Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Persistence diagrams as morphological signatures of cells

A method to measure and compare cells within a population

Bleile, Yossi Bokor; Koehl, Patrice; Rehfeldt, Florian

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Persistence diagrams as morphological signatures of cells: A method to measure and compare cells within a population

Yossi Bokor Bleile[1,2*], Patrice Koehl[3], Florian Rehfeldt[4]

1 Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark
2 School of Mathematics and Statistics, The University of Sydney, Sydney, NSW, Australia
3 Department of Computer Science, University of California, Davis, California, America
4 Experimental Physics I, University of Bayreuth, Bayreuth, Germany

* yossib@math.aau.dk (YBB)

Abstract

Cell biologists study in parallel the morphology of cells with the regulation mechanisms that modify this morphology. Such studies are complicated by the inherent heterogeneity present in the cell population. It remains difficult to define the morphology of a cell with parameters that can quantify this heterogeneity, leaving the cell biologist to rely on manual inspection of cell images. We propose an alternative to this manual inspection that is based on topological data analysis. We characterise the shape of a cell by its contour and nucleus. We build a filtering of the edges defining the contour using a radial distance function initiated from the nucleus. This filtering is then used to construct a persistence diagram that serves as a signature of the cell shape. Two cells can then be compared by computing the Wasserstein distance between their persistence diagrams. Given a cell population, we then compute a distance matrix that includes all pairwise distances between its members. We analyse this distance matrix using hierarchical clustering with different linkage schemes and define a purity score that quantifies consistency between those different schemes, which can then be used to assess homogeneity within the cell population. We illustrate and validate our approach to identify sub-populations in human mesenchymal stem cell populations.

Author summary

Cells are the basic unit of life. Understanding how they grow, divide, die, and change shape is of central importance in many other areas of the life sciences. In this paper, we focus on the concept of shape and, more specifically, on how to compare the shapes of two cells. We characterise this shape with the cell contour supplemented by the position of its nuclei. We use topological data analysis to define a signature of that shape, generated from its persistence diagram, a structure that reflects the relative position of the nucleus with respect to segments of the contours. We compute the distance between two cells as the Wasserstein distance between their shape signature. Using this distance, we analyse populations of cells to help identify members with unusual shapes (usually referred to as outliers) as well as sub-populations. We validate

our approach to identify sub-populations within human mesenchymal stem cell populations that are known to be heterogeneous.

1   Introduction

Cells are the basic unit of life. Understanding how they grow, divide, die, and change shape is of central importance for immunology, cancer biology, pathology, tissue and organ morphogenesis during development, as well as for many other areas in the life sciences. In this paper, we focus on the concept of shape. The shape of a cell is defined by the geometrical constraints of the space it occupies and is determined by the external boundaries and positions of the internal components. The shape is the result of the mechanical balance of forces exerted on the cell membrane by intra-cellular components and the extra-cellular environment. It is a geometric property controlled by a variety of biochemical pathways. Cell biologists study in parallel the morphology of cells (their geometry) with the regulation mechanisms that modify this morphology. These studies are benefiting from recent advances in microscopy and image processing techniques. Current microscopes provide 2D images that make it possible to study cellular shapes, or more precisely 2D projections of cellular shapes. The question remains as to how to measure and compare those shapes. This paper focusses on a new technique for performing those analyses.

Our proposed method for 2D shape comparisons is motivated by a seminal paper by Engler et al. that demonstrated that the mechanical properties (Young's elastic modulus E) of the extracellular matrix direct the differentiation of human mesenchymal stem cells (hMSCs) [1]. While up- and down-regulation of genes and transcription factors takes up to several days or even weeks, experiments focused on the first 24 hours of hMSCs after seeding on a substrate showed a significant impact of matrix rigidity on the structural formation of acto-myosin stress fibers and quantified that by an order parameter S that could be used as an early morphological descriptor of mechano-directed stem cell differentiation [2]. Although this analysis was based on the filamentous structure of the cytoskeleton and its pattern formation, we aim to use the global cell morphology, in particular, the outline of the cellular cortex in two dimensions. Importantly, the hMSCs used in all these studies are primary cells, collected from the bone marrow of human individuals, and not an immortalised cell line. This leads to an intrinsic variety of the cell population that is expected to be further impacted by potential sub-populations of bone marrow fibroblasts (roughly 5%) [3, 4]. Our aim is to see if geometry alone allows us to identify those sub-populations within a sample of cells collected from the bone marrow.

A 2D shape is defined as a domain D in $\mathbb{R}^2$, delimited by its boundary, ∂D, often referred to as the contour of D. In all our applications, we will take the contour to be a piecewise smooth or polygonal Jordan curve, that is, a simple closed curve in $\mathbb{R}^2$. There are multiple geometric representations of such 2D shapes, leading to different methods for their characterisations. We briefly review three such representations.

In the ***digital image*** representation, common to most real applications, raw data is provided in the form of 2D images (see Figure 1A). In essence, the data to be understood and compared is a collection of pixels. Traditional methods of comparing such images usually proceed in three steps. They first define a set of well-chosen landmarks or key points on the surfaces of the shapes, then assign "signatures" to these key points (coordinates in a parameterising domain), and finally determine a map maximising the correspondence of signatures (for a review, see [5]). With the increase in computing power and the large number of image data sets that are generated, these ideas are often studied in the context of deep learning, where the key points and signatures are learnt from large data sets. Deep learning has become the
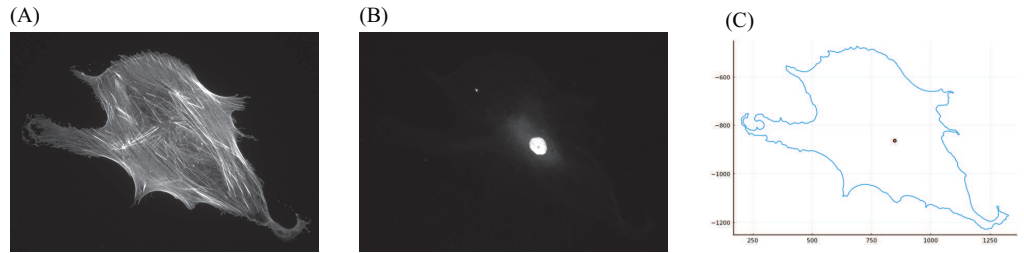
Fig 1. (A) Fluorescence microscopy image of a human mesenchymal stem cell (hMSC). (B) Fluorescence microscopy image of the corresponding nucleus. (C) Plot of the corresponding contour of that cell with the centre of the cell shown as a dot.

predominant method used in 2D image analysis (see [6] for a review of applications to the analysis of medical images). However, its applicability requires access to large data sets. In many cases, limited numbers of images are available, either because they are expensive to produce or because they model a rare phenomenon. This is the case for the stem cell images considered in this paper. In addition, deep learning remains something of a black-box procedure for classification. Cell biologists seek to understand the interplay between the geometry of a cell and the biochemical processes that are responsible for this geometry. They need a finer and more mechanistic understanding of the processes that drive shape, requiring mathematical approaches.

