scientific reports

OPEN

Persistent homology analysis distinguishes pathological bone microstructure in non-linear microscopy images

Ysanne Pritchard[1✉], Aikta Sharma[4,6], Claire Clarkin[4], Helen Ogden[1,3,5], Sumeet Mahajan[2,3,7] & Rubén J. Sánchez-García[1,3,5,7]

We present a topological method for the detection and quantification of bone microstructure from non-linear microscopy images. Specifically, we analyse second harmonic generation (SHG) and two photon excited autofluorescence (TPaF) images of bone tissue which capture the distribution of matrix (fibrillar collagen) structure and autofluorescent molecules, respectively. Using persistent homology statistics with a signed Euclidean distance transform filtration on binary patches of images, we are able to quantify the number, size, distribution, and crowding of holes within and across samples imaged at the microscale. We apply our methodology to a previously characterized murine model of skeletal pathology whereby vascular endothelial growth factor expression was deleted in osteocalcin-expressing cells (OcnVEGFKO) presenting increased cortical porosity, compared to wild type (WT) littermate controls. We show significant differences in topological statistics between the OcnVEGFKO and WT groups and, when classifying the males, or females respectively, into OcnVEGFKO or WT groups, we obtain high prediction accuracies of 98.7% (74.2%) and 77.8% (65.8%) respectively for SHG (TPaF) images. The persistence statistics that we use are fully interpretable, can highlight regions of abnormality within an image and identify features at different spatial scales.

A wide range of imaging techniques are used to investigate and analyse bone tissue to quantify the effects on structure and morphology as a result of genetic alterations or disease, both for research and diagnostic purposes. Non-linear microscopy imaging techniques such as two photon excitation fluorescence (TPEF) and second harmonic generation (SHG) are well-suited to imaging biological tissue in their native state without the need to label components with dyes or stains to generate image contrast, yet can provide chemical and structural details of samples. These label-free non-linear techniques use a near infrared pulsed laser source to expose samples instantaneously to high-intensity light while keeping the average (total) power within the damage threshold of the tissue. Both TPEF images and SHG images are greyscale, and each image can be encoded as an $n \times m$ matrix with integer pixel intensities between 0 (black) and 255 (white). In TPEF[1,2], molecules are excited to a higher energy level by near simultaneous absorption of two photons, causing the excited molecules to 'fluoresce' spontaneously i.e. emit a single photon, with more energy than one of the absorbed photons. We use TPEF images of the natively fluorescent (autofluorescent) molecules in bone, which we refer to as two photon auto-fluorescence images (TPaF), which provide detail of the overall morphology. In SHG imaging, the second harmonic wave is only generated by media that are non-centrosymmetric i.e. lack structural inversion symmetry in the plane[3], such as crystalline quartz, myosin, tubulin and fibrillar collagen[4,5]. Therefore, SHG is ideal for selectively imaging collagen fibres and their distribution in human skin, tissues, tendons and bones[4]. Collagen is the most abundant protein in human tissue and is a predominant component of the extracellular matrix in tissues. The role of the extracellular matrix in health conditions and disease processes is increasingly becoming established and is now often studied by using SHG imaging[6,7]. In bone tissue, SHG specifically images type I collagen fibres and their distribution, that is, the bulk of the bone matrix. Together, two photon autofluorescence (TPaF) and second

[1]School of Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, UK. [2]School of Chemistry, University of Southampton, Southampton SO17 1BJ, UK. [3]Institute for Life Sciences, University of Southampton, Southampton SO17 1BJ, UK. [4]School of Biological Sciences, University of Southampton, Southampton SO17 1BJ, UK. [5]The Alan Turing Institute, London NW1 2DB, UK. [6]Mechanical Engineering, University College London, London WC1E 7JE, UK. [7]These authors contributed equally: Sumeet Mahajan and Rubén J. Sánchez-García. ✉email: ysanne.pritchard@soton.ac.uk

harmonic generation (SHG) form a set of complementary techniques to image the chemical and structural changes, providing a label-free imaging readout of the bone microstructural properties (characterised by the absence of TPaF and SHG signals).

Correspondingly to these imaging techniques, there is a need for automated and quantitative methods that can effectively analyse the structure in these images in a biologically meaningful way. Topological data analysis (TDA) refers to a collection of topology-based methods that are able to quantify 'shape' within data, so are a natural choice to capture structural differences in the samples without requiring large sample sizes and can provide interpretable summaries in the context that each statistic is a measure of specific features in the images (Table 1). This is in contrast to data-driven techniques such as machine learning, which often lack interpretability and require a large training data set for similar classification tasks, something typically infeasible in animal or human research.

We define the term 'micro-holes' on three levels: mathematically, they are connected regions of black pixels in a binary image. For microscopy imaging, micro-holes are regions with low signal, that is, pixels that convert to background in the binary image (where the TPaF signal is autofluorescence, the SHG signal is largely collagen in these samples). At the biological level, for the bone samples we analyse, micro-holes encompass a range of structures including osteocyte lacunae and vascular canals (see[7,8] for the details). Using a powerful and versatile TDA technique called persistent homology, we can analyse the distribution of micro-holes in cortical bone samples i.e. we can quantify the number, size, connectivity, crowdedness and organisation of the micro-holes in the sample.

Persistent homology[9,10] is a topological method which tracks the evolution of connected pieces, loops and higher-dimensional holes of a 'continuous shape', i.e. a topological space[11], represented as a discrete structure such as a simplicial or cubical complex, as a scale parameter varies (Fig. 2b). This gives a persistence diagram of points (Fig. 2c), which is a plot of births (scale parameter value at which the feature appears) against deaths (scale parameter value at which the feature disappears, namely the pieces join or the loops or higher-dimensional 'holes' fill). To combine persistent homology with classification algorithms, the format of persistence diagrams needs to be adapted into feature vectors, which can be done through the use of persistence statistics and summary functions[12–15], persistence images[16], or persistence landscapes[17,18], with the former being our choice as we can tailor specific persistent summaries to bone microstructure measures of interest. TDA methods, including persistent homology, are being increasingly applied to biomedical image data including the cortical thickness of brain from magnetic resonance (MR) images in autistic patients[19]; chronic obstructive pulmonary disease[20]; identifying regions of interest in images of the colon[21]; the classification of diabetic retinopathy images[12]; hepatic
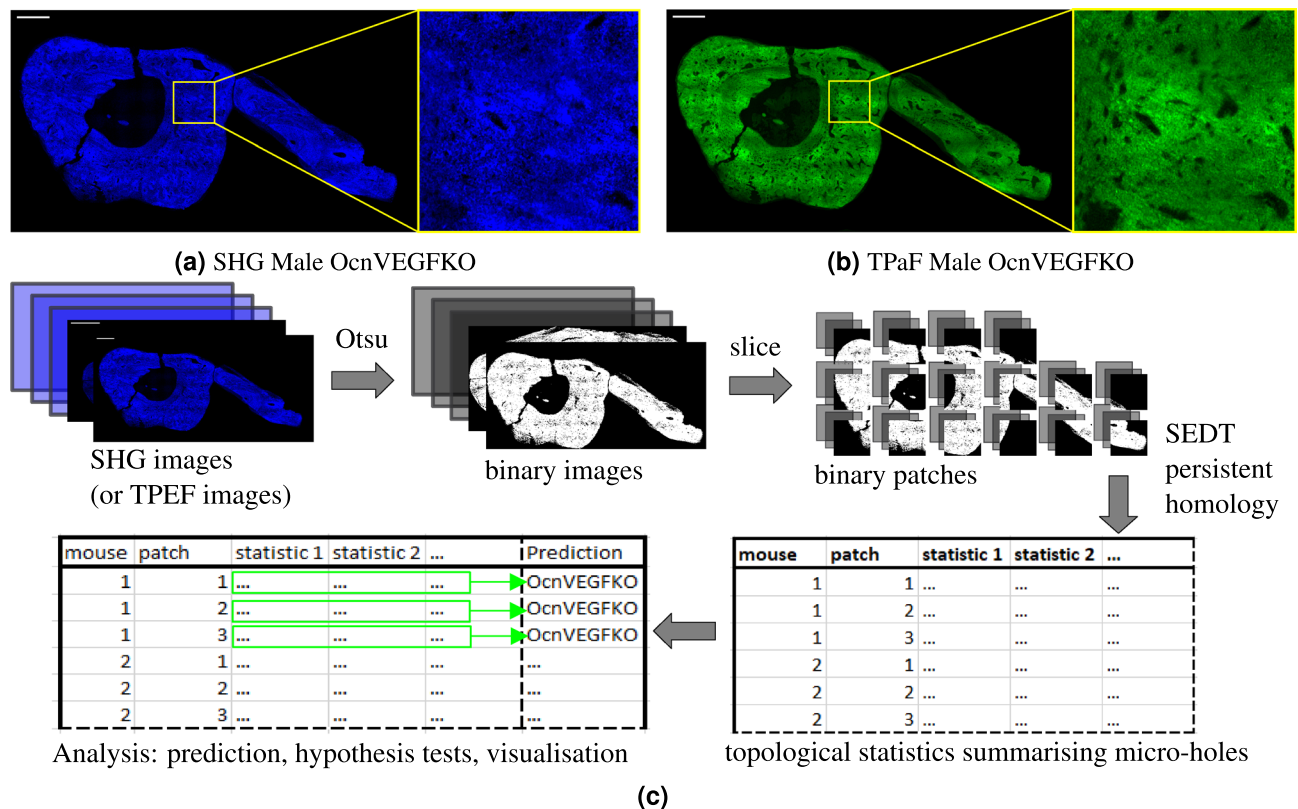

Figure 1. Non-linear microscopy images adapted from[7] and methodology schematic. Tibiofibular junction (TFJ) bone section of a male Osteocalcin-specific *Vegf* knockout (OcnVEGFKO) with SHG (**a**) and TPaF (**b**) images from a custom multiphoton microscope[7], with zoomed regions. Scale bars are 250 μm. Schematic in (**c**) gives a high level overview of the method process.

lesions[22]; skin lesions[23]; and endoscopy images of the stomach[24]. Previous TDA applications to biomedical images relate to X-ray, Computed Tomography (CT) and MRI techniques. As far as we know, our work is the first to use persistent homology for TPaF and SHG microscopy images of biological samples.

