OF THE BREAST CANCER IMMUNE MICROENVIRONMENT: A FEASIBILITY STUDY[*]

Andrew Aukerman,[†] *Mathieu Carrière,*[‡] *Chao Chen,*[§] *Kevin Gardner,*[¶]*Raúl Rabadán*[‖] *and Rami Vanguri*[**]

Abstract. Persistent homology is a powerful tool in topological data analysis. The main output, persistence diagrams, encode the geometry and topology of given datasets. We present a novel application of persistent homology to characterize the biological environment surrounding breast cancers, known as the tumor microenvironment. Specifically, we will characterize the spatial arrangement of immune and malignant epithelial (tumor) cells within the breast cancer immune microenvironment. Quantitative and robust characterizations are built by computing persistence diagrams from quantitative multiplex immunofluorescence, which is a technology which allows us to obtain spatial coordinates and protein intensities on individual cells. The resulting persistence diagrams are evaluated as characteristic biomarkers predictive of cancer subtype and prognostic of overall survival. For a cohort of approximately 700 breast cancer patients with median 8.5-year clinical follow-up, we show that these persistence diagrams outperform and complement the usual descriptors which capture spatial relationships with nearest neighbor analysis. Our results thus suggest new methods which can be used to build topology-based biomarkers which are characteristic and predictive of cancer subtype and response to therapy as well as prognostic of overall survival.

1 Introduction

Descriptors computed with tools from topological data analysis (TDA), such as persistence diagrams [EH08, ZC05] and Mapper [SMC07], have shown strong analytical power in many real world biological data. Examples include, but are not limited to, neuronal structures [LWA$^+$17, KDS$^+$18], cardiac trabeculae [GCZ$^+$13, WCW$^+$17], brain images [PHC$^+$11, LKC$^+$12], breast images [WKL$^+$21] and genomics data [NLC11, CCR13, RCK$^+$17]. These methods capture multi-scale geometric and structural patterns of data with guaranteed robustness against potential noise introduced in measurement [CSEH07, CSEHM10] and in upstream preprocessing steps [BCOS16]. They provide a principled way to systematically

[*] This research was partially supported by NIH Grants T15LM007079-28 and 5R01CA253368.

[†] *Department of Pathology & Cell Biology, Columbia University*, aa4542@cumc.columbia.edu

[‡] *Department of Systems Biology, Columbia University*, mc4660@cumc.columbia.edu

[§] *Department of Biomedical Informatics, Stony Brook University*, chao.chen.1@stonybrook.edu

[¶] *Department of Pathology & Cell Biology, Columbia University*, klg2160@cumc.columbia.edu

[‖] *Department of Systems Biology, Columbia University*, rr2579@cumc.columbia.edu

[**] *Department of Pathology & Cell Biology, Columbia University*, r.vanguri@columbia.edu

quantify complex biomedical systems. Furthermore, state-of-the-art discriminative models (i.e., classifiers) [CCO17, HKNU17, KHF16] and unsupervised models (i.e., clustering methods) [LCO18] have been recently introduced, and are able to effectively connect topological features with clinical/biological outcomes of interest.

We present a new application of topological data analysis to the characterization of the spatial organization of immune cells surrounding breast tumors, known as the breast cancer immune microenvironment, using persistence diagrams. Despite tremendous advancements in cancer screening, diagnostic methods and treatment, breast cancer remains the second leading cause of cancer death in women with projections of 270,000 new cases and approximately 42,000 deaths from invasive breast cancer in 2019 [SMJ19]. By characterizing the interplay of cells which comprise the breast cancer immune microenvironment, we can characterize the response of the patient immune system to the tumor, which is important in determining response to therapy. Predictors of response to therapy are a critical, unmet need in breast cancer [DXLB16], and can aid in the development of novel potential therapeutic targets. We show how persistence diagrams work towards fulfilling this need.

Cancer research and characterization of spatial cell arrangement. In the past decade, a major focus of cancer research has been on the interplay between the tumor and immune environments, referred to as the *tumor-immune microenvironment* [BAAN17]. By characterizing host-specific functional anti-tumor immune responses and their correlations to cancer subtype and overall survival, patient specific immunotherapeutic targets can be identified [PKS$^+$16] with higher precision. To achieve the goal, it is necessary to characterize the complex spatial arrangement between cancer cells and a mixture of different immune cells, e.g., T-cells and macrophages, both of which play a versatile biological role and are believed to be crucially relevant to initiation and regulation of the immune response. This task involves two important steps: cell detection and characterization.

Thanks to the rapid development of imaging technology and deep learning methods, we are able to detect not only locations, but also types of different cells within a slide of tumor biopsy sample from a cancer patient. By staining the slide using immunohistochemical (IHC) markers, we are able to tag different types of cells with different stains, i.e., colors bounded with different protein biomarkers. Using a brightfield image scanner, we convert the stained slide into a whole slide image in which various cells can be identified by their respective stains [PFCW19, KB17]. The identification of cells is referred to as *phenotyping*. Advanced deep learning methods [FAG$^+$19, AFH$^+$19] have been developed to unmix the stains to detect and identify cells. This approach, called multiplex IHC, is scalable but less precise as noise is introduced due to the additional deep learning cell detector. Alternatively, we may use quantitative multiplex immunofluorescence (qmIF), which stains different cells with different fluorescent stains and detect them using lenses with specific filters. The qmIF approach is highly reliable, albeit costly in material and time.

Once cells of different types are detected, we need to quantitatively characterize their spatial arrangements in order to evaluate correlations with various outcomes of interest. There are two major challenges. First, the spatial arrangement is highly heterogeneous across different patients and even within a single tissue sample. Second, stain intensity is relative, and phenotype thresholds must be manually determined. Discerning true signal from background is not always clear, and is currently done in relation to other tissue samples. Nonetheless, qmIF imaging provides rich data for study; see Figure 1 for an example of raw image data.

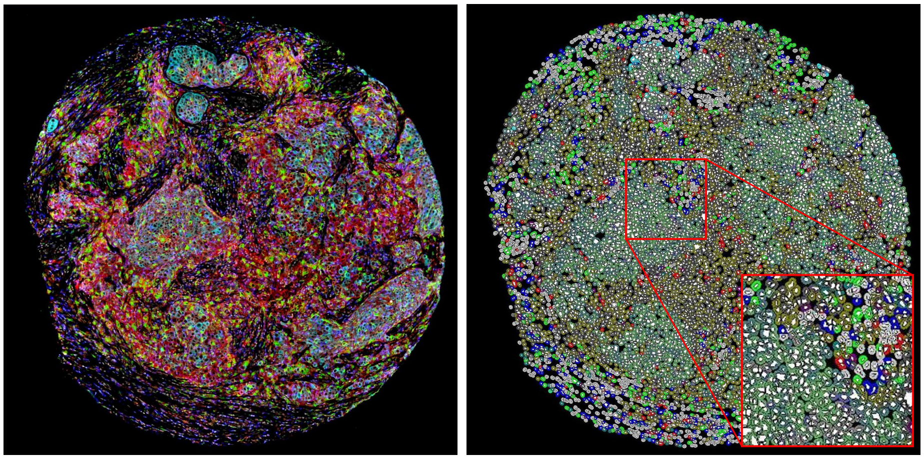

Figure 1: An example input data. Left: The raw microscopic image of a stained tissue sample. The sample is approximately 1 × 1 mm^2 large. The image is 2,000 × 2,000 pixels, 0.5 × 0.5 μ^2 per pixel. A sample usually contains 3,000 to 5,000 cells. Right: The processed results. Cells are identified by localizing their nuclei with a special stain (shown as white regions). The phenotype of each cell can be identified by the stain intensity of its cytoplasm and nucleus: T-cells are tagged with CD8 (blue), macrophages are tagged with CD68 (green), tumor cells are tagged with pancytokeratin (cyan). Any cell may additionally be tagged with PD-L1 (red). The cells are abstracted into point clouds with different stain intensities, as shown in Figure 3.

Related work. Previous methods [GMH+18, SBSO16] focus on using nearest neighbor distances from cells of one type (obtained by thresholding the stain intensities) to cells of a second type. Unfortunately, this thresholding-based approach lacks the ability to model stain concentration variations, and thus is sensitive to noise. Moreover, it can only characterize fixed neighborhoods around the cells and is oblivious to larger and more complex cell arrangements.

Persistent homology has recently been used to characterize cellular architecture in pathology images in [LSB+19], where these descriptors were shown to successfully detect and quantify circular cell structures corresponding to glands. In contrast, our work operates on coordinates of phenotyped cells and deals with the global characterization of complex interactions between these cellular phenotypes.

