Title: Lung Topology Characteristics in patients with Chronic Obstructive Pulmonary Disease

Authors: Francisco Belchi[1†], Mariam Pirashvili[1], Joy Conway[2 3], Michael Bennett[3 4], Ratko Djukanovic[3 4 ‡], Jacek Brodzki[1 * §]

Affiliations:

[1]Mathematical Sciences, University of Southampton, UK.

[2]Faculty of Health Sciences, University of Southampton, UK.

[3]NIHR Southampton Respiratory and Critical Care Biomedical Research Centre.

[4]Clinical and Experimental Science, Faculty of Medicine, University of Southampton, UK.

*To whom correspondence should be addressed: J.Brodzki@soton.ac.uk

† ORCID Id: 0000-0001-5863-3343

‡ ORCID Id: 0000-0001-6039-5612

§ ORCID Id: 0000-0002-4524-1081

One Sentence Summary (A brief summary of the main result of the paper, without excessive jargon): We computed new topological descriptors of the bronchial tree revealed through chest CT scans of patients with chronic obstructive pulmonary disease (COPD) to create new radiomic features that stratify the patient cohort in agreement with the GOLD guidelines for COPD and can

distinguish between inspiratory and expiratory scans. These results form a proof of concept that topological methodology applied to lung images offers a new, clinically meaningful way to discover a finer classification of COPD, increasing the possibilities for more personalized treatment.

Abstract: Quantitative features that can currently be obtained from medical imaging do not provide a complete picture of Chronic Obstructive Pulmonary Disease (COPD). In this paper, we introduce a novel analytical tool based on persistent homology, that extracts quantitative features from chest CT scans to describe the geometric structure of the airways inside the lungs. We show that these new radiomic features stratify COPD patients in agreement with the GOLD guidelines for COPD and can distinguish between inspiratory and expiratory scans. These CT measurements are very different to those currently in use and we demonstrate that they convey significant medical information. The results of this study are a proof of concept that topological methods can enhance the standard methodology to create a finer classification of COPD and increase the possibilities of more personalized treatment.

Introduction

Chronic obstructive pulmonary disease (COPD) is a progressive lung disease, affecting more than 200 million people worldwide. COPD is the fourth leading cause of death in the world and is projected to be the third leading cause of death by 2020. There were more than 3 million deaths from COPD in 2012 worldwide. The global burden on health resources as a result of COPD is expected to rise [1, 2]. COPD is characterized by chronic inflammation of the bronchi and the lung parenchyma, resulting in varying degrees of obstructive bronchitis and emphysema due to remodeling of the airways and destruction of the alveoli, respectively. Although its pathology is heterogeneous, in functional terms, all forms of COPD result in loss of lung function, which is usually quantified by measuring the forced expiratory volume in 1 second (FEV1) and the Forced Vital Capacity (FVC). While these spirometry measures are widely used in clinical practice, both to diagnose and stratify COPD by severity, they have important limitations, the main being that they are integrative measurements which, therefore, do not take into account the highly heterogeneous regional pathological changes of COPD [3]. Furthermore, FEV1 correlates weakly with clinical outcomes and health status [4, 5, 6].

For the needs of COPD, lung function measurements are increasingly complemented by imaging methods as a means of visually quantifying regional ventilation and perfusion abnormalities, gas trapping, emphysema, and airway remodeling [3]. High-resolution computed tomography (HRCT) scans are the most widely used form of imaging, with MRI and nuclear medicine increasingly, but still less commonly, used. Technical advances have resulted in dramatic reductions in radiation dose of CTs, allowing repeat imaging in longitudinal studies. Assessment of bronchial wall and cross-section thickness is comparable to histological quantification and also enables estimation of the degree of small airways disease that are not directly visualized by CT [7]. Of note, CT imaging allows for detection of lung pathology, such as smoking-related inflammation of the small, distal bronchi (bronchiolitis), years before airflow limitation is detected by spirometry [8]. For example, CT-detected emphysema, assessed by the 15% percentile (Perc15) technique is able prospectively to identify rates of lung function decline, even in individuals in whom spirometry does not detect airway obstruction [9].

Common CT measurements in COPD research include lung attenuation area, mean lung density, airway wall area percentage, Perc15, lung volume, airway wall thickness and airway lumen area [10]. There is, however, significant room for development of radiomic features derived by data-characterization algorithms applied to large sets of quantitative features extracted from medical images and, thereby, uncover characteristics that cannot be appreciated by the naked eye. In the current study, we have applied, to our knowledge for the first time ever, the technique of persistent homology to process lung CT data. We took advantage of the computational tool of persistent homology [11, 12, 13] to create topological descriptors which capture the complexity of the lung structure; this also enabled computation of a measure of similarity between images. Using this approach, our study has introduced a novel set of descriptors computable from a chest CT scan, focusing on characteristics that are very different from those used at present. Specifically, we started by considering three new radiomic features: upwards complexity, which quantifies the way branches stretch upwards, the length of the bronchial tree visible in an inspiratory CT scan, and the number of bifurcations in the same tree. We then showed that these three numerical values are very closely related and any of them can stratify the inspiratory CT scans of our cohort into groups that agree with those given by the GOLD guidelines of COPD. Of note, these stratification results are better than those obtained by other CT measurements, like the emphysema score and the volume of the lumen. Apart from the upwards complexity, we also computed two additional numerical values related to the way branches stretch upwards. Using these, we could clearly distinguish between inspiratory and expiratory CT scans. Additionally, we observed that we can also classify our cohort into healthy individuals and COPD patients by quantifying and classifying the topological structure of the space between the lung periphery and the visible airways in an inspiratory CT scan. Finally, we developed a computable characteristic that describes how the branches in the bronchial tree curve towards one another and showed that this radiomic feature correlates with lung function more strongly when the computations are done using the expiratory CT scans rather than the inspiratory CT scans, a phenomenon that is also seen when using standard CT measurements.

We propose that the relation between lung diseases and the shape of the bronchial tree, including properties such as trajectory changes, are of value to advancing our understanding of the mechanisms of COPD. We also propose that further research that applies this method in prospective, longitudinal studies and interventional trials is justified.

Results

The overall aim of this study was to develop a set of new radiomic features that can distinguish between healthy non-smokers as well as healthy smokers and patients with COPD. For this purpose, the following four study participant groups defined by smoking status and spirometry given by the GOLD guidelines [6] were studied: healthy non-smokers and healthy smokers (both judged as healthy by spirometry showing FEV1>80% of predicted and FEV1/FVC>0.75), mild COPD patients, consisting of GOLD stage 1 (with FEV1≥80% of predicted and FEV1/FVC<0.70) and moderate COPD patients, consisting of GOLD stage 2 (50%≤ FEV1<80% of predicted and FEV1/FVC<0.70). See Materials and Methods for cohort details and data used.

In this paper, we made use of Topological Data Analysis (TDA), with emphasis on persistent homology, for the computation of our new radiomic features. In Supplementary Materials, we explain what persistent homology is and how it works. In Materials and Methods, we explain the way we use this TDA tool to obtain each of our geometric signatures.

Directional complexity

For this computation, we began by extracting a graph representing the bronchial tree from each inspiratory CT scan (see Materials and Methods). Starting from the top of the scan, we recorded the height at which a segment of the bronchial tree changes direction and starts pointing upwards or downwards. We computed a geometric summary of this information using TDA as described in Materials and Methods. This consisted of a single numerical output we call upwards complexity, which was obtained by counting the number of times a particular branch changes its trajectory to start stretching upwards and sum this number over all branches in the bronchial tree. Upwards complexity allowed us to stratify the inspiratory CT scans of our cohort into COPD groups that agree with those given by the GOLD guidelines. A boxplot illustrating this group-separation can be found in Figure 1A and details of the pairwise Kolmogorov-Smirnov (KS) tests can be found in Table 1A.

We also studied how the branches in the bronchial tree bend in other directions, obtaining a different number for each direction. This directional complexity in directions other than upwards did not improve the group-separation results obtained by the combination of upwards complexity and bronchial tree length (introduced in the next section), hence our focus on the latter two measurements. More details on this are given in Materials and Methods.

Length of the bronchial tree and number of branching points

To complement directional complexity, we measured the length of the entire bronchial tree observable in an inspiratory CT scan. The length of the bronchial tree was estimated from a graph representing the bronchial tree in the CT (see Materials and Methods) using the number of vertices in this graph as a proxy for the length of the bronchial tree. Using this measure, we could again stratify the inspiratory CT scans of our cohort into groups that agree with those given by the GOLD guidelines. For these group-separation results, see Figure 1B for the boxplot and Table 1B for the results of the pairwise KS test. In particular, notice how the bronchial tree length separates the group of moderate COPD patients from all other three groups.

It is remarkable that the combination of upwards complexity and the length of the bronchial tree were able to distinguish all groups, except for that of healthy smokers (HS), from those of healthy non-smokers (HNS) or mild COPD patients (Mild). This can be checked by observing how for all comparisons except for HSvsHNS and HSvsMild, either in Table 1A or in Table 1B. Of note, these findings may indicate that the group of healthy smokers is heterogeneous and intersects with the healthy non-smokers at one end and with the mild COPD patients at the other. This could not be established using FEV1 (% of predicted) and the ratio FEV1/FVC.

To investigate whether any part of the lung may be contributing more to the above findings, we computed the length of the bronchial tree starting from different airway generations. This showed that such thresholding does not improve the separation presented in Figure 1B and Table 1B regardless of the generation from which the computation began. In a separate computation, an almost identical separation to the one in Figure 1B and Table 1B was reproduced by using the total number of points where airways branch out instead of the length of the bronchial tree.

Relationship between directional complexity and bronchial tree length

When investigating the directional complexity in any given direction and the length of the bronchial tree, these two measures were found to be strongly related (see Figure 2A for an illustration of this using upwards complexity). Moreover, as shown in Figure 2A, this close relation was maintained across the four groups in the cohort.