A second representation of 2D shapes, which we refer to as **_shape as planar contour_**, is based on the curve describing the outer boundary of the shape (see Figure 1C). This is well suited to applications focused on the geometric configuration of a shape, where factors such as the colour or grey level of the interior are not relevant or available. Methods to model the similarity between two shapes given as planar contours have been based on defining a distance between two curves in the plane. The proposed distances include the Hausdorff and Frechet distances [7]. Other techniques are based on the Poisson equation [8], integral invariants [9], and an elastic shape distance on the energy required to elastically deform one boundary contour to the other [10, 11].

Methods based on shape as planar contour do not directly consider the interior of a shape, possibly discarding relevant information. A third approach, **_shape as planar region,_** compares shapes using surface correspondences that take into account both the contour and the interior of the shape. Measures of similarity based on the distortion energy of a 2-dimensional correspondence taking one shape to another have been based on conformal [12–14] and quasi-conformal mappings [15–17]. These are of particular interest when aligning landmarks, special points of interest that lie on the boundary or in the interior of the shape. The Uniformization Theorem implies that conformal maps can be found that align up to 3 boundary landmarks in each of a pair of disk type shapes, or one in the interior and one on the boundary. Quasi-conformal maps allow the alignment of any number of landmarks [15, 17], and can also be used for shape alignment when there are holes in the interior of a shape [16]. When applied to studying cell shapes, they make it possible to take into account the positions of the nucleus, of actin filaments, and of reticulum endoplasmic in the interior of a cell, which are of special interest because they are visible in microscopy images.

Paraphrasing a recent review paper by D. Chitwood and colleagues, 'Shape is data and data is shape' [18]. As described above, shape is a signature of biological objects such as cells discussed above, that are significant for their biological functions. As such, the shape characteristics are integral parts of the data that represent these biological objects. Reversely, there is a geometric structure within data that is referred

as the shape of data. Analysing the shape of data has become an essential section of data science, known as *Topological Data Analysis*, or in short as TDA. TDA has its roots in the pioneering works of Robins [19], Edelsbrunner et al [20] and Zomorodian and Carlsson [21] in persistent homology and became popular with the publication of a landmark paper by G. Carlsson [22]. Since this paper was published, it has become ubiquitous to data science, with many applications in biology (see, for example, the review mentioned above, [18], and references therein illustrating applications in structural biology, evolution, cellular architecture, and neurobiology). TDA is particularly useful when the data are represented in the form of a graph, or network. As such, it proceeds by connecting data points to form a geometric complex structure whose topological behaviour is then used to analyse the data. Coming back to the fact that the shape is data, a shape can be characterised through TDA. Using, for example, the Euler characteristic transform to study the morphology of barley seeds [23].

In this paper, we introduce a new method for analysing the morphology of a cell that falls into the second category described above, namely with the cell represented with its contour with one additional point C, taken to be the center of mass of the cell nucleus. From TDA, we use *persistent homology* to obtain a summary of the morphological features of the cell contour. We use the persistence of sub-level sets of the radial distance function from C and compute the corresponding persistence diagram (see the next section for a primer on persistent homology applied to analysing cell contours). As the contour of each cell is a closed, non-self-intersecting curve, we know that it consists of a single connected component and a single 1-cycle. These two cycles correspond to a persistent cycle with infinite life (called *essential* cycles) in dimension 0 and dimension 1, respectively. Hence, we combine the information from these two persistent cycles by pairing the birth of the essential connected component with the birth of the essential 1-cycle. A pair of cells is then compared by computing the *2-Wasserstein distance* between their *persistence diagrams*, providing a measure of similarity between the two cells. We can then apply various clustering techniques to these similarity scores, to identify homogeneous populations of cells.

The paper is organised as follows. The next section introduces the concept of persistence homology applied to analysing the morphology of a cell, the construction of the persistence diagram of a cell contour, and the computation of the Wasserstein distance between two persistence diagrams. The Materials and Methods section gives information on the experimental data and implementations of the methods mentioned above. The Results section discusses the applications of this new method for identifying sub-populations among samples of human mesenchymal stem cells collected from bone marrow which may contain some bone marrow fibroblasts [3, 4]. We conclude with a discussion of future applications of persistence homology for comparing cell shapes.

2 Theory: persistence homology applied to analysing cell contours

2.1 Persistent Homology on Contours

Given a microscopy image of a fixed and immuno-stained cell, we use a graph G to represent the boundary in 2 dimensions. This graph is a list of ordered vertices (pixel locations), V, with edges, E, between neighbouring vertices. Note that G is connected and every vertex has degree 2, so G consists of precisely one cycle. We extract morphological information using the persistence of connected components of the sub-level sets of a radial function from the centroid of the nucleus.

For a graph G, we say that two vertices v_1, v_2 are in the same *equivalence class*, or

connected component, if there is a path γ from v_1 to v_2. For each connected component of G, we choose a representative vertex v and denote the set of vertices v' connected to v by $[v]$. We call the set $\{[v] \text{ for } v \in G\}$ the *connected components* of G.

To use persistent homology, we need to define a filtration on G.

Definition 1 (Sub-level sets and sequence of graphs). *Let f be a function from the vertices V of a graph G to $\mathbb{R}$, and fix $a \in \mathbb{R}$. The* sublevel set $G_a := f^{-1}((-\infty, a])$ *is the subgraph consisting of the set V_a of vertices v with $f(v) \leq a$ and the set of edges E_a between any pair of neighbouring vertices that are in both V_a. Note that for any*

$$a \leq b \in \mathbb{R}$$

we have

$$f^{-1}((-\infty, a]) \subseteq f^{-1}((-\infty, b]),$$

and the sub-level sets form a sequence of nested graphs.

Remark 1. The above definition of sub-level sets is cell-wise constant, rather than piecewise-linear one. The distance of a point on an edge to the centre of the function is not the standard Euclidean distance in $\mathbb{R}^2$, but instead the maximum of the distances of the two vertices. This is not an issue, as the difference in these two values is bounded.

2.2 Persistence Diagrams

Given a nested sequence of graphs $G_0 \subseteq G_1 \subseteq \ldots \subseteq G_\alpha$ (in general $G_\alpha = G$ the full graph), we can track the changes in connected components of the graphs as the filtering parameter varies. Consider some G_β, and let $C_\beta := \left\{ [v_j]^\beta \right\}_{j=1}^{n_i}$ be the set of connected components in G_β. For each connected component of G_β we choose a canonical representative vertex, namely the vertex with the lowest function value. We say that a connected component $[v_j]$ is *born* at time β if there is no vertex in $[v_j]$ it is in $C_{\beta-1}$. We say $[v_j]$ *dies* at γ if in G_γ, $[v_j]$ becomes path connected to a component born before v_j. For any pair $\beta \leq \gamma$ we obtain a map $\mathfrak{A}_\beta^\gamma : C_\beta \to C_\gamma$, which is induced by the inclusion $\iota_\beta^\gamma : G_\beta \to G_\gamma$.