Note that once persistence diagrams have been computed from phenotyped cells, there are several ways to use them for subsequent analysis, by either defining scalar products, or kernels [RHBK15, KHF16, LY18], or explicit vectorization methods [AEK+17, Bub15]. In this article, we choose to use kernel methods based on the sliced Wasserstein kernel [CCO17]. We utilize kernel methods mostly because (universal) kernels are known to enjoy several useful theoretical properties for statistical testing [GHS+05, GBSS05], which often lead to better results than using simpler vectorization techniques.

Contributions. In this article, *we propose the first topological analysis of tumor immune microenvironment.* More specifically, we provide empirical evidence that persistence diagrams are suitable descriptors by experimentally demonstrating the following points:

- First, stain concentration levels, or stain intensities, that are usually used by practitioners to filter cells, are natural candidates for defining *filtrations* (in the TDA vocabulary) from which persistence diagrams can be computed. This way, the whole range of stain intensities is taken into account instead of thresholding. We hypothesize that the stain intensity is biologically meaningful and the resulting persistence diagrams will be more predictive than just using cell coordinates from thresholding. In particular, the stability of persistence diagrams is essential for controlling the noise and perturbations that often occur when measuring stain intensities.
- Second, persistence diagrams are able to capture topological and structural features that are characteristic of the arrangement of the cells. This is because the structures encoded by persistence diagrams are robust to spatial deformation and other types of noise introduced in detection, which prevents the analysis from being biased by measurement errors, contrarily to other descriptors used in the literature.

Our study, although preliminary, demonstrates the potential of persistence homology to be a novel tool to characterize the tumor immune microenvironment. With rich computation and learning tools available for persistence-derived features, we are confident that topological characterization will lead to powerful predictive and prognostic cancer biomarkers.

Plan of the article. We introduce our biological data, and briefly recall the basics of topological data analysis in Section 2. Then, we explain our methods for computing and running statistical tests on persistence diagrams in Section 3. Finally, we conclude and summarize future investigations and open questions in Section 4.

2 Data and Background

In this section, we introduce our biological data (Section 2.1), and briefly recall the rationale for nearest neighbor analysis (Section 2.2) and topological data analysis (Section 2.3).

2.1 Biological Data

We analyze a large cohort of patients with extensive 8.5 years of follow-up. For each tissue sample, qmIF imaging was obtained with a panel of immune markers for phenotyping the tumor immune microenvironment, including: CD8 (cytotoxic T-cells), CD68 (macrophages) and pancytokeratin (tumor cells). Then, a commercial software package (HALO, Indica Labs) was used to perform nuclear segmentation, cytoplasmic definition, and stain quantification. Cell phenotypes are assigned based on manual thresholds applied to individual stain intensities. See Figure 1 for the conventional threshold-based phenotype analysis.

Patient Cohort. Our raw data is comprised of high-throughput tissue microarrays (TMA) consisting of 1mm × 1mm cores of tissue. The TMA were assembled with tissues from a cohort of 900 patients that underwent tumor resection following a diagnosis of breast cancer at Pitt County Memorial Hospital (now Vidant Hospital) in Greenville, North Carolina. Patient samples and clinicopathological data were collected under an IRB approved protocol at the Brody School of Medicine, East Carolina University [BSP+19]. The cohort is uniquely valuable for research as there is median 8.5 year follow-up data which allows for predictive and prognostic evaluation for topological biomarkers using patient attributes and clinical outcomes.

Quantitative Multiplex Immunofluorescence. Unlike traditional immunohistochemistry, qmIF enables simultaneous staining of multiple markers in a single piece of tissue. We use the Ultivue UltiMapper I/O PD-L1 assay consisting of the following markers: CD8 (cytotoxic T-cells), CD68 (macrophages), PD-L1 (an immunosuppressive protein), pancytokeratin (epithelial cells), and DAPI (DNA marker) for identification of cell nuclei. In our data, positively stained epithelial cells via pancytokeratin are considered to be tumor cells. Every cell in the tissue is designated with a PD-L1 status being either positive or negative corresponding to above or below threshold stain intensity. All staining thresholds are adaptively determined to enhance signal (consistent with a positive staining pattern assessed visually) to background. The result of the phenotyping analysis is a text file for each tissue sample consisting of entries listing information about each cell location, including the manual phenotyping result and raw stain intensities. Each tissue sample consists of 3,000-5,000 cells.

2.2 Nearest Neighbor Analysis

Nearest neighbor analysis is commonly performed with qmIF data [GMH+18]. We measure the nearest neighbor distance between cells of different phenotypes. For any two phenotypes t_1 and t_2, we denote their corresponding point/cell sets P_{t_1} and P_{t_2}. For any cell p of phenotype t_1, $p \in P_{t_1}$, its nearest neighbor distance to P_{t_2} is

$$d(p, P_{t_2}) = \min_{q \in P_{t_2}} d(p, q),$$

in which $d(p, q)$ is the Euclidean distance between p and q. The mean and standard deviation of $d(p, P_{t_2})$ over all p's in P_{t_1} are calculated, and referred to as the nearest neighbor distance features for the pair of phenotypes (t_1, t_2). Note that we ignore distances below 0.05 microns to avoid errors due to overlapping cells. We apply the same process for all pairs of phenotypes and use the corresponding means and deviations as features of biomarkers, potentially predictive of triple-negative status and prognostic of overall survival.

2.3 Topological Data Analysis

In this article, we aim at characterizing the spatial arrangement of phenotypes using persistence diagrams, which are common descriptors of topological data analysis. In this section we describe the basics of persistent homology and persistence diagrams. A thorough treatment of persistence can be found in several computational topology and algebraic topology textbooks such as [EH10, CdSGO16, Oud15].

Persistent homology. The aim of persistent homology is to encode the topological information contained in a dataset X through the lens of a *filter function* $f : X \to \mathbb{R}$. This is achieved by considering the *sublevel sets* of f: $F_\alpha = \{x \in X : f(x) \leq \alpha\}$. The family of sublevel sets $\mathcal{F} = \{F_\alpha\}_{\alpha\in\mathbb{R}}$ defines a *filtration*, i.e., a family of subsets of X that are nested with respect to the inclusion: $F_\alpha \subseteq F_\beta$ if $\alpha \leq \beta$. The idea of persistence is to track the topological changes occurring in the filtration as the sublevel set threshold α increases from $-\infty$ to $+\infty$. For instance, each time a *topological structure* such as a connected component, a handle or a void, appears in the sublevel set, we use the corresponding threshold as the so-called *birth time* for this structure. Similarly, each time a structure disappears in the sublevel set (think for instance of a handle being filled in after data points inside the handle were added to the sublevel set), we use the corresponding threshold as the *death time*. This tracking is eventually encoded in a *persistence diagram*, that we denote by $D(f)$, which is a set of dots in the Euclidean plane $\mathbb{R}^2$, each dot representing a topological structure whose birth and death times can be retrieved from the coordinates of the dot.

Persistence on images. In Figure 2, we provide an example of persistent homology computation performed on an image taken from the MNIST [LBBH98] dataset. We use the opposite of the pixel intensity as the filter function, so that it increases from black to white. Given a specific filter function value, the black pixels displayed in the top row of Figure 2 are those constituting the sublevel set. One can see that at values b and d, handles are created in the union of black pixels, and they are eventually filled in at value e, for which the corresponding sublevel set includes all pixels. Other examples on our biological data are also displayed in Figures 5 and 6.

Stability of persistence diagrams. One of the most useful properties of persistence diagrams is their *stability*: persistence diagrams computed from similar images must be similar themselves w.r.t. the so-called *Wasserstein distances* between them.

Definition 2.1 ([CdSGO16, CSEH07])**.** *The p-*Wasserstein distance *d_p between two persistence diagrams D, D' is defined as:*

$$d_p(D, D')^p = \inf_\gamma \sum_{q\in D\cup\Delta} \|q - \gamma(q)\|_\infty^p,$$

where Δ is made of an infinite number of copies of the diagonal $\{(x, x) : x \in \mathbb{R}\}$ and γ ranges over all matchings between $D \cup \Delta$ and $D' \cup \Delta$.

When the sum in Definition 2.1 is replaced by a maximum, the Wasserstein distance becomes the so-called *bottleneck distance* d_∞. Using this distance, one can state the *stability property* of persistence diagrams, which shows that the Wasserstein distance between persistence diagrams is upper bounded by the distance (in the $\|\cdot\|_\infty$ norm) between filter functions.

