

1 **CartoCell, a high-throughput pipeline for accurate 3D
2 image analysis, unveils cell morphology patterns in
3 epithelial cysts.**

4 Jesús A. Andrés-San Román^{1,2,†}, Carmen Gordillo-Vázquez^{1,2,†}, Daniel Franco-
5 Barranco^{3,4,†}, Laura Morato^{1,2}, Antonio Tagua^{1,2}, Pablo Vicente-Munuera⁶, Ana M.
6 Palacios¹, María P. Gavilán⁷, Valentina Annese^{1,2}, Pedro Gómez-Gálvez^{1,8,9,*}, Ignacio
7 Arganda-Carreras^{3,4,5,*}, Luis M. Escudero^{1,2,*}.

8

9

10 *¹Instituto de Biomedicina de Sevilla (IBiS), Hospital Universitario Virgen del
11 Rocío/CSIC/Universidad de Sevilla and Departamento de Biología Celular, Facultad
12 de Biología, Universidad de Sevilla. 41013 Seville, Spain*

13 *²Biomedical Network Research Centre on Neurodegenerative Diseases
14 (CIBERNED), 28029 Madrid, Spain*

15 *³Department of Computer Science and Artificial Intelligence, University of the
16 Basque Country (UPV/EHU). 20018 San Sebastian, Spain*

17 *⁴Donostia International Physics Center (DIPC). 20018 San Sebastian, Spain*

18 *⁵Ikerbasque, Basque Foundation for Science. 48009 Bilbao, Spain*

19 *⁶Laboratory for Molecular Cell Biology, University College London. London, UK*

20 *⁷Centro Andaluz de Biología Molecular y Medicina Regenerativa (CABIMER), Junta
21 de Andalucía /CSIC/Universidad de Sevilla/Universidad Pablo de Olavide, 41092
22 Seville, Spain.*

23 *⁸ MRC Laboratory of Molecular Biology, Cambridge Biomedical Campus, Francis
24 Crick Avenue, Cambridge, UK*

25 *⁹Department of Physiology, Development and Neuroscience, University of
26 Cambridge, UK*

27

28

29

30

1 **ABSTRACT**

2 Decades of research have not yet fully explained the mechanisms of epithelial self-
3 organization and 3D packing. Single-cell analysis of large 3D epithelial libraries is
4 crucial for understanding the assembly and function of whole tissues. Combining 3D
5 epithelial imaging with advanced deep learning segmentation methods is essential for
6 enabling this high-throughput analysis. We introduce CartoCell, a deep learning-based
7 pipeline that uses small datasets to generate accurate labels for hundreds of whole
8 3D epithelial cysts. Our method detects the realistic morphology of epithelial cells and
9 their contacts in the 3D structure of the tissue. CartoCell enables the quantification of
10 geometric and packing features at the cellular level. Our Single-cell Cartography
11 approach then maps the distribution of these features on 2D plots and 3D surface
12 maps, revealing cell morphology patterns in epithelial cysts.

13

1 INTRODUCTION

2 Analysis of epithelial tissue properties at the cellular level has enabled advancement
3 in the understanding of different cellular phenomena during morphogenesis.
4 Traditionally, most of these works have been approached in two dimensions (2D) after
5 the analysis of the apical surfaces of monolayer epithelia. However, the need to
6 understand the 3D morphology of epithelial cells to study organogenesis (Bryant &
7 Mostov, 2008; Davidson, 2012; Lecuit & Lenne, 2007), cell migration (Okuda & Sato,
8 2022), branching formation (Okuda et al., 2018), tumorigenesis (Messal et al., 2019)
9 or wound healing (Ioannou et al., 2020) has become clear in recent years. An example
10 is the discovery that epithelial cells can present very complex geometries due to the
11 exchange of neighbors along the apico-basal axis. These cell shapes have been called
12 scutoids, and it has been shown that they have a role in morphogenesis as well as in
13 the connectivity and biophysical properties of tissues, cushioning and minimizing cell
14 surface tension and leading to a balanced energetic state (Gómez et al., 2021;
15 Gómez-Gálvez et al., 2018, 2022; Gómez-Gálvez, Vicente-Munuera, et al., 2021; Lou
16 et al., 2022; Prabhakara et al., 2022; Rupprecht et al., 2017). Scutoids represent a
17 new paradigm, but they set a challenge for the quantitative analysis of the complex
18 epithelial 3D packing requiring a very accurate reconstruction of 3D epithelial tissues
19 from microscopy images to allow capturing precise cell shapes and neighboring
20 relationships.

21 In the last few years, deep learning has become the state-of-the-art solution for the
22 analysis of biomedical images (Laine et al., 2021; Meijering, 2020; Moen et al., 2019).
23 Deep learning is a subdomain of machine learning that makes use of large (or so-
24 called *deep*) artificial neural networks to solve a wide variety of tasks. As opposed to
25 conventional algorithms, before they can be used, deep learning methods (or models)
26 need to be *trained*. In other words, models can *learn* from a set of examples how to
27 solve a specific task. Once trained, the models can be directly applied to new samples,
28 what is usually called prediction or *inference*. In the particular case of image
29 segmentation, the training dataset is commonly formed by a set of raw images and
30 their corresponding ground-truth annotations or *labels*. This type of learning
31 framework, with both raw and label images available, is known as supervised learning.
32 Furthermore, realistic 3D reconstruction of epithelial cells requires assigning each
33 individual cell a unique label, in a process called *instance segmentation*.

1 The 3D instance segmentation of microscopy data is a difficult task, especially in the
2 presence of a dense concentration of cells and anisotropic voxel resolution, as it is
3 common in volumetric images of epithelial tissue. State-of-the-art learning-based
4 methods tackle these challenges using a top-down strategy, by first training a deep
5 neural network (DNN) to predict representations of the objects of interest (cells in our
6 case), and then extracting individual instances from those representations using
7 different post-processing methods. Common representations include cell masks or
8 boundaries (Ciresan et al., 2012; Falk et al., 2018; Wei et al., 2020; Wolny et al., 2020),
9 distance or flow maps (Schmidt et al., 2018; Stringer et al., 2021; Weigert et al., 2020),
10 or a combination of some of the latter (Lin et al., 2021; Z. Yan et al., 2018). On top of
11 those representations, cell instances are then calculated usually by means of
12 watershed (Cousty et al., 2009; Meyer, 1994) or graph-partitioning methods (Bailoni
13 et al., 2019; Kappes et al., 2011; Wolf et al., 2018). Other techniques have shown
14 success in segmenting cell nuclei and tracking cell lineage (de Medeiros et al., 2022;
15 He et al., 2022), but they do not have the high level of accuracy in cell shape required
16 to obtain detailed geometric and topological information at the cellular level.