We validate our methodology using previously published TPaF and SHG images of a murine osteocalcin-specific *Vegf* knockout (OcnVEGFKO) bone exhibiting increased cortical porosity[8] in CT images and pathological extracellular matrix organisation[7] (Fig. 1). Our topological analysis (summarised in Fig. 4) summarises the micro-holes providing insight into micro-hole organisation and structure (Table 1, Fig. 5). These statistics can be refined to include only features at specific scales of biological interest (Fig. 4c), and can be compared across patches for intra-sample analysis (Fig. 4d). The statistical significance of our topological summaries is tested directly using permutation hypothesis tests, as shown by Table 2. These tests show significant differences in persistence statistics of image patches (Table 2). Finally, we train a support vector classifier to predict whether unseen patches of samples are from the WT (control) or OcnVEGFKO (test) categories. The performance results, shown in Table 3, confirm the strong predictive power of the topological features, particularly from SHG images.

Adapting persistent homology to analyse biologically or clinically relevant features in non-linear microscopy images expands the set of available tools that capture skeletal morphology. The topological approach that we present is not limited to this bone cell specific OcnVEGFKO or mouse models, and can be used in biological sample images where the analysis of micro-holes or hole-like structures may be required, including in a diagnostic context.

Our article is organised as follows. A detailed description of our methodology is given in section "Methods", including the specimen preparation, SHG and TPaF imaging, pre-processing of the images, a short introduction to persistent homology, a description of the filtration called the signed Euclidean distance transform (SEDT) that we use as input for the persistent homology, the persistence statistics per quadrant, the permutation hypothesis tests, and the classification using a support vector machine. In section "Results", we present the results of our topological methodology applied to mice bone images, including the outcome of the hypothesis tests, the interpretation of the most meaningful persistence statistics between groups of mice, and the results of the classification. The article finishes with a discussion and some conclusions in sections "Discussion" and "Conclusion".

Methods

Specimen preparation.
The data presented herein were acquired previously on samples used in the studies by Goring et al. and Sharma et al.[7,8]. The tibiae were derived from 16 week-old littermate matched mice evenly split between male and female, and OcnVEGFKO and WT[7,8]. Briefly[8], the tibiae were fixed in 4% paraformaldehyde for 48 hours, dehydrated in 70% ethanol and embedded in poly-methyl methacrylate, prior to taking 5 μm thick sections from the tibiofibular junction (TFJ). The use of animal tissue was carried out in compliance with the Animals (Scientific Procedures) Act 1986 and the regulations set by the UK home office, as in[7,8].

SHG and TPaF microscopy imaging.
Two photon excitation autofluorescence (TPaF) and (circularly polarised) second harmonic generation (SHG) images of each bone section were previously acquired by Sharma et al.[7] using a custom multiphoton microscopy setup. For full details of the imaging system and acquisition methodology see[7]. Example images are shown in Fig. 1. The images are 2D greyscale images with integer pixel intensities between 0 (black) and 255 (white). The TPaF images show the mineralised matrix and the SHG images capture the collagen fibres. Per imaging technique, there are 12 images of the 2–3 mm bone sections described in section "Specimen preparation". Note that the breaks in the bone are irrelevant for our microstructure analysis.

Preprocessing images.
We thresholded all images of the mouse samples to binary using Otsu's method[25], which divides the image pixels into two classes (black and white) by maximising the between-class variance. We trimmed the binary images of empty borders, padded them to an integer multiple of the desired patch shape, and split each image into 300 × 300 pixel patches, which is equivalent to a 183 μm × 183 μm (3 s.f.) view[7]. Note the number of patches is not an input but is determined by patch size and how many fully background patches are discarded. For these images, this scale allows for the largest meaningful features to be present in a patch. The patches were taken with a stride of 300 pixels to ensure they did not overlap, and we discarded any patches that were entirely background. This gave us a total of 475 patches and 476 patches from the original 12 TPaF and SHG images respectively, which were treated as separate sets of images in the analysis process. Note that edge patches have varying amounts of bone and, due to the irregular sample shape and to differences in the size of the bone sections, each individual image may produce a different number of patches. For the SHG (TPaF) images respectively there were 93 (93) female KO patches, 103 (101) female WT patches, 163 (160) male KO patches, and 117 (121) male WT patches.

Topological data analysis method.
Persistent homology tracks the evolution of connected pieces, loops and higher-dimensional 'holes' by calculating the homology groups of an increasing sequence of complexes (called a filtration) built using a scale parameter (Fig. 2b). We used cubical homology (section "Persistent homology of cubical complexes") and a signed Euclidean distance transform (SEDT) filtration (Fig. 3, section "Signed Euclidean distance transform") on pre-processed binary image patches (Fig. 4a, section "Preprocessing images"). The quadrants of the resulting persistence diagrams (Fig. 4b) quantify specific features, such as separate micro-holes, connected networks of micro-holes, and loops in bones surrounded by micro-holes (Fig. 4b, Table 1). Using this, we calculated summary statistics per patch (Fig. 5, section "Persistence statistics"), which quantify the quantity, size and organisation of the micro-holes within each patch, and can be visualised on the whole sample to highlight regions of interest (Fig. 4d).

All calculations were done in python 3 on a laptop with 5 cores and 8GB RAM running Windows 10. Image data and dataframes were handled in PIL, numpy, and pandas, persistent homology was calculated using the homcloud package, and for classification we used sklearn. The figures were created using matplotlib, seaborn and statannot. All computer code, documentation and an example computation are available on github[26]. Calculating the persistent homology, the longest part of the computation, took under 2 seconds per patch, and can be parallelised.

Persistent homology of cubical complexes. The first step to apply a topological method to a data set is to build a 'shape' (formally, a topological space) from the data points and some measure of similarity or distance between them. We build a topological space by gluing together basic building blocks (typically triangles or squares, and their higher-dimensional counterparts) with the data points as vertices. For images (pixels on a rectangular grid), using squares, cubes, etc., is the most natural option, resulting in what we call a cubical complex. More precisely, a cubical complex[27] is a topological space made of any number of n-cubes (0-cubes are points, 1-cubes are edges, 2-cubes are squares, 3-cubes are cubes, etc.), for different n, glued together along their faces (the vertices of an edge, the sides of a square or a cube, etc.) and embedded in $\mathbb{R}^d$ (for d large enough). A cubical complex is a combinatorial model of a topological space, or, in our case, an image. See Fig. 2b for some examples, and[27] for a formal definition.

Suppose that we want to encode a greyscale image as a cubical complex, taking into account both the pixel positions and their intensities (0 black to 255 white). A standard way is using a levelset filtration: we only include pixels below a certain threshold intensity level, and always join adjacent pixels (pixels whose coordinates differ by exactly 1), filling in squares (or cubes, for 3-dimensional images) when all the edges are present. As we increase the threshold intensity level, we include more and more pixels and obtain a sequence of increasingly larger complexes, called a filtration (see Fig. 2b for an example). There are other ways to construct a filtration of cubical complexes from an image and we use the Signed Euclidean Distance Transform (SEDT) filtration, explained in section "Signed Euclidean distance transform".

Before describing the SEDT filtration, we will briefly and informally discuss persistent homology (for a formal treatment, see e.g.[9,28,29]). The homology groups $H_k(X)$ of a topological space X (such as a cubical complex) detect the number of k-dimensional 'holes' in X, for $k = 0, 1, 2, \ldots$ For $k = 0$, $H_0(X)$ detects the connected 'pieces' that make up X, while $H_1(X)$ detects the cycles (loops) in X (see Fig. 2b). The rank of the homology groups (the number of k-dimensional 'holes') are called the Betti numbers, written β_k, and uniquely determine $H_k(X)$ for each k (assuming coefficients in a field F, which is $F = \mathbb{R}$ in our case). For full details on the homology of cubical complexes, see[27]. Persistent homology extends homology from a single topological space to a filtration (a nested, increasing sequence of topological spaces). It calculates the homology groups across the filtration, keeping track of when a feature is born (appears) or dies (disappears). The result is encoded in a persistence diagram, such as the one shown in Fig. 2c. Note that, since we analyse 2D images, we only consider homology for dimensions $k = 0, 1$, as the homology is zero for larger k. Each persistent diagram is a plot of pairs of points (b_i, d_i), where b_i, d_i are respectively the birth and death time, of the ith k-dimensional feature in H_k. A feature can only die after it is born, hence $b_i < d_i$ and all points shown in a persistent diagram must be above the diagonal. It can be shown that all features eventually die (at most when the connectivity between data points is maximal), except in H_0 where a final connected 'piece' of X remain indefinitely (shown with the symbol ∞ in our persistent diagrams). In Fig. 2, we show a sample greyscale image, levelset filtration, and resulting persistent homology diagram, for illustration.