Theorem 2.2 ([CCSG$^+$09, CSEH07])**.** *Given a topological space X and two continuous functions $f, g : X \to \mathbb{R}$, the following inequality holds:*

$$d_\infty(D(f), D(g)) \leq \|f - g\|_\infty \tag{1}$$

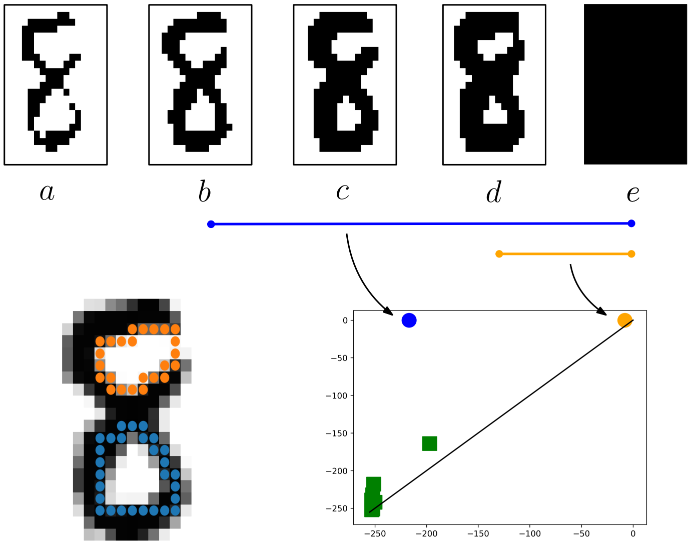

Figure 2: Example of a persistence diagram (lower right) computed on an image taken from the MNIST [LBBH98] dataset (lower left) using the opposite of the pixel stain intensity whose sublevel sets are displayed in the top row. Green squares represent connected components while the blue and orange circles represent handles, whose representative cycles are displayed on the original image.

This result is particularly relevant for stain intensities as the presence of several biological factors and potential biases in experiments can lead to noise in the data. However, Theorem 2.2 ensures that as long as the amplitude of the noise is bounded, the corresponding persistence diagrams, as well as the different statistics that one can compute from them (see Section 3 below), are meaningful and reliable.

Note that similar stability results can be obtained with p-Wasserstein distances, with different upper bounds [CSEHM10, Oud15].

3 Methods and Results

In this section, we detail our methods to compute and analyze persistence diagrams from point clouds representing cells with different stain intensities. More specifically, we show how to discretize the cell domain into an image with stain intensity-valued pixels, from which we calculate the corresponding persistence diagrams (in homological dimensions zero and one) in Section 3.1. Then, we show how to run statistical tests between different populations based on persistence diagrams using Hilbert space embeddings with the *Sliced Wasserstein kernel* [CCO17] in Section 3.2. Finally, we discuss results for different patient groups (patients with different molecular subtypes, patients that survived after 8.5 years vs. deceased) in Section 3.3.

3.1 Persistence Diagrams of Cells with Stain Intensity Values

In this section, we explain how persistence diagrams are computed on our point clouds representing cells so as to make use of the associated stain intensities.

Point clouds. As mentioned earlier, the image data for each patient is summarized in a point cloud, where the points represent cells, and have four associated stain intensities, corresponding to the CD8, CD68, PD-L1, and pancytokeratin (tumor) stains (see Section 2.1). Each patient also has two binary labels corresponding to overall survival and whether the cancer subtype is triple-negative. After removing samples with bad quality or missing labels, our final dataset is comprised of 671 point clouds. See Figure 3 for an example of such point clouds, where we only kept the cells with stain intensities above a certain threshold to ease visualization. One can see from these point clouds that different topological structures seem to emerge depending on the stain being considered: structures can be either isolated components corresponding to the scattered spots of cells exhibiting large stain intensity values (such as pancytokeratin (tumor) in Figure 3) or small cycles corresponding to regions where there are no cells with large stain intensity (such as CD8 in Figure 3). The lack of any discernible structure is also a possible feature if the stain intensity is diffuse across the whole tissue (such as PD-L1 in Figure 3).

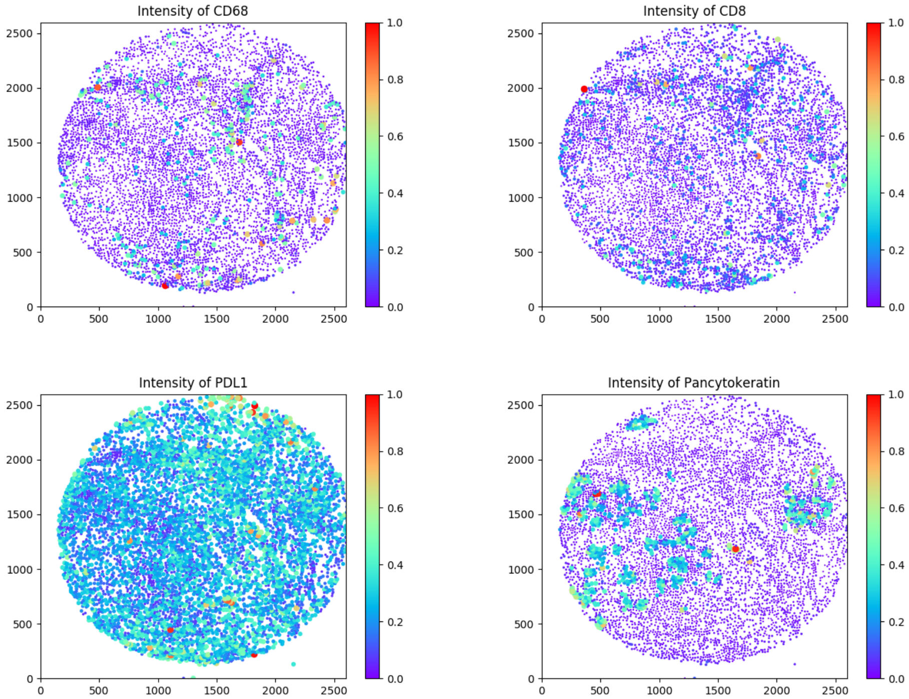

Figure 3: Illustration of the point clouds corresponding to the different stains (cell color and size is proportional to stain intensity to ease visualization). One can see that the different stain intensities induce different geometric patterns.

Persistence Diagrams. It is common in topological data analysis to use Vietoris-

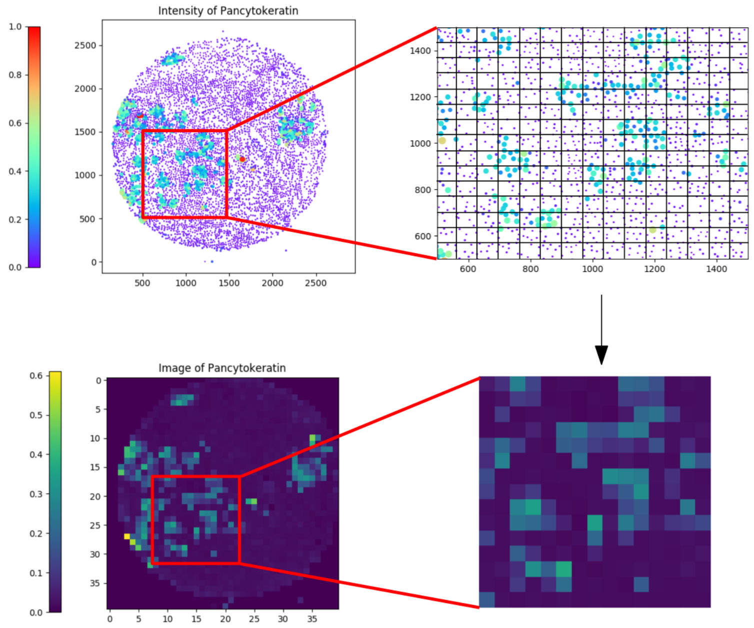

Figure 4: Discretization process turning a point cloud with stain intensity values into an image. We start with the full point cloud with the corresponding stain intensity values (upper left). Note that we only show cells above a certain stain intensity threshold to ease visualization. The cells are then placed into pixels of a grid drawn on top of the plane (upper right). These pixels with the corresponding stain intensity values are then turned into an image (bottom row), by summing the stain intensities in each pixel.

Rips, Cech or Alpha filtrations [CCSG+09] when dealing with point clouds. However, these filtrations would be agnostic to stain intensities, only providing information about the shape of the whole point cloud, which may not be sufficient to successfully encode the spatial and geometrical relationships between phenotypes.

In order to take the stain intensities into account when computing topological descriptors, we propose an image-based filtration. We first discretized the plane into a grid of 40×40 pixels. Next, we binned the stain intensity values on this grid, and sum up the stain intensities in each bin, or pixel, so as to obtain an image. Note that the choice of resolution (i.e., the number of pixels) has to be carefully done: if the number of pixels is too small, one might not be able to see and compute the topological structures, but on the other hand, a resolution that is too large would induce artifacts, in the sense that all cells would be isolated, and no interesting topology could be computed. Our resolution of 40×40 pixels was manually chosen and seemed to be the best tradeoff on our data. See Figure 4 for an illustration of this process. Note also that it would be interesting to use Nadaraya-Watson kernel-based estimators (see Chapter 6 in [HTF03]) to smooth the stain intensities of the pixels, but we left this possibility for future work.