Comparison with other analytical methods

Having observed significant separation between study groups using our new radiomic methods on inspiratory CT scans, we looked for similar differences between subject groups when using other CT measurements. Specifically, we quantified emphysema using the standard measure of percentage area of low attenuation and we approximated the volume of the airway lumen as the number of voxels inside the airways (see the red airway structure in Figure 3A). This showed that the differences between subject groups identified by our radiomic features were much more significant than the differences identified by the emphysema score and the volume of the lumen (compare the boxplots Figure 1C and Figure 1D with those in Figure 1A and Figure 1B, and the numerical results in Table 1C and Table 1D with those in Table 1A and Table 1B). Indeed, using the volume of the lumen, we did not find any difference between subject groups with a p-value<0.05, and the emphysema score only found two such differences – namely, that between healthy smokers and moderate COPD patients (|KS|=0.55, p=9.88$\cdot 10^{-3}$) and that between healthy non-smokers and moderate COPD patients (|KS|=0.48, p=4.16$\cdot 10^{-2}$). These separations were weaker and less significant than the separation of the same groups obtained using the bronchial tree length (Figure 1B and Table 1B).

Relation to height

When assessing standard lung function measurements, the values are typically normalized by the individual's height. To study the effect of height on some of our new radiomic features, we divided the upwards complexity of each participant by the person's height. We then examined which participant groups separated more clearly by either applying or not the height normalization. To this end, we compared 2-sample Kolmogorov–Smirnov tests. Normalizing upwards complexity exhibited a clearer separation in 3 cases (using the notation in Figure 1, those cases are the comparisons Mod-HNS, HS-Mod and Mod-Mild) and a less clear separation in 2 cases (Mild-HNS, HS-Mild) (compare Figure 1E and Table 1E to Figure 1A and Table 1A, respectively).

We repeated the same normalization with the length of the bronchial tree and found that in 2 instances (HNS-Mild, HS-Mod), not normalizing by height provided a more clear separation between groups, and normalizing did not improve the clarity of separation in any instance (compare Figure 1F and Table 1F to Figure 1B and Table 1B, respectively). This suggests that, unlike standard spirometry, the bronchial tree length may capture bronchial structure information relevant in COPD in a way that is independent of height.

Comparison of expiratory and inspiratory phase CTs

For 30 participants (8 healthy non-smokers, 9 healthy smokers, 8 mild COPD and 5 moderate COPD), both inspiratory and expiratory CT scans were obtained. This provided an opportunity to demonstrate first that our methodology can not only distinguish between healthy individuals and COPD patients but also detects differences in structure between the inspiratory and expiratory phases of the breathing cycle. Furthermore, we showed that the amount to which branches in the bronchial tree curve towards one another correlates with lung function more strongly in expiratory CT scans than in inspiratory ones. The stronger correlation with expiratory scans is in keeping with our previous findings using standard CT measurements that mean lung density (MLD) during expiration correlated better with reduced lung function that inspiratory MLD [14].

To address the first point, we considered again the height at which branches in the bronchial tree start or stop stretching upwards, as used in the computation of upwards complexity. We used the same input to compute a different topological summary (see Materials and Methods), which allowed us to compare the scans of different participants and plot them together. The output of our computations were two values per CT scan, which we used as coordinates of a point in the plane, see Figure 4B. This showed a clear separation between the inspiratory and expiratory CT scans.

To quantify how branches curve towards one another in the bronchial tree graph, we introduced another radiomic feature, called *branch-to-branch proximity,* which. This was done by virtually thickening the visible airways and recording the thickness at which the airways begin to touch (see Figure 5 and Figure 6 for an illustration). Using this approach, we found that the branch-to-branch proximity observed in the expiratory phase correlated more strongly with FEV1 (% of predicted) than the branch-to-branch proximity observed in the inspiratory phase (compare Figure 2B and Figure 2C). Again, this was consistent with standard CT measurements, which also correlate better with FEV1 (% of predicted) when measured during expiration [10, 14].

Small airways

Having shown that we can compute clinically meaningful topological features of the airways by using their tree structure, we showed that the shape of the space separating the lung periphery from the airways visible in an inspiratory CT scan is also related to the development of COPD. Our CT scan data consists of voxels which are cubes 0.7mm long in each direction, giving a spatial resolution of 2.1mm. This makes the small airways, which are defined as those with a diameter <2mm, invisible in a CT scan. However, it is well known that small airways dysfunction plays a key role in COPD. Hence, what happens in the void between the visible airways and the lung periphery is crucial. To overcome the relatively low resolution of the standard CT scan, we found a topological way to quantify and classify the structure of that void and showed how the resulting radiomic feature can distinguish between healthy individuals and COPD patients. This was achieved by placing virtual balls centered within the visible structure (airways and lung periphery) and allowing them to expand until they fully occupied the space. This procedure uses thickening in a similar way to Figure 5, see Materials and Methods for details. The output is a pair of real numbers (x, y) representing each CT scan. In Figure 4A, we represent each CT scan as a point in the plane with the corresponding coordinates (x, y). This approach placed the healthy smoking and non-smoking individuals into one group that was distinct from the mild and moderate COPD patients who formed another group (see Figure 4A).

Discussion

Since its creation, persistent homology, a key tool of TDA, has developed rapidly, both in terms of its mathematical foundations [15, 16, 17] and possible applications. Persistent homology has been used in fields as diverse as digital imaging [18, 19], sensor networks coverage [20], materials science [21, 22], molecular modelling [23, 24, 25], signal processing [26, 27] and virus evolution [28]. In this study, we took advantage of the ability of persistent homology to offer alternative ways of measuring global properties of complex objects through the use of topology.

In respiratory imaging, this method represents a completely new way of taking measurements of the bronchial tree. In contrast to existing methods, such as measuring the dimensions of the airway walls and lumen, which look at airway branches individually, our approach condenses the topological properties of the entire bronchial tree into a small number of unique characteristics for each individual. Along with previous studies [29, 30], this creates major new opportunities for persistent homology to be used more widely in medicine, in particular in clinical radiology. This new methodology is especially applicable to studying the lung at the population level because of the manner in which it represents the complexity of the airway tree through a single low-dimensional data point that represents the entire bronchial tree. By collecting these data over a large number of subjects and combining them with other imaging, physiological and measurements of pathobiological biomarkers, we could build a picture of how variations in the topological nature of the bronchial tree impact on the pathophysiology of people with a variety of respiratory diseases such as COPD, asthma and idiopathic pulmonary fibrosis (IPF). This is likely to have significant translational impact as a valuable tool for use in deep-phenotyping, which is central in stratified medicine and precision medicine [31].

We have created a set of novel radiomic features which capture the overall complexity of the lung structure and enable a quantitative comparison of the CT scan images. We have demonstrated that these properties are important in the context of a common respiratory disease. These measurements provided a more complete picture of differences between the four groups in this study than standard CT measurements. In particular, there was a significant relationship between upwards complexity, the length of the bronchial tree, and COPD severity.

Additionally, our comparison between inspiratory and expiratory phases can have various applications. The manner in which tissue inside the lung expands and contracts throughout the breathing cycle is known to be an indicator of disease, for example gas trapping in COPD. Comparison of the inspiratory and expiratory scans can therefore be exploited as a means of making localized measurements of disease [32]. Our proposed technique makes it possible to study how the shape of the bronchial tree changes during the breathing cycle and offers the potential to be a new method for the identification of localized areas of disease, such as gas trapping. Similarly, indices that are the subject of current, clinical, pulmonary CT research also include the Parametric Response Mapping (PRM) technique, which uses co registration of paired inspiratory and expiratory scans to compare areas of low attenuation on a voxel to voxel basis [33, 34]. Our methodology may further inform the PRM technique and could be the subject of future research.

Our approach to persistent homology is similar to that employed in [30] to study cerebral vasculature, but our topological summaries, such as the directional complexity or branch-to-branch proximity, are different and have not been used before. As explained in Supplementary Materials, the output from persistent homology calculation is summarized in the so-called barcode. In the study of cerebral vasculature, Bendich et al [30] simplified the analysis by retaining the 100 longest bars from which summaries were produced using the Principal Component Analysis. In contrast, in our study this approach did not work as the number of bars can vary significantly between patients, which can be seen in Figure 1A where some participants have about 160 bars in the upwards complexity barcode, whereas others exhibit only about 20 bars. For this reason, in our study we retained the entire barcode without thresholding. From this input data, we computed a measure of similarity between scans of individual patients. This can be used for visualization or to train classification models in a way similar to the approach by Adcock et al. on hepatic lesions [29].

We also created a new topological characteristic to circumvent the relatively low spatial resolution of CT scans. This is one of the main mathematical novelties of the paper as it provides a first instance where topology has been used to infer the structure of the object under study. We achieved this by incorporating to the computation the boundary of the lung lobes. This technique can be applied to any kind of imaging, for example, to 3-dimensional Magnetic Resonance Angiography images of the arterial tree within the brain [30], where the new characteristics developed here can be used to enhance the analysis if the meninges are used in the same way we used the outer layer of the lobes in our study.

Of note, we made use of persistent homology in degrees 0, 1 and 2 in different ways to obtain different kinds of clinical insight. In degree 0, it was used to define the directional complexity (a number that can distinguish severity groups) and to characterize the distinction between the inspiratory and expiratory CT scans. Using persistent homology in degree 1, we showed that CT measurements correlate with FEV1 (% of predicted) more strongly during the expiratory phase than in the inspiratory phase, which could be expected based on similar observations in past CT studies [10, 14]. As stated before, the degree 2 was used to overcome the limitation of the low spatial resolution of CT scans by including information of the outer boundary of the lobes. This led to a much clearer visualization of the difference between the healthy and COPD participants, which could not be recovered using the topological characteristics in degrees 0 and 1. Thus, our methodology provides one of the first significant uses of the second-degree persistent homology in applications. The other uses of degree 2 up to date are summarized in [19, 35, 36].