Signed Euclidean distance transform. We analyse the micro-holes in the SHG and TPaF images (which we refer to as 'microstructure') of bone samples using persistent homology with a signed Euclidean distance transform (SEDT) filtration (Figs. 3, 4). The SEDT filtration allows us to encode the different structural features of an image as points in specific quadrants of the resulting persistence diagrams. This filtration has been successfully used to study the structure of bead packing, sand packing, limestone and sandstone samples on micro CT images in[30],
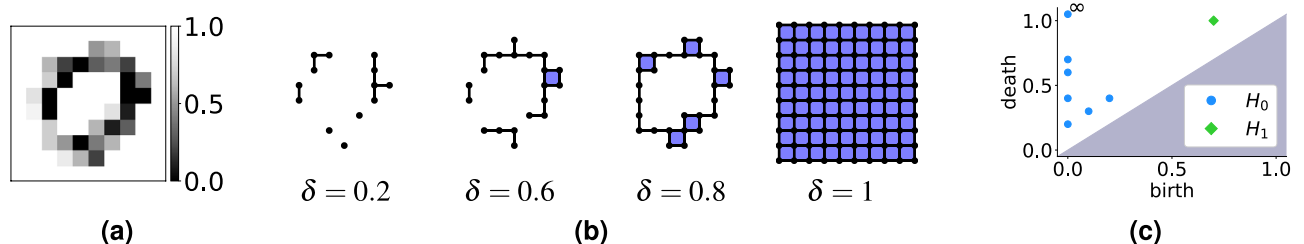

Figure 2. Cubical levelset filtration of a greyscale image. (**a**) Example greyscale image with pixel intensity values between 0 (black) and 1 (white), as in the colour bar. (**b**) Some ordered steps of the cubical levelset filtration of the image in (**a**). As the threshold parameter δ increases from 0 to 1, pixels (of intensity lower than δ) are included as points (0-cubes), with two adjacent pixels joined by edges (1-cubes), and 4 adjacent pixels joined by squares (2-cubes). Note the ring we see in the image is captured by the third step of the filtration shown. (**c**) Persistence diagram for H_0 and H_1 of the image (**a**) with respect to the levelset filtration described in (**b**). The blue dots (H_0) correspond to connected pieces, which 'die' as they join one another. The final connected piece never 'dies', so has infinite persistence (marked as '∞'). The green diamond (H_1) corresponds to a loop (**b**, panel 3) capturing the ring structure in the image (**a**). This loop disappears ('dies') as we add more pixels (**b**, panel 4).

and to analyse the flow of fluids and trapping of bubbles in sandstone samples in 3D X-ray CT scans[31]. In medical imaging, the SEDT filtration has been recently adapted to tumour images by Moon et al.[32] to study the relation between topological shape features in tumour progression and survival risks. Further, a similar (Manhattan) distance transform filtration approach has been used in[33] to predict transepidermal water loss from skin images.

The input of the signed Euclidean distance transform (SEDT) is a binary (black/white) image, rather than a greyscale image. It takes the binary image and assigns to every pixel in both phases (white, or black, pixels) the shortest Euclidean distance to the opposite phase of the image. Pixels in the white ('bone') phase are given a positive distance, and pixels in the black ('micro-hole') phase are given a negative distance, as in Fig. 3a. We use persistent homology with a signed Euclidean distance transform filtration as in[31], that is, the filtration where we include pixels from the most negative distance to the most positive distance, with respect to the SEDT. The pixels with the most negative distance values are the micro-hole centres and are included first in the filtration, while the pixels with the greatest positive distances are bone pixels furthest from any micro-holes and are included last in the filtration. In Fig. 3, we show an example of the persistent homology of a binary image with respect to the SEDT filtration in detail. In Fig. 4, we explain the process taking a patch of a greyscale image to binary, before calculating the persistent homology using an SEDT filtration, and showing the quadrants of the persistence diagrams and how the features captured in each quadrant relate back to the micro-holes in the binary image.

Let us discuss a few properties of the SEDT persistent homology which are relevant to our analysis (cf. section "Persistence statistics" and Table 1). We can divide the persistent diagram into four quadrants (positive/negative birth/death), which we label 1 to 4 from the top right, proceeding anti-clockwise (see Fig. 3c). Since all points are above the diagonal, there are no points in quadrant 4. All black pixels (and thus all micro-holes, which we define formally as connected regions of black pixels in the micro scale binary images of the samples) are born by $\delta = 0$ (let us call δ the parameter in the SEDT filtration), hence there are no H_0 points in quadrant 1. There can be H_0 points in quadrant 3, for connected micro-holes born at different pixels (see Fig. 3c) and merged later (at most by $\delta = 0$, when all black (micro-hole) pixels are included). More irregular (e.g. less 'circular') micro-holes can be born at different pixels (at each local minimum of the SEDT distance within the same micro-hole). By $\delta = 0$, all micro-holes have been included: the number of H_0 points in quadrant 2 (including the point labelled ∞) equals the number of micro-holes, and each of the remaining points in quadrant 3 corresponds to a unique point (micro-hole) in quadrant 2. Moreover, the birth of each of these points equals the radius of the largest (open) circle that can be inscribed in the micro-hole, which follows from the definition of the SEDT. We refer to this quantity as the *size* of the micro-hole. Hence, the larger the micro-hole, the earlier it is born (more negative δ), with the largest micro-hole appearing first and persisting indefinitely in the filtration (labelled '∞' in the persistent diagram). Although not entirely obvious, the introduction of white (bone) pixels cannot create new connected pieces: a new white pixel will join its closest micro-hole(s), possibly merging two or more micro-holes. All in all, the persistence diagram for the 0th homology group contains the disconnected micro-holes in quadrant 2 and the connected micro-holes in quadrant 3, with empty quadrants 1 and 4. Finally, note that the death value for micro-holes in quadrant 2 of H_0 is half of the shortest distance from the micro-hole to its nearest micro-hole (the distance between two micro-holes A and B is the minimum distance between a pixel in A and a pixel in B), as both micro-holes expand by including the white pixels between them and thus it is a measure of micro-hole 'crowdedness'. To visualise the features from the quadrants of the persistence diagrams under the SEDT filtration, see Fig. 4b. The birth pixels and optimal volumes[34] of the features have been plotted on the images for H_0 quadrant 2 (green), and H_1 quadrant 1 (red), using homcloud.

For the 1st homology group (H_1), any loop born before $\delta = 0$ would be a loop of black (micro-hole) pixels, containing white (bone) pixels, so it will not die until some $\delta > 0$. Therefore, there are no H_1 points in quadrant 3, but there could be points in quadrant 2 (micro-hole loops). Points in quadrant 1 of H_1 correspond to loops in the bone between micro-holes, that is, white (bone) pixels connecting black regions (micro-holes) in a ring (loop) fashion (see Fig. 3), with larger loops indicating regions with sparser micro-hole distribution.

For reproducibility, we include the following technical note. The 0th persistence diagram contains a point with the most negative birth value (known as 'essential birth') and infinite death (labelled '∞' in Fig. 3c), which naturally falls in quadrant 2 and corresponds to the largest micro-hole (connected region of black pixels) in the binary image patch. These points were omitted when calculating the statistics, since for most patches, the largest negative distance was a point in the background region surrounding the bone sample.

Persistence statistics. For each image patch, we calculate the persistence statistics for quadrant 2 of the 0th persistence diagram, and for quadrant 1 of the 1st persistence diagram, as these are the most informative quadrants for our analysis (see discussion in section "Signed Euclidean distance transform"). Namely, for each quadrant of interest, we considered the following summary statistics, calculated per image patch, per homological dimension (H_0 and H_1), and for both birth and death distributions: mean, median, quartiles, interquartile range (IQR) and standard deviation. We also calculated, for each image patch and dimension, the number of points n in the persistent diagram, the total persistence $P = \sum_{i=1}^{n}(d_i - b_i)$, and the persistent entropy[35,36] $\sum_{i=1}^{n} -\frac{p_i}{P} \log\left(\frac{p_i}{P}\right)$. The latter is the Shannon entropy[37] of the persistent values $d_i - b_i$ as a probability distribution (divided by the total persistence P), and measures the 'diversity' of the persistence[23], that is, how close is the uniform distribution (where all features are equally persistent).

Each of the chosen persistent statistics has a direct interpretation in terms of the micro-hole (bone) structure, as shown in Table 1. Moreover, we can restrict our attention to features within a specific scale of interest, for example micro-holes of a specific size (see Fig. 4c). Further, by focusing on a persistent statistic of interest, we can give image patches a score, highlighting regions of interest within the sample (see Fig. 4d).
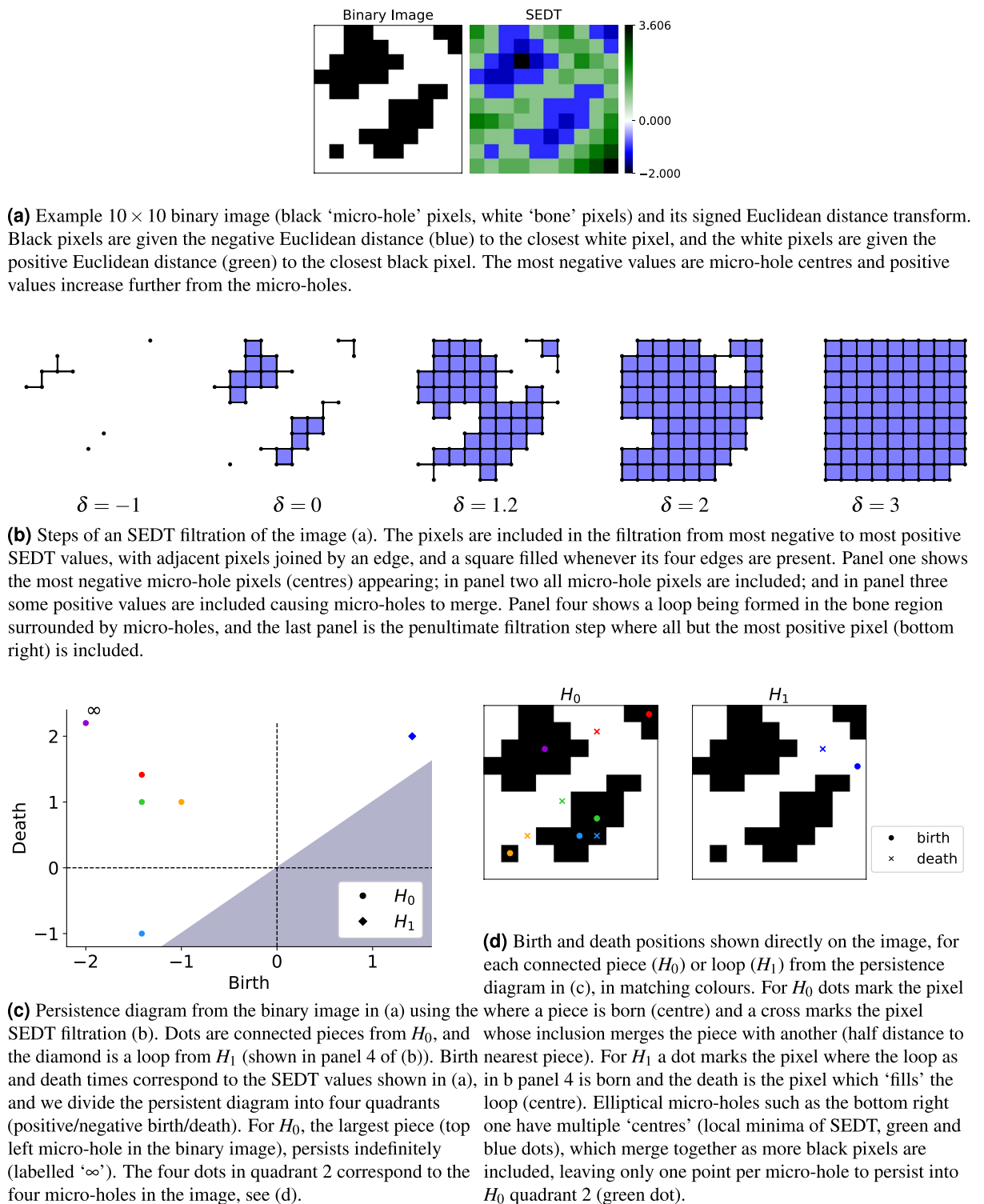
(a) Example 10×10 binary image (black 'micro-hole' pixels, white 'bone' pixels) and its signed Euclidean distance transform. Black pixels are given the negative Euclidean distance (blue) to the closest white pixel, and the white pixels are given the positive Euclidean distance (green) to the closest black pixel. The most negative values are micro-hole centres and positive values increase further from the micro-holes.