Remark 2. The map $\mathfrak{A}_\beta^\gamma : C_\beta \to C_\gamma$ is obtained from the inclusion $\iota_\beta^\gamma : G_\beta \to G_\gamma$ by

$$\mathfrak{A}_\beta^\gamma ([v]) := \left[\iota_\beta^\gamma(v) \right],$$

which is a well-defined map.

The births and deaths of the connected components can be visualised in a *persistence diagram*.

Definition 2 (Persistence Diagram). *Let f be a function from a graph G to $\mathbb{R}$, and let $\mathfrak{G} = \{G_a\}_{a \in \mathbb{R}}$. Let $C = \bigcup_{a \in \mathbb{R}} C_a$ be the set of connected components across the sequence of graphs $\mathfrak{G}$. The* persistence diagram, $\mathfrak{D}(\mathfrak{G})$ *of $\mathfrak{G}$ is the multiset of points $(b_j, d_j) \in \mathbb{R}^2$, where b_j is the birth time of $[v_j] \in C$, and d_j its death time. A point with $d_j = \infty$ is called an* essential point, *and the corresponding equivalence class an* essential class.

We can also define these filtrations and persistence diagrams algebraically, including persistence modules, as in [24].

2.3 Example

The *input contour* C (see Figure 2A), with the center of the nucleus marked, forms a graph. Using the center as a reference point, we construct a *radial distance function* to the graph as follows: for vertices, we use the standard Euclidean distance to the center of the nucleus, and for edges, we take the maximum of the distances of their two endpoints. Vertices and edges whose radial distances are below a certain threshold (or 'time step'), form a sub-graph of C (Figure 2B). The *persistence diagram* (Figure 2D), captures the changes in the connected components of the sequence or filtration of subgraphs of C obtained at increasing time values.
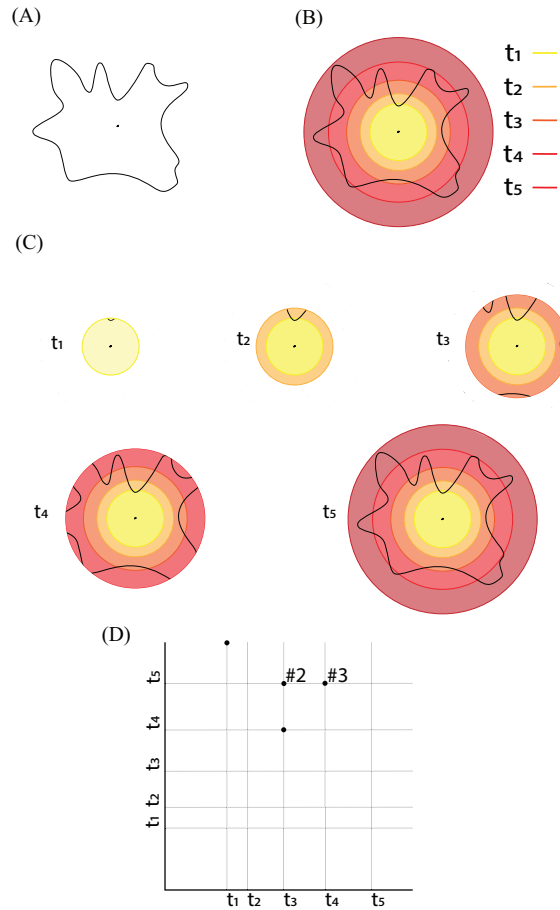
Fig 2. A) ***The input data:*** a cell contour and the center of its nucleus marked; the latter serves as the base point for the radial distance function. B) ***The radial distance function:*** The complete cell contour forms a graph G. The edges of this graph are measured relative to the cell center by computing the largest Euclidean distance between the center and the endpoints of the edge: the corresponding measure is the radial distance function with respect to the center. Edges whose radial distance function is below a given cutoff value (or 'time step'), illustrated as concentric circles around the center, define a sub-graph of the whole contour. C) ***Graph filtration:*** Examples of subgraphs for five different time steps. The different graphs obtained at increasing values of time form a filtration of the graph G. D) ***The persistence diagram*** captures the topological properties of the graph filtration. The points marked as '#2' and '#3' indicate that the corresponding points have multiplicity 2 and 3, respectively, in the persistence diagram.

The relationship between the sequence of subgraphs and the persistence diagram is as follows. At t_1, we see the birth of a single connected component, which has infinite life and corresponds to the point (t_1, ∞) in the diagram (where ∞ is represented by being at the top of the diagram). At t_2, there are no changes (no birth or death events). At t_3, 3 connected components are born. At t_4, a component born at t_3 merges with another component (and hence dies), which corresponds to the point (t_3, t_4). We also see the birth of 3 components. At t_5, we have a single connected component, formed by the remaining 2 components born at t_3 merging with the component born at t_1, corresponding to the multiplicity 2 point (t_3, t_5), and all 3

components born at t_4 merge with the original component as well, corresponding to the multiplicity 3 point (t_4, t_5).

As a multi-set of points, the persistence diagram is

$$\mathcal{D} = \{(t_1, \infty), (t_3, t_4), (t_3, t_5), (t_3, t_5), (t_4, t_5), (t_4, t_5), (t_4, t_5)\},$$

and, since we are only considering the connected components, we call this a *dimension* 0 persistence diagram.

As we are using graphs to represent each contour, we can also consider the information captured by the cycles in the subgraph filtration. Each contour is a simple, closed curve in $\mathbb{R}^2$, and hence the corresponding graph G contains a single cycle. Furthermore, this cycle appears only in the filtration when the *last* vertex appears. While it is an important descriptor of the *size* of the contour, it is inefficient to capture this information in a *dimension* 1 persistence diagram. Hence, we modify our dimension 0 diagram as follows, so that we capture this information: we pair the birth of the essential class in dimension 0 with the birth of the essential class in dimension 1. In this case, the set of points in the persistence diagram becomes

$$\mathcal{D} = \{(t_1, t_5), (t_3, t_4), (t_3, t_5), (t_3, t_5), (t_4, t_5), (t_4, t_5), (t_4, t_5)\}.$$

Remark 3. Readers familiar with persistent homology and persistence diagrams will notice that this is a nonstandard modification. Due to the nature of the contours, performing this *essential pairing* allows us to more efficiently represent and compare the topological descriptors.

2.4 Comparing two persistence diagrams using the Wasserstein distance

A persistence diagram provides a summary of the changes in the connected components as we progress along the sequence of graphs. Let us consider two sequences of graphs

$$\mathfrak{G}^1 = G_0^1 \to G_1^1 \to \ldots G_{\alpha_1}^1$$

and

$$\mathfrak{G}^2 = G_0^2 \to G_1^2 \to \ldots G_{\alpha_2}^2,$$

corresponding to two cell contours, with their associated persistence diagrams $D_1 = \mathfrak{D}(\mathfrak{G}^1)$, $D_2 = \mathfrak{D}(\mathfrak{G}^2)$. We define the distance between the cell contours as the distance between D_1 and D_2, where the distance is the *Wasserstein distance*, defined below.