17 Despite the benefits observed from these supervised approaches, their main
18 drawback is the large number of annotated samples needed to establish a training
19 dataset and obtain reliable performance (Razzak et al., 2018). Preparing and
20 processing such a large amount of data manually or semi-automatically is usually
21 tedious and time-expensive. This problem arises both from the acquisition time of high-
22 resolution images, as well as from the labeling of raw images performed by experts in
23 the field, which is usually the main bottleneck of the protocol. To address this issue, a
24 common strategy consists in using data augmentation, i.e. synthetically increasing the
25 size of the training data by morphological and intensity transformations or noise
26 addition (Perez & Wang, 2017; Shorten & Khoshgoftaar, 2019). However, data
27 augmentation may not be sufficient to realistically recreate the diversity of image data
28 to be processed. A much less exploited alternative to speed up the segmentation
29 protocol would consist in the use of low-resolution images instead, which are acquired
30 and annotated at a considerably faster pace. Certainly, this option would be ideal if the
31 quality of the output segmented cells remains comparable to that obtained with high-
32 resolution images.

1 In this article, we image, process, and analyze whole Madin-Darby Canine Kidney
2 (MDCK) 3D epithelial cell cultures. Despite their simplicity, these cysts have previously
3 been used as a suitable model system to study the establishment of cell polarity and
4 cell junctions (Bryant & Mostov, 2008; Dukes et al., 2011; Martín-Belmonte et al.,
5 2008; Vidal-Quadras et al., 2017; Yonemura, 2014), epithelial morphogenesis and
6 physiology (Alfonso-Pérez et al., 2022; Guo et al., 2008; Herranz & Martín-Belmonte,
7 2022; Imai et al., 2015; O'Brien et al., 2002; Wells et al., 2013; Yu et al., 2007), tumor
8 progression (Fessenden et al., 2018; Sakurai et al., 2012; Schmeichel & Bissell, 2003;
9 L. Yan et al., 2021), and for exploring the constraints on epithelial tissue
10 morphogenesis (Beck et al., 2022). MDCK cysts have provided a lot of useful
11 information to study more complex systems helping to understand the self-
12 organization in organoids, embryoids and the early stages of mammal development
13 (Beck et al., 2022; Carleton et al., 2022; Metzger et al., 2018).

14 Importantly, we use a small training dataset of high-resolution images to then obtain
15 a large training dataset of low-resolution images that are automatically segmented.
16 The structure of the method is a top-down pipeline that makes use of a DNN
17 architecture with multiple cell representations and watershed post-processing to
18 initially segment the epithelial cells as 3D instances. These instances are refined by a
19 second post-processing step: a 3D Voronoi algorithm that provides more realistic 3D
20 epithelial boundaries where cells are in close contact to each other. The algorithm is
21 based on tiling the space between a set of (Voronoi) seeds by proximity, without
22 leaving any gaps among the generated compartments (Voronoi, 1908). These
23 compartments are called Voronoi cells. Honda and colleagues showed that the
24 Voronoi compartmentalization of a 2D space, after using the cell nuclei as seeds, fitted
25 the pattern of cellular contacts found in epithelial surfaces (Honda, 1978). In 3D, the
26 implementation of Voronoi approaches has been used to simulate the shapes of
27 globular cells (Gómez-Gálvez, Anbari, et al., 2021), and here, it is key to increase the
28 cell segmentation quality of our pipeline.

29 In short, we have developed an accessible and fast tool to investigate the complex
30 organization of epithelial tissues. The production of a large number of samples with an
31 accurate segmentation has opened up a new way of 3D high-throughput analysis. The
32 representation of the extracted features values in each cell provides maps of the cysts
33 at single-cell resolution. Due to the similarities with the practices of making and using

1 maps, we called our approach Single-cell Cartography, and our high-throughput
2 segmentation method, CartoCell. The simple observation of these maps reveals the
3 presence of cell morphology patterns where cells are distributed following geometric
4 cues. These patterns illustrate how different the cells within the same cyst really are,
5 and how cells with similar characteristics have the tendency to cluster together in
6 specific zones of the cysts. Importantly, the large number of processed individual cells
7 permits us to quantify the frequency of the patterns and even to find hidden traits of
8 organizational features within the 3D structure of the tissue.

9

10 **RESULTS**

11 **CartoCell, a high-throughput pipeline for segmentation of 3D epithelial cysts**

12 The realistic analysis of whole epithelial tissues at the cell level is a critical point to a
13 bottom-up understanding of how tissues self-organize during development. In this
14 work, by means of deep learning and image processing strategies, we have developed
15 CartoCell, an automated pipeline (**Fig. 1**) to segment and analyze hundreds of
16 epithelial cysts at different stages (**Table S1**) with minimal human intervention.
17 CartoCell is subdivided into five consecutive phases.

18 In Phase 1, a small dataset of 21 cysts, stained with cell outlines markers, was
19 acquired at high-resolution in a confocal microscope (**Fig. 1**, 7 cysts of 4 days and 14
20 cysts of 7 days, **STAR Methods**). Next, the individual cell instances were segmented
21 using LimeSeg ((Machado et al., 2019), **STAR Methods**), a semi-automatic
22 segmentation plugin of Fiji (Schindelin et al., 2012). The final high-resolution label
23 images were the output of a curation process aided by a custom Matlab code (**STAR**
24 **Methods**). In particular, we implemented a Matlab graphic user interface to facilitate
25 manual deletion, insertion, fusion, and proper profiling of cell instances and lumen
26 segmentations (**STAR Methods**). On average, we estimate that the segmentation and
27 curation process took 3 to 5 complete working days, of one person, per cyst. The
28 high-resolution images from Phase 1 provide the accurate and realistic set of data
29 necessary for the following steps (see **Discussion**).

30 In Phase 2, both high-resolution raw and label images were down-sampled to create
31 our initial training dataset (**Fig. 1**). Specifically, the image volumes were reduced to
32 match the resolution of the images acquired in Phase 3 (see **STAR Methods** for

1 details). Using that dataset, a first DNN was trained. The DNN employed was a custom
2 stable 3D residual U-Net (Franco-Barranco et al., 2022), 3D ResU-Net from now on
3 (**Fig. 1** and **Fig. S1** related to **Fig. 1**). We will refer to this first model as “model M1”
4 (**Fig. 1, STAR Methods**).

5 In Phase 3, a large number of low-resolution stacks of multiple epithelial cysts was
6 acquired (**Fig. 1**). This was a key step to allow high-throughput analysis of samples,
7 since it greatly reduces the acquisition time (**STAR Methods**). Here, we extracted the
8 single-layer and single-lumen cysts by cropping them from the complete stack (**Fig. 1**
9 and **STAR Methods**). This way, we obtained a set of 293 low-resolution images,
10 composed of 84 cysts at 4 days, 113 cysts at 7 days and 96 cysts at 10 days (**Fig. S2**
11 related to **Fig. 1**). Next, we applied our trained model M1 to those images and post-
12 processed their output to produce (i) a prediction of individual cell instances (obtained
13 by marker-controlled watershed), and (ii) a prediction of the mask of the full cellular
14 regions (**Fig. 1, Fig. S1** related to **Fig. 1**, and **STAR Methods**). At this stage, the
15 output cell instances were generally not touching each other, which is a problem to
16 study cell connectivity in epithelia. Therefore, we applied a 3D Voronoi algorithm to
17 correctly mimic the epithelial packing (Honda, 1978; Voronoi, 1908). More specifically,
18 each prediction of cell instances was used as a Voronoi seed, while the prediction of
19 the mask of the cellular region defined the bounding territory that each cell could
20 occupy (**Fig. 1, Fig. S3** related to **Fig. 1**, and **STAR Methods**). The result of this phase
21 was a large dataset of low-resolution images and their corresponding accurate labels.