(b) Steps of an SEDT filtration of the image (a). The pixels are included in the filtration from most negative to most positive SEDT values, with adjacent pixels joined by an edge, and a square filled whenever its four edges are present. Panel one shows the most negative micro-hole pixels (centres) appearing; in panel two all micro-hole pixels are included; and in panel three some positive values are included causing micro-holes to merge. Panel four shows a loop being formed in the bone region surrounded by micro-holes, and the last panel is the penultimate filtration step where all but the most positive pixel (bottom right) is included.

(c) Persistence diagram from the binary image in (a) using the SEDT filtration (b). Dots are connected pieces from H_0, and the diamond is a loop from H_1 (shown in panel 4 of (b)). Birth and death times correspond to the SEDT values shown in (a), and we divide the persistent diagram into four quadrants (positive/negative birth/death). For H_0, the largest piece (top left micro-hole in the binary image), persists indefinitely (labelled '∞'). The four dots in quadrant 2 correspond to the four micro-holes in the image, see (d).

(d) Birth and death positions shown directly on the image, for each connected piece (H_0) or loop (H_1) from the persistence diagram in (c), in matching colours. For H_0 dots mark the pixel where a piece is born (centre) and a cross marks the pixel whose inclusion merges the piece with another (half distance to nearest piece). For H_1 a dot marks the pixel where the loop as in b panel 4 is born and the death is the pixel which 'fills' the loop (centre). Elliptical micro-holes such as the bottom right one have multiple 'centres' (local minima of SEDT, green and blue dots), which merge together as more black pixels are included, leaving only one point per micro-hole to persist into H_0 quadrant 2 (green dot).

Figure 3. Cubical SEDT filtration of a binary image and persistent homology diagram.

Hypothesis tests. We determine the statistical significance of the persistence statistics described in section "Persistence statistics" in our data set of image patches using permutation hypothesis tests[38]. Namely, we test the null hypothesis that two groups (e.g. female test (OcnVEGFKO) v. female control (WT)) have identical distributions for the given persistence statistics (e.g. mean birth in H_0 quadrant 1). In the following sections, OcnVEGFKO is referred to as 'test' and WT as 'control', as the method is independent of the mouse model. The result is a pseudo p-value, which is the probability that the difference in the mean statistic between the two shuffled groups is at least as large as the difference we observe, if the null hypothesis that the two groups have the
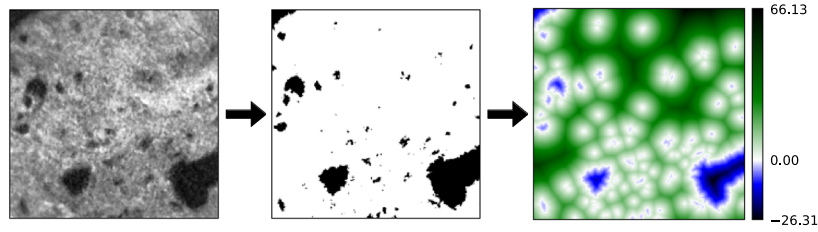
(a) A 300×300 pixel greyscale patch, equivalent to $183\,\mu m \times 183\,\mu m$ ($3s.f.$) view [7] from the TPaF image in Fig. 1 (left), converted to binary with an Otsu threshold (middle), before the signed Euclidean distance transform is calculated (right), with the micro-hole phase shown in blue and the bone phase in green.
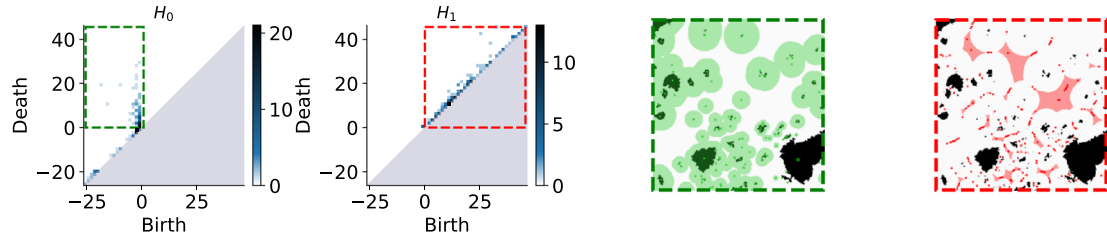
(b) Persistence diagrams (left) for H_0 and H_1 as heatmaps of points for the image (a) under the SEDT filtration. Quadrant 2 H_0 (outlined in green) contains a point per micro-hole, while in quadrant 1 of H_1 (outlined in red) each point represents a loop in the bone surrounded by micro-holes. The two panels on the right show birth positions and optimal volumes [34] of features from the two quadrants of the persistence diagrams (left), in corresponding colours.
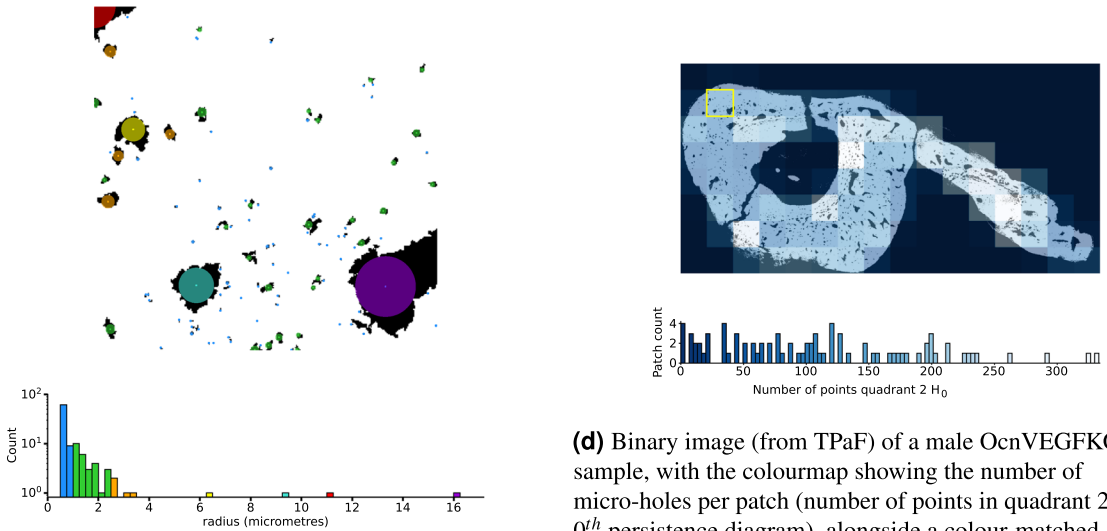
(c) A binary image patch from (a), with the micro-hole birth values from H_0 quadrant 2 converted to μm and used as radii of inscribed circles, coloured by radius to match the histogram. The largest radius is 26 pixels. By using the (converted) birth values, we can restrict to specific micro-hole sizes.

(d) Binary image (from TPaF) of a male OcnVEGFKO bone sample, with the colourmap showing the number of micro-holes per patch (number of points in quadrant 2 of the 0th persistence diagram), alongside a colour-matched histogram of the values per patch. The darker blue regions contain fewer micro-holes, and the lighter the region, the more micro-holes were found in the patch. The image patch from (a), (b) and (c) is outlined in yellow. Note that a peak at 0 from patches without bone has been removed from the histogram.

Figure 4. Persistent homology micro-hole structure analysis in a greyscale microscopy image.

same distribution is true. If this probability is very small, we reject the hypothesis that the distributions are the same for the two groups.