We chose the image-based filtration instead of other alternatives because it was easier and more intuitive to deal with resolution/scale for this study. Indeed, when handling a point cloud with attached values, one usually has two other choices: (1) Use scalar field analysis [CGOS11], which is based on δ-neighborhood graphs. However, tuning the parameter δ is difficult and depends on the geometric characteristics (such as reach and radius of convexity) on the Riemannian manifold the data is supposed to be sampled from; it is thus quite hard to estimate. Also, the theoretical approximation results of scalar field analysis are only valid for nested pairs of Vietoris-Rips filtrations, which are notoriously more difficult to compute in positive homology dimensions. (2) A simpler option is to directly build a Vietoris-Rips or Alpha complex, and filter it with the intensities using lower-star filtrations; however, it also requires to either define a neighborhood scale parameter δ, which is difficult to estimate (for Vietoris-Rips complexes), or to work with Delaunay triangulations, which might introduce biases due to the coordinates of the cells and independent from the point cloud values. Compared with these alternatives, we found it much easier to use an image-based filtration, as we can naturally control the scale by tuning the image/grid resolution.

Finally, we used persistent homology (see Section 2.3) to produce persistence diagrams out of our stain intensity-based images, by filtering the pixels with the opposite of the stain intensity (so that pixels with large stain intensity appear first). Note that points with death time 0 corresponds to topological structures that disappeared when adding the pixels with stain intensity 0, i.e., the pixels corresponding either to the cells not belonging to the corresponding phenotype or to pixels with no associated cells. These points should thus not be considered characteristic of the corresponding phenotype. See Figure 5 for examples of such persistence diagrams. Correlations between tissue morphology shown in the images and the persistence diagrams are observed. For example, the distance to the diagonal of points in persistence diagrams of homological dimension 0 and the number of points in persistence diagrams of homological dimension 1 seem to be correlated to the aggregation of cells with large stain intensity.

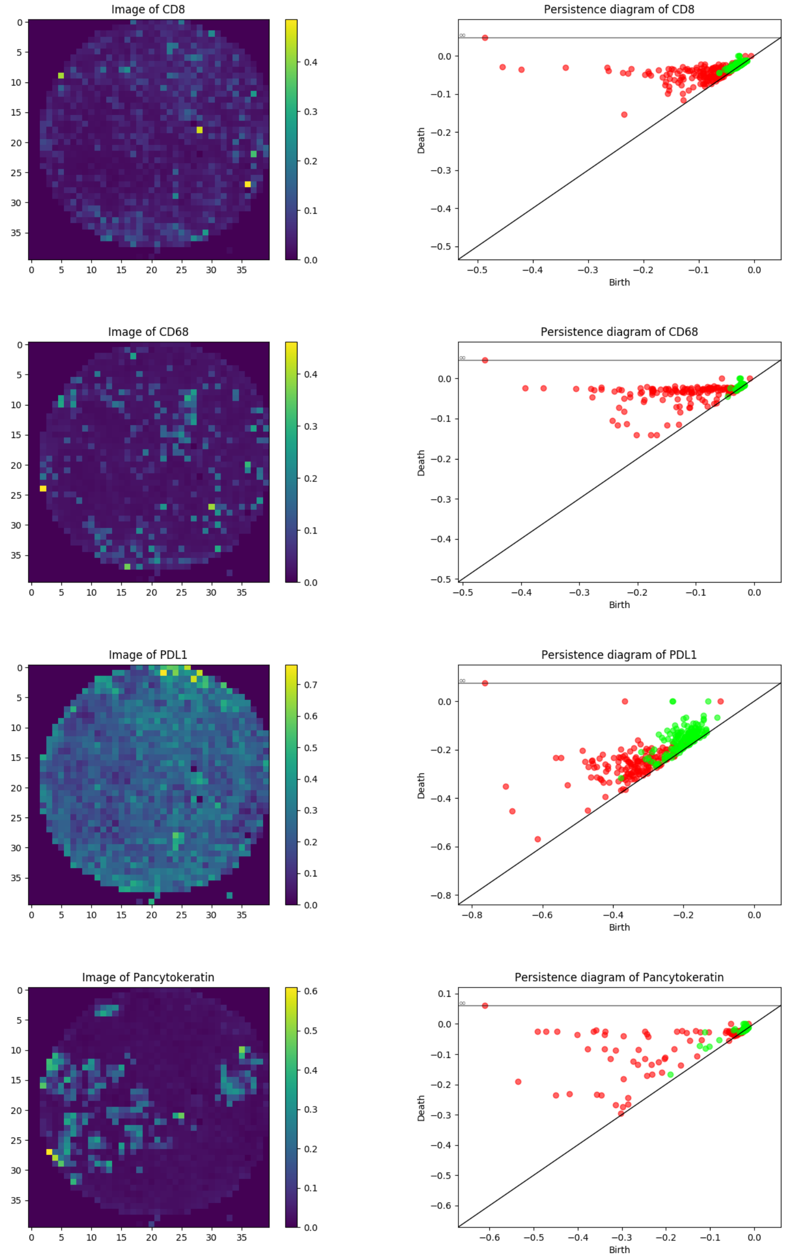

Figure 5: Examples of images with stain intensity-based pixels computed from point clouds (left) and their corresponding persistence diagrams (right). Points in homological dimension 0 are displayed in red and points in homological dimension 1 are displayed in green. From top to bottom: stains of CD8, CD68, PD-L1 and pancytokeratin.

Pairs of phenotypes. As mentioned in Section 1, characterizing the interactions, or co-localizations, between pairs of phenotypes might be as important, if not more, as characterizing them alone. Hence, we also computed persistence diagrams out of images with pixels colored by the average of pairs of phenotypes. This can be thought of as a similar but quite more general measure of co-localization than the one given by nearest neighbors (see Section 2.2). Indeed, the standard nearest neighbor analysis basically ranks the cells with respect to the distance to their closest neighbors. In terms of persistence, this ranking can be retrieved from the pixel filtration values: the lower they are, the more the corresponding pixels are likely to contain cells that co-localize from the two phenotypes. However, persistence diagrams also encode the interactions between the topological structures that are born from these co-localization spots. See Figure 6 for examples of such persistence diagrams. One can see from these images that the topological structures that are present in the image of a pair of phenotypes roughly include those of each phenotype alone, and that the structures that co-localize are emphasized.

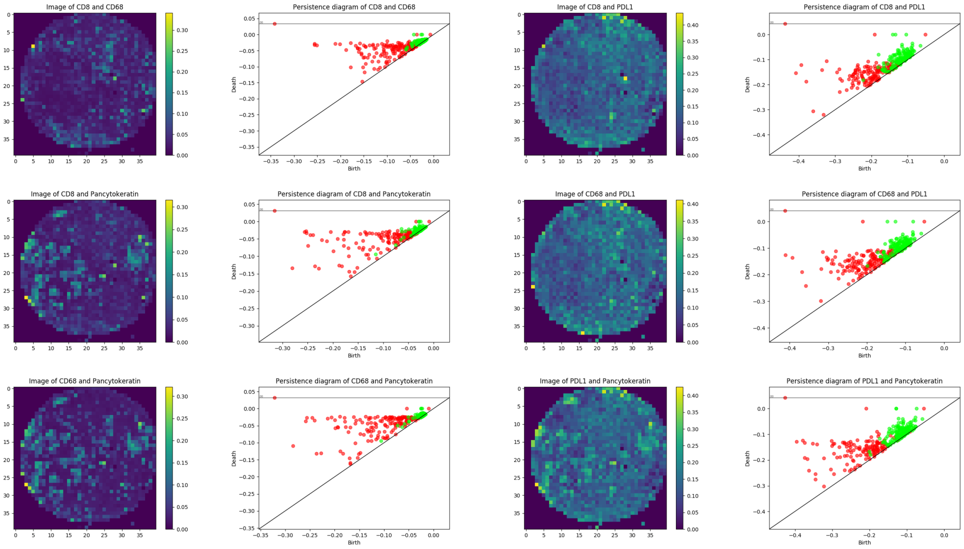

Figure 6: Examples of images and associated persistence diagrams computed from pairs of phenotypes/stains. Points in homological dimension 0 are displayed in red and points in homological dimension 1 are displayed in green.