22 In Phase 4, a new 3D ResU-Net model (“model M2”, from now on) was trained on
23 the “*training low-resolution dataset*”, composed of the newly produced large dataset
24 of low-resolution raw images and its paired label images along with the “*training down-
sampled dataset*” (**Fig. 1**, and **STAR Methods**). This was a crucial step, since the
25 performance of deep learning models is highly dependent on the number of training
26 samples.

27 In Phase 5, model M2 was applied to new low-resolution cysts and their output was
28 post-processed as in Phase 3, thus achieving high-throughput segmentation of the
29 desired cysts (**Fig. 1** and **Fig. S2** related to **Fig. 1**).

30

31

32

1 **Optimization of the method**

2 Once the CartoCell pipeline was defined, we performed an automatic screening of
3 parameters of the 3D ResU-Net to optimize the quality of the prediction of models M1
4 and M2 (**STAR Methods**). To this aim, we elaborated a test set with 60 low-resolution
5 new cysts (20 cysts at 4d, 20 at 7d and 20 at 10d of development), not used for any
6 of the previous training steps (**STAR Methods**). These cysts were semi-automatically
7 segmented and manually curated to obtain their ground-truth labels (**STAR Methods**).
8 The parameter search for the M1 and M2 models aimed to ensure the highest quality
9 of segmentation, based on the comparison between the prediction of the models and
10 the ground-truth labels of the test set (**Table S2** and **STAR Methods**). In addition, this
11 optimization also demonstrated that M2, a model trained with a large number of low-
12 resolution images worked better than M1, a model trained with a sparse but perfectly
13 segmented dataset (**Table S3**).

14

15 **Performance evaluation and comparison with other methods**

16 To test CartoCell against current alternatives, we compared the performance of our
17 segmentation pipeline with that provided by the state-of-the-art approaches StarDist
18 3D (Weigert et al., 2020), Cellpose (Stringer et al., 2021), and PlantSeg (Wolny et al.,
19 2020) (**Fig. S3** related to **Fig. 1** and **STAR Methods**). In all cases, the small down-
20 sampled dataset from Phase 2 was used as the training set and the same 60 new
21 cysts were used as the test set (**STAR methods**). Moreover, for the sake of analyzing
22 method robustness and stability, each method was trained ten times under the same
23 conditions. Segmentation metrics were thus provided on average over those ten
24 repetitions (**Table S3**). In summary, CartoCell compares favorably with state-of-the-
25 art alternatives, especially after retraining our model with the new dataset in Phase 4.

26

27 **Management of results: final curation before biological analysis**

28 As a result of this full process, we got 353 segmented cysts (293 low-resolution
29 from Phase 3 + 60 tests). Nevertheless, to obtain a detailed quantitative analysis of
30 3D epithelial packing, we performed a semi-automatic final curation to correct small
31 defects in the segmentation (**Fig. S4** related to **Fig. 1** and **STAR Methods**). As an
32 outcome, we obtained our ground-truth dataset with accurate feature values used in

1 the biological analyzes (**Fig. 2**, **Table S1** and **Table S4**). On average, each imperfect
2 cyst took 12 ± 6 minutes to be curated. Thanks to our ground-truth dataset, we could
3 compare the values of the biological features extracted before and after the final
4 curation step, thus measuring their impact on the results. Nevertheless, the user may
5 opt for skipping this semi-manual step and keep a fully-automatic processing pipeline.
6 In our specific case, we automatically selected for analysis the cysts released after
7 Phase 5 whose epithelial monolayer was completely tiled by cells, i.e., without gaps
8 produced by an under-segmentation. Namely, 307 “closed cysts” out of the 353
9 segmented cysts (**Fig. 1** and **Fig. S4** related to **Fig. 1**). Considering all the geometric
10 characteristics extracted from the closed cysts, the mean relative error was $5.6 \pm 4.1\%$.
11 However, in the case of connectivity characteristics, the differences were greater
12 (mean value $13.3 \pm 13.9\%$) (**Fig. S4** related to **Fig. 1**, **Table S4** and **STAR Methods**,
13 and **Discussion**).

14

15 **Epithelial cysts adopt different shapes in 3D culture**

16 The pipeline that we have developed is able to realistically reconstruct the whole cyst
17 and its lumen. In this way, it allowed us to identify the real shape of the complete 3D
18 epithelial structure. We found that the full set of processed cysts (our “*ground-truth*
19 *dataset*”) can present a high heterogeneity in terms of shape (**Fig. S2** related to **Fig.**
20 **1**). We analyzed the 3D structure of the total number of single-lumen cysts pointing to
21 two geometrical features: axes lengths and solidity (a curvature index, **Fig. S5** related
22 to **Fig. 2**, **STAR Methods**). We considered two axes to be similar when the difference
23 in their lengths was inferior to 10%. According to these considerations, we established
24 five types of shapes (**Fig. 2A**, **Fig. S2** related to **Fig. 1**, and **STAR Methods**). When
25 the three axes of symmetry were similar in length, cysts were classified as spheres
26 (4.0%). When 2 of the 3 axes were similar, but one differed, they were called
27 spheroids. Spheroids divided into prolate (18.1%) when the different axis was the
28 major one, and oblate when it was the minor one (34.3%). Cysts with three axes of
29 different lengths were classified as ellipsoids (39.9%). Finally, we categorized cysts
30 with negative curvature (3.7%) when solidity was less than 0.9, independently of axes
31 values (**Fig. 2A**, **Fig. S5** related to **Fig. 2** and **STAR Methods**). We also quantified the
32 frequency of each type of shape at 4, 7 and 10-day cysts. We found that at 4 and 7
33 timepoints the most frequent shapes were ellipsoids with the percentage of oblate

1 cysts increasing with the time of culture (**Fig. 2B**). At 10 days there was a 37.9% of
2 ellipsoids and a 44.8% of oblate cysts.