To calculate the pseudo p-value, we randomly assign image patches to two groups (of the same size as the two groups we are comparing) and calculate the difference between the means of the given statistic for the two randomly selected groups. By repeating this step a large number of times ($n = 10{,}000$ in our tests), we obtain a distribution of differences in means. The pseudo p-value is the proportion of the shuffled differences in means which fall above the observed difference in our initial groups. That is, the probability of having observed that

H_*	Q	Persistent statistic	Interpretation
H_0	2	Number of points	Number of micro-holes
H_0	2	Mean birth	Mean micro-hole radius (of largest inscribed circle)
H_0	2	Median birth	Median micro-hole radius
H_0	2	25p. birth	Minimum radius of largest 25% of micro-holes
H_0	2	IQR birth	Range in radii of middle 50% of micro-holes
H_0	2	St. dev. birth	Spread of micro-hole radii
H_0	2	Mean death	Average micro-hole crowding (mean of inter-micro-hole distance)
H_0	2	Median death	Median inter-micro-hole distance
H_0	2	25p. death	Maximum of smallest 25% of the inter-micro-hole distances
H_0	2	IQR death	Range in the middle 50% of the inter-micro-hole distances
H_0	2	St. dev. death	Spread of inter-micro-hole distances
H_0	2	Persistent entropy	Diversity of micro-hole lifetimes
H_0	2	Number micro-holes $r \geq 2$	Number of micro-holes with radius greater than or equal to 2
H_0	2	Number micro-holes $r < 2$	Number of micro-holes with radius less than 2
H_1	1	Number of points	Number of bone regions surrounded by micro-holes
H_1	1	Mean birth	Mean distance from micro-holes at which loop is formed
H_1	1	Median birth	Middle of the distances from micro-holes at which loop is formed
H_1	1	25p. birth	Distance 25% through distances from micro-holes when loop is formed
H_1	1	IQR birth	Range of middle 50% of distances from micro-holes when loop is formed
H_1	1	St. dev. birth	Spread of distances from micro-holes at which loop is formed
H_1	1	Mean death	Mean of the radii of regions surrounded by micro-holes
H_1	1	Median death	Radius of middle-sized bone region surrounded by micro-holes
H_1	1	25p. death	Max radius of smallest 25% of bone regions surrounded by micro-holes
H_1	1	IQR death	Range in radii of middle 50% of bone regions surrounded by micro-holes
H_1	1	St. dev. death	Spread of radii of bone regions surrounded by micro-holes
H_1	1	Persistent entropy	Diversity of lifetimes for bone regions surrounded by micro-holes

Table 1. Interpretation of persistence statistics. An advantage of using persistent homology with a SEDT filtration is its interpretability. Here we list the persistent statistics that we use and their morphological interpretation. We use 'Q.' for quadrant, 'p.' for percentile, 'st. dev.' for the standard deviation, IQR for the interquartile range, and micro-hole radius or size refers to the radius of the largest inscribed (open) circle. The inter-micro-hole distance is half the distance from one micro-hole to the next nearest micro-hole.

difference between two randomly selected groups under the hypothesis that the two groups have identical distributions. This is an estimate of the p-value (thus called pseudo p-value), since computing all possible permutations is infeasible. In particular, a pseudo p-value of zero should be interpreted as a very small non-zero value.

We have also included paired permutation hypothesis tests for each of the 26 statistics to determine if there are significant differences between the statistic calculated on the TPEF and SHG images. We test the null hypothesis that the statistic has identical distributions from the TPEF and SHG imaging types. Each statistic is paired per image patch per sample and are shuffled ($n = 10{,}000$ times) between the TPEF and SHG statistic column. For each shuffle we calculate the mean of the shuffled differences normalised by the standard deviation of the shuffled differences. We use a two-tailed test and the p-value is the proportion of shuffled differences more extreme than the measured difference.

When comparing a large number of persistence statistics using hypothesis tests, we should account for the multiple comparison problem. Given that we do not intend to claim biological discoveries based on a single statistic, the effect that a single false positive has for our experiment is not overly damaging. Thus, for this data, we chose a less-strict measure by using a Benjamini–Hochberg p-value adjustment[39] which controls the false discovery rate (expected proportion of all positives that are false) $\mathrm{FDR} = \mathbb{E}\left(\frac{\mathrm{FP}}{\mathrm{FP+TP}}\right)$, where TP and FP are the number of true positives and false positives respectively.

Classification. In addition to the hypothesis tests (section "Hypothesis tests"), we assessed the usefulness of the persistent statistics for classification purposes by training a classifier on the images to predict whether unseen patches are from our test (OcnVEGFKO) or control (WT) set. That is, we addressed the question whether all or some of the persistent statistics that we considered are enough to distinguish the test condition images from the control images.

We used a support vector classifier (SVC)[40] with a radial basis function (RBF) kernel, which performed better than the more standard linear kernel for our dataset of image patches. Cross-validation[41] was used to split the data into training and test sets, specifically a 10-fold cross-validation stratified into groups where both the training and test sets contain a number of patches from each of the four categories (male OcnVEGFKO, male

WT, female OcnVEGFKO, female WT). The resulting prediction accuracy, precision, recall and F1 score[42] on the unseen test sets are given in Table 3.

Results

Separately for each image type (SHG and TPaF), we took 476 patches from the 12 images, and using hypothesis testing, we compared the persistent statistics (section "Persistence statistics") for the four groups: male OcnVEG-FKO, male WT, female OcnVEGFKO, female WT. Note that we divided the samples not only into WT (control) and OcnVEGFKO (test), but also by sex, due to the known differences per sex of this condition on bone[7,8]. In this section, we present the results of the hypothesis tests (Table 2), which show where the persistent statistics are significantly different between groups of mice, interpret these differences in the persistence statistics (Fig. 5), and show the performance results of the support vector classifiers (Table 3).

As discussed (section "Signed Euclidean distance transform"), each micro-hole is represented by a (birth, death) point in quadrant 2 of the H_0 persistent diagram, with the birth the largest distance from the bone matrix (the maximal radius of a circle inscribed in the micro-hole) and the death is (half) the distance at which it merges into other pieces originating from surrounding micro-holes (so the death tells us about the local crowding behaviour). Hence, as in Table 1, the following 13 statistics that summarise this quadrant can summarise the microstructure of the holes: number of points (number of micro-holes), persistent entropy (diversity of persistence), mean and median birth (averages of micro-hole radius), mean and median death (averages of micro-hole crowding), 25th percentile of births (minimum radius of largest 25% of micro-holes), 25th and 75th percentiles
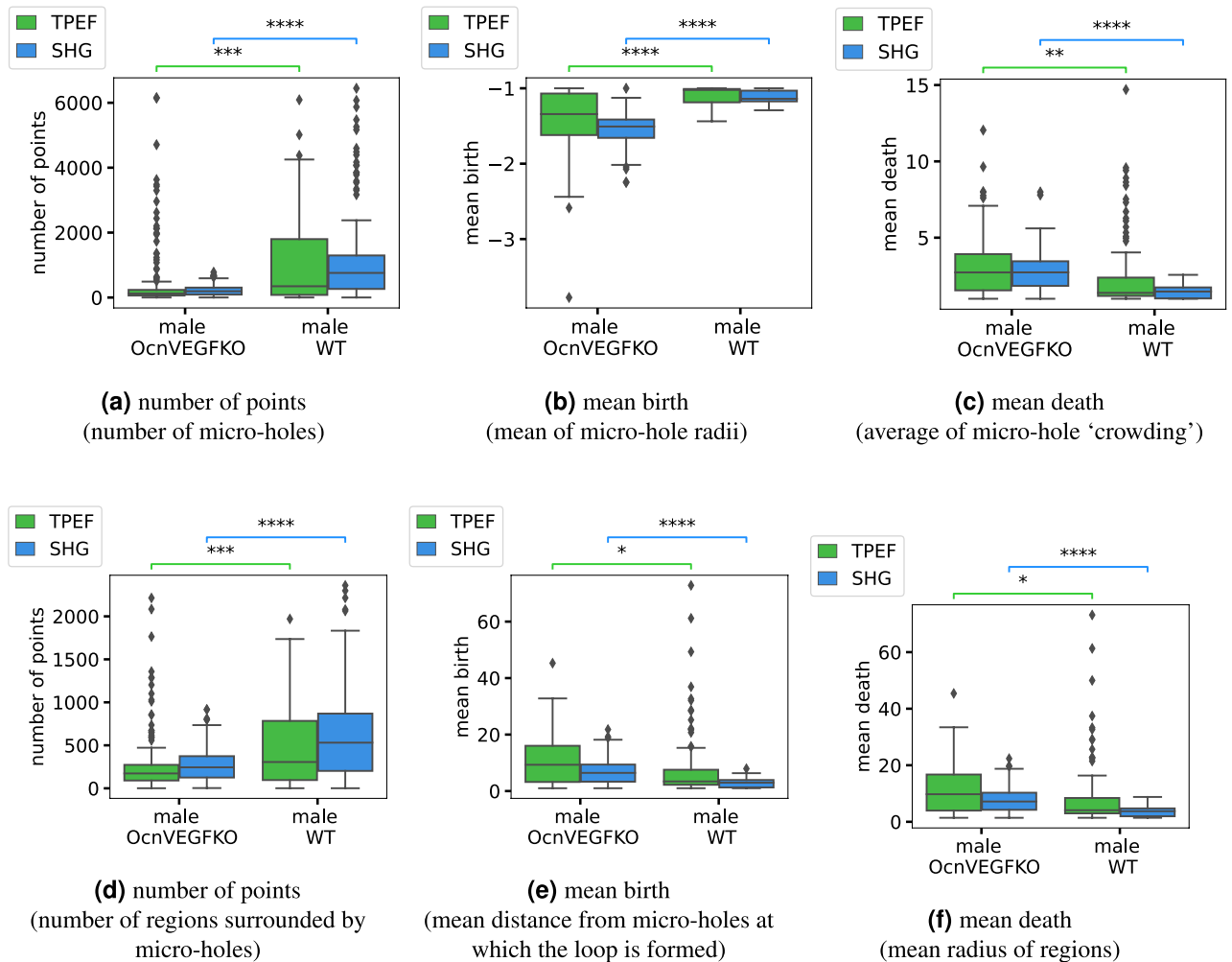

(a) number of points
(number of micro-holes)

(b) mean birth
(mean of micro-hole radii)

(c) mean death
(average of micro-hole 'crowding')

(d) number of points
(number of regions surrounded by
micro-holes)

(e) mean birth
(mean distance from micro-holes at
which the loop is formed)

(f) mean death
(mean radius of regions)

Figure 5. Box plots of persistence statistics per image patch, with three (**a–c**) from quadrant 2 of H_0 which summarise the micro-holes and three (**d–f**) persistence statistics from quadrant 1 of H_1, which summarise the loops in the filtration that are regions of bone surrounded by micro-holes. The box plots show the median and quartiles, with outliers marked as points outside of 1.5 times the interquartile range. We annotate their significance using the adjusted p-values comparing male OcnVEGFKO to male WT on each imaging type (TPaF or SHG). Here the stars indicate significance levels: 'ns' if it is not significant $0.05 < p \leq 1$, '*' if $0.01 < p \leq 0.05$, '**' if $0.001 < p \leq 0.01$, '***' if $0.0001 < p \leq 0.001$, and '****' if $p \leq 0.0001$. We include only the male patches and two of eight significance annotations for readability; for the full p-value results, see Table 2.