Imagine that there are N farms that serve N markets, and assume balance, that is, that each farm produces enough fruits and vegetables as needed by one market. A company in charge of the distribution of the produce from the farms to the market will take into account the individual cost of transport from any farm to any market to find an 'optimal transportation plan', namely an assignment of farms to markets that leads to a minimal total cost for the transport. The seemingly simple problem can be traced back to the work of Monge in the 1780s [25]. What makes it so interesting is that its solution includes two essential components. First, it defines the assignment between farms and markets, enabling the registration between those two sets. Second, and more relevant to us, it defines a distance between the set of farms and the set of markets, with such distance being referred to as the Monge distance, the Wasserstein

distance, or the earth mover's distance, depending on the field of applications. Formally, if F is the set of farms and M the set of markets, and if we define $C(i,j)$ the cost of transport between farm i and market j, the assignment problem refers to finding a bijection f between F and M that minimises

$$U = \sum_{i \in F} C(i, f(i)). \tag{1}$$

Note, f can be seen as a permutation of $\{1, \ldots, N\}$. As mentioned above, the optimal U_{min} is a distance between F and M. This is the distance we use to compare two cell contours based on their persistence diagram.

As described above, a persistence diagram is defined by a set of points. Let S_1 (resp. S_2) be the set of points associated with D_1 (resp. D_2):

$$S_1 = \{X_1, \ldots, X_N\}$$
$$S_2 = \{Y_1, \ldots, Y_N\}$$

Note that we assume first that the two sets have the same number of points. We define the cost matrix C be to a power of the Euclidean distance, i.e.,

$$C(x_i, y_j) = ||x_i - y_j||^p$$

The p-Wasserstein distance between S_1 and S_2 is then:

$$W_p(S_1, S_2) = \left(\min_f \sum_{x_i \in S_1} ||x_i - f(x_i)||^p \right)^{1/p}$$

The formalism defined above assumes that the two sets of points S_1 and S_2 considered have the same size, that is, there are as many points in D_1 as there are points in D_2. There is no reason that this is the case. In the more general case, S_1 contains N_1 points and S_2 contains N_2, with $N_1 > N_2$, without loss of generality. This problem, however, can easily be reduced to the balanced case presented above by adding $N_1 - N_2$ pseudo, or 'ghost' points in S_2 that the two corresponding sets have the same cardinality. The distance between a point is S_1 and one of these pseudo-points can be chosen arbitrarily. One option is to position the "ghost" points on the diagonal of D_2.

In the following, we will use the 2-Wasserstein distance to compare two cell contours via their persistence diagrams.

3 Materials and Methods

3.1 Human Mesenchymal Stem Cells

Adult human mesenchymal stem cells (hMSCs) were purchased from Lonza (catalogue $\#PT - 2501$) and cultured in low glucose DMEM (Gibco, $\#1885 - 023$) supplemented with 10% FBS (Sigma-Aldrich, Ref. $F7524$), and 1% penicillin/streptomycin (Gibco, $\#15140122$) in regular tissue culture treated flasks (greiner Bio-One, $75cm^2$, $\#658175$) at $37°$ C and 5.0% CO_2. Cells were kept subconfluent at low density all the time and passaged and split every two or three days using trypsin incubation of 3 min for detachment after a washing step with PBS (Gibco, $\#14190144$). Cells were seeded on ibidi μ-Dishes (35 mm, high, ibiTreat, Cat.No: $\#81156$) at a density of 500 cells cm^{-1} to maintain a sufficient number of isolated cells for observation and grown for 24 hours under identical culture conditions. The cells were then washed once with PBS and

chemically fixed for 5min in a 10% solution of formaldehyde (Sigma-Aldrich, 252549) in PBS. Next, cells were permeabilized with TritonX (Sigma-Aldrich, T 9284) and extensively washed with PBS. Filamentous actin was stained using fluorescent Phalloidin-Atto 550 (ATTO-TEC GmbH, AD $550-81$) and the nucleus was visualised using a DNA-intercalating dye (Invitrogen, Hoechst #33342).

3.2   Unbiased Microscopy

The fixed cells were imaged on an inverted fluorescence microscope (Zeiss AxioObserver, Oberkochen, Germany) using a 20x objective (Zeiss, Plan-Neofluar, 440340-9904) and recorded by a sCMOS camera (Andor Zyla, 4.2P USB3.0) using two filter sets (blue (Zeiss Filterset 49) and red (AHF, F46-008)) for the stained nucleus and actin, respectively. For unbiased data acquisition, the samples were inspected using the nucleus channel first and selecting cells that were isolated (no other nucleus in the field of view) and had a healthy-looking nondeformed nucleus. Multiple nuclei, oddly shaped nuclei as well as any oddly shaped nuclei were excluded to avoid recording cell outlines from abnormal cells. Subsequently, the actin channel of the cell was recorded to complete the data set for each cell. In this way, three individual data sets were recorded from three individual ibidi μ-Dishes.

3.3   Image Processing and Contour Generation

We used the FilamentSensor2.0 tool [26] to perform the image processing and extract the contour of each cell. Here, we used the features 'Include Area-Outline' to export the contour from the binarized image of the cells. The *center* of the cell is obtained from the *center of mass* from the aligned microscopy image of the nucleus. Here, we thresholded the nucleus in Fiji [27] using the 'Otsu' method, before outlining it and determining the x- and y-coordinates of the centre of mass.

3.4   Contour Analysis: computing the distance between 2 cells

After extracting the contour from each image and identifying the centre of the nucleus, we convert it to the graph representation G. Recall that every vertex in G is of degree 2, and G contains a single cycle. Let $V = \{v_i\}_{i=1}^n$ be the set of vertices of G, ordered clockwise around the contour. Then every edge e of G is of the form (v_i, v_{i+1}), where $v_{n+1} = v_1$. Before we obtain our sequence of graphs $\mathfrak{G}$, We *clean* our graph representation G of C by replacing any set of consecutive edges $\{(v_i, v_{i+1}), \dots, (v_{j-1}, v_j)\}$ which are colinear with the edge (v_i, v_j) and removing the vertices v_k for $i < k < j$.

Remark 4. Consider a contour C, and let G be the original graph representation and G' the graph after it has been cleaned. As the metrics on the edges of G, G' are defined as the maximum of the values on the 2 vertices, the sequences of graphs $\mathfrak{G}$ and $\mathfrak{G}'$ generated by these metrics on G and G' respectively will have different topological features. In particular, connected components may be born *later*, by the removal of vertices that are closer to the base point of the radial distance function. These changes in values are bounded, and hence, by the stability of persistence diagrams [28], the distance between the respective persistence diagrams is also bounded. Although it is possible to generate contours where this cleaning process leads to large bounds on the distance between the persistence diagrams, the geometric features that lead to this are not of concern in our application. Hence, we prioritise computational efficiency and proceed with the cleaned graphs.