In summary, this study has shown that our analytical method can extract information from CT scans to provide a new perspective on lung structure. Because this method can be readily applied to large CT datasets, we propose that it is of value for clinical research. Further studies are needed to assess its prognostic value in longitudinal and interventional studies.

Materials and Methods

All the procedures explained in the Results section can be formalized and computed efficiently through the tool of persistent homology, which is described in detail in the Supplementary Materials.

Study design and participants

The imaging data used for this study were acquired from two previous imaging studies performed in Southampton (manuscripts in preparation). Both studies focused on COPD and had identical inclusion and exclusion criteria. In both studies, participants were recruited into two COPD groups; GOLD stage 1 disease (FEV1/FVC ratio < 0.70 and FEV1≥80% of predicted) and GOLD stage 2 disease (FEV1/FVC ratio < 70% and FEV1 50-79% of predicted), referred to as mild and moderate COPD, respectively. Both healthy smoker and healthy non-smoker groups had no clinical evidence of obstructive airways disease, and had spirometry results of FEV1/FVC ratio >0.75 and FEV1>80% of predicted. Both studies were approved by the Southampton and West Hampshire local research ethics committee (LREC number: 11/SC/0319 and 09/H0502/91). In total, 64 participants were assessed (18 healthy non-smokers, 19 healthy smokers, 14 COPD GOLD-1 and 13 COPD GOLD-2).

MSCT imaging

Multi-Slice Computed Tomography (MSCT) scans were performed on a Siemens Sensation 64 CT scanner (Siemens Medical Solutions, Erlangen, Germany) using a high-resolution algorithm, with detector thickness 0.75 mm, pitch 1.0, effective mAs 90 and a tube voltage of 120kV. The high-resolution algorithm was chosen to ensure the best visualization of the airway tree [37], and the scanning was performed at suspended full inspiration and expiration. The images were reconstructed using a slice thickness of 0.75 mm, a reconstruction increment of 0.5 mm, and a sharp reconstruction algorithm. Additional reconstructions were also performed using several soft reconstruction kernels, including B30f and B35f, which were chosen to suit the recommended protocol in the Apollo analysis software (Vida Diagnostics, Iowa, USA).

MSCT analysis

The Apollo software (Vida Diagnostics, Iowa, USA) was used to perform the analysis of the multi-slice computed tomography scans. This software was designed to semi-automatically analyze pulmonary MSCT imaging data, including segmentation of the lungs, the airway tree and the lobes (see Figure 3A). For the needs of the current study, only the lung, lobes and airway tree were of interest. In many cases, the Apollo software was able to achieve the desired results entirely automatically, but for some participants it was necessary to manually edit the results of the lobe segmentation to ensure that they were defined as accurately as possible.

Custom software written in Matlab (R2015b, MathWorks, Natick, MA 01760-2098, US) was used to extract the specific details of the center lines and branch points of the airway tree from the .XML data output by the Apollo software. An example of the extracted center lines, along with the segmented airway tree is shown in Figure 3B, where the center lines are colored according to generation number and are, for simplicity, plotted between the branch points only. For all of the analysis in this paper, the complete center line information was used, which captured the true shape of the airway branches. In particular, there can be up to 264 extra points describing the shape of the bronchial tree between two branch points.

Along with the branch center lines, binary masks were exported representing the airway tree and lobes for each participant.

Persistent homology

Persistent homology [11, 12, 13] has been designed to provide numerical information about the key features of an object under study at a range of scales, which can be regarded as variable resolution at which the object is viewed. It is commonly used as a 3-step process: first, a simplified approximation of the object is built, which grows as the scale parameter r is varied. In this study, we made use of two different approximations: one based on alpha complexes (see Supplementary Materials) and one based on a notion of height function, explained in the directional complexity section (below) and expanded in the Supplementary Materials. We computed topological characteristics of the chosen approximation, using all scales at once. These are numerical invariants obtained by computing homology groups H_n of the approximation, and tracing the life-span of features as they appear and disappear with the changing scale. For each degree $n \geq 0$, this information is represented in the form of a collection of intervals with multiplicities, called the degree-n barcode explained in detail in the Supplementary Materials. These intervals have the form $[r_1, r_2)$ for different values of the changing parameter r, and are also known as *bars*, hence the name barcode for a collection of these.

Intuitively, degree-0 gives information about the evolution of the connected components along the sequence of growing representations of the object under study. Similarly, degree 1 indicates the evolution of the loops or holes, and degree 2 captures the evolution of cavities or voids, etc. We compared the resulting barcodes using pseudo-distance functions, the Wasserstein and Bottleneck pseudo-distances being examples with important stability properties; see Supplementary Materials.

In summary, we used persistent homology to take a growing approximation of an object, compute the associated degree-n barcode for some $n \geq 0$, and compare the corresponding barcodes of different objects using the Bottleneck or Wasserstein distances.

Directional complexity

We quantified the amount of changes in trajectory in a particular direction by defining a notion of directional complexity on the 3D graph representation of the bronchial tree described in the MSCT analysis subsection (above). To measure upwards complexity, we slid a horizontal plane downwards (see example in Figure 7A). At any given distance h from the top of the imaginary box containing the bronchial tree, X_h was defined as the part of the tree that sits above the plane at that position. In this way, we obtained an approximation of the bronchial tree that converged to the original tree as we increased the distance h from the top (see Figure 7A).

The degree-0 barcode corresponding to this sequence of growing graphs has the following interpretation: a bar of the form $[h_1, h_2)$ in this barcode indicates that there is a connected component C in the graph X_{h_1} which is not present in X_h for any $h < h_1$. Additionally, the following holds for h_2 but it does not hold for any $h < h_2$: in the graph X_{h_2}, the component represented by C will merge with another component of X_{h_2} which was present in X_h for some $h < h_1$.

In Figure 7B, we represent each bar of the form $[h_1, h_2)$ in the degree-0 barcode as a vertical line, with the starting point at distance h_1 from the top and end point at distance h_2 from the top. In this representation, every bar corresponds to a branch changing trajectory to start stretching upwards. We called upwards complexity the number of vertical lines in such a representation, *i.e.*, the number of upwards changes of trajectory of the airways.

To compute directional complexity in other directions, we rotated the bronchial tree, slid the plane top to bottom and counted the number of finite bars in the corresponding degree-0 barcode.

As mentioned in the Results section, directional complexity in other directions did not improve the group-separation results obtained by the combination of upwards complexity and bronchial tree length, hence our focusing on the upwards direction. For instance, by rotating 10˚, 220˚ and 0˚ around the X, Y and Z axes, respectively, following the right-hand rule, directional complexity produced no group separation at all. However, using instead the angles 20˚, 40˚ and 0˚, respectively, the group-separation results given by directional complexity were very similar to those of the bronchial tree length.

To generate these barcodes, we used the publicly available software package TDATools [38]. To compute the barcode of one of these graphs in a 3D box, we used the function 'rca1mfscm' of this package, which requires the definition of a function F on each vertex and edge of the graph. For instance, to compute the upwards complexity, we assigned to each vertex its distance to the top of the box, and to each edge, the maximum of the values of F attained at the two vertices it connects. Note that all barcodes in this study were computed with coefficients in the field of two elements, $\mathbb{Z}_2$.

Length of the bronchial tree

The length of the bronchial tree was estimated from the 3D graph representation of the bronchial tree described in the MSCT analysis subsection above. We used the number of vertices in this graph (that include not only the branch points but also the many vertices connecting consecutive branch points) as a proxy for the length of the bronchial tree. As stated in the Results section, a separate computation with only the branch points was performed and the results were similar to those in Figure 1B and Table 1B.

Small airways

For the computation of the representation in Figure 4A, we started with a 3D array of binary voxels representing the luminal surface of the airways together with the surface of the lobes as in Figure 3A. For each binary voxel image, we constructed a point cloud in $\mathbb{R}^3$ by including the coordinates of every voxel with value 1 and then built the alpha complex filtration (see Supplementary Materials) on these points. The degree-2 barcode of this filtration gave information about how the airways fill the cavity of the lobes. The alpha complex filtrations and their barcodes were computed using the GUDHI library [39].

Next, we computed the bottleneck distances between all the degree-2 barcodes. This gave a measure of distance between the lung scans by proxy, giving us a pseudo-metric on the set of lungs. The bottleneck distances were computed using the Hera software [40]. Due to computational constraints, we made use of the software's approximate bottleneck calculation. If one supplies a relative error, then the software computes an approximate distance which satisfies the inequality

$$\left|d_{exact} - d_{approx}\right| \Big/ d_{exact} < relative\ error,$$

where d_{exact} is the exact bottleneck distance and d_{approx} is the computed approximation, as described in the documentation of [40]. We used a relative error of 10^{-4}. After measuring the pairwise distances between all barcodes, we used Multi-Dimensional Scaling (MDS) to obtain a 2D representation shown in Figure 4A.

Expiratory CT analysis

For 30 participants (8 healthy non-smokers, 9 healthy smokers, 8 mild COPD and 5 moderate COPD), both inspiratory and expiratory CT scans were obtained. Recall that the upwards complexity was computed as the number of vertical lines in Figure 7B. This was, in turn, the number of bars in the degree-0 barcode constructed by considering the part of the bronchial tree graph that sits on top of a horizontal plane that we slide downwards.

In order to compute the representation shown in Figure 4B, we used the same degree-0 barcode in a different way. We compute such barcodes for both the inspiratory and expiratory bronchial tree graphs and compared those 60 barcodes (corresponding to the inspiratory and the expiratory bronchial tree of the 30 patients) using the Wasserstein2 distance (see Supplementary Materials). The Wasserstein computations were done with the software package Hera [40]. After calculating the distances between all these barcodes, the final representation in Figure 4B was obtained using a 2D MDS projection.

In a separate computation, we quantified how branches bend towards one another on the tree graph (See Figure 5 and Figure 6 for an intuitive illustration of this computation). We used the alpha complex filtration (see Supplementary Materials) built on the nodes of this graph. Next, we computed the degree-1 barcode, which consisted of points of the form (r_1, r_2). Finally, we defined the branch-to-branch proximity as the sum of the numbers $r_2 - r_1$ corresponding to all such points. We performed this for both the inspiratory and expiratory tree graphs of each participant. The alpha complex filtrations and their barcodes were computed using the GUDHI library [39].