H_*	Q	Persistent statistic	All	Male only:		Female only:		KO only:		WT only:	
			SHG	KO v WT		KO v WT		Male v Female		Male v Female	
			v TPaF	SHG	TPaF	SHG	TPaF	SHG	TPaF	SHG	TPaF
H_0	2	Number of points	0.628	**0.000**	**0.000**	**0.001**	0.585	**0.000**	0.092	0.764	0.523
H_0	2	Pers. entropy	**0.015**	**0.000**	**0.001**	0.528	0.243	**0.000**	**0.009**	0.138	0.586
H_0	2	Mean birth	**0.000**	**0.000**	**0.000**	**0.000**	0.104	**0.000**	**0.000**	**0.009**	1.000
H_0	2	Median birth	**0.000**	**0.000**	**0.037**	1.000	1.000	**0.000**	**0.026**	1.000	1.000
H_0	2	25p. birth	**0.000**	**0.000**	**0.000**	**0.001**	0.104	**0.000**	**0.000**	**0.010**	0.523
H_0	2	IQR birth	**0.000**	**0.000**	**0.000**	**0.001**	0.104	**0.000**	**0.000**	**0.008**	0.557
H_0	2	St. dev. birth	**0.007**	**0.000**	**0.000**	**0.000**	0.585	**0.000**	**0.000**	0.764	0.430
H_0	2	Mean death	**0.000**	**0.000**	**0.009**	0.992	0.365	**0.000**	**0.000**	**0.000**	0.384
H_0	2	Median death	0.870	**0.000**	**0.004**	0.776	0.104	**0.000**	**0.000**	**0.000**	0.384
H_0	2	25p. death	**0.000**	**0.000**	0.135	0.528	0.104	**0.000**	0.091	0.058	0.384
H_0	2	75p. death	**0.001**	**0.000**	**0.002**	0.992	0.980	**0.000**	**0.000**	**0.006**	0.586
H_0	2	IQR death	**0.000**	**0.000**	**0.002**	0.992	0.980	**0.000**	**0.000**	**0.007**	0.658
H_0	2	St. dev. death	**0.000**	**0.000**	**0.033**	0.992	0.104	**0.000**	**0.019**	**0.000**	0.384
H_0	2	Number holes $r \geq 2$	**0.000**	0.158	**0.000**	**0.000**	**0.003**	0.149	**0.019**	**0.000**	0.384
H_0	2	Number holes $r < 2$	0.344	**0.000**	**0.000**	**0.000**	0.585	**0.000**	0.091	0.808	0.523
H_1	1	number of points	**0.000**	**0.000**	**0.000**	0.144	0.718	**0.000**	0.085	0.331	0.430
H_1	1	Pers. entropy	**0.000**	**0.000**	**0.001**	0.995	0.585	**0.000**	**0.019**	0.197	0.982
H_1	1	Mean birth	**0.000**	**0.000**	**0.037**	0.797	0.392	**0.000**	**0.019**	**0.000**	0.430
H_1	1	Median birth	**0.000**	**0.000**	0.182	0.776	0.585	**0.000**	0.132	**0.000**	0.384
H_1	1	25p. birth	**0.000**	**0.000**	0.471	0.470	0.640	**0.000**	**0.001**	**0.000**	0.384
H_1	1	IQR birth	**0.000**	**0.000**	**0.017**	0.995	0.585	**0.000**	0.105	**0.000**	0.700
H_1	1	St. dev. birth	**0.000**	**0.000**	**0.001**	0.991	0.104	**0.000**	0.177	**0.000**	0.700
H_1	1	Mean death	**0.000**	**0.000**	**0.037**	0.981	0.392	**0.000**	**0.019**	**0.000**	0.430
H_1	1	Median death	**0.000**	**0.000**	0.170	0.776	0.585	**0.000**	0.141	**0.000**	0.384
H_1	1	25p. death	**0.000**	**0.000**	0.478	0.797	0.640	**0.000**	**0.002**	**0.001**	0.384
H_1	1	75p. death	**0.000**	**0.000**	**0.037**	0.960	0.585	**0.000**	**0.026**	**0.000**	0.523
H_1	1	IQR death	**0.000**	**0.000**	**0.017**	0.991	0.585	**0.000**	0.098	**0.000**	0.700
H_1	1	St. dev. death	**0.000**	**0.000**	**0.001**	0.992	0.104	**0.000**	0.166	**0.000**	0.700

Table 2. Benjamini–Hochberg adjusted pseudo p-values (3 d.p.) from permutation tests for 28 test statistics with paired tests to compare the imaging types and tests between OcnVEGFKO (labelled KO above) and WT males and female groups using either the SHG or TPaF images. To see if there is a significant difference from the imaging technique we paired the TPaF and SHG image patches per statistic. Also, for SHG and TPaF separately we tested the hypothesis that the distributions are the same for males between OcnVEGFKO (test) and WT (WT); females between OcnVEGFKO and WT; OcnVEGFKO between male and female; and WT between male and female. All p-values in bold are significant at the 5% level. The adjustments were made separately for the statistics from TPaF and SHG images in the OcnVEGFKO/WT M/F group comparisons. Here H_* is the homology group, Q is the quadrant and we have included statistics from the H_0 quadrant 2 (separate micro-holes) and from H_1 quadrant 1 (loops in bone surrounded by micro-holes). We write p. for percentile, r for radius and st. dev. is the standard deviation. For the interpretation of each persistent statistic, see Table 1.

Images	Groups	Features	Accuracy	Precision	Recall	F1 score
SHG	Males	10 (20)	**0.987** (0.986)	**0.986** (0.984)	**0.994** (0.993)	**0.989** (0.989)
TPaF	Males	10 (20)	0.742 (0.740)	0.810 (0.800)	0.736 (0.747)	0.766 (0.768)
SHG	Females	10 (20)	**0.778** (0.773)	**0.785** (0.777)	**0.765** (0.765)	**0.767** (0.762)
TPaF	Females	10 (20)	0.658 (0.663)	0.704 (0.713)	0.556 (0.553)	0.609 (0.611)

Table 3. Performance results (for OcnVEGFKO vs WT mice image patches) of an SVC (support vector classifier) with an RBF (radial basis function) kernel ($C = 3$) with stratified 10-fold cross validation, using 10 or 20 persistence features (see main text) to predict whether an image patch in the test set is from a OcnVEGFKO or WT mouse. Classifiers with 10 features use only quadrant 2 of H_0, whereas 20 features adds additional features from quadrant 1 of H_1, with no improvement. All given measures are mean measures over 100 repeats. Numbers in bold show the best performance results (SHG images, 10 features).

of death distribution (crowding bounds of smallest 25% and largest 25% of micro-holes), the interquartile range of births and deaths (radius/crowding difference for middle 50% of micro-holes) and the standard deviation of births/deaths (spread of radius and crowding). Analogous persistent statistics for quadrant 1 of H_1, which captures loops in the bone matrix surrounded by micro-holes, quantify the number and size of uninterrupted regions surrounded by micro-holes. Finally, to show the number of features at a given scale, we also include the number of micro-holes of radius less than 2, and the number of micro-holes of radius greater than or equal to 2, calculated using the births from quadrant 2 of H_0.

In Fig. 5, we show three example persistent statistics for H_0 and three persistent statistics for H_1, which allow us to quantify the differences between the male OcnVEGFKO and WT groups with respect to specific interpretable topological (persistent) features. For the full significant differences per statistic see Table 2 and for plots of other statistics and the female results please see the supplemental information. The statistics shown are divided into male OcnVEGFKO and WT groups, so there are patches from bone tissue samples from 3 mice, each represented by its respective box plot. Firstly, in Fig. 5a, the male OcnVEGFKO mice have fewer points, hence fewer micro-holes, but the micro-holes are much larger as they have more negative birth times in Fig. 5b, i.e. the micro-hole centres are farther from the bone pixels. The top 25% of micro-holes have a radius larger than the 25th percentile for births and the male OcnVEGFKO set have consistently larger radii across the patches. Similarly, the interquartile range (IQR) of births shows the middle 50% of micro-hole sizes have a much greater range of sizes in the OcnVEGFKO male patches compared to the other groups whose separate micro-holes are much more consistent in size. There is a decrease in persistent entropy for the OcnVEGFKO males, which indicates that features are more diverse, as persistent entropy increases with homogeneity of persistence. Importantly, the OcnVEGFKO male patches are also less crowded, which is shown by the increase in mean death time in Fig. 5c, as micro-holes persist to greater distances from the micro-hole before colliding. This is further supported by OcnVEGFKO male patches showing significant increases for deaths in the median, standard deviation, IQR and the 75th percentile. It should be noted that, for the SHG images only, there are a few significant differences between female OcnVEGFKO and WT patches, including the number of points, mean birth, IQR, standard deviation and 25th percentile for birth.