3

4 **Single cell geometric analysis reveals cell morphology patterns in the MDCK**
5 **cysts.**

6 Reconstructing the 3D outlines of all the cells allowed the precise quantification of a
7 large number of geometrical and connectivity characteristics (**Table S1** and **Fig. S5**
8 related to **Fig. 2**). We designed a strategy to visualize these data in two ways: 3D
9 maps of the surface of the reconstructed cysts and 2D plots that represent the position
10 of all the cells analyzed within the cysts (**Fig. 2C**). We named this approach Single-
11 cell Cartography. In the case of the 3D maps, we obtained a readout of seven cell
12 geometric features by plotting their normalized values using a color palette (**Fig. 2D**,
13 **F**, **Fig. S5** related to **Fig. 2** and **STAR Methods**). Regarding the “cell height” there
14 was a clear gradient “top to bottom” with the cells with lower values on the top of the
15 cyst, and a progressive increase of their height towards the base (**Fig. 2D**).
16 Importantly, detailed examination of the entire surface of the cysts revealed a
17 subpattern: just in the center of the bottom region, the cells were shorter (**Fig. 2D**).
18 These complex cell morphology patterns were consistent among a high number of
19 cysts and appeared at different time points (**Fig. 2D**). To confirm that the pattern was
20 general, we leveraged high-throughput analysis and visualized the height of each cell
21 of all processed cysts. We plotted the “cell height” data (with the color code used in
22 the 3D cysts) on a scatter-polar diagram considering the angle and radius of the cell
23 centroids with respect to the centroid of the whole cyst (**Fig. 2E**). In this way, we were
24 able to visualize the distribution of the “cell height” feature in all cells, from all cysts, at
25 the same time. The plots confirmed the patterns observed in the individual cysts:
26 “shorter” cells (yellower colors) were located at the top and center of the bottom of the
27 cysts, while “taller” cells (pink-purple colors) were found at the periphery of the lateral
28 and bottom part of the cysts. A similar cell morphology pattern, although not that
29 evident, was observed in the distribution of “cell basal area”, “cell volume”, and “cell
30 surface area” values (**Fig. S6** related to **Fig. 2**). Furthermore, we found a different
31 pattern involving the distribution of the “cell apical area” values (**Fig. 2F**, **G**). In this
32 case, cells with a bigger apical area were enriched on the top and on the bottom of
33 the cysts. Meanwhile, cells with the smaller apical area were located in the middle

1 region of the cysts. We also found that “cell solidity” or “cell aspect ratio” characteristics
2 did not show any pattern (**Fig. S6** related to **Fig. 2**).

3

4 **The cell morphology pattern of some features can correlate with the shape of**
5 **the whole cysts**

6 Our high-throughput approach revealed that the MDCK cysts can present different
7 shapes and also intrinsic cell patterns. To test if the cell morphology patterns can be
8 affected in some way by the global shape of the cysts, we plotted the values of the
9 features against the position of the centroid of the cells in the Z-axis of the cyst. In this
10 way, we can quantify differences in populations of cells between the three more
11 abundant categories: ellipsoids, spheroids oblate and spheroids prolate. In the case
12 of “cell height”, there was a clear and robust gradient from “shorter” to “taller” cells,
13 from the top towards the bottom on the three types of shapes analyzed (**Fig. 2H, I**).
14 However, a more detailed analysis of the bottom-left side of the graphs (corresponding
15 to the shorter cells in the base of the cysts) revealed significant differences on the 10-
16 day cysts (**Fig. 2I, J**) but not on 4-day and 7-day cysts (**Fig. S7** related to **Fig. 2, Table**
17 **S5 and STAR Methods**). We also obtained differences in the case of the “basal area”
18 feature, but again, only on 10-day cysts (**Fig. S7** related to **Fig. 2 and Table S5**).
19 Conversely, we did not find differences at any time point with the “cell apical area”
20 feature (**Fig. S7** related to **Fig. 2 and Table S5**). Our results suggest that the shape of
21 the whole cyst could correlate with changes in cell morphology patterns (see
22 **Discussion**).

23

24 **The emergence of cell packing patterns in the cysts.**

25 Motivated by the finding of cell morphology patterns in the distribution of the values
26 of cell geometric features, we examined the presence of particular arrangements
27 linked to the connectivity of the cells. To this aim, we obtained Single-cell Cartography
28 representations of the distribution of scutoids (Gómez-Gálvez et al., 2018) in the cysts
29 (**Fig. 2K-M**). In this case, the high heterogeneity in the number of scutoids per cyst
30 and their distribution did not enable the identification of any clear pattern using the 3D
31 reconstructions of the cysts (**Fig. 2K**). Then, we plot the total number of cells and
32 analyze the distribution of their position along the XY-axes (**Fig. 2L**) and the Z-axis

1 (Fig. 2M) of the cysts. We did not find differences in the distributions in the first case.
2 However, our analysis detected a significant increase of the proportion of scutoids
3 from top to bottom of the cysts (Fig. 2M) that was not observed in non-scutoidal cells
4 (Chi-Square test, **Table S5** and **STAR Methods**). Our results suggest that cells pack
5 following self-organization patterns in the MDCK cysts.

6

7 DISCUSSION

8 In this work, we present CartoCell, a high-throughput segmentation framework that,
9 coupled with our Single-cell Cartography approach, provides new solutions on the
10 study of 3D complex epithelia. The combination of both tools allows the analysis of
11 hundreds of whole epithelial cysts at the cellular level. We depict the values of any
12 morphological or connectivity parameter at cellular resolution in two ways: using
13 heatmaps of the feature values over 3D reconstructions of each cyst, and with 2D plots
14 of the feature values and the spatial distribution of all the analyzed cells (>20,000
15 cells).

16 The generation of a large and sufficiently general training dataset of 3D segmented
17 epithelia is the bottleneck of high-throughput analysis of epithelial 3D packing. The
18 repositories providing segmented 3D epithelia that could be used as a training dataset
19 are very few and case-specific (von Chamier et al., 2021; Wolny et al., 2020). For that
20 reason, CartoCell starts with the accurate annotation of a small number of high-
21 resolution samples (Phase 1, 21 cysts, Fig. 1). Despite being time-consuming, this
22 step facilitates the interactive curation process since the actual shape of the individual
23 cells are easily identified. Only later, the reduction of resolution is possible while
24 maintaining the precision in the identification of cell outlines (Phases 2 and 3, Fig. 1).
25 Importantly, the downsampling resolution must match that of the low-resolution Z-
26 stacks acquired in batches of multiple cysts simultaneously. Here, we leverage our
27 deep learning approach to carry out two main steps. First, the small down-sampled
28 dataset is used to train our custom DNN, and subsequently infer a segmentation over
29 hundreds of cysts. Furthermore, thanks to the 3D Voronoi post-processing (Fig. 1 and
30 Fig. S3 related to Fig. 1), the largest proportion of the segmented cells and their
31 outlines are realistically predicted (Table S4). Although these cysts present some little
32 imperfections in their segmentation, they are fundamental for the second step,

1 consisting of retraining our DNN using hundreds of the previously segmented cysts
2 (Phase 4, **Fig. 1**). The use of this high number of cysts, despite not being perfectly
3 segmented, adds generality to the model, providing highly reliable results (**Table S4**)
4 even on completely newly acquired cysts (Phase 5, **Fig. 1**).