Working with the cleaned graph G_X for each cell X, we filter G_X (see Definition 1 and Section 2.3), and obtain a persistence diagram D_X (Definition 2). Then we construct a distance matrix M, using the 2-Wasserstein distance between the persistence diagrams D_X, D_Y as the distance between two cells X, Y.

3.5 Clustering cells based on their contour

Clustering is the task of regrouping cells such that those that belong to the same group, referred to as a cluster, are more similar to each other than to those in other clusters. The similarity between two cells is set to be the 2-Wasserstein distance between the persistence diagrams of their contours (see above). The clustering of the cells is then performed using the agglomerative hierarchical clustering analysis, or HCA. The is a bottom-up approach in which each cells starts in its own cluster, and pairs of clusters are merged iteratively until all cells belong to the same cluster. The whole procedure defines a clustering tree. While the distance between two cells is clearly defined above, a key element is to define the distance between two clusters. When two clusters A and B are sets of elements, the distance between A and B is then defined as a function of the pairwise distances between their elements. Four common choices of linkage are:

- ***Average linkage***: the distance between two clusters is the arithmetic mean of all the distances between the objects of one and the objects of the other:

$$d(A,B) = \sum_{a \in A} \sum_{b \in B} \frac{d(a,b)}{|A||B|}$$

  where $|\cdot|$ stands for cardinality. Average linkage, also called UPGMA, is the default linkage for most HCA implementations.

- ***Single linkage***: the distance between two clusters is the smallest distance between the objects in one and the objects in the other.

$$d(A,B) = \min\{d(a,b), a \in A, b \in B\}$$

- ***Complete linkage***:the distance between two clusters is the largest distance between the objects in one and the objects in the other.

$$d(A,B) = \max\{d(a,b), a \in A, b \in B\}$$

- ***Ward's linkage*** accounts for the variances of the clusters to be compared. For a cluster A, the variance $SSE(A)$ is defined as:

$$SSE(A) = \sum_{a \in A} d(a, m(A))^2$$

  where d is the underlying distance used to compare two objects and $m(A)$ is either the centroid (if it can be computed) or medioid of the cluster (the medioid is the point in A that has the least total distance to the other points in A). The Ward distance between two clusters A and B is then:

$$d(A,B) = SSE(A \bigcup B) - (SSE(A) + SSE(B))$$

The choice of the linkage can have a significant influence in the clustering found by HCA: for example, simple linkage only looks locally at cluster distance and as such may lead to elongated clusters, while reversely complete linkage will have a tendency

to generate more compact clusters. There is no consensus as to which linkage to use for a specific data set; this is, in fact, an active area of research.

To avoid possible biases associated with the choice of linkage, we will use all four options in our analyses, performing HCA with [29]. However, this requires a way to compare the results of one option with the others. We chose our own concept of purity to perform such a comparison, defined as follows. Let C_1 be one cluster identified with HCA with a linkage method L_1. It is possible that C_1 may not be identified as its own cluster within the tree T_2 generated with another linkage method L_2. To assess how well T_2 recognises C_1, we follow the following algorithm:

1) We choose first a seed, S_1, i.e. an object that belongs to C_1. We initialise a list of obects $O = \{S_1\}$.

2) We identify the leaf of T_2 corresponding to S_1, and add to the list O the object that has the same parent P_1 in T_2 as S_1.

3) We find the parent P_2 of P_1 and add to O all objects that are in the sub tree of T_2 starting from P_2. Wet set $P_1 \leftarrow P_2$.

4) We repeat step 3 until O contains all objects in C_1

If the results with the linkage L_2 map exactly to the results with the linkage L_1, O will be equal to C_1. However, in general, O will be bigger because it will include objects that are found by L_2 to be similar to objects in C_1 that were not identified by L_1. The *purity* $P(C_1/L_2)$ of C_1 with respect to L_2 is then defined as:

$$P(C_1/L_2) = \frac{N - |O|}{N - |C_1|} \tag{2}$$

where $|\cdot|$ stands for cardinality and N is the total number of objects. Note that P is between 0 and 1. The closer P is to one, the more consistent the two linkage strategies L_1 and L_2 are with respect to C_1.

4 Results and discussion

With the advent of imaging techniques associated with advanced microscopes, cell biology has become quantitative. It is now common to study even large populations of cells by analysing their morphological features captured in an image. For example, those morphological features may be measured from two populations of the same cell types, with one population treated with chemical or physical constraints, while the other is not treated and serves as a control population. The effects of the treatment are then quantified by measuring changes in the features in the two populations (see, for example, [30–33]). Identifying which morphological feature is relevant and measuring those features in the images are fields of study by themselves (see [33] for a review). However, there are two other main difficulties that cannot be ignored in such studies. First, as with any experimental techniques, there are possible artefacts coming from the sample itself (dead cells, cells undergoing apoptosis, dividing cells, etc.), the cell-fixing process and subsequent staining, or even the imaging and/or image processing steps of the analysis. Detecting cells that were affected by such artefacts, usually referred to as *outlier cells*, is a time-consuming process if performed manually, especially with large populations of cells, and might sometimes be subjectively influenced by the human experts. Second, the population of cells itself may be heterogeneous (e.g. primary cells collected from a patient), leading to *sub-populations*. In this section, we report how our method for comparing the shapes

of hMSC cells using persistence homology applied to the cell contours can help identify both unusual cell shapes as well as possible sub-populations. hMSC cells are known to exist as heterogeneous populations (see, for example, [34]).

We analysed one set of hMSCs, $X1$, with the experimental setup and analysis pipeline described in Section 3. The whole procedure and results are discussed in Section 4.1.

4.1 $X1$

The set $X1$ consists of 136 cells. These cells have already been selected based on manual inspection, as described in Section 3.2. To further analyse the homogeneity of this set of cells, we computed all pairwise distances between the cell contours using the persistence homology technique described above. The corresponding distance matrix is visualised as a heat map in Figure 3. The column/row of mostly bright yellow suggests that there is one cell that differs significantly from the others. This cell is shown Figure 4. Clearly, this cell is oddly shaped: it is long and thin, with three long filipods, significantly different from the expected shape of a hMSC (see Figure 1 and Figure 7). Such a shape is usually considered an outlier.
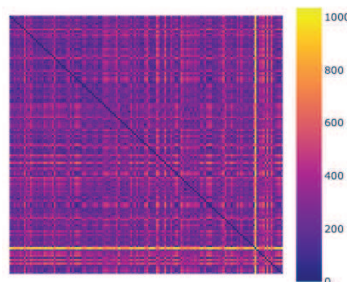

Fig 3. Heat map of the distance matrix for $X1$. There is a cell that has distinctly higher than average distances to the other cells, indicated by the row/column of mostly bright yellow. Generated with [35].
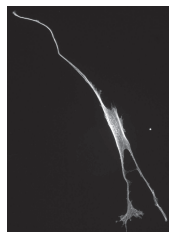

Fig 4. Image of the unusual cell shape (`X1-031`) identified in Figure 3 (image processed using [36].)