Supplementary Materials

The Supplementary Materials expand on the Materials and Methods of the manuscript, and contain the background information on persistent homology. It is divided in 7 subsections: Persistent homology, Persistence modules and barcodes, Simplicial complexes, Comparing persistence diagrams, Stability, Height filtration and Alpha complexes. The Supplementary Materials also include two figures which illustrate some of the constructions in these sections.

References and Notes

[1] C. D. Mathers and D. Loncar, "Projections of Global Mortality and Burden of Disease from 2002 to 2030," *PLoS Medicine,* vol. 3, no. 11, p. e442, Nov 2006.

[2] R. Lozano, M. Naghavi, K. Foreman, S. Lim, K. Shibuya, V. Aboyans, J. Abraham, T. Adair, R. Aggarwal and S. Y. Ahn et al, "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010," *The Lancet,* vol. 380, no. 9859, p. 2095–2128, Dec 2012.

[3] H. O. Coxson, J. Leipsic, G. Parraga and D. D. Sin, "Using Pulmonary Imaging to Move Chronic Obstructive Pulmonary Disease beyond FEV1," *American Journal of Respiratory and Critical Care Medicine,* vol. 190, pp. 135-144, 2014.

[4] D. E. Doherty, "A Review of the Role of FEV1in the COPD Paradigm," *COPD: Journal of Chronic Obstructive Pulmonary Disease,* vol. 5, pp. 310-318, jan 2008.

[5] P. W. Jones, "Health Status and the Spiral of Decline," *COPD: Journal of Chronic Obstructive Pulmonary Disease,* vol. 6, pp. 59-63, jan 2009.

[6] "Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2017," [Online]. Available: http://goldcopd.org.

[7] P. A. de Jong, N. L. Müller, P. D. Paré and H. O. Coxson, "Computed tomographic imaging of the airways: relationship to structure and function," *European Respiratory Journal,* vol. 26, pp. 140-152, jul 2005.

[8] A. Sayiner, C. Hague, A. Ajlan, J. Leipsic, L. Wierenga, N. M. Krowchuk, N. Ceylan, A. Sayiner, D. D. Sin and H. O. Coxson, "Bronchiolitis in young female smokers," *Respiratory Medicine,* vol. 107, pp. 732-738, may 2013.

[9] F. A. A. M. Hoesein, B. de Hoop, P. Zanen, H. Gietema, C. L. J. J. Kruitwagen, B. van Ginneken, I. Isgum, C. Mol, J. van Klaveren and R., A. E. Dijkstra, H. J. M. Groen, H. M. Boezen, D. S. Postma, M. Prokop and J.-W. J. Lammers, "CT-quantified emphysema in male heavy smokers: association with lung function decline," *Thorax,* vol. 66, pp. 782-787, apr 2011.

[10] X. Xie, P. A. de Jong, M. Oudkerk, Y. Wang, N. H. T. Hacken, J. Miao, G. Zhang, G. H. de Bock and R. Vliegenthart, "Morphological measurements in computed tomography correlate with airflow obstruction

in Chronic Obstructive Pulmonary Disease: systematic review and meta-analysis," *European Radiology,* vol. 22, pp. 2085-2093, 2012.

[11] H. Edelsbrunner, D. Letscher and A. Zomorodian, "Topological persistence and simplification," *Discrete Comput. Geom.,* vol. 28, pp. 511-533, 2002.

[12] G. Carlsson and A. Zomorodian, "Computing persistent homology," *Discrete Comput. Geom.,* vol. 33, pp. 249-274, 2005.

[13] H. Edelsbrunner and J. Harer, "Persistent homology -- a survey," in *Surveys on discrete and computational geometry,* vol. 453, Amer. Math. Soc., Providence, RI, 2008, pp. 257-282.

[14] R. A. O'Donnell, C. Peebles, J. A. Ward, A. Daraker, G. Angco, P. Broberg, S. Pierrou, J. Lund, S. T. Holgate, D. E. Davies, D. J. Delany, S. J. Wilson and R. Djukanovic, "Relationship between peripheral airway dysfunction, airway obstruction, and neutrophilic inflammation in COPD," *Thorax,* vol. 59, pp. 837--842, Oct 2004.

[15] G. Carlsson and V. de Silva, "Zigzag persistence," *Found. Comput. Math.,* vol. 10, pp. 367-405, 2010.

[16] D. Cohen-Steiner, H. Edelsbrunner and J. Harer, "Extending Persistence Using Poincaré and Lefschetz Duality," *Found. Comput. Math.,* vol. 9, pp. 79-103, 2009.

[17] F. Belchí and A. Murillo, "A∞ persistence," *Appl. Algebra Engrg. Comm. Comput.,* vol. 26, pp. 121-139, 2015.

[18] V. Robins, P. J. Wood and A. P. Sheppard, "Theory and Algorithms for Constructing Discrete Morse Complexes from Grayscale Digital Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 33, pp. 1646-1658, 2011.

[19] J. A. Perea and G. Carlsson, "A Klein-Bottle-Based Dictionary for Texture Representation," *Int. J. Comput. Vision,* vol. 107, pp. 75-97, #mar# 2014.

[20] V. de Silva and R. Ghrist, "Coverage in sensor networks via persistent homology," *Alg. & Geom. Top.,* vol. 7, pp. 339-358, 2007.

[21] M. Kramár, A. Goullet, L. Kondic and K. Mischaikow, "Quantifying force networks in particulate systems," *Physica D: Nonlinear Phenomena* , vol. 283, pp. 37-55, 2014.

[22] R. MacPherson and B. Schweinhart, "Measuring shape with topology," *J. Math. Phys.,* vol. 53, pp. 073516, 13, 2012.

[23] P. K. Agarwal, H. Edelsbrunner, J. Harer and Y. Wang, "Extreme elevation on a 2-manifold," *Discrete Comput. Geom.,* vol. 36, pp. 553-572, 2006.

[24] V. Kovacev-Nikolic, P. Bubenik, D. Nikolić and G. Heo, "Using persistent homology and dynamical distances to analyze protein binding," *Stat. Appl. Genet. Mol. Biol.,* vol. 15, pp. 19-38, 2016.

[25] M. Gameiro, Y. Hiraoka, S. Izumi, M. Kramar, K. Mischaikow and V. Nanda, "A topological measurement of protein compressibility," *Japan Journal of Industrial and Applied Mathematics,* vol. 32, pp. 1-17, 2015.

[26] S. Emrani, T. Gentimis and H. Krim, "Persistent homology of delay embeddings and its application to wheeze detection," *IEEE Signal Processing Letters,* vol. 21, pp. 459-463, 2014.

[27] K. A. Brown and K. P. Knudson, "Nonlinear statistics of human speech data," *Internat. J. Bifur. Chaos Appl. Sci. Engrg.,* vol. 19, pp. 2307-2319, 2009.

[28] J. M. Chan, G. Carlsson and R. Rabadan, "Topology of viral evolution," *Proc. Natl. Acad. Sci. USA,* vol. 110, pp. 18566-18571, 2013.

[29] A. Adcock, G. Carlsson and D. Rubin, "Classification of hepatic lesions using the matching metric," *Comput. Vis. Image Und.,* vol. 121, pp. 36-42, 2014.

[30] P. Bendich, J. S. Marron, E. Miller, A. Pieloch and S. Skwerer, "Persistent homology analysis of brain artery trees," *Ann. Appl. Stat.,* vol. 10, pp. 198-218, 2016.

[31] C. M. Delude, "Deep phenotyping: The details of disease," *Nature,* vol. 527, pp. S14--S15, nov 2015.

[32] Y. Nagatani, K. Murata, M. Takahashi, N. Nitta, Y. Nakano, A. Sonoda, H. Otani, H. Okabe and E.

Ogawa, "A new quantitative index of lobar air trapping in chronic obstructive pulmonary disease (COPD): Comparison with conventional methods," *European Journal of Radiology,* vol. 84, no. 5, pp. 963 - 974, 2015.

[33] M. Kirby, Y. Yin, J. Tschirren, W. C. Tan, J. Leipsic, C. J. Hague, J. Bourbeau, D. D. Sin, J. C. Hogg and H. O. Coxson et al., "A Novel Method of Estimating Small Airway Disease Using Inspiratory-to-Expiratory Computed Tomography," *Respiration,* vol. 94, no. 4, p. 336–345, 2017.

[34] C. J. Galbán, M. K. Han, J. L. Boes, K. A. Chughtai, C. R. Meyer, T. D. Johnson, S. Galbán, A. Rehemtulla, E. A. Kazerooni and F. J. Martinez et al., "Computed tomography–based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression," *Nature Medicine,* vol. 18, no. 11, pp. 1711-1715, Oct 2012.

[35] J. A. Perea, "Persistent homology of toroidal sliding window embeddings," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2016.

[36] Y. Lee, S. D. Barthel, P. Dłotko, S. M. Moosavi, K. Hess and B. Smit, "Quantifying similarity of pore-geometry in nanoporous materials," *Nature Communications,* vol. 8, p. 15396, May 2017.

[37] H. O. Coxson, "Quantitative Computed Tomography Assessment of Airway Wall Dimensions: Current Status and Potential Applications for Phenotyping Chronic Obstructive Pulmonary Disease," *Proceedings of the American Thoracic Society,* vol. 5, pp. 940-945, 12 2008.

[38] J. Harer, R. Bar-On, N. Strawn, C. Tralie, P. Bendich, A. Pieloch and J. Slaczedek, "TDAtools," 2014. [Online]. Available: https://github.com/ksian/ML2015FP/tree/master/3TDATools.

[39] The_GUDHI_Project, "GUDHI User and Reference Manual," 2015. [Online]. Available: http://gudhi.gforge.inria.fr/doc/latest/.