5 Regarding the usability of CartoCell by the community, here we provide an open-
6 source, well-documented and easy-to-use (without programming skills) segmentation
7 tool that can be used for any lab immediately (see **Code Availability** section for details
8 and tutorials of CartoCell and Single-cell Cartography). In addition, all the DNN
9 segmentation models generated in this work have been made publicly available.
10 CartoCell outperforms the state-of-the-art alternatives when our DNN model is trained
11 on a small low-resolution dataset (Phase 2, **Fig. 1**). The results are substantially better
12 when it is retrained on the larger low-resolution dataset produced by our own pipeline
13 (Phases 3-5, **Fig. 1**, see **Table S4**). Users can start their own analysis in Phase 5
14 without the laborious manual annotation of a new training dataset (**Fig. 1**). The only
15 requirement to use our pretrained models is other MCDK cysts images to acquire
16 low-resolution images with the same markers (anti-β-catenin and phalloidin, see
17 **STAR Methods**) that decorate the cell outlines and adjust the resolution during the
18 image acquisition (**STAR Methods**). Additionally, as we checked in our own samples,
19 the geometric parameters extracted by CartoCell without the final
20 curation/proofreading step are very reliable: a mean relative error of 5.6% compared
21 with the extracted features from the ground-truth cysts (**Fig. S4** related to **Fig. 1, Table**
22 **S4** and **STAR Methods**). However, the differences were larger in the case of
23 connectivity characteristics, suggesting that the final curation step is necessary for
24 these types of features. In light of these results, other labs could use the completely
25 automatic version of CartoCell to analyze in a fast way the geometric features of a
26 large number of samples. Nevertheless, the accurate detection of the epithelial
27 packing and connectivity of the tissues is an increasingly complex task that may
28 require the final curation/proofreading step to obtain accurate results. In that sense,
29 the use of CartoCell crucially reduces the proofreading time from 3-5 days to just $12 \pm$
30 6 minutes per cyst (**Fig. S4** related to **Fig. 1, Table S4** and **STAR Methods**).

31 Accounting with such a large number of samples is the key to quantifying the cell
32 morphology patterns in these epithelial structures. We uncover two different "cell
33 morphology" patterns within the cysts (**Fig. 2**). First, in the case of the "cell height",

1 “cell basal surface”, “cell volume” and “cell surface area” features, the cells present a
2 clear increase in the values from top to lateral bottom. Then, cells with lower values
3 also appear in the bottom-center of the cyst (**Fig. 2D, E** and **Fig. S6** related **Fig. 2**). A
4 different pattern can be easily distinguished when comparing the polar histograms of
5 those characteristics, with that of “cell apical area” (comparison between **Fig. 2E** and
6 **Fig. 2G**). In this second case, cells with a larger apical area are distributed at the
7 lateral top and lateral bottom of the cysts. Importantly, we also found features that do
8 not show any spatial pattern on the cyst (**Fig. S6** related to **Fig. 2**), suggesting that
9 different geometric features are independent of others. Our approach also reveals that
10 the complexity of the cyst can reach even the level of the packing and connectivity of
11 the cells. Here we show the example of the scutoids. Although the 3D maps of the
12 presence of scutoids do not reveal any clear pattern (**Fig. 2K**), the power of the high-
13 throughput approach reveals a clear accumulation of scutoids on the bottom side of
14 the cysts when compared with non-scutoidal cells (**Fig. 2M**).

15 At the tissue level, we show that cysts can adopt a variable range of shapes beyond
16 being symmetric spheres (Beck et al., 2022; Cerruti et al., 2013) (**Fig. 2A-B, Fig. S2**
17 related to **Fig. 1** and **Table S1**). This finding reveals a degree of complexity of the
18 MDCK cysts that allows us to study in detail the interplay between the shape of the
19 whole structure and the individual cell morphology. Indeed, we find that in 10-day cysts
20 (when the size of the cyst and the number of cells increase) it is possible to find a
21 correlation between some cellular geometric patterns and the shape of the cysts (**Fig.**
22 **2H-J, Fig. S7** related to **Fig. 2** and **Table S5**). Our Single-cell Cartography
23 methodology demonstrates that the first hints of asymmetry can emerge even in
24 tissues where there is not cell differentiation. Essentially, our method sheds light on a
25 very basic degree of variation at both cell and global level, that was not deeply
26 described before in cyst cultures. Taken together, our results reinforce the usefulness
27 of this simple system to study 3D morphogenesis and help to answer complex
28 questions such as “How do cells with different characteristics self-organize themselves
29 in a tridimensional epithelial tissue?” or “Are cells with different shapes physiologically
30 equivalent?”. In addition, the extracted morphological and connectivity cell information
31 could be used to feed biophysical models and force inference analysis (Dahl-Jensen
32 & Grapin-Botton, 2017; Gómez-Gálvez et al., 2021; Gómez-Gálvez et al., 2022),
33 providing valuable knowledge of cell mechanics.

1 Our high-throughput analysis presents several advantages related to the versatility
2 and efficiency of the method. By working with low-resolution images, it saves time
3 during image acquisition by allowing the capture of several samples in parallel.
4 Moreover, it accelerates image processing (segmentation and feature extraction) and
5 reduces photobleaching of the samples, opening the possibility of applying CartoCell
6 to *in vivo* imaged epithelia to study tissue development and its dynamical events. In
7 fact, our pipeline could be combined with cell lineage analysis methods that use nuclei
8 segmentation and tracking to study how different cell lineages remodel their geometric
9 and connectivity features during tissue development (de Medeiros et al., 2022; He et
10 al., 2022). For other epithelial systems or other cell membrane markers, the strategy
11 that we present here can be adapted. In this case, to automatically segment a large
12 number of low-resolution samples, the users should obtain a new dataset of high-
13 resolution images and follow our pipeline (Phase 1 to Phase 5, **Fig. 1**). This
14 generalization of our approach could be the solution to overcome several challenges
15 in the field of organoids, embryoids or even tissue engineering (Huch et al., 2017;
16 Laurent et al., 2017). CartoCell and the Single-cell Cartography methodology can
17 unveil hidden patterns in a simple and visual manner, what is pivotal to improve the
18 study of the self-organization of complex epithelial tissues where cells are in close
19 contact with each other (Carleton et al., 2022; Dahl-Jensen & Grapin-Botton, 2017;
20 Metzger et al., 2018; Takasato et al., 2016; Treacy et al., 2023). In a biomedical
21 context, the possibility of analyzing a very large number of samples is ideal for testing
22 the reproducibility of epithelial organoids cultures and performing detailed
23 comparisons between physiological and pathological conditions. Furthermore, high-
24 throughput drug testing on animal or human epithelial organoids could take advantage
25 of our approach to automatically analyze the effect of every drug against a target
26 disease at the cellular level (Beck et al., 2022; Drost & Clevers, 2018; Lo et al., 2020).