We perform clustering of cell contours in X1 using the Wasserstein distance between their associated persistence diagram, in the presence Figure 5, and in the absence Figure 6) of the 'outlier' `X1-031` identified above. We used HCA, with four different linkages: average, complete, single, and Ward.
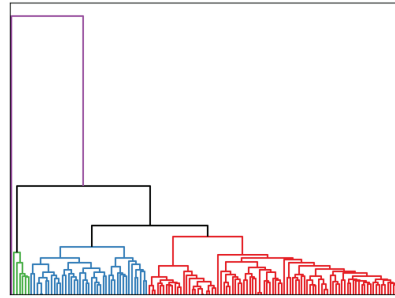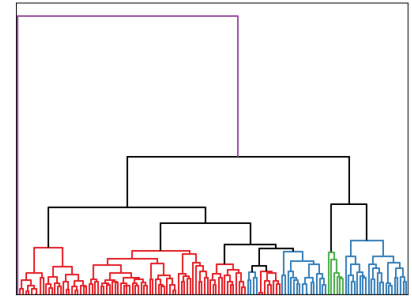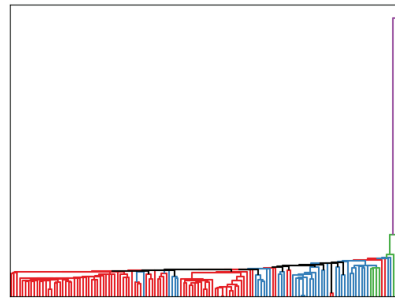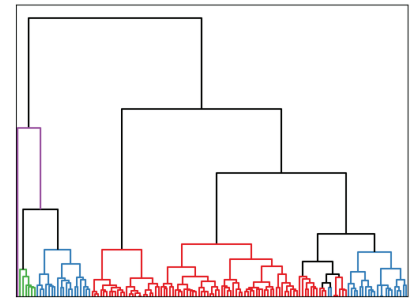
Fig 5. Dendrograms for $X1$, the colours correspond to 4 clusters obtained using average linkage. (A) Average linkage. (B) Complete linkage. (C) Single linkage. (D) Ward linkage. In each of these, there is an outlier, with the corresponding leaf coloured purple. Generated with [37, 38].

As expected, cell `X1-031` is identified as its own cluster with all four linkages in Figure 5. This cell has a unique shape that differentiates it from other hMSC cells. Although there are many possible reasons for this behaviour, `X1-031` is considered an outlier.
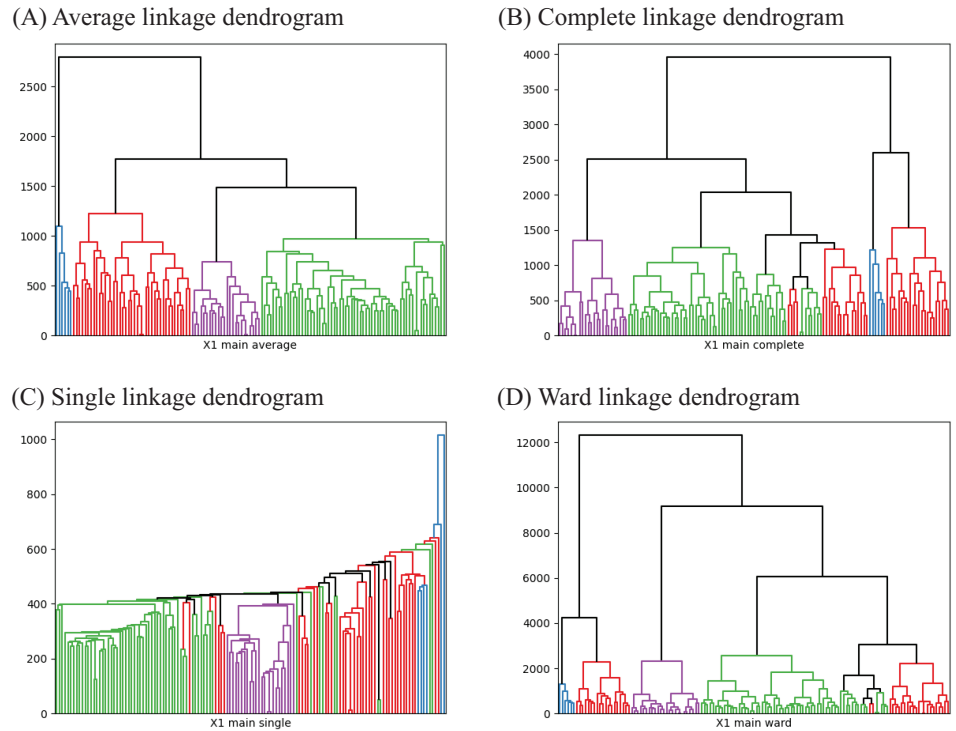
Fig 6. Dendrograms for $X1$ main, the colours correspond to 4 clusters obtained using average linkage. (A) Average linkage. (B) Complete linkage. (C) Single linkage. (D) Ward linkage. In each of these, there is a consistent sub-population, coloured purple. Generated using [37, 38].

The clustering of the set X1 with this outlier removed identifies subgroups among X1. However, those subgroups seem to differ under different choices of the linkage for HCA (Figure 6). This behaviour is not unexpected, as different linkage schemes capture different geometries for the cluster (see Section 3.5). It is common to focus on only one linkage scheme, usually the average linkage, and ignore the others. Our approach is different. We use all four linkages and assess their consistency, as illustrated in Figure 6. We start with the average linkage scheme and cut the associated dendrogram to get four clusters. These four clusters are referenced as A (in red), with ($n = 86$) elements, B (in blue) with ($n = 7$) elements, C (in green, ($n = 22$)), and D (in purple, ($n = 24$)). We then consistently colour the dendrograms for all linkage schemes based on those clusters A, B, C, and D. As expected, there are differences. However, some consistencies are observed. For example, we note that cluster D (in purple) is grouped together across all 4 linkage schemes. To confirm this visual consistency, we computed a purity score (see Section 3.5) of the clusters obtained with the average linkage in all four linkage schemes. The purity score quantifies how 'pure' a group of objects is within a dendrogram. It is computed by first identifying the subtree within the dendrogram that contains all objects within that group. If this subtree only contains this group, it is deemed pure and the purity score is set to 1. If instead this subtree contains other objects, its purity is reduced. When the subtree is the whole tree, the purity score is reduced to 0. The purity scores of clusters A to D are reported in Table 1, while examples of cells for each clusters are shown in Figure 7.

Linkage	Cluster			
	A (red, $n = 86$)	B (blue, $n = 7$)	C (green, $n = 22$)	D (purple, $n = 24$)
average	1.0	1.0	1.0	1.0
complete	0.0	1.0	0.732	1.0
single	0.021	0.0	0.056	1.0
ward	0.0	1.0	0.690	1.0

Table 1. Purity score of the 4 clusters obtained with the average linkage for $X1$ main, see Figure 6. The colour and size of each cluster is in parentheses.