Robustness. From a theoretical point of view, the stability property that persistence diagrams enjoy (see Section 2.3 and Proposition 2.2) is very advantageous. Indeed, it is well-known that any nearest neighbor analysis is sensitive to measurement errors: even a slight mistake in the measurement of stain intensity can induce different phenotype assignments for the cells, and thus different outputs from a nearest neighbor analysis. Since we do not depend on thresholding to compute persistence diagrams, we avoid this issue. On the other hand, the stability theorem for persistence diagrams ensures that any measurement error only has a small effect, provided that the error is small itself.

3.2 Statistics on Persistence Diagrams

In this section, we provide details about the statistical methods we used to assess the efficiency of persistence diagrams as characteristic and predictive biological descriptors.

Kernel-based Statistical Tests. In order to formally assess the statistical power of persistence diagrams with respect to the groups of interest, such as survived vs. deceased, or triple-negative cancer subtype vs. other subtype, we need to be able to run statistical tests on distributions of persistence diagrams. Several recent works have looked at this question from a theoretical point of view [KHN$^+$15, RT17, VJM18]. In this article, we focus on *Kernel Mean Embeddings* [GBR$^+$12], that is, we characterize a sample of a distribution $\mathcal{D}$ of persistence diagrams $\hat{\mathcal{D}}_n = \{D_1, \dots, D_n\}$ by embedding the diagrams in a Hilbert space $\mathcal{H}$ with a continuous map Φ, and by taking the mean (in the Hilbert space) of this sample: $\Phi(\hat{\mathcal{D}}_n) := \frac{1}{n}\sum_{i=1}^n \Phi(D_i)$.

Now, given two samples $\hat{\mathcal{D}}_n$ and $\hat{\mathcal{D}}'_n$, one can compute the statistic:

$$\mathrm{MMD}(\hat{\mathcal{D}}_n, \hat{\mathcal{D}}'_n) := \|\Phi(\hat{\mathcal{D}}_n) - \Phi(\hat{\mathcal{D}}'_n)\|_{\mathcal{H}},$$

also called the *maximum mean discrepancy*, and use it to perform statistical tests in order to check whether $\mathcal{D}$ and $\mathcal{D}'$ are the same. This statistic has been shown to be a good proxy, with quantified approximation bounds, to its continuous version $\|\Phi(\mathcal{D}) - \Phi(\mathcal{D}')\|_{\mathcal{H}}$ in [GBR$^+$12], where $\Phi(\mathcal{D})$ is defined as $\mathbb{E}_{D\sim\mathcal{D}}[\Phi(D)]$.

Choice of the embedding function. It might not be totally clear how to choose such a map Φ for embedding persistence diagrams. This can actually be done quite easily with the use of *kernels*:

Definition 3.1. *Let $\mathcal{D}_{N,L}$ be the space of persistence diagrams with at most N points included in $[-L, L]^2$. A* kernel *is a pairwise function $k : \mathcal{D}_{N,L} \times \mathcal{D}_{N,L} \to \mathbb{R}$ such that the matrix $K = ((k(D_i, D_j)))_{1\leq i,j\leq n}$ is positive semi-definite for any family of persistence diagrams $D_1, \dots, D_n \in \mathcal{D}_{N,L}$.*

A useful result of kernel methods actually relates kernels to embeddings in Hilbert spaces:

Proposition 3.2. *Let k be a kernel on $\mathcal{D}_{N,L}$. Then, there exists a Hilbert space $\mathcal{H}_k$ and a map Φ_k such that, for any $D, D' \in \mathcal{D}_{N,L}$, one has $k(D, D') = \langle\Phi(D), \Phi(D')\rangle_{\mathcal{H}_k}$.*

In other words, any kernel matrix can be interpreted as a Gram matrix in an implicit (and potentially infinite-dimensional) Hilbert space. Moreover, the statistic MMD can be easily computed from k with:

$$\begin{aligned}
\mathrm{MMD}(\hat{\mathcal{D}}_n, \hat{\mathcal{D}}'_n)^2 &= \left\langle \frac{1}{n}\sum_{i=1}^{n}\Phi(D_i) - \frac{1}{m}\sum_{j=1}^{m}\Phi(D'_j),\ \frac{1}{n}\sum_{i=1}^{n}\Phi(D_i) - \frac{1}{m}\sum_{j=1}^{m}\Phi(D'_j)\right\rangle_{\mathcal{H}_k} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{u=1}^{n}\langle\Phi(D_i),\Phi(D_u)\rangle_{\mathcal{H}_k} + \frac{1}{m^2}\sum_{j=1}^{m}\sum_{v=1}^{m}\langle\Phi(D'_i),\Phi(D'_v)\rangle_{\mathcal{H}_k} \\
&\qquad - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\langle\Phi(D_i),\Phi(D'_j)\rangle_{\mathcal{H}_k} \\
&= \frac{1}{n^2}\|K\|_1 + \frac{1}{m^2}\|K'\|_1 - \frac{2}{nm}\|\tilde{K}\|_1,
\end{aligned}$$

where K, K' and $\tilde{K}$ are the kernel matrices computed on $\mathcal{D}\times\mathcal{D}$, $\mathcal{D}'\times\mathcal{D}'$, and $\mathcal{D}\times\mathcal{D}'$ respectively. Note however that it has been shown in [GBR+12] that MMD is a biased statistic—in practice, we compute the *unbiased* MMD, defined as:

$$\begin{aligned}
\mathrm{MMD}_u(\hat{\mathcal{D}}_n, \hat{\mathcal{D}}'_n)^2 = \frac{1}{n(n-1)}\sum_{\substack{i=1\\ u\neq i}}^{n}\langle\Phi(D_i),\Phi(D_u)\rangle_{\mathcal{H}_k} + \frac{1}{m(m-1)}\sum_{\substack{j=1\\ v\neq j}}^{m}\langle\Phi(D'_i),\Phi(D'_v)\rangle_{\mathcal{H}_k} \\
- \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\langle\Phi(D_i),\Phi(D'_j)\rangle_{\mathcal{H}_k}
\end{aligned}$$

Now it only remains to pick a kernel for persistence diagrams. Several choices have been proposed in recent works [AEK+17, Bub15, CCO17, KHF16, RHBK15], and we will focus on one called the *Sliced Wasserstein kernel* k_{SW} [CCO17] in this work, since it has been shown to be one of the most efficient approaches in different statistical tasks [CCO17]. Its definition is based on the *Sliced Wasserstein distance* SW between persistence diagrams, which is defined (informally) as the integral over all possible lines of the 1-Wasserstein distance (see Section 2.3) computed between projections of these diagrams onto a line going through the origin. In practice, one does not compute this integral exactly but rather sample a fixed number of lines, finding the average Wasserstein distance between the corresponding projections. We refer the interested reader to [CCO17] for a precise definition of this distance, and we merely recall the definition of the associated kernel:

Definition 3.3 ([CCO17])**.** *Let $D, D' \in \mathcal{D}_{N,L}$ and $\sigma > 0$. The* Sliced Wasserstein kernel *is defined as:*

$$k_{\mathrm{SW}}(D, D') = \mathrm{e}^{-\frac{\mathrm{SW}(D,D')}{2\sigma^2}},$$

where SW *denotes the Sliced Wasserstein distance between persistence diagrams.*

One can easily see that k_{SW} can be interpreted as a Gaussian kernel, with its only parameter σ being the corresponding bandwidth.

Characteristic kernels. There is a specific class of kernels in the literature that is of particular interest when it comes to statistical tests: the so-called *characteristic* kernels [SFL11, SGS18].

Definition 3.4. *A kernel k is called* characteristic *if its corresponding map Φ_k is injective on distributions, i.e., for any pair of distributions $\mathcal{D}$ and $\mathcal{D}'$, one has:*

$$\|\Phi(\mathcal{D}) - \Phi(\mathcal{D}')\|_{\mathcal{H}_k} = 0 \Longrightarrow \mathcal{D} = \mathcal{D}'$$

Obviously, any statistical test based on a kernel requires it to be characteristic in order to be theoretically backed-up. Even though it is not clear whether the Sliced Wasserstein kernel is characteristic or not, there exists a strategy to build a characteristic kernel out of another one, that was first presented in [KHN$^+$15], and that we use again in this work:

Theorem 3.5 ([KHN$^+$15])**.** *Let k be a kernel on $\mathcal{D}_{N,L}$ whose associated map Φ_k is continuous and injective and whose associated Hilbert space $\mathcal{H}_k$ is separable. Then the kernel $\tilde{k} := \mathrm{e}^k$ is a characteristic kernel.*

Theorem 3.5 is actually a consequence of a more general theorem that is valid on any compact metric space (the fact that $\mathcal{D}_{N,L}$ is compact with respect to the first Wasserstein distance between persistence diagrams was proved in [KHN$^+$15]). Moreover, it has been shown in [CCO17] that the map $\Phi_{k_{\mathrm{SW}}}$ associated to k_{SW} is continuous and injective. Finally, since it is also known that $\mathcal{D}_{N,L}$ is separable [MMH11], it follows that the Hilbert space associated to k_{SW} is separable as well, as the completion of the span of a separable space. Hence the following result:

Proposition 3.6. *The kernel $\tilde{k}_{\mathrm{SW}} := \mathrm{e}^{k_{\mathrm{SW}}}$ is characteristic.*

All of the statistical analysis presented in the following section has been performed with the kernel $\tilde{k}_{\mathrm{SW}}$, which we call the *characteristic Sliced Wasserstein kernel.*

Comparison with NN features. Concerning the features given by nearest neighbor analysis, i.e., the means and variances of the distribution of Euclidean distances to the closest neighbors (see Section 2.2), we use kernel-based statistical tests based on the MMD computed with a standard linear kernel (which is known to be characteristic). Moreover, we also test the independence between persistence diagrams and nearest neighbor features in order to check whether these two types of features are complementary or not. Kernel methods can also be used to run independence tests based on the Hilbert-Schmidt criterion [GBSS05]. The so-called *Hillbert-Schmidt Independence Criterion* (HSIC for short) [GBSS05] is:

$$\mathrm{HSIC}(\hat{\mathcal{D}}_n^X, \hat{\mathcal{D}}_n^Y) = \frac{1}{(n-1)^2}\mathrm{tr}\left(K_X \cdot H \cdot K_Y \cdot H\right),$$

where $\hat{\mathcal{D}}_n^X$ (resp. $\hat{\mathcal{D}}_n^Y$) is a sample of size n from a distribution $\mathcal{D}_X$ (resp. $\mathcal{D}_Y$) in a space X (resp. Y), K_X (resp. K_Y) is the kernel matrix associated to $\hat{\mathcal{D}}_n^X$ (resp. $\hat{\mathcal{D}}_n^Y$), $H = \mathbf{I}_n - \frac{1}{n}\mathbf{1}$, $\mathbf{I}_n$ is the identity matrix of size n, and $\mathbf{1}$ is the $n \times n$ matrix containing only ones. The quantity HSIC is known to be a good estimator of the cross-covariance operator between

$\Phi(\mathcal{D}_X)$ and $\Psi(\mathcal{D}_Y)$, where Φ (resp. Ψ) is the feature map associated to K_X (resp. K_Y), and is known to be zero if and only if the distributions $\mathcal{D}_X$ and $\mathcal{D}_Y$ are independent (provided that both kernels are characteristic, see Theorem 4 in [GBSS05] [1]). Another measure of this cross-covariance operator, which provides a value that can thought of as a generalization of the usual covariance between random variables, is the so-called *constrained covariance* coefficient [GSB+05], defined as:

$$\mathrm{COCO}(\hat{\mathcal{D}}_n^X, \hat{\mathcal{D}}_n^Y) = \frac{1}{n}\sqrt{\|(H \cdot K_X \cdot H) \cdot (H \cdot K_Y \cdot H)\|_2}.$$

Again, this coefficient is zero if and only if the distributions $\mathcal{D}_X$ and $\mathcal{D}_Y$ are independent for characteristic kernels.

3.3 Results

In this section, we provide the experimental results obtained on our data using the *characteristic Sliced Wasserstein kernel* $\tilde{k}_{\mathrm{SW}}$ presented in Section 3.2 for persistence diagrams and a standard linear kernel for the NN features. In lieu of building a classifier, we instead focus on evaluating the statistical significance between groups of patients. This is due to the lack of tissue area in the tissue microarrays, which limits robust measurements of immune population densities typically available in whole-slide images used for diagnosis.

The kernel bandwidth σ of $\tilde{k}_{\mathrm{SW}}$ was selected manually as the median of all pairwise sliced Wasserstein distances between persistence diagrams, as described in [FSCF16]. We conduct two types of statistical tests; in the first one, we use the MMD statistic to assess whether persistence diagrams can successfully distinguish interesting subgroups in the data, and in the second one, we use the HSIC and COCO statistics to assess how independent persistence diagrams are from NN features. In both cases, p-values are approximated with $10 \cdot 10^3$ random permutations, which are either permutations of the subgroup labels (for the MMD statistic), or permutations of the rows of the kernel matrices (for the HSIC statistic). Moreover, the p-values were adjusted with Bonferroni corrections in order to control the familywise error rate.

Triple-negative subtype. In this first experiment, we separate the patients with respect to their cancer subtype. More specifically, we aim at distinguishing between patients with triple-negative breast cancer and those with other subtypes. Triple-negative breast cancer is especially interesting due to its high ability to provoke an immune response, or immunogenecity, among subtypes. However, triple-negative breast cancer patients typically have poor prognosis due to the lack of response to hormonal or receptor-status therapy. By better understanding the immune profiles associated with triple-negative breast cancers and the association with treatment response (i.e. overall survival), it could be possible to design targeted immunotherapies [LLJW18].

We show in Figure 7 (left) the p-values obtained with persistence diagrams, and the ones computed with NN features, for each (pair of) phenotypes. It can be seen from

[1] The cited result is actually proved for the so-called *universal kernels* but we leave this subtlety aside in the context of this work since it has no effect on our analysis.

this plot that the p-values obtained with persistence diagrams are always better than those given by NN features. We find that the NN metrics are not always significant, and this was further verified with the full NN distribution shapes. On the other hand, persistence diagrams demonstrated consistency of the p-values including CD8-involved pairs, indicating they reveal topology beyond that quantified by the NN algorithm.

Survival. In this second experiment, we now aim at distinguishing between patients that were alive at the latest follow-up after diagnosis. Although this includes causes unrelated to the breast cancer morbidity and associated treatment, such as dying of natural causes or other disease, this is still a good measure of overall disease-free survival. The corresponding p-values are displayed in Figure 7 (right). It can be seen that the p-values corresponding to persistence diagrams are in general much lower than those corresponding of NN features, especially in PD-L1 involved pairs. PD-L1 combinations are relatively rare and, as explained at the end of Section 3.1, NN features are sensitive to noise and the counting statistics on the number of phenotype pairs. Characterizing the spatial interactions of PD-L1 expression, however, would provide valuable insight into the possible immuno-repressive patterns in the tumor immune microenvironment.

We see, on the other hand, stability of persistence diagrams providing statistically significant measures. This makes diagrams a more robust descriptor than NN alone at the same statistical power. Similarly, it is clear from the distribution of values that persistence diagrams are more stable descriptors than NN features, picking up topology relating to PD-L1.

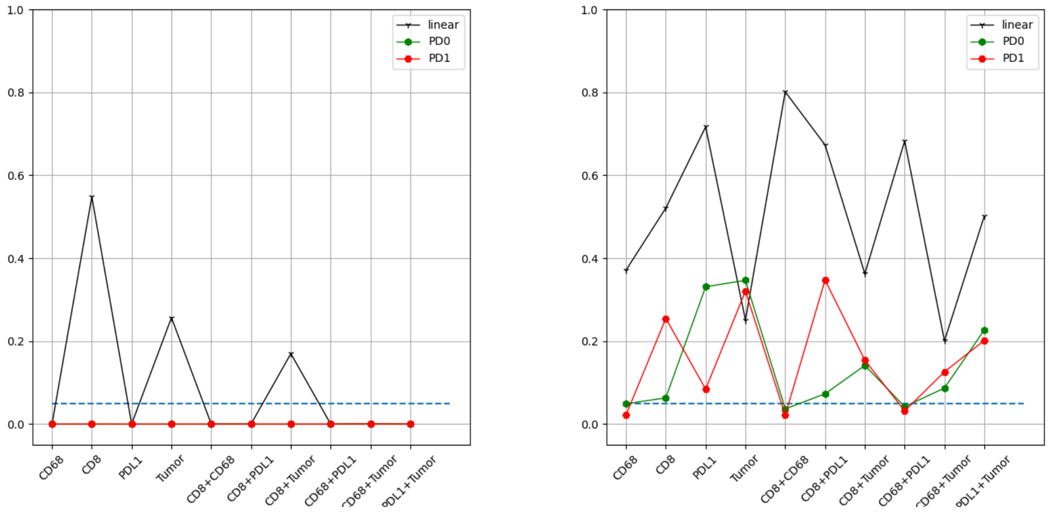

Figure 7: P-values computed for the MMD statistical test for NN features (black), 0-dimensional persistence diagrams (green) and 1-dimensional persistence diagrams (red), for different (pairs of) phenotypes. We also indicate the significance level 0.05 with a dashed blue horizontal line.