27

1 **STAR METHODS**

2

3 **MDCK cyst cell culture**

4 Type II MDCK (Madin-Darby canine kidney) cells were maintained in minimum
5 essential medium (MEM) containing GlutaMAX (Gibco) and supplemented with 10%
6 fetal bovine serum (FBS), 100 U/ml penicillin and 100 µg/ml streptomycin, in a 5% CO₂
7 humidified incubator at 37°C. For cyst formation, MDCK cells (2500 cells/well) were
8 suspended in complete medium containing 2% Matrigel (Corning, Life Sciences). Cell
9 suspension was plated in a 4-well culture slide (Corning, Life Science) on a thin layer
10 coating of 100% Matrigel. The plates were kept at 37°C in a humidified atmosphere of
11 5% CO₂ for 4, 7 or 10 days and the medium was changed every 2 days.

12

13 **Immunostaining and confocal imaging**

14 Cysts grown on 4-well chamber slides (Corning, Life Science) were fixed with 4%
15 paraformaldehyde in PBS and permeabilized with 0.5% TritonX-100 in Dubelcco's
16 Phosphate Buffered Saline (DPBS, Sigma-Aldrich) for 15 min at room temperature
17 (RT). After blocking with a solution of 0.02% Saponin (Sigma) and 3% BSA
18 (Applichem) in DPBS for 2h at RT, cysts were incubated overnight at 4°C with anti-β-
19 catenin antibody (1:1000 in DPBS-0.02% Saponin-3% BSA; rabbit, Sigma-Aldrich).
20 The following day, the cysts were washed with DPBS-0.02% Saponin-3% BSA
21 solution (3x, 5 min each) and incubated for 90 min at RT in this solution plus anti-rabbit
22 conjugated to Alexa Fluor (1:800, Molecular Probes) and phalloidin-Alexa-Fluor 647
23 (0.08 µM, Invitrogen). After washing with DPBS, plastic chambers were removed from
24 microscope slides and coverslips were mounted onto slides using Fluoromount-G
25 mounting solution (SouthernBiotech).

26 Cysts were imaged using a Nikon Eclipse Ti-E laser scanning confocal microscope.
27 High-resolution Z-stack confocal images were captured using a dry x40 objective (0.95
28 NA) and variable zoom (3.5-5.5), with a step size of 0.5 µm per slice and a scan speed
29 of 0.25 ms, from top to bottom. Then, they were exported as 1024 × 1024 pixels .nd2
30 files with an XY resolution ranging between 11.54-16.50 pixels per micron. Low
31 resolution images were captured (from top to bottom) using x20 oil objective (0.75 NA),

1 step size of 0.7 μm per slice, scan speed of 0.5 ms and exported as 1024×1024 pixels
2 .nd2 files with an XY resolution of 1.6248 pixels per micron.

3

4 **Custom DNN architecture: 3D ResU-Net**

5 Building upon the state of the art, we have designed 3D ResU-Net, a stable 3D
6 residual U-Net (Franco-Barranco et al., 2022) to segment epithelial cysts at the cell
7 level. The architecture is presented in **Fig. S1** related to **Fig. 1**. More specifically, 3D
8 ResU-Net is formed by full pre-activation residual blocks (two 3×3 convolutional
9 layers with a shortcut as shown in **Fig. S1** related to **Fig. 1**), with 52 filters in the first
10 level and adding 16 more at each level, dropout of 0.1 at each block. Down-sampling
11 (max-pooling) operators are performed only in 2D, since the input volumes are
12 anisotropic. The total number of trainable parameters is 1.3M.

13 The network received raw cyst images as input and outputs three different channels:
14 i) *cell masks*, with the probability that a voxel belongs to an individual cell, ii) *contour*,
15 containing the probabilities of cell outlines, and iii) *cell region*, representing the
16 foreground probability of the complete cyst.

17

18 **Network optimization**

19 To find the best solutions with our custom 3D ResU-Net, we made an exhaustive
20 search of hyperparameters (**Table S2**) and training configurations, exploring different
21 loss functions, optimizers, learning rates, batch sizes, and data augmentation
22 techniques. In particular, we minimized the binary cross-entropy (BCE) loss using the
23 Adam optimizer, with a learning rate of 0.0001, a batch size value of 2 and using a
24 patch size of $80 \times 80 \times 80$ voxels. We used a Tesla P40 GPU card to train the network
25 until convergence, i.e., for 1300 epochs with a patience established at 50 epochs
26 monitoring the validation loss and picking up the model that performs best in the
27 validation set (2 samples of “*training high-resolution dataset*” were used for model M1
28 and model M2 validation). Moreover, we applied on-the-fly data augmentation with
29 random rotations, vertical, horizontal and Z-axis flips and brightness distortions.

30

31

1 **Images preprocessing**

2 Before accessing the network, all raw images were preprocessed for contrast
3 homogenization using Fiji (Schindelin et al., 2012) macros. In this preprocessing, a
4 contrast adjustment was performed using the 'enhance contrast' function with 0.3% of
5 saturated pixels. Additionally, an 8-bit transformation is applied to them.

6 In the case of high-resolution images, a down-sampling was applied using Fiji
7 macros, transforming the variable images resolutions (11.54-16.50 pixels per micron
8 in XY and 0.5 μm between Z slices) to have the pixel size of the low-resolution images
9 (1.6248 pixels per micron in XY and 0.7 μm between Z slices). This was done by
10 calculating a correction factor that multiplies the size of the original image:

$$\begin{aligned} 11 \quad & [newSize_x, newSize_y, newSize_z] \\ 12 \quad & \rightarrow \left[Size_x \cdot \left(\frac{lowRes_{XY}}{highRes_{XY}} \right), Size_y \cdot \left(\frac{lowRes_{XY}}{highRes_{XY}} \right), Size_z \cdot \left(\frac{lowRes_z}{highRes_z} \right) \right] \end{aligned}$$

13 Both preprocessing Fiji macros are available at the public repository of the laboratory
14 (see **Code availability section**). Note that for label down-sampling the resize
15 command "Size" must have the interpolation method set as "None" and the option of
16 "average when down-sampling" disabled. For raw images, however, the interpolation
17 method should be set as "Bilinear" and the option of "average when down-sampling"
18 ticked. In the case of the low-resolution images, before applying the aforementioned
19 automatic contrast enhancement process, a manual preprocessing was performed in
20 which each cyst was cropped. This procedure was carried out with Fiji by drawing the
21 region of interest (ROI) using the "rectangle" tool and cropping using the "Crop" or
22 "Duplicate" command.

23

24 **Inference and postprocessing**

25 Each patch of 80x80x80 voxels was processed by the network to reconstruct the
26 output to the original cyst image size applying a padding of 16x16x16. In this way, we
27 avoided border effects on every patch, as reported in (Franco-Barranco et al., 2022).
28 The DNN produced three different outputs representing the probabilities of the
29 individual cell masks, contours and the whole cell region. We binarized the first two
30 outputs based on a fixed set of threshold values (0.2 was experimentally found to work

1 best) and created instance seeds to be fed to a marker-controlled watershed
2 algorithm.