In addition, for quadrant 1 of the 1st persistence diagram (Fig. 5d–f), the male OcnVEGFKO patches contain fewer loops formed by bone matrix regions around micro-holes, as there is a lower number of points (Fig. 5d), which fits with the expectation given there are fewer micro-holes. These loops have a higher mean birth (Fig. 5e), so are born at greater distance from the micro-holes and similarly an increased mean death (Fig. 5f) means they are filled in at higher distance values, so are larger. Overall, this shows that the micro-holes in the OcnVEGFKO male patches are less crowded, as the bone matrix areas between micro-holes are fewer, larger and less interrupted by smaller features.

We show the Benjamini–Hochberg adjusted p-values for the permutation tests for our choice of persistence statistics in Table 2. For each imaging type (SHG, and TPaF), we test the hypothesis that the distributions are the same for: males between OcnVEGFKO and WT; females between OcnVEGFKO and WT; OcnVEGFKO between male and female; and WT between male and female. The p-values show many significant differences between the male OcnVEGFKO and WT image patches, and between the OcnVEGFKO male and female patches. However, between the female OcnVEGFKO and WT patches, the statistics show only 5 significant differences for the SHG (and none for TPaF). The vast majority of statistics are also significantly different between the OcnVEGFKO male and female patches, both for the SHG and TPaF images. For the SHG images, there are also some significant differences between the male and female WT patches that are not present in the TPaF images, consistent with what we observe in Fig. 5. The SHG images capture the collagen structure of the samples, so the stronger significant differences and improved classification performance shown in Table 3 indicate that the differences are due to the collagen distribution in the bone matrix (see section "Discussion").

In Table 2, the first p-value column gives the results of paired permutation tests, comparing the difference between an SHG patch and the corresponding TPaF patch, for each statistic. The majority of these p-values are significant at the 5% level, demonstrating that the choice of imaging technique results in significantly different persistent statistics. In these paired tests, the statistics relating to the sizes of the micro-holes are significantly different between the imaging techniques. This is to be expected, as the SHG images are detecting micro-holes in the collagen structure, in comparison to micro-holes in the autofluorescence of the TPaF image patches. However, the number of micro-holes was not found to be significantly different between the imaging techniques per patch, indicating a more consistent count of micro-holes between TPaF and SHG images.

By training a classifier on the persistence statistics, we can predict whether an unseen image patch is typical of a condition. Here, we compare image patches from OcnVEGFKO (test) mice to WT (control) mice, per sex. We used a support vector classifier (section "Classification") with 10 of the persistence statistics, per quadrant: number of points, persistent entropy, and the mean, median, standard deviation of births and deaths as well as the 25th percentile of births and the 75th percentile of deaths. We omit the number of micro-holes with specific radii as they sum to the number of micro-holes, and similarly we also omit the interquartile ranges (IQR) as they are the difference between the quartiles. In addition, the 25th percentile of deaths is omitted due to fewer significant differences (see Table 2). The results (Table 3) show a good separation between OcnVEGFKO and WT samples in the males, and reasonably strong results for females, given that there are far fewer significant differences. Further, when including features from quadrant 1 of H_1, results shown in brackets, there is no improvement, which suggests that the features from quadrant 2 of H_0 contain enough topological information on their own.

Discussion

We have presented a topology-based workflow to quantify morphological features, namely microstructure of the holes in the bone, in non-linear microscopy images. We extracted fully interpretable topological statistics from each image patch (Table 1), which quantify biologically relevant characteristics such as the number, radii and crowding of micro-holes throughout the samples. We demonstrated our methodology on mice bone samples and found significant differences in topological statistics between samples from OcnVEGFKO (test) and WT (control) images (Table 2), in line with previous observations[8]. Furthermore, we were able to accurately classify unseen image patches, using a support vector machine trained with a subset of the topological statistics. On this data, we see improved classification performance on patches from males in comparison to females (Table 3), stemming from the stronger significant differences between male OcnVEGFKO and WT (Table 2). The predictions for the SHG images are notably more accurate, which demonstrates the benefits of using a specialised microscopy technique, such as SHG, which captures the collagen structure.

The main strengths of our methodology are its interpretability, objectivity, and flexibility. Indeed, we were able to directly relate observable, biologically relevant micro-hole characteristics to specific topological statistics and determine whether differences were statistically significant. Namely, using an SEDT filtration, we encoded geometric information, such as radius or inter-micro-hole distances, using topological summaries of persistent homology. Our workflow is automated and objective, relying only on the choice of patch size, which reduces the potential for human error or bias. The persistent homology methodology itself is highly flexible, and can be customised to analyse other morphological characteristics of interest, by tailoring the filtration and choice of persistent statistics, and it can also be used to separate and summarise features of specific sizes (see Fig. 4c or Table 2).

By taking patches of the images, we can analyse the localised microstructure of the samples (such as the average minimal radius of micro-holes in region), at the cost of ignoring the macrostructure (the overall 'shape' of the sample), which is irrelevant to our structural analysis. (This is not a real limitation, as our topological workflow can be applied to whole sample images, and the topological summaries adapted to global morphological features of interest.) By using image patches, we also increase the number of images, and, together with the engineered topological features, we make classification tasks feasible on small sample sizes, as demonstrated in our mice data set (Table 3). Moreover, the patch-based approach allows for clear, intuitive visualisations of the topological statistics on the (binary) whole sample image. These visualisations can be used to indicate atypical or interesting regions within a single sample, and allow direct comparison of full sample images across a set of samples, which can highlight consistent regional effects or anomalous samples. Also, it should be noted that the background outside of the sample will register as a single micro-hole per patch containing background, which we mitigated by trimming images to reduce the size of the background component.

There are several existing methods to quantify porosity in bone tissue, including the method in[8,43] applied to SR CT images of the same samples in[8]. Porosity is a volumetric measure of how porous a sample is. Despite analysing micro-holes in 2D images of 3D 'pore' structures, the topological method on SHG images that we have presented is comparable to this 'geometric' method on SR CT image data. However, our topological method offers a distinct advantage over existing methods in that it is automated, and encodes a more diverse summary of different microstructure characteristics, including micro-hole crowdedness, which can be used in assessing a wider range of morphological characteristics in bone samples.

An important limitation when using persistent homology is that it only uses one scale parameter (the SEDT values in our analysis), as multi-parameter persistent homology is problematic[44]. In particular, the original pixel intensities are not used beyond thresholding into a binary image. This is necessary for the SEDT filtration, which allows us to separate (and thus summarise) the individual micro-holes (black pixels). Our method, therefore, relies on a good separation of the images to a binary format of bone matrix and micro-hole phase. Note that the existing geometrical method we mentioned above also relies on converting greyscale images into binary.

Finally, in terms of computational cost, the calculation of the persistent statistics took a couple seconds per image patch, and can be run in parallel.

Conclusion

We have provided a new method for the detection and quantification of micro-hole morphology on microscopy images. Its input is any greyscale image, which is converted into a binary image, and a patch size, and it uses a topological method to summarise micro-holes (regions without signal, or connected pieces of black pixels in the binary image) morphology per image patch. Namely, it detects individual micro-holes (H_0 points) and their radius (largest inscribed circle, H_0 birth), as well as the distance to the closest micro-hole (H_0 death), in an automatic way. Additionally, we can also summarise the signal region morphology, by detecting signal regions surrounded by micro-holes (H_1 points), and the formation of loops around signal regions (H_1 birth), and the radius of such regions (H_1 death). Our topological method is automatic, requiring only a greyscale image and patch size, robust, as persistent homology is known to be robust against noise (a small perturbation of the input results in a small perturbation of the output[45]), and interpretable, in the sense that each topological summary is linked to a micro-hole morphological characteristic (Table 1).

We demonstrate how micro-hole morphology analysis using our topological method is sufficient to detect significant differences, and classify with high predictive power, on transgenic mice bone TPaF and SHG images, despite a small sample size. The use of SHG images provided us with a detailed analysis of micro-holes in the collagen structure, which can be directly compared to those in the autofluorescence of TPaF images. There is a clear benefit to analysing the SHG images, when we look at the comparative strength of the results between the SHG and TPEF images. SHG images are typically used to study collagen structure[46], mostly focusing on analysing

or extracting fibres. It would be interesting to combine our topological methods with existing fibre analysis in the context of collagen structure in SHG images, but this is beyond the scope of the present article.

Our topological method provides a new tool in bone image analysis, and is applicable to other samples, materials and imaging techniques where we want to quantify 'micro-hole' (lack of signal regions) morphology. We have provided ready-to-use python code for the entire workflow. Note that our topological summaries can be combined with other image analysis features, in particular to improve the predictive performance of a classifier. In that sense, our method can also be seen as high-quality morphological feature engineering for image analysis.

Our results demonstrate the usefulness, versatility, and potential of topological analysis methods in non-linear microscopy imaging. By validating this method on bone samples, we broaden the bone analysis toolkit by adding a flexible topological workflow, and new microstructure topological features for bone. We expect our, and similar, topological methods to be useful in other biological research settings on tissue samples beyond bone, and for a wide range of biological applications. In the future, we plan to adapt our topological method to analyse other microscale morphological characteristics beyond the structure of micro-holes in the image data, by modifying the filtration and customising the topological summaries. One important avenue for future work is to explore the potential for use as a diagnostic aid, and tailor this method to specific diseases which alter bone structure at the microscale within human samples.