[40] M. Kerber, D. Morozov and A. Nigmetov, "Hera," 7 2017. [Online]. Available: https://bitbucket.org/grey_narn/hera.

[41] P. Gabriel, "Unzerlegbare Darstellungen I,," *Manuscr. Math.,* vol. 6, pp. 71-103, 1972.

[42] R. Ghrist, "Barcodes: the persistent topology of data," *Bull. Amer. Math. Soc. (N.S.),* vol. 45, pp. 61-75, 2008.

[43] M. Kerber, D. Morozov and A. Nigmetov, "Geometry Helps to Compare Persistence Diagrams," in *2016 Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, 2016.

[44] D. Cohen-Steiner, H. Edelsbrunner and J. Harer, Stability of Persistence Diagrams, Symposium on computational geometry, 2005.

Acknowledgments:

Funding: This work was partially supported by the EPSRC grant EPSRC EP/N014189/1, and by the National Institute for Health Research (NIHR) through support of the NIHR Southampton Respiratory Biomedical Research Unit, specifically funding Dr. Michael Bennett and the acquisition of CT data. Professor Djukanovic is a NIHR Senior Investigator. **Author contributions:** J. B., M. B. and R.D. designed the study with contributions from F. B. and M. P. The CT data were created and preprocessed by M.B and J. C. The computations were designed by F. B., J. B., and M. P., and

carried out by F. B. and M. P. The results were analyzed by F. B., M. B., R. D. and J. B. The manuscript was written by F. B.,M. B., R. D., and J. B. The Supplementary Materials were written by M. P., F. B, and J. B. **Competing interests:** The authors declare no competing financial interests. **Data and materials availability:** The data that support the findings of this study are available on request from the corresponding author (J. B.). The data are not publicly available due to them containing information that could compromise research participant privacy.

Figures:

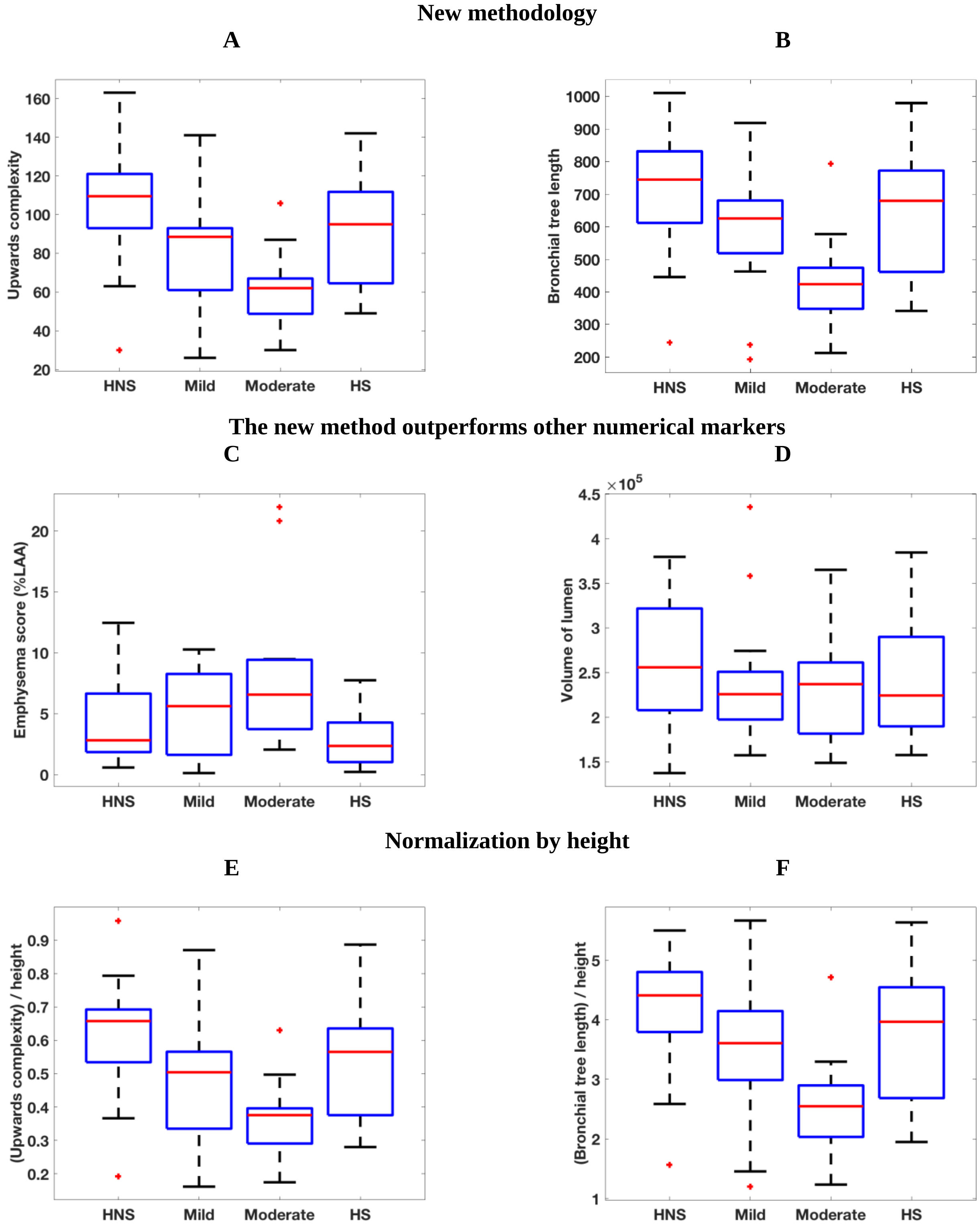

Figure 1 Differences between severity groups given by 6 radiomic features. In the boxplots, **HNS** = healthy non-smokers, **Mild** = mild COPD patients, **Mod** = moderate COPD patients and **HS** = healthy smokers. The + signs denote outliers. The 6 radiomic features studied are **(A)**

upwards complexity (see Materials and Methods for details), **(B)** bronchial tree length, **(C)** emphysema score (as percentage of low attenuation area), **(D)** volume of the airways (computed as the number of voxels inside the red airway structure in Figure 3A), **(E)** upwards complexity divided by participant's height, **(F)** bronchial tree length divided by participant's height. The combination of radiomic features A and B can distinguish all groups except for HS from HNS or from Mild, which outperforms the combination of methods C and D.

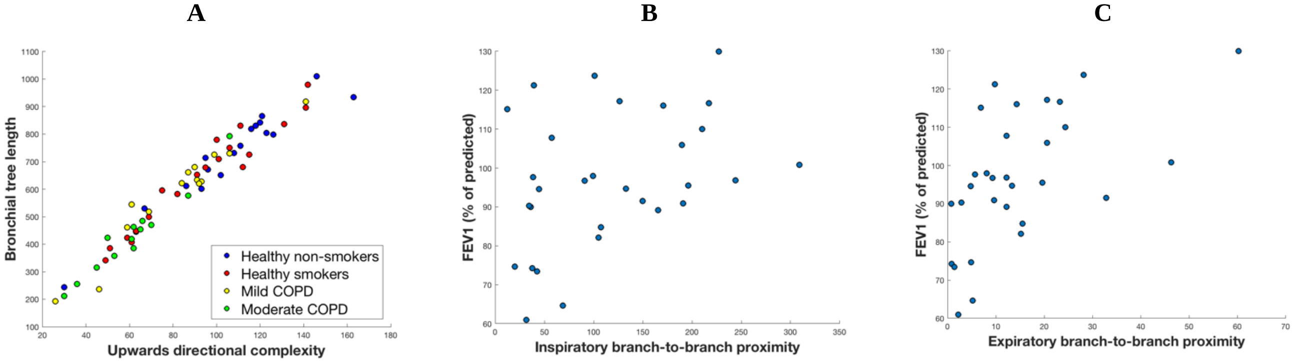

Figure 2 Analysis of topological characteristics. **(A)** Correlation between upwards complexity and bronchial tree length (Pearson correlation coefficient ρ=0.97, p-value p=1.56 $\cdot$ 10^{-41}). Similar results were obtained using directional complexity in other directions. **(B)** Correlation between the inspiratory branch-to-branch proximity, which quantifies how branches of the inspiratory bronchial tree bend towards one another, and FEV1 (% of predicted) (ρ=0.38, p=0.040). **(C)** Expiratory counterpart of (b) (ρ=0.57, p=0.001). Notice that the correlation is stronger and more significant for expiratory scans than for inspiratory scans.

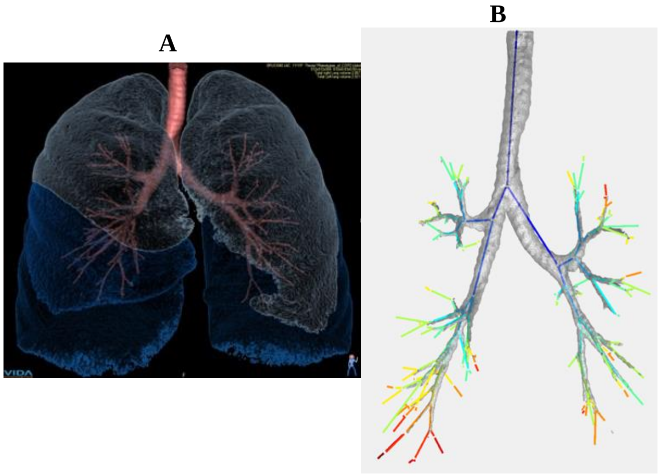

Figure 3 MSCT analysis. **(A)** The Apollo software (Vida Diagnostics, Iowa, USA) being used to segment the lobes from the MSCT scan. **(B)** Illustration of the extracted branch center lines, along with the segmented airway tree for one of the participants. The center lines are colored according to generation number. Note that for the purposes of illustration, the center lines are plotted between the branch points only. For all of the analysis described in this paper, the complete center line information was used, which captured the true shape of the airways as in Figure 7.