As mentioned above, cluster D (purple) is visually homogeneous within all four linkage schemes: this is confirmed as its purity scores remain equal to 1. Cells in this cluster have compact shapes and a prominent nucleus, as expected from cells that have been plated on glass. The same types of cells was distinguished as a sub-population FC by Haaster *et al.* [34]. In contrast, cluster A (red) is much less consistent within the different linkage schemes, with purity scores close to 0 (with the obvious exception of the average linkage). Visually, cells belonging to cluster A are more heterogeneous, with a star-shaped or a triangular shape (first row of Figure 7). This group of cells maps with the sub-population RS identified by Haaster *et al.*. Cells belonging to cluster B are significantly more elongated. Their purity score is high with the exception of the single linkage scheme but this could just be anecdotic as there are only 7 cells in this cluster. They may correspond to elongated, fibroblastic-like, spindle-shaped cells, identified as SS cells by Haaster *et al.*. Cells in cluster C are mostly compact, similar to those in cluster D, but usually bigger. The purity scores of cluster C are close to 1, indicating that they form a group with homogeneous shapes. They were likely identified as belonging to the sub-population FC by Haaster *et al.*.
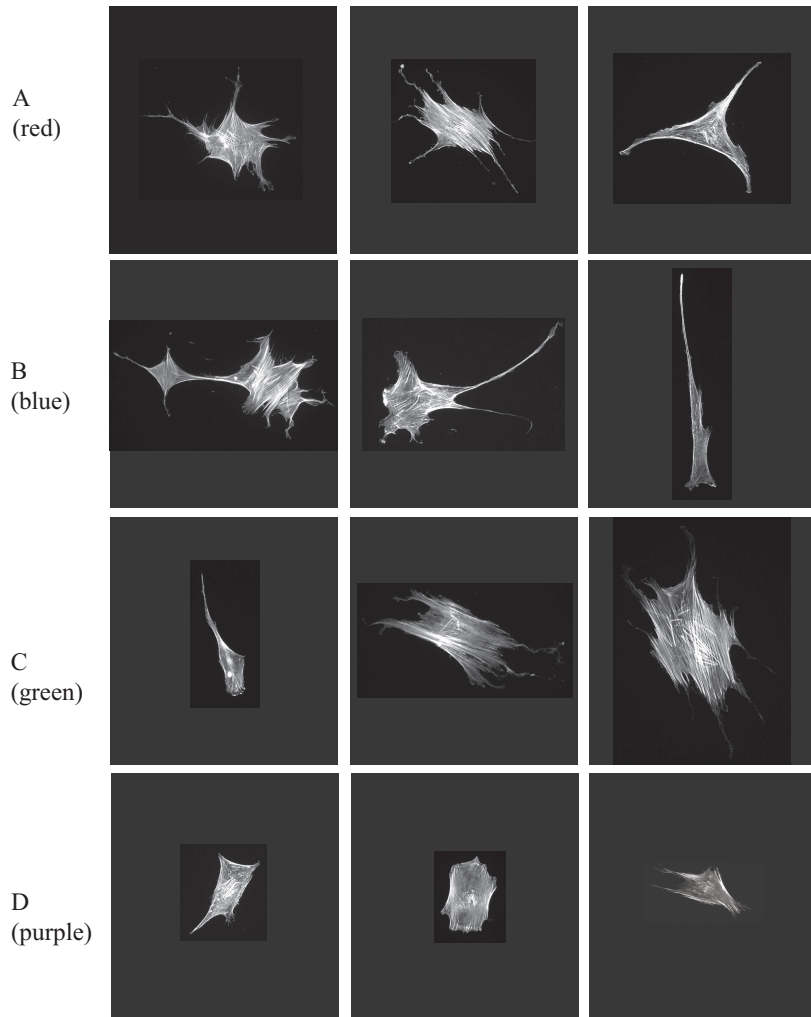
Fig 7. Example cells from each cluster of the set $X1$ (after removal of the outlier, see text for details). Those clusters are identified with HCA and average linkage scheme (see Figure 6). All cell images are shown at the same magnification level. Images were processed using [36].

5 Conclusion

Cell biologists commonly study in parallel the morphology of cells with the regulation mechanisms that affect this morphology. In the case of stem cells for example, the shapes they assumed when plated on substrate with different rigidity are expected to define morphological descriptors of mechano-directed differentiation. The heterogeneous nature of cell population is, however, a major difficulty when studying cell shape based on images from images from digital microscopes. It is common to manually assess first all the images associated with a population of cells under study in order to identify "outliers", i.e. cells with unusual shapes that raise questions on their nature (i.e. these cells could be associated with contamination) or on the presence of experimental artefacts. The aim of the present study was to propose an alternative, automated method to help with this manual assessment. We have developed a new method for analysing cell shapes that is based on three elements:

- **A description of cell shapes using persistence homology**. The shape of a cell is defined from its contour and the position of its nucleus. We compute a filtration of the edges defining the contour, using the radial distance to the nucleus as a filter. This filtration is used to define a persistence diagram that serves as a signature of the cell contour.

- **A distance between two cells**. This distance is the Wasserstein distance between the persistence diagrams of their contours.

- **A measure of homogeneity of cell subgroups**. We perform hierarchical clustering on cell shapes using the distance defined above, with four different linkage schemes. We define a purity score for subgroups of cells within the dendrograms associated with those clustering. This purity score reflects homogeneity.

We have tested our method on hMSC cells that are known to be heterogeneous. We have shown that it automatically identifies unusual cells that can then be deemed outlier or not, as well as sub-populations that are consistent with previous analyses of sub-populations of hMSCs [34].

There are many morphometric parameters that could have been included to complement our topological data analysis, such as cell area, aspect ratios, ellipticity, curvature of the contours, ... It is our intent to complement our analyses with a more comprehensive set of morphological signatures of cell shapes. In addition, all those parameters, including the persistence diagrams presented in this paper, are computed based on 2D images. Cells are 3D objects and ultimately should be studied as such. The concepts we have introduced in this paper extend to the analyses of 3D surfaces. We will explore this in further studies.

References

1. Engler AJ, Sen S, Sweeney HL, Discher DE. Matrix Elasticity Directs Stem Cell Lineage Specification. Cell. 2006;126:677–689. doi:10.1016/j.cell.2006.06.044.

2. Zemel A, Rehfeldt F, Brown AEX, Discher DE, Safran S. Optimal matrix rigidity for stress-fibre polarization in stem cells. Nature Physics. 2010;6:468-473. doi:10.1038/NPHYS1613.

3. Costa LA, Eiro N, Fraile M, Gonzalez LO, Saá J, Garcia-Portabella P, et al. Functional heterogeneity of mesenchymal stem cells from natural niches to culture conditions: implications for further clinical uses. Cellular and Molecular Life Sciences. 2021;78:447–467. doi:10.1007/s00018-020-03600-0.