Independence. Finally, we check the independence measures (computed with the

HSIC and COCO statistics) between persistence diagrams and nearest neighbor features. We show the computed values in Figure 8. One can see that the p-values are always small, indicating some dependencies between NN features and persistence diagrams, which is expected since, even though persistence diagrams encode different information than NN features, the construction of both types of features are similar. Moreover, the COCO coefficients indicate, for each (pair of) phenotypes, how dependent persistence diagrams and NN features are.

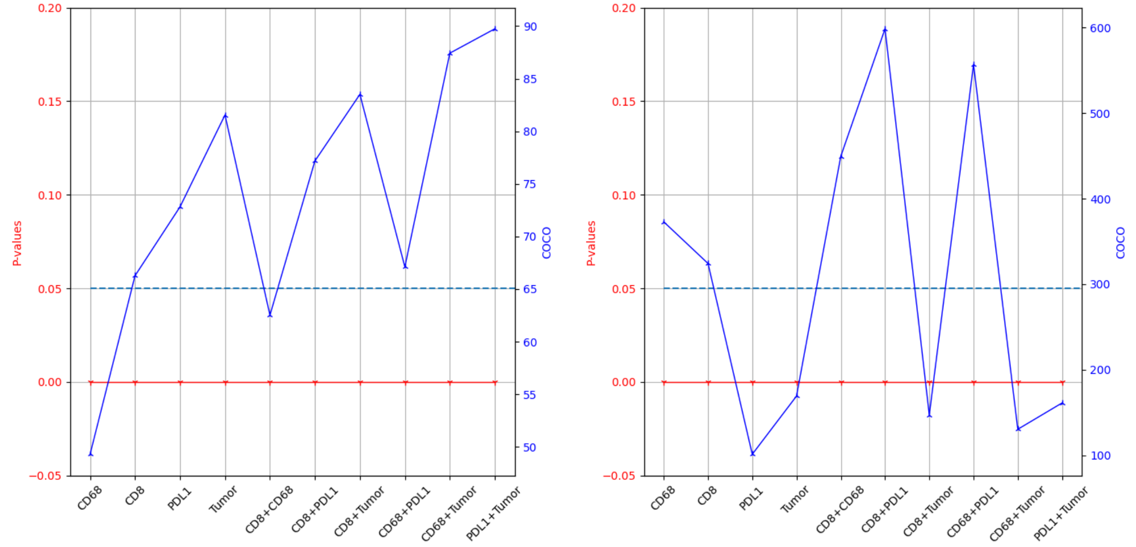

Figure 8: P-values computed for the HSIC statistical test and COCO coefficients between NN features and persistence diagrams in dimension 0 (left) and 1 (right). We also indicate the significance level 0.05 with a dashed blue horizontal line.

4 Open Questions and Future Work

We show a novel approach for the application of topological data analysis techniques to cancer characterization through the analysis of qmIF data using persistent homology. We evaluated our method on a unique cohort of 671 patients using high-throughput tumor microarrays with a median 8.5 year follow-up. Our preliminary analyses show that features derived from persistent homology between groups of patients stratified by survival and triple negative status are statistically significant and are complementary to the state-of-the-art nearest neighbor approach. This indicates that the persistent homology features can be used as a complementary biomarker.

Open questions. Our preliminary study is by no means comprehensive, and many questions remain open.

- In this article, we only focus on verifying the statistical significance of the topological signal arising from cell arrangement observed via multiplex IF. Our results indicate potential in a discriminative model (e.g., a classifier with topological features

[KHN+15, CCO17]). This work should be extended to develop classifiers, which would require careful design of the learning module and featurization of persistence diagrams.

- Our analysis only considered single phenotypes and pairs of phenotypes. However, a more complete characterization should handle interactions between more than two phenotypes, despite greatly increasing the number of persistence diagrams computed for each patient. Moreover, there is no single solution on how to combine the different stain intensities. In this work, we merely took the average between normalized stain intensities, even though it would be interesting to weight the filtrations given by stain intensities in order to take the range of stain intensity values into account. The weight coefficients could even be learned so as to avoid a brute force search, using for instance recent works on differentiability of persistence diagrams for learning [BGND+19, CNBW19, HFSC19, PSO18].
- Multiple stain intensities actually fits into the multiparameter persistence framework, see [CZ09, HOST19], where data is filtered by several filtrations at the same time. Our approach of taking linear combinations of stain intensities actually amounts to draw lines in this multiparameter space and compute usual persistence along these lines, which is the approach that is also advocated in recent works [CFK+19, LW15]. However, multiparameter persistence is a current area of research, and invariants have been obtained in recent works, at least for bifiltrations, that is, filtrations with two parameters [BCB18, BL18, CO16]. Even though they are harder to encode than persistence diagrams, it might be interesting to apply these results in our context.

References

[AEK+17] 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), 2017.

[AFH+19] Shahira Abousamra, Danielle Fassler, Le Hou, Yuwei Zhang, Rajarsi Gupta, Tahsin Kurc, Luisa F. Escobar-Hoyos, Dimitris Samaras, Beatrice Knudson, Kenneth Shroyer, Joel Saltz, and Chao Chen. Weakly-supervised deep stain decomposition for multiplex ihc images. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2019.

[BAAN17] Samantha Burugu, Karama Asleh-Aburaya, and Torsten O Nielsen. Immune infiltrates in the breast cancer microenvironment: detection, characterization and clinical implication. *Breast Cancer*, 24(1):3–15, 2017.

[BCB18] Magnus Botnan and William Crawley-Boevey. Decomposition of persistence modules. *arXiv*, nov 2018.

[BCOS16] Mickaël Buchet, Frédéric Chazal, Steve Y Oudot, and Donald R Sheehy. Efficient and robust persistent homology for measures. *Computational Geometry*, 58:70–96, 2016.

[BGND$^+$19] Rickard Brüel-Gabrielsson, Bradley Nelson, Anjan Dwaraknath, Primoz Skraba, Leonidas Guibas, and Gunnar Carlsson. A topology layer for machine learning. *arXiv*, may 2019.

[BL18] Magnus Botnan and Michael Lesnick. Algebraic stability of zigzag persistence modules. *Algebraic and Geometric Topology*, 18(6):3133–3204, oct 2018.

[BSP$^+$19] Jung Byun, Sandeep Singhal, Samson Park, IK Dae, Ambar Caban, Nasreen Vohra, Eliseo Perez-Stable, Anna Napoles, and Kevin Gardner. Transcription regulatory networks associated with luminal master regulator expression and breast cancer survival, 2019.

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

[CCO17] Mathieu Carrière, Marco Cuturi, and Steve Oudot. Sliced Wasserstein kernel for persistence diagrams. In *International Conference on Machine Learning*, volume 70, pages 664–673, jul 2017.

[CCR13] Joseph Minhow Chan, Gunnar Carlsson, and Raul Rabadan. Topology of viral evolution. *Proceedings of the National Academy of Sciences*, 110(46):18566–18571, 2013.

[CCSG$^+$09] Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas Guibas, and Steve Oudot. Proximity of persistence modules and their diagrams. In *International Symposium on Computational Geometry*, page 237, 2009.

[CdSGO16] Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. *The structure and stability of persistence modules*. Springer International Publishing, 2016.

[CFK$^+$19] René Corbet, Ulderico Fugacci, Michael Kerber, Claudia Landi, and Bei Wang. A kernel for multi-parameter persistent homology. *Computers & Graphics: X*, 2:100005, dec 2019.

[CGOS11] Frédéric Chazal, Leonidas Guibas, Steve Oudot, and Primoz Skraba. Scalar field analysis over point cloud data. *Discrete & Computational Geometry*, 46(4):743–775, 2011.

[CNBW19] Chao Chen, Xiuyan Ni, Qinxun Bai, and Yusu Wang. A topological regularizer for classifiers via persistent homology. In *International Conference on Artificial Intelligence and Statistics*, pages 2573–2582, 2019.

[CO16] Jérémy Cochoy and Steve Oudot. Decomposition of exact pfd persistence bimodules. *arXiv*, may 2016.

[CSEH07] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, jan 2007.

[CSEHM10] David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Yuriy Mileyko. Lipschitz functions have l p-stable persistence. *Foundations of computational mathematics*, 10(2):127–139, 2010.

[CZ09] Gunnar Carlsson and Afra Zomorodian. The theory of multidimensional persistence. *Discrete and Computational Geometry*, 42(1):71–93, jul 2009.

[DXLB16] Xiaofeng Dai, Liangjian Xiang, Ting Li, and Zhonghu Bai. Cancer hallmarks, biomarkers and breast cancer molecular subtypes. *Journal of Cancer*, 7(10):1281, 2016.

[EH08] Herbert Edelsbrunner and John Harer. Persistent homology-a survey. *Contemporary mathematics*, 453:257–282, 2008.

[EH10] Herbert Edelsbrunner and John Harer. *Computational topology: an introduction.* American Mathematical Soc., 2010.