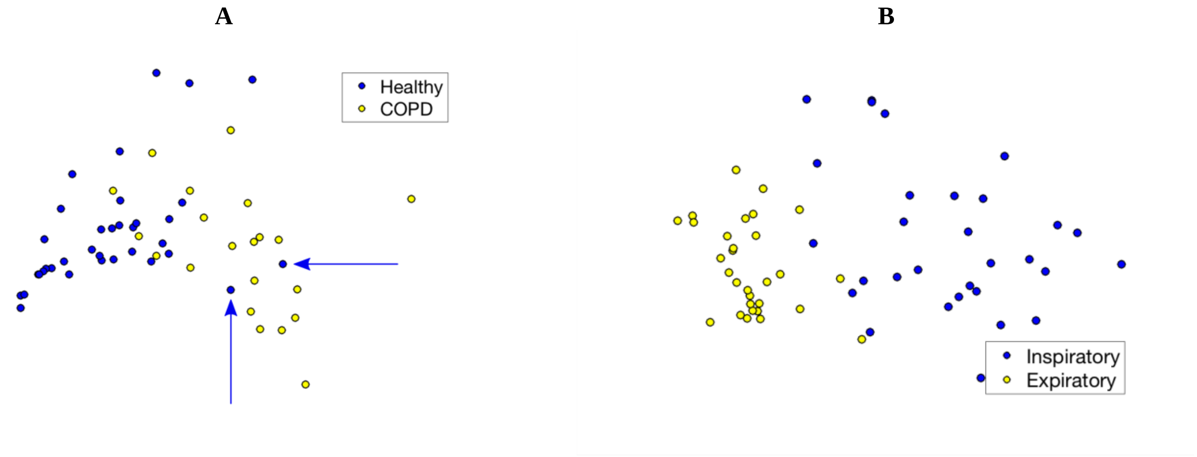

Figure 4 Spatial representation of similarities between lungs. These are obtained by describing the shape of each lung through a set of topological characteristics called barcodes (see Materials and Methods for details) and computing distances between the barcodes of individual subjects. In the legends, Healthy=healthy smokers and non-smokers, COPD=mild and moderate COPD patients. **(A)** This representation uses degree-2 persistent homology of inspiratory data to infer the shape of the airways inside the cavity of the lobes and it shows a clear separation between Healthy and COPD groups. Note two dots indicated by arrows: they represent healthy smokers which our algorithm places among the COPD patients, indicating a potential undiagnosed problem. **(B)** This representation takes into account how the airways bend upwards and shows that this topological feature clearly separates the inspiratory and expiratory stages of the bronchial tree. This analysis was not performed for the expiratory phase because the information about the lobe structure was not available.

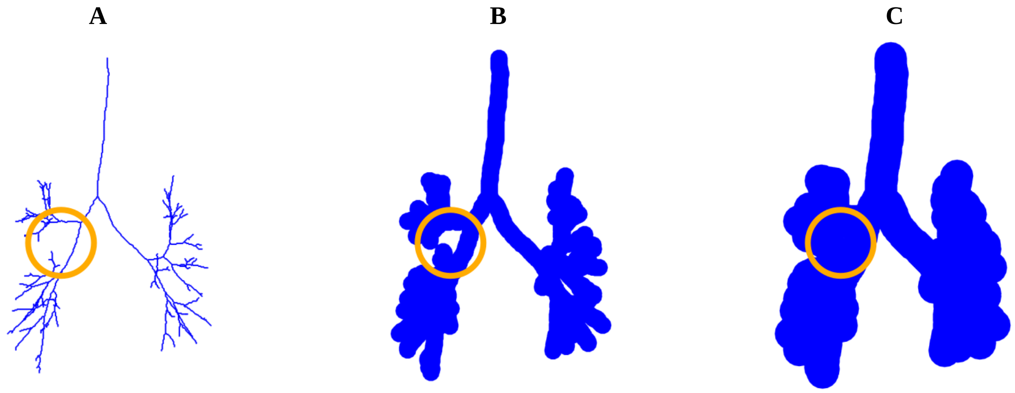

Figure 5 Computing of *branch-to-branch proximity*. Consider the graph representing the bronchial tree as explained in Materials and Methods **(A)**. This graph is called a tree since it contains no loops, *i.e.*, no branches that bifurcate and then merge. Of note, there are many nodes (up to 264) between any two consecutive bifurcations, so the nodes appear dense in the graph representation. Centered at each node of this graph, we virtually set a ball of a fixed radius, thickening the construction. As we keep thickening more and more, by increasing the radius of those balls, at some point we will find that some branches merge, creating a loop **(B)**. We record the radius r_1 at which this happens. For a large enough radius r_2, though, this loop will be filled in **(C)**. If a merging of branches creates a loop that appears for the value r_1 of the radius and disappears at r_2, we represent this merging as the positive number $r_2 - r_1$. Summing up all these terms, we obtain a number we call *branch-to-branch proximity.*

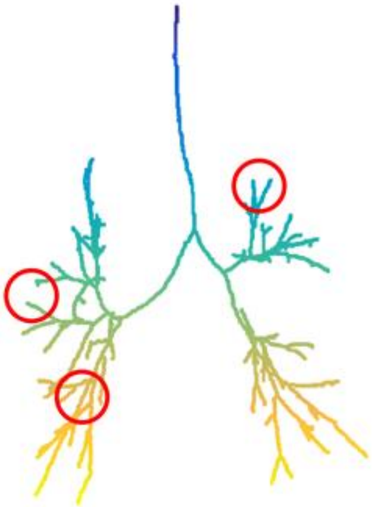

Figure 6 Calculations show that the lung function is better when more branches bend towards one another in the expiratory bronchial tree (such as the branches in the two circles on the left, in contrast with those in the circle on the right). See Figure 2C.

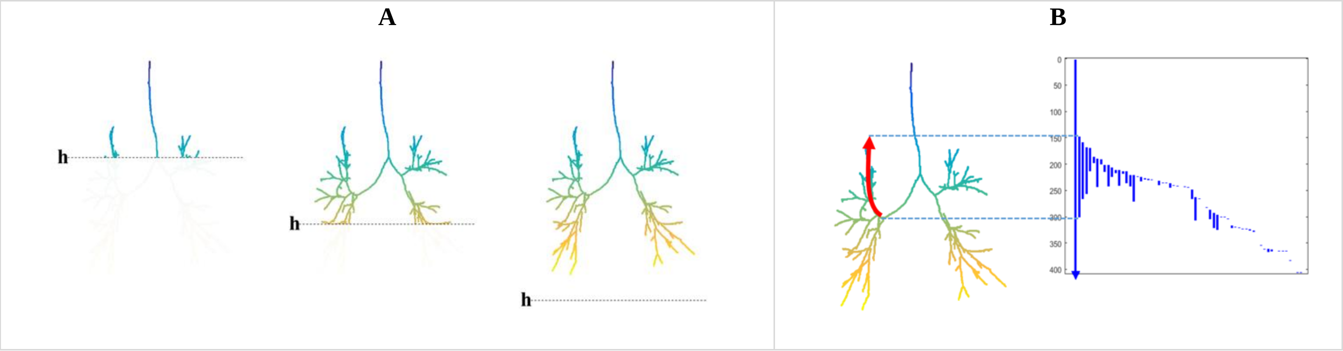

Figure 7 Explanation of upwards complexity. The color gradient indicates height. **(A)** To study upwards complexity, we slide a horizontal plane downwards. If we denote by X_h the part of the tree that sits above the horizontal plane at distance h from the top of the image, then $X_h \subseteq X_{h'}$ whenever $h \leq h'$, obtaining a sequence of nested graphs approximating the bronchial tree more accurately as we increase h. **(B)** The right part of the panel shows the degree-0 barcode of the sequence of nested graphs in (A). In this picture, the correspondence between bars in the barcode and branches that change trajectory upwards becomes apparent. In particular, the length of a bar indicates for how long a branch follows that upwards trajectory.

A

|KS| p-value	**HNS**	**Mild**	**Mod**	**HS**
HNS		*0.51*	*0.70*	*0.29*
Mild	2.14E-02		*0.49*	*0.33*
Mod	4.95E-04	5.25E-02		*0.53*
HS	3.71E-01	2.79E-01	1.56E-02	

B

|KS| p-value	**HNS**	**Mild**	**Mod**	**HS**
HNS		*0.48*	*0.76*	*0.25*
Mild	3.24E-02		*0.63*	*0.24*
Mod	1.26E-04	4.59E-03		*0.61*
HS	5.23E-01	6.57E-01	3.44E-03	

C

|KS| p-value	**HNS**	**Mild**	**Mod**	**HS**
HNS		*0.29*	*0.48*	*0.31*
Mild	4.34E-01		*0.29*	*0.40*
Mod	4.16E-02	5.63E-01		*0.55*
HS	2.70E-01	1.15E-01	9.88E-03	

D

|KS| p-value	**HNS**	**Mild**	**Mod**	**HS**
HNS		*0.29*	*0.31*	*0.21*
Mild	4.69E-01		*0.18*	*0.17*
Mod	3.83E-01	9.73E-01		*0.29*
HS	7.71E-01	9.51E-01	4.56E-01	

E

|KS| p-value	**HNS**	**Mild**	**Mod**	**HS**
HNS		*0.48*	*0.76*	*0.40*
Mild	3.24E-02		*0.56*	*0.29*
Mod	1.26E-04	1.67E-02		*0.58*
HS	7.90E-02	4.22E-01	5.66E-03	

F

|KS| p-value	**HNS**	**Mild**	**Mod**	**HS**
HNS		*0.42*	*0.76*	*0.29*
Mild	8.90E-01		*0.63*	*0.26*
Mod	1.26E-04	4.59E-03		*0.56*
HS	3.47E-01	6.00E-01	3.43E-02	

Table 1 Differences between severity groups given by 6 radiomic features. For each radiomic feature, we show the table with the pairwise Kolmogorov-Smirnov that compares **HNS** = healthy non-smokers, **Mild** = mild COPD patients, **Mod** = moderate COPD patients and **HS** = healthy smokers. The values in ***italics*** in yellow shaded boxes indicate the absolute value of the KS score and the values in **roman type** in blue shaded boxes indicate p-values. The 6 radiomic features analyzed are **(A)** upwards complexity (see Materials and Methods for details), **(B)** bronchial tree length, **(C)** emphysema score (as percentage of low attenuation area), **(D)** volume of the airways (computed as the number of voxels inside the red airway structure in Figure 3A), **(E)** upwards complexity divided by participant's height, **(F)** bronchial tree length divided by participant's height. The combination of radiomic features A and B can distinguish all groups except for HS from HNS or from Mild, which outperforms the combination of methods C and D.