3 As a final post-processing step, we ran the Voronoi algorithm along the three
4 dimensions, using the resulting instances of the watershed algorithm as Voronoi
5 seeds, and the cell region output (binarized using Otsu thresholding (Otsu, 1979)), as
6 the bounded region to be occupied by Voronoi cells (**Fig. S3** related to **Fig. 1**). Thus,
7 the unoccupied intercellular space of the cell region was filled by the nearest individual
8 cells (Voronoi seeds).

9

10 Comparison with the state of the art

11 Three state-of-the-art methods were tested against our protocol, being these
12 methods PlantSeg (Wolny et al., 2020), Cellpose (Stringer et al., 2021) and StarDist
13 3D (Weigert et al., 2020). For a more robust comparison, each one of the methods
14 was executed 10 times using default values and the small low-resolution raw dataset
15 as input. Each of the trained models was evaluated on the same test set to obtain
16 measures of the error in the results yielding the table (**Table S3**).

17 Cellpose training was performed locally by specifying the use of the GPU (Graphics
18 Processing Unit) and the non-use of a pretrained model following the training
19 documentation (<https://cellpose.readthedocs.io/en/latest/train.html>). Inference was
20 performed by following the instructions given in the command line documentation
21 (<https://cellpose.readthedocs.io/en/latest/command.html>) and using the diameter
22 suggested by the Cellpose GUI.

23 StarDist 3D training was performed in Google Colab using the official
24 ZeroCostDL4Mic (von Chamier et al., 2021) implementation
25 (<https://github.com/HenriquesLab/ZeroCostDL4Mic/wiki>) using default values except
26 for the following parameters: patch size, which was changed to 48 and patch height,
27 which was changed to 32 for convenience given the size of the images to be used.

28 PlantSeg training was performed locally following the training documentation
29 instructions (<https://github.com/hci-unihd/plant-seg>) using as default configuration the
30 3D Unet example for confocal imaging (https://github.com/wolny/pytorch-3dunet/blob/master/resources/3DUnet_confocal_boundary/train_config.yml)
31 replacing patch size to [32, 64, 64] as [z, x, y] for convenience given the size of the

1 images used and the minimum values allowed by PlantSeg. Further to the training of
2 the network, the PlantSeg GUI has modifiable parameters for the postprocessing. This
3 part was performed in a custom way trying to optimize the watershed (done with
4 Simple ITK) output obtained from the probability maps predicted by the network. The
5 parameters used were: Under-/Over- Segmentation Factor=0.75, Run Watershed
6 2D=False, CNN Prediction Threshold=0.113, Watershed Seeds Sigma=1.0,
7 Watershed Boundary Sigma=0.4, Superpixels Minimum Size=1, Cell Minimum
8 Size=5.

9

10 **Training and test datasets acquisition**

11 “*High-resolution label images*” were obtained after segmentation of the “*high-*
12 *resolution raw images*” (21 cysts) using LimeSeg (Machado et al., 2019), a plugin of
13 Fiji (Schindelin et al., 2012) for 3D segmentation, based on surface elements
14 (“Surfels”). This software was sourced from a set of seeds, manually placed over the
15 volumetric image to localize every single cell. These seeds grow until the cell outlines
16 are identified by detection of intensity gradient changes. The output of LimeSeg was
17 processed using an in-house Matlab program (2021a MathWorks) to detect and curate
18 imperfections during cysts segmentation (see **Proofreading of segmented cysts**
19 **section**). A down-sampled version of these 21 segmented cysts (see **Images**
20 **preprocessing section**) was used for training the model M1 (Phase 2, **Fig. 1**): 19
21 cysts composing the training dataset, and the remaining 2, making up the validation
22 dataset.

23 After running by default our high-throughput pipeline, we used the trained model M2
24 (Phase 5, **Fig. 1**) to infer and subsequently segment the “*test raw images*” (60 cysts
25 acquired at low resolution). These segmented images were manually curated using
26 our in-house Matlab proofreading program, obtaining the “*test label images*”

27

28 **Proofreading of segmented cysts**

29 A custom program developed in Matlab and available in the public repository of the
30 laboratory (see **Code availability section**) was designed for the proofreading of
31 segmented cysts. The software includes a user-friendly graphical user interface (GUI)
32 that allows to remove, modify, merge or create labels by drawing on the two-

1 dimensional slices of the image stack, also allowing the interpolation between labels
2 on different slices for faster curations. Both cell and lumen labels can be modified
3 using the GUI, which also has specific tools for each of them to ensure a proper
4 visualization.

5 In view that our biological study was developed on single-layer and single-lumen cysts,
6 the proofreading software relies on a segmentation error detection tool specific to our
7 purpose. The GUI displays the cell IDs of cells that do not contacting the apical and/or
8 basal surface of the cyst.

9 The software was designed to work quickly on batches of cysts. Once the cyst stops
10 displaying errors in the GUI and is marked by the user as fixed, the next cyst will be
11 displayed in the GUI to be corrected.

12 The procedure we carried out for the cyst curation started with the creation of 3
13 folders: One of them containing the batch of labels predicted by our pipeline, another
14 one, the batch of raw images and a third one reserved for the curated labels. The
15 software merges the raw images and the labels, and displays the result in a GUI along
16 with information on possible segmentation errors. An expert reviewed the displayed
17 image by carefully comparing each label with the staining of the cell membrane of the
18 raw image, and adjusting the labels until a perfect segmentation was achieved.

19

20 Features extraction and cyst shape classification

21 Using an in-house Matlab code, we quantified a set of geometrical and topological
22 parameters of the segmented epithelial cysts (**Table S1**) as is graphically described in
23 **Fig. S5** related to **Fig. 2**. We carried out a classification of cysts depending on the
24 morphology and differences between axes lengths (**Fig. 2A**). We considered that two
25 axes lengths were different if they differed more than 10%. We classified all cysts into
26 5 groups: 1. Sphere, when the lengths of the three axes of symmetry were similar. 2.
27 Oblate, when two axes lengths were similar and the different one was the shortest axis
28 length. 3. Prolate, when two axes lengths were similar and the different one was the
29 longest axis length. 4. Ellipsoid, when the three axes lengths were different. 5.
30 Negative curvature, when the solidity (volume / convex volume) of the cyst was inferior
31 to 0.9.

32

1 Quantification of biological features errors in non-curated segmented cysts

2 We extracted the features of both manually curated cysts (ground-truth) and the
3 output of our high-throughput segmentation pipeline (without proofreading). Some of
4 the segmented cysts without curation presented under-segmentation that promoted
5 gaps in the segmented tissue. This defective segmentation was called "cyst opening".
6 These gaps prevented the identification of the lumen of the "open cysts" automatically,
7 and thus some biological features could not be extracted. The 13% (46 cysts) of the
8 automatically segmented cysts presented this defect, and they were not used in the
9 comparison of the biological features values (**Fig. S4** related to **Fig. 1**). For the
10 remaining 87% of cysts (307 cysts), features were automatically extracted and
11 compared with the features extracted from manually curated cysts.