4. Phinney DG. Functional heterogeneity of mesenchymal stem cells: Implications for cell therapy. Journal of Cellular Biochemistry. 2012;113:2806–2812. doi:10.1002/jcb.24166.

5. Chen L, Rottensteiner F, Heipke C. Feature detection and description for image matching: from hand-crafted design to deep learning. Geo-spatial Information Science. 2021;24:58–74. doi:10.1080/10095020.2020.1843376.

6. Zhou SK, Greenspan H, Davatzikos C, Duncan JS, Van Ginneken B, Madabhushi A, et al. A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises. Proceedings of the IEEE. 2021;109:820–838. doi:10.1109/JPROC.2021.3054390.

7. Alt H. The Computational Geometry of Comparing Shapes. In: Albers, S., Alt, H., Näher, S., editors. Efficient Algorithms. Lecture Notes in Computer Science. 2009;5760:235–248. doi:10.1007/978-3-642-03456-5_16

8. Gorelick L, Galun M, Sharon E, Basri R, Brandt A. Shape Representation and Classification Using the Poisson Equation. IEEE Trans Pattern Anal Mach Intell. 2006;28:1991–2005. doi:10.1109/TPAMI.2006.253.

9. Manay S, Cremers D, Hong BW, Yezzi AJ, Soatto S. Integral Invariants for Shape Matching. IEEE Trans Pattern Anal Mach Intell. 2006;28:1602–1618. doi:10.1109/TPAMI.2006.208.

10. Bauer M, Bruveris M, Michor PW. Overview of the geometries of shape spaces and diffeomorphism groups. Journal of Mathematical Imaging and Vision. 2014;50:60–97. doi:10.1007/s10851-013-0490-z.

11. Dogan G, Bernal J, Hagwood CR. A Fast Algorithm for Elastic Shape Distances Between Closed Planar Curves. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015;4222–4230. doi:10.1109/CVPR.2015.7299050.

12. Sharon E, Mumford D. 2D shape analysis using conformal mapping. International Journal of Computer Vision. 2006;70:55–75. doi:10.1007/s11263-006-6121-z.

13. Feiszli M, Mumford D. Shape representation via conformal mapping. In: Bouman CA, Miller EL, Pollak I, editors. Computational Imaging V. 2007;6498:133–141. doi:10.1117/12.716028.

14. Lui LM, Zeng W, Yau ST, Gu X. Shape analysis of planar objects with arbitrary topologies using conformal geometry. In: Daniilidis K, Maragos P, Paragios N, editors. Computer vision - ECCV 2010;6315:672–686. doi:10.1007/978-3-642-15555-0_49.

15. Jones GW, Mahadevan L. Planar morphometry, shear and optimal quasi-conformal mappings. Proc R Soc A. 2013;469:20120653. doi:10.1098/rspa.2012.0653.

16. Nian X, Chen F. Planar domain parameterization for isogeometric analysis based on Teichmüller mapping. Computer Methods in Applied Mechanics and Engineering 2016;311:41–55. doi:10.1016/j.cma.2016.07.035.

17. Choi GPT, Mahadevan L. Planar morphometrics using Teichmüller maps. Proc Math Phys Eng Sci. 2018;474:20170905. doi:10.1098/rspa.2017.0905.

18. Amézquita EJ, Quigley MY, Ophelders T, Munch E, Chitwood DH. The shape of things to come: Topological data analysis and biology, from molecules to organisms. Developmental Dynamics. 2020;249:816–833. doi:10.1002/dvdy.175.

19. Robins V. Computational Topology at Multiple Resolutions: Foundations and Applications to Fractals and Dynamics. University of Colorado; 2000.

20. Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. Discrete & Computational Geometry. 2002;28:511–533. doi:10.1007/s00454-002-2885-2.

21. Zomorodian A, Carlsson G. Computing persistent homology. In: Proceedings of the twentieth annual Symposium on Computational Geometry; 2004;347–356. doi:10.1145/997817.997870.

22. Carlsson G. Topology and data. Bulletin of the American Mathematical Society. 2009;46:255–308. doi:10.1090/S0273-0979-09-01249-X.

23. Amezquita E, Quigley M, Ophelders T, Landis JB, Koenig D, Munch E, et al. Measuring hidden phenotype: quantifying the shape of barley seeds using the Euler characteristic transform. in silico Plants. 2022;4:1:15. doi:10.1093/insilicoplants/diab033

24. Chazal F, de Silva V, Glisse M, Oudot S. The Structure and Stability of Persistence Modules. Springer Briefs in Mathematics. 2016.

25. Monge G. Mémoire sur la théorie des déblais et des remblais. Histoire de l'Académie Royale des Sciences et Mémoires de Mathématique et de Physique tirés des registres de cette Académie. 1781;1784:666–704.

26. Hauke L, Primeßnig A, Eltzner B, Radwitz J, Huckemann SF, Rehfeldt F. FilamentSensor 2.0: An open-source modular toolbox for 2D/3D cytoskeletal filament tracking. PLOS ONE. 2023;18:e0279336:1–21. doi:10.1371/journal.pone.0279336.

27. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an open-source platform for biological-image analysis. Nature Methods. 2012;9:676–682. doi:10.1038/nmeth.2019.

28. Skraba P, Turner K. Wasserstein Stability for Persistence Diagrams. arXiv: Algebraic Topology. 2020. doi:10.48550/arXiv.2006.16824.

29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830. doi:10.5555/1953048.2078195.

30. Feng Y, Mitchison TJ, Bender A, Young DW, Tallarico JA. Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds. Nature Reviews Drug Discovery. 2009;8:567–578. doi:10.1038/nrd2876.

31. Pennisi E. 'Cell painting' highlights responses to drugs and toxins. Science. 2016;352:877–878. doi:10.1126/science.352.6288.877.

32. Caicedo JC, Singh S, Carpenter AE. Applications in image-based profiling of perturbations. Current Opinion in Biotechnology. 2016;39134–142. doi:10.1016/j.copbio.2016.04.003.

33. Caicedo JC, Cooper S, Heigwer F, Warchal S, Qiu P, Molnar C, et al. Data-analysis strategies for image-based cell profiling. Nature methods. 2017;14:849–863. doi:10.1038/nmeth.4397.

34. Haasters F, Prall WC, Anz D, Bourquin C, Pautke C, Endres S, et al. Morphological and immunocytochemical characteristics indicate the yield of early progenitors and represent a quality control for human mesenchymal stem cell culturing. Journal of anatomy. 2009;214:759–767. doi:10.1111/j.1469-7580.2009.01065.x.

35. Inc PT. Collaborative data science; 2015. Available from: `https://plot.ly`.

36. Rasband WS. ImageJ; 1997-2018. Available from: `https://imagej.nih.gov/ij/`.

37. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020;17:261–272. doi:10.1038/s41592-019-0686-2.

38. Hunter JD. Matplotlib: A 2D graphics environment. Computing in Science & Engineering. 2007;9(3):90–95. doi:10.1109/MCSE.2007.55.