Supplementary Materials:

Materials and Methods

Persistent homology

The main mathematical method used in this paper to analyze the lung CT scans is called persistent homology [11, 12, 13]. It has been designed as a computational way to capture the shape of objects depending on the scale at which they are viewed. To understand the basic idea of persistence, imagine a given set of high resolution images of a human face. If one zooms in, one can capture tiny details of the face, but one may not be able to recognize the person in the photo. Zooming out, one sees less detail, but it will be easier to see the person in the photograph. Continuing the process of zooming out, eventually all details are lost. It is clear that the choice of zooming scale depends on the kind of information we are hoping to recover. We can avoid making a particular choice of zooming scale, and instead study all possible scales at once. This is the approach of persistent homology to study data – it presents us with information about the shape of our data at a range of scales, controlled by a parameter r. For small values of r, we see single points; *as* r increases, connections between points begin to emerge, creating an approximate shape of the data.

The data set to be analyzed is typically thought to be a discrete subset S sampled from a metric space. To understand the structure of the set, and so capture the information it contains, one creates approximations of S by simple shapes K_r, called simplicial complexes, for a range of values of the scale parameter r. We define the notion of simplicial complex in Section 3 below. As r increases, the corresponding complexes will grow and their structure will also change. Persistent homology will exhibit the evolution of these approximations. Intuitively, homology in degree zero describes the components of the set, in degree one it uncovers the existence of non-trivial loops at a particular scale, while degree two identifies voids or cavities.

This topological information is represented in the form of a set of intervals or bars with multiplicities, called the barcode (see Figure 8). The long bars, which represent features that persist over a wide range of values of the scale parameter, represent significant features of the underlying space the data was sampled from, while short bars typically (but not always) represent noise. Two barcodes can be compared by computing their distance, which provides a measure of similarity. Thanks to the stability theorem, this comparison is robust with respect to noise and small-scale perturbations. We now give details of this process.

Persistence modules and barcodes

A starting point of persistent homology is the notion of a persistence module V, which is a family of vector spaces over some field $\mathbb{F}$ and linear maps of the form:

$$V_0 \xrightarrow{f_0} V_1 \xrightarrow{f_1} \cdots \xrightarrow{f_{n-1}} V_n,$$

in which composing the consecutive maps starting from some V_i to V_j we get linear maps $f_{i,j}: V_i \to V_j$, for any $0 \leq i \leq j \leq n$. In particular, $f_{i,i}$ is the identity map and $f_{i,i+1} = f_i$. A particularly simple persistence module, denoted by $I_{[i,j-1]}$, which plays a special role in the theory, is obtained as follows. Fix i and j, such that $0 \leq i \leq j-1 \leq n$ and consider the following persistence module

$$0 \to \cdots \to 0 \to \mathbb{F} \xrightarrow{id} \cdots \xrightarrow{id} \mathbb{F} \to 0 \to \cdots \to 0$$

where $\mathbb{F}$ is the ground field considered as a 1-dimensional vector space over $\mathbb{F}$. The first and the last nontrivial terms appear at the places i and $j-1$ respectively. More complex examples are obtained by taking sums of these simple modules in the following sense. If V and V' are persistence modules, then their direct sum $V \oplus V'$ is the persistence module V'', where

$$V_i'' = V_i \oplus V_i', \quad \text{and} \quad f_i'' = \begin{pmatrix} f_i & 0 \\ 0 & f_i' \end{pmatrix}.$$

A theorem by Gabriel [41] states that any persistence module is a direct sum of persistence modules of the form $I_{[i,j]}$. Hence, a persistence module V can be fully characterized by a finite set $D(V)$, called the persistence diagram, which contains a point $(i,j) \in \mathbb{R}^2$ $(0 \le i \le j < n)$ for every summand of the form $I_{[i,j-1]}$ appearing in the decomposition of V (and a point of the form $(i,\infty) \in \mathbb{R} \times (\mathbb{R} \cup \{\infty\})$ for every summand of the form $I_{[i,n]}$). Each point in $D(V)$ appears with multiplicity equal to the number of copies of the corresponding summand. For technical reasons, all points in the diagonal $\{(x,y) \in \mathbb{R}^2 | \, y = x\}$ are added to $D(V)$ as well.

A barcode is a graphical representation of V equivalent to the persistence diagram $D(V)$. It is a collection of intervals with multiplicities [42]. The barcode of V consists of one interval (or *bar*) of the form $[i,j)$ for every off-diagonal point (i,j) in $D(V)$, which describes the range of values of the scale parameter over which a particular feature persists. The multiplicity of an interval is that of its corresponding point in $D(V)$. Figure 8 shows an example of a barcode.

Simplicial complexes

The persistence modules most commonly used in topological data analysis arise from filtered simplicial complexes, whose combinatorial nature is very suitable for computations.

A simplicial complex K with vertex set S is a family of nonempty, finite subsets of S. Subsets of S of $p+1$ elements are called p-simplices. A p-simplex is represented as a list of its vertices $[v_0, \dots, v_p]$. In a simplicial complex K, one requires that all elements v of S form 0-simplices $[v]$ in K, and if $\sigma \in K$ and $\emptyset \neq \tau \subset \sigma$, then $\tau \in K$. We usually consider the case when S is finite. A simplicial complex K has the associated space $|K|$, called the geometric realization, which can be regarded as a triangulated polyhedron in an appropriate Euclidean space. The combinatorial structure of K can be used to define the so-called p^{th} homology $H_p(K)$ of $|K|$ for all $p \geq 0$. To define $H_p(K)$, we first define the space of p-chains $C_p(K)$ to be the vector space consisting of all finite sums $\sum_\sigma a_\sigma \sigma$, where σ runs through all p-simplices, and a_σ is an element of a ground field $\mathbb{F}$. Typically, coefficients a_σ are taken from a finite field $\mathbb{Z}_\mathrm{p}$ of integers modulo p, for instance $\mathbb{Z}_2 = \{0,1\}$, which was also used in our computations. These vector spaces are connected by the boundary homomorphism $\partial: C_p(K) \to C_{p-1}(K)$. This map is defined on p-simplices $\sigma = [v_0, \dots, v_p]$ by

$$\partial(\sigma) = \sum_k (-1)^k [v_0, \dots, \widehat{v_k}, \dots, v_p],$$

and then extended by linearity. Here the symbol $\widehat{v_k}$ means that the corresponding element v_k is omitted. The next step is to define a vector space $Z_p(K)$ of p-cycles of K, which consists of all vectors $v \in C_p(K)$ such that $\partial(v) = 0$, and a vector space $B_p(K)$ of p-boundaries of K, which consists of all $v \in C_p(K)$ such that $z = \partial(w)$, for some $w \in C_{p+1}(K)$. It is important to note that $\partial \circ \partial = 0$, that is, performing this operation twice sends every simplex to zero. This guarantees that $B_p(K)$ forms a vector subspace of $Z_p(K)$. Hence, it makes sense to define the quotient space, which is called the p^{th} homology of K:

$$H_p(K) = {}^{Z_p(K)}\Big/_{B_p(K)}.$$

The dimension of H_p (K) is known as the p^{th} *Betti number* β_p, of $|K|$. Intuitively, β_0 computes the number of connected components of the geometric realization of K. Likewise, β_1 computes the number of 1-dimensional holes, β_2 computes the number of 2-dimensional holes, etc.

If L is also a simplicial complex on the set of vertices T, such that T is a subset of S and any simplex σ of L is also a simplex of K, then L is called a subcomplex of K and we write $L \subset K$. It follows that

$C_p(L) \subset C_p(K)$. If z is a p-cycle in L, it is also a p-cycle of K and if z is a p-boundary in L, it is also a p-boundary in K. Hence, there is a well-defined map $f: H_p(L) \to H_p(K)$, $p \geq 0$, which is called the induced map. In general, the induced map is not injective, even though $Z_p(L) \subset Z_p(K)$.

A filtered complex K is a nested sequence of subcomplexes

$$K_0 \subset K_1 \subset \cdots \subset K_n.$$

Choosing a homology degree $p \geq 0$, we can write all homology groups and induced maps as a sequence

$$H_p(K_0) \xrightarrow{f_0} H_p(K_1) \xrightarrow{f_1} \cdots \xrightarrow{f_{n-1}} H_p(K_n)$$

that forms a persistence module. The *degree-p barcode* of $|K|$ is defined as the barcode of this persistence module.

Comparing persistence diagrams

There is a number of ways to compare persistence diagrams [43]. If X and Y are persistence diagrams, then the bottleneck distance between X and Y is defined by

$$d_B(X,Y) = \underset{\gamma}{inf}\ \underset{x}{sup}\|x - \gamma(x)\|_\infty$$

where γ runs through all bijections from $X \to Y$, while x runs through all points of X and for a point of the form $z = (a,b) \in \mathbb{R} \times (\mathbb{R} \cup \{\infty\})$, one has $\|z\|_\infty = max\{|a|,|b|\}$, and $\|(a,\infty) - (b,\infty)\|_\infty = |a - b|$. The q^{th} Wasserstein distance $(q \geq 1)$ is defined by

$$d_{W_q}(X,Y) = \underset{\gamma}{inf}\left(\sum_x \|x - \gamma(x)\|_\infty^q\right)^{\frac{1}{q}}.$$

These expressions define pseudo-metrics, as it is possible to create distinct persistence diagrams for which either of these distances is zero.