[FAG$^+$19] Danielle J Fassler, Shahira Abousamra, Rajarsi Gupta, Chao Chen, Maozheng Zhao, David Paredes-Merino, Syeda Areeha Batool, Beatrice Knudsen, Luisa Escobar-Hoyos, Kenneth R Shroyer, Dimitris Samaras, Tahsin Kurc, and Joel Saltz. Deep learning-based image analysis methods for brightfield-acquired multiplex immunohistochemistry images, 2019. under review.

[FSCF16] Seth Flaxman, Dino Sejdinovic, John Cunningham, and Sarah Filippi. Bayesian learning of kernel embeddings. In *32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016)*, pages 182–191. AUAI Press, 2016.

[GBR$^+$12] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[GBSS05] Arthur Gretton, Olivier Bousquet, Alexander Smola, and Bernhard Schölkopf. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In *International Conference on Algorithmic Learning Theory (ALT 2005)*, pages 63–77. Springer-Verlag, 2005.

[GCZ$^+$13] Mingchen Gao, Chao Chen, Shaoting Zhang, Zhen Qian, Dimitris Metaxas, and Leon Axel. Segmenting the papillary muscles and the trabeculae from high resolution cardiac ct through restoration of topological handles. In *International Conference on Information Processing in Medical Imaging*, pages 184–195. Springer, 2013.

[GHS$^+$05] Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

[GMH$^+$18] Robyn Gartrell, Douglas Marks, Thomas Hart, Gen Li, Danielle Davari, Alan Wu, Zoe Blake, Yan Lu, Kayleigh Askin, Anthea Monod, et al. Quantitative analysis of immune infiltrates in primary melanoma. *Cancer immunology research*, 6(4):481–493, 2018.

[GSB$^+$05] Arthur Gretton, Alexander Smola, Olivier Bousquet, Ralf Herbrich, Andrei Belitski, Mark Augath, Yusuke Murayama, Jon Pauls, Bernhard Schölkopf, and Nikos Logothetis. Kernel constrained covariance for dependence measurement. In *International Conference on Artificial Intelligence and Statistics*, 2005.

[HFSC19] Xiaoling Hu, Li Fuxin, Dimitris Samaras, and Chao Chen. Topology-preserving deep image segmentation. In *the Thirty-third Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[HKNU17] Christoph Hofer, Roland Kwitt, Marc Niethammer, and Andreas Uhl. Deep learning with topological signatures. In *Advances in Neural Information Processing Systems*, pages 1634–1644, 2017.

[HOST19] Heather Harrington, Nina Otter, Hal Schenck, and Ulrike Tillmann. Stratifying multiparameter persistent homology. *SIAM Journal on Applied Algebra and Geometry*, 3(3):439–471, jan 2019.

[HTF03] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer-Verlag, 2003.

[KB17] Jessica Kalra and Jennifer Baker. Multiplex immunohistochemistry for mapping the tumor microenvironment. In *Signal Transduction Immunohistochemistry*, pages 237–251. Springer, 2017.

[KDS$^+$18] Lida Kanari, Paweł Dłotko, Martina Scolamiero, Ran Levi, Julian Shillcock, Kathryn Hess, and Henry Markram. A topological representation of branching neuronal morphologies. *Neuroinformatics*, 16(1):3–13, 2018.

[KHF16] Genki Kusano, Yasuaki Hiraoka, and Kenji Fukumizu. Persistence weighted Gaussian kernel for topological data analysis. In *International Conference on Machine Learning*, volume 48, pages 2004–2013, jun 2016.

[KHN$^+$15] Roland Kwitt, Stefan Huber, Marc Niethammer, Weili Lin, and Ulrich Bauer. Statistical topological data analysis - a kernel perspective. In *Advances in Neural Information Processing Systems*, pages 3070–3078, 2015.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LCO18] Théo Lacombe, Marco Cuturi, and Steve Oudot. Large scale computation of means and clusters for persistence diagrams using optimal transport. In *Advances in Neural Information Processing Systems*, pages 9770–9780, 2018.

[LKC$^+$12] Hyekyoung Lee, Hyejin Kang, Moo K Chung, Bung-Nyun Kim, and Dong Soo Lee. Persistent brain network homology from the perspective of dendrogram. *IEEE transactions on medical imaging*, 31(12):2267–2277, 2012.

[LLJW18] Zhixian Liu, Mengyuan Li, Zehang Jiang, and Xiaosheng Wang. A comprehensive immunologic portrait of triple-negative breast cancer. *Translational oncology*, 11(2):311–329, 2018.

[LSB$^+$19] Peter Lawson, Andrew B Sholl, J Quincy Brown, Brittany Terese Fasy, and Carola Wenk. persistent homology for the quantitative evaluation of architectural features in prostate cancer histology. *Scientific reports*, 9, 2019.

[LW15] Michael Lesnick and Matthew Wright. Interactive visualization of 2D persistence modules. *arXiv*, dec 2015.

[LWA$^+$17] Yanjie Li, Dingkang Wang, Giorgio A Ascoli, Partha Mitra, and Yusu Wang. Metrics for comparing neuronal tree shapes based on persistent homology. *PloS one*, 12(8):e0182184, 2017.

[LY18] Tam Le and Makoto Yamada. Persistence Fisher kernel: a Riemannian manifold kernel for persistence diagrams. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2018)*, pages 10027–10038. Curran Associates, Inc., 2018.

[MMH11] Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, dec 2011.

[NLC11] Monica Nicolau, Arnold J Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270, 2011.

[Oud15] Steve Oudot. *Persistence theory: from quiver representations to data analysis.* American Mathematical Society, 2015.

[PFCW19] Edwin Roger Parra, Alejandro Francisco-Cruz, and Ignacio Ivan Wistuba. State-of-the-art of profiling immune contexture in the era of multiplexed staining and digital analysis to study paraffin tumor tissues. *Cancers*, 11(2):247, 2019.

[PHC$^+$11] Deepti Pachauri, Chris Hinrichs, Moo K Chung, Sterling C Johnson, and Vikas Singh. Topology-based kernels with application to inference problems in alzheimer's disease. *IEEE transactions on medical imaging*, 30(10):1760–1770, 2011.

[PKS$^+$16] Lajos Pusztai, Thomas Karn, Anton Safonov, Maysa M Abu-Khalaf, and Giampaolo Bianchini. New strategies in breast cancer: immunotherapy. *Clinical Cancer Research*, 22(9):2105–2110, 2016.

[PSO18] Adrien Poulenard, Primoz Skraba, and Maks Ovsjanikov. Topological function optimization for continuous shape matching. In *Computer Graphics Forum*, volume 37, pages 13–25. Wiley Online Library, 2018.

[RCK$^+$17] Abbas Rizvi, Pablo Cámara, Elena Kandror, Thomas Roberts, Ira Schieren, Tom Maniatis, and Raul Rabadan. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nature Biotechnology*, 35(6):551–560, may 2017.

[RHBK15] Jan Reininghaus, Stefan Huber, Ulrich Bauer, and Roland Kwitt. A stable multi-scale kernel for topological machine learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[RT17] Andrew Robinson and Katharine Turner. Hypothesis testing for topological data analysis. *Journal of Applied and Computational Topology*, 1(2):241–261, dec 2017.

[SBSO16] Alexandra Signoriello, Marcus Bosenberg, Mark Shattuck, and Corey O'Hern. Modeling the spatiotemporal evolution of the melanoma tumor microenvironment. In *APS Meeting Abstracts*, 2016.

[SFL11] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

[SGS18] Carl-Johann Simon-Gabriel and Bernhard Schölkopf. Kernel distribution embeddings: universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44):1–29, 2018.

[SMC07] Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson. Topological methods for the analysis of high dimensional data sets and 3D object recognition. In *Eurographics Symposium on Point-Based Graphics*, pages 91–100, 2007.

[SMJ19] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(1):7–34, 2019.

[VJM18] Mikael Vejdemo-Johansson and Sayan Mukherjee. Multiple testing with persistent homology. *arXiv*, dec 2018.

[WCW$^+$17] Pengxiang Wu, Chao Chen, Yusu Wang, Shaoting Zhang, Changhe Yuan, Zhen Qian, Dimitris Metaxas, and Leon Axel. Optimal topological cycles and their application in cardiac trabeculae restoration. In *International Conference on Information Processing in Medical Imaging*, pages 80–92. Springer, 2017.

[WKL$^+$21] Fan Wang, Saarthak Kapse, Steven Liu, Prateek Prasanna, and Chao Chen. Topotxr: A topological biomarker for predicting treatment response in breast cancer. In *International Conference on Information Processing in Medical Imaging*, pages 386–397. Springer, 2021.

[ZC05] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.