Data availability

References

1. Murphy, D. B. & Davidson, M. W. Two-photon excitation fluorescence microscopy. In *Fundamentals of Light Microscopy and Electronic Imaging*, chap. 14, 307–329 (Wiley, 2012). https://doi.org/10.1002/9781118382905
2. Lakowicz, J. R. *Principles of Fluorescence Spectroscopy* 3rd edn. (Springer, 2006).
3. New, G. *Introduction to Nonlinear Optics* (Cambridge University Press, 2011).
4. Bueno, J. M., Ávila, F. J. & Artal, A. P. Second harmonic generation microscopy: A tool for quantitative analysis of tissues. *Microsc. Anal.* (2016). https://doi.org/10.5772/63493
5. Fine, S. & Hansen, W. P. Optical second harmonic generation in biological systems. *Appl. Opt.* **10**, 2350–2353. https://doi.org/10.1364/AO.10.002350 (1971).
6. Jones, M. G. *et al.* Nanoscale dysregulation of collagen structure–function disrupts mechano-homeostasis and mediates pulmonary fibrosis. *eLife* **7**, e36354. https://doi.org/10.7554/eLife.36354 (2018).
7. Sharma, A. *et al.* Multiscale molecular profiling of pathological bone resolves sexually dimorphic control of extracellular matrix composition. *Dis. Models Mech.* **14**. https://doi.org/10.1242/dmm.048116 (2021).
8. Goring, A. *et al.* Regulation of the bone vascular network is sexually dimorphic. *J. Bone Miner. Res.* **34**, 2117–2132. https://doi.org/10.1002/jbmr.3825 (2019).
9. Edelsbrunner, H., Letscher, D. & Zomorodian, A. Topological persistence and simplification. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, 454–463. https://doi.org/10.1109/SFCS.2000.892133 (2000).
10. Collins, A., Zomorodian, A., Carlsson, G. & Guibas, L. J. A barcode shape descriptor for curve point cloud data. *Comput. Graph.* **28**, 881–894. https://doi.org/10.1016/j.cag.2004.08.015 (2004).
11. Hatcher, A. *Algebraic Topology* (Cambridge University Press, 2002).
12. Garside, K., Henderson, R., Makarenko, I. & Masoller, C. Topological data analysis of high resolution diabetic retinopathy images. *PLoS One* **14**, e0217413. https://doi.org/10.1371/journal.pone.0217413 (2019).
13. Henderson, R., Makarenko, I., Bushby, P., Fletcher, A. & Shukurov, A. Statistical topology and the random interstellar medium. *J. Am. Stat. Assoc.* **115**, 625–635. https://doi.org/10.1080/01621459.2019.1647841 (2020).
14. Atienza, N., Gonzalez-Diaz, R. & Soriano-Trigueros, M. A new entropy based summary function for topological data analysis. *Electron. Notes Discrete Math.* **68**, 113–118. https://doi.org/10.1016/j.endm.2018.06.020 (2018).
15. Biscio, C. A. N. & Møller, J. The accumulated persistence function, a new useful functional summary statistic for topological data analysis, with a view to brain artery trees and spatial point process applications. *J. Comput. Graph. Stat.* **28**, 671–681. https://doi.org/10.1080/10618600.2019.1573686 (2019).
16. Adams, H. *et al.* Persistence images: A stable vector representation of persistent homology. *J. Mach. Learn. Res.* **18**, 1–35 (2017).
17. Bubenik, P. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* **16**, 77–102 (2015).
18. Bubenik, P. & Dłotko, P. A persistence landscapes toolbox for topological statistics. *J. Symb. Comput.* **78**, 91–114. https://doi.org/10.1016/j.jsc.2016.03.009 (2017).
19. Chung, M. K., Bubenik, P. & Kim, P. T. Persistence diagrams of cortical surface data. In *Information Processing in Medical Imaging. Lecture Notes in Computer Science* (eds Prince, J. L. *et al.*) 386–397 (Springer, 2009). https://doi.org/10.1007/978-3-642-02498-6_32.
20. Belchi, F. *et al.* Lung topology characteristics in patients with chronic obstructive pulmonary disease. *Sci. Rep.* **8**, 1–12. https://doi.org/10.1038/s41598-018-23424-0 (2018).
21. Nakane, K., Takiyama, A., Mori, S. & Matsuura, N. Homology-based method for detecting regions of interest in colonic digital images. *Diagn. Pathol.* **10**, 36. https://doi.org/10.1186/s13000-015-0244-x (2015).
22. Adcock, A., Rubin, D. & Carlsson, G. Classification of hepatic lesions using the matching metric. *Comput. Vis. Image Underst.* **121**, 36–42. https://doi.org/10.1016/j.cviu.2013.10.014 (2014).
23. Chung, Y.-M., Hu, C.-S., Lawson, A. & Smyth, C. Topological approaches to skin disease image analysis. In *2018 IEEE International Conference on Big Data (Big Data)*, 100–105. https://doi.org/10.1109/BigData.2018.8622175 (2018).
24. Dunaeva, O., Edelsbrunner, H., Lukyanov, A., Machin, M. & Malkova, D. The classification of endoscopy images with persistent homology. In *2014 16th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, 565–570. https://doi.org/10.1109/SYNASC.2014.81 (2014).
25. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66. https://doi.org/10.1109/TSMC.1979.4310076 (1979).
26. Pritchard, Y. Topological-bone-analysis. https://github.com/yp4g14/topological-bone-analysis
27. Kaczynski, T., Mischaikow, K. & Mrozek, M. *Computational Homology. Applied Mathematical Sciences* (Springer, 2004).

28. Ghrist, R. Barcodes: The persistent topology of data. *Bull. Am. Math. Soc.* **45**, 61–76. https://doi.org/10.1090/S0273-0979-07-01191-3 (2007).
29. Zomorodian, A. & Carlsson, G. Computing persistent homology. *Discrete Comput. Geom.* **33**, 249–274. https://doi.org/10.1007/s00454-004-1146-y (2005).
30. Robins, V., Saadatfar, M., Delgado-Friedrichs, O. & Sheppard, A. P. Percolating length scales from topological persistence analysis of micro-CT images of porous materials. *Water Resour. Res.* **52**, 315–329. https://doi.org/10.1002/2015WR017937 (2016).
31. Herring, A. L., Robins, V. & Sheppard, A. P. Topological persistence for relating microstructure and capillary fluid trapping in sandstones. *Water Resour. Res.* **55**, 555–573. https://doi.org/10.1029/2018WR022780 (2019).
32. Moon, C., Li, Q. & Xiao, G. Predicting survival outcomes using topological features of tumor pathology images. arXiv:2012.12102 [cs, stat] (2020)
33. Koseki, K. *et al.* Assessment of skin barrier function using skin images with topological data analysis. *npj Syst. Biol. Appl.* **6**, 1–9. https://doi.org/10.1038/s41540-020-00160-8 (2020).
34. Obayashi, I. Volume-optimal cycle: Tightest representative cycle of a generator in persistent homology. *SIAM J. Appl. Algebra Geom.* **2**, 508–534. https://doi.org/10.1137/17M1159439 (2018).
35. Rucco, M., Castiglione, F., Merelli, E. & Pettini, M. Characterisation of the idiotypic immune network through persistent entropy. In *Proceedings of ECCS 2014, Springer Proceedings in Complexity* (eds Battiston, S. *et al.*) 117–128 (Springer International Publishing, 2016). https://doi.org/10.1007/978-3-319-29228-1_11.
36. Atienza, N., Gonzalez-Díaz, R. & Soriano-Trigueros, M. On the stability of persistent entropy and new summary functions for topological data analysis. *Pattern Recognit.* **107**, 107509. https://doi.org/10.1016/j.patcog.2020.107509 (2020).
37. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x (1948).
38. Wilcox, R. R. 8—Comparing two independent groups. In *Applying Contemporary Statistical Techniques* (ed. Wilcox, R. R.) 237–284 (Academic Press, 2003). https://doi.org/10.1016/B978-012751541-0/50029-8.
39. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **57**, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x (1995).
40. James, G., Witten, D., Hastie, T. & Tibshirani, R. Support vector machines. In *An Introduction to Statistical Learning: With Applications in R, Springer Texts in Statistics* (eds James, G. *et al.*) 337–372 (Springer, 2013). https://doi.org/10.1007/978-1-4614-7138-7_9.
41. James, G., Witten, D., Hastie, T. & Tibshirani, R. Resampling methods. In *An Introduction to Statistical Learning: With Applications in R, Springer Texts in Statistics* (eds James, G. *et al.*) 175–201 (Springer, 2013). https://doi.org/10.1007/978-1-4614-7138-7_5.
42. James, G., Witten, D., Hastie, T. & Tibshirani, R. Classification. In *An Introduction to Statistical Learning: With Applications in R, Springer Texts in Statistics* (eds James, G. *et al.*) 127–173 (Springer, 2013). https://doi.org/10.1007/978-1-4614-7138-7_4.
43. Núñez, J. *et al.* Regional diversity in the murine cortical vascular network is revealed by synchrotron X-ray tomography and is amplified with age. *Eur. Cells Mater.* **35**, 281–299. https://doi.org/10.22203/eCM.v035a20 (2018).
44. Otter, N., Porter, M. A., Tillmann, U., Grindrod, P. & Harrington, H. A. A roadmap for the computation of persistent homology. *EPJ Data Sci.* **6**, 17. https://doi.org/10.1140/epjds/s13688-017-0109-5 (2017) (**1506.08903.**).
45. Cohen-Steiner, D., Edelsbrunner, H. & Harer, J. Stability of persistence diagrams. *Discrete Comput. Geom.* **37**, 103–120. https://doi.org/10.1007/s00454-006-1276-5 (2007).
46. Ouellette, J. N. *et al.* Navigating the collagen jungle: The biomedical potential of fiber organization in cancer. Bioengineering **8**. https://doi.org/10.3390/bioengineering8020017 (2021).

Acknowledgements

Author contributions

A.S. and C.C. sourced, prepared and imaged the samples, Y.P., R.S.G., H.O. and S.M. conceived the project, R.S.G., H.O. and S.M. supervised the project, Y.P. completed the calculations and prepared the first draft of the manuscript. All authors reviewed and contributed to the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-28985-3.

Correspondence and requests for materials should be addressed to Y.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.