Stability

The stability theorem for persistent homology, due to Cohen-Steiner, Edelsbrunner and Harer [44], is easier to state in terms of tame functions on triangulable spaces, that is on spaces which can be represented as a simplicial complex.

Let X be a triangulable topological space and let $f: X \to \mathbb{R}$ be a real-valued function on X. A homological critical value of f is a real number a, for which there exists an integer p such that for all sufficiently small $\varepsilon > 0$ the map $H_p(f^{-1}(-\infty, a-\epsilon]) \to H_p(f^{-1}(-\infty, a+\epsilon])$ induced by inclusion is not an isomorphism. So, the number a corresponds to the value at which the homology of sub-level sets changes. A function f is tame if it has a finite number of homological critical values and the homology groups $H_p(f^{-1}(-\infty, a-\epsilon])$ are finite dimensional for all $p \geq 0$ and $a \in \mathbb{R}$. Typical examples of such functions are Morse functions on compact manifolds and piece-wise linear functions on finite simplicial complexes. For a real number r, one sets $V_r = H_p(f^{-1}(-\infty, r])$. If $r < t$, we have $f^{-1}(-\infty, r] \subset f^{-1}(-\infty, t]$ and therefore we can consider the induced linear map $V_r \to V_s$, which is an isomorphism, if the interval $[r, s]$ contains no homological critical value of f. Hence, varying r, one obtains a finite number of distinct vector spaces V_{r_i}, leading to a persistence module

$$V_{r_0} \to V_{r_1} \to \cdots \to V_{r_n}.$$

In particular, we have a corresponding persistence diagram $D(f)$. The classical stability theorem reads as follows [44]:

Theorem 1. *Let X be a triangulable topological space with continuous tame functions $f, g: X \to \mathbb{R}$. Then the persistence diagrams satisfy*

$$d_B\big(D(f), D(g)\big) \leq \|f - g\|_\infty.$$

In other words, persistence diagrams are stable under possibly irregular perturbations of the function used to create the diagram. In our particular case, this theorem ensures that imprecisions of measurement, such as small differences in the alignment of lungs when the scans were taken, will not lead to a drastic change in the resulting barcodes. There are similar results in terms of Wasserstein distances [44].

Height filtration

In data analysis, a given data set can typically be approximated in several different ways by a family of simplicial complexes. In choosing a suitable representation, one is guided by the properties of the set and computational efficiency. Such a representation is fixed by choosing a real-valued tame function f and computing its sublevel sets $f^{-1}(-\infty, t]$ as in the section on Stability above.

For instance, given a 3D object X, a commonly used function $f: X \to \mathbb{R}$ is that which sends each point $(x, y, z) \in X$ to its "height", i.e. its third coordinate, z. In the paper, to compute the upwards complexity of the graph representation X of a bronchial tree, we use a function very similar to this: for each vertex v in X, we define $f(v)$ as the vertical distance from v to the highest point in the CT scan, and for each edge e in X connecting two vertices v and v', we define
$f(e) = max\{f(v), f(v')\}$.

For functions like these, the bars in the degree-0 barcode have a clear interpretation as changes in trajectory. In the case of upwards complexity, those are airway trajectories that change to face upwards.

Alpha complexes

Another construction we use are 3D alpha complexes, which can substantially reduce the computational complexity. To describe this construction, first let us say a word about *Voronoi diagrams*. Given a set S of points in Euclidean space $\mathbb{R}^n$, one defines convex polytopes V_s, $s \in S$ called Voronoi cells, which consist of all points $x \in \mathbb{R}^n$ such that $dist(x, s) \leq dist(x, s')$ for any other $s' \in S$. The subsets V_s give a tessellation of $\mathbb{R}^n$.

Given a finite set of points $S \subset \mathbb{R}^n$ and a real number $r \in \mathbb{R}^n$, one defines the region $R_s(r) = \bar{B}_s(r) \cap V_s$, where $\bar{B}_s(r)$ is the closure of the ball of radius r centered at s. Now we can form the α-complex (or alpha complex) K_r as follows: a subset $\sigma \subset S$ is called an α-simplex if

$$\bigcap_{s \in \sigma} R_s(\sigma) \neq \emptyset.$$

See Figure 9 for an illustration of this construction. Varying r, one obtains a finite family of nested α-complexes

$$K_{r_0} \subset K_{r_1} \subset \cdots \subset K_{r_n}.$$

This is a typical example of a filtered complex. Hence, one can apply the machinery of persistent homology. In particular, we have the corresponding persistence diagrams. This geometric construction can also be phrased in terms of tame functions as before, and thus fits into the same general framework.

In the Materials and Methods section in the paper, we applied the alpha complex filtration to two sets of points in $\mathbb{R}^3$. On the one hand, we used the vertices of the 3D graph representation of the bronchial tree described in the Materials and Methods subsection called MSCT analysis. The degree-

1 barcode of the alpha complex filtration on this collection of vertices provided additional information about the complexity of the branching structure of the airways. On the other hand, we also used a 3D array of binary voxels representing the luminal surface of the airways together with the surface of the lobes as in Figure 3A of the paper. For each binary voxel image, we constructed the point cloud in $\mathbb{R}^3$ by including the coordinates of every voxel with value 1. The degree-2 barcode of the alpha complex filtration on this set of points gave information about how the airways fill the cavity of the lobes.

Figures

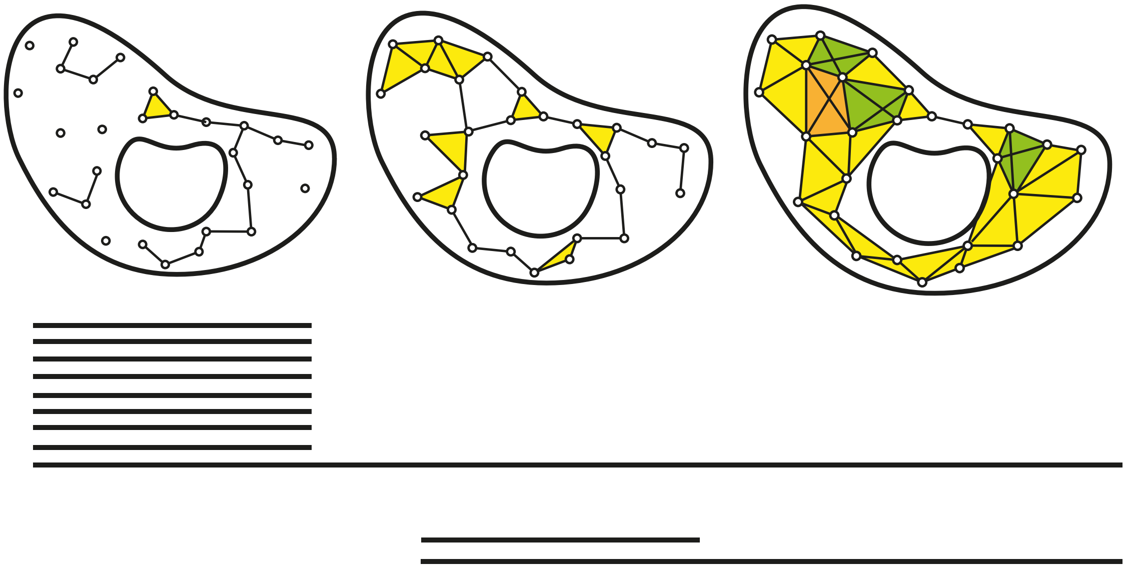

Figure 8 A point cloud is sampled from a deformed annulus in the plane. The sequence of pictures from left to right shows simplicial representations of the set at different scales. In the picture on the left, there are nine components, represented by the horizontal bars below the picture. As the scale increases, all the components combine into one as we move from the left to the middle picture, and this persists for all remaining scales. There are no loops in the left picture, but two loops emerge at the middle scale, represented by the two bars at the bottom. Increasing the scale parameter from the middle to the right picture causes one of those loops to disappear, while the other one persists. Thus, the topological signal is that we have a ‘shape consisting of one piece with a hole in it’.

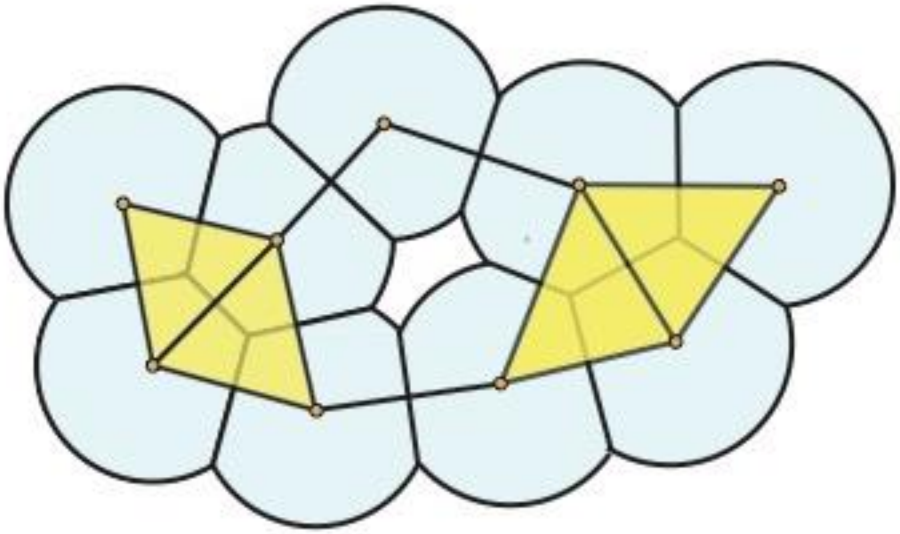

Figure 9 An example of a system of Voronoi cells constructed for a particular value of the scale parameter on a subset of the plane (shown in blue). Superimposed is the alpha complex that represents the structure the set at this scale. The topological signal here is that, at this scale, the points were sampled from a deformed annulus.

References

The reference list can be found in the main manuscript.