12 For each cyst, measurements of every feature were compared by computing the
13 relative error calculated as $\text{relative error} = \frac{|\text{predicted} - \text{ground truth}|}{\text{ground truth}}$. In the particular
14 case of percentage of scutoids, we could not calculate the relative error because in
15 some cases this feature represented a 0%, resulting in indetermination. Therefore, for
16 the calculation of its relative error, we defined the complementary of this feature (100%
17 - percentage of scutoids) such that we did not find any cyst with the 100% of cells
18 being scutoids. Finally, we calculated the mean and standard deviation of the errors
19 for every feature (**Table S4**).

20

21 Single-cell Cartography representation

22 We performed an analysis of the spatial distribution of features from more than
23 20,000 cells from 353 segmented cysts using our Single-cell Cartography tools
24 available in the public repository of the laboratory (see **Code availability section**).
25 Different types of representations arose from the use of these tools:

26 *Computer rendering of 3D cysts (Fig. 2D, F and K)*. We displayed a 3D visualization
27 of the segmented cysts and, using a gradient of color over the cell surfaces. We can
28 plot the normalized value of individual cell features, using our custom MATLAB
29 function Paint3D. A batch processing of cysts allowed the creation of large sheets with
30 the previously described three-dimensional representations of cysts on which to
31 perform a visual pattern analysis.

1 *Polar plots (Fig. 2E, G and Fig. S6 related to Fig.2)*. We used two types of two-
2 dimensional polar plots. Polar scatter plots and polar histograms were used to
3 represent the relative spatial position (**Fig. 2C**) and frequencies of a normalized cell
4 feature for all cells of all cysts simultaneously (**Fig. 2E, G and Fig. S6 related to Fig.**
5 **2**). For the creation of these two-dimensional polar plots (both polar scatter plot and
6 polar histograms) we proceeded as follows: The polar coordinate center for each cyst
7 was set at the centroid of the cyst. The radius was normalized from 0 to 1, being 1 the
8 distance to the farthest cell centroid from the centroid of the cyst. The colatitude angle
9 (representing the height on the vertical axis) was calculated with respect to the
10 horizontal plane passing through the cyst at the centroid, thus having positive angles
11 for cells above the cyst centroid and negative angles for cells below the cyst (**Fig. 2C**).
12 The azimuthal angle (which rotates around the vertical axis) was ignored since the
13 scope of the study was to search for patterns along the vertical axis. Disregarding this
14 angle led to a two-dimensional representation. This approach consisted of 5 polar
15 scatter plots and 4 polar histogram plots. First, a general polar scatter plot was shown
16 in which all cells were represented (**Fig. 2E, G and Fig. S6 related to Fig.2**). The value
17 of the features was represented by a color gradient as in the previous case. The rest
18 of the plots were dedicated to different ranges of the normalized feature to be studied:
19 0-0.25, 0.25-0.50, 0.50-0.75, 0.75-1. Each of the ranges was analyzed with a polar
20 scatter plot and a polar histogram plot showing, normalized, the distribution of cells
21 along the colatitude. In this way, we were able to visualize all the values of a particular
22 cell feature distributed along the cysts vertical axis.

23 *Normalized cell spatial data (Fig. 2H, I, L, M and Fig. S7 related to Fig. 2)*. For each
24 cell, the following data were represented: the Z-position of the cell centroid with
25 respect to the centroid of the lowest cell in the cyst (**Fig. 2H, and M**); the distance from
26 the cell centroid to the vertical (Z) axis passing through the centroid of the cyst (**Fig.**
27 **2L**) and the value of the cellular characteristic to be studied regarding its spatial
28 position (**Fig. 2H, I, L and M**). The cell features were normalized regarding the
29 maximum and minimum value of the feature in the whole cyst.

30

31

32

1 Evaluation metrics

2 To evaluate our results, we used common metrics to measure instance segmentation
3 performance in 2D and 3D images, which are calculated by matching the ground-truth
4 and prediction segmentation masks with an Intersection Over Union (IoU) value over
5 a certain threshold. In particular, we show values that required at least 30%, 50% and
6 75% IoU with the ground-truth for a detection to be a true positive (TP) (**Table S3**).
7 More specifically, we used the following metrics:

8 *Precision*, defined as

$$9 \quad precision = \frac{TP}{TP + FP}$$

10 where TP and FP are the number of true and false positives, respectively.

11 *Recall*, defined as

$$12 \quad recall = \frac{TP}{TP + FN}$$

13 where FN is the number of false negatives.

14 *Accuracy*, defined as

$$15 \quad accuracy = \frac{TP}{TP + FP + FN}$$

16 *F1 or F-score*, defined as

$$17 \quad F1 = \frac{2 \times precision \times recall}{precision + recall}$$

18 *Panoptic quality*, a unified metric to express both segmentation and recognition quality,
19 defined as in Equation 1 of (Kirillov et al., 2018)

$$20 \quad panoptic = \frac{\sum_{(p,g) \in TP} IoU(p, g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}$$

21 where p and g are the predicted and ground-truth segments, respectively. Therefore,
22 $\frac{1}{|TP|} \sum_{(p,g) \in TP} IoU(p, g)$ is the average of matched segments, and $\frac{1}{2}|FP| + \frac{1}{2}|FN|$ in the
23 denominator penalizes segments without matches.

24

25

26

1 **Statistical analysis**

2 At each time-point sampled (4 days, 7 days and 10 days cysts), at least seven
3 independent cultures were carried out. For comparisons of the features values on
4 certain cell populations between different categories of cyst shapes, we used a
5 univariate statistical protocol. First, samples were evaluated for normal distribution and
6 similar variance by using the Shapiro-Wilk test and the two-sample F-test,
7 respectively. If samples followed a normal distribution and similar variance, we
8 employed the two unpaired Student's t-test; whereas data had a normal distribution
9 but not equal variance, we used the two-tailed Welch test. Finally, when data not
10 adjusting to a normal distribution, we employed the non-parametric Mann-Whitney U
11 test. Data were represented in a bar graph as mean \pm SD (standard deviation) and
12 $p \leq 0.05$ was considered statistically significant (**Fig. 2J** and **Fig. S7** related to **Fig. 2**).
13 “*”, “**” and “***” indicating $p \leq 0.05$, $p \leq 0.01$ and $p \leq 0.001$ respectively (**Table S5**). In
14 a different statistical analysis, we tested cell spatial distribution similarity from bottom
15 to top (in Z-axis) or from cyst centroid to outside (in XY axis) of in the proportion of
16 scutoids and non-scutoidal cells (**Fig. 2L-M**). Following the guidelines from (Gómez-
17 Gálvez et al., 2022; Sánchez-Gutiérrez et al., 2016), we used the chi-square test for
18 the trend across all samples to determine if there is a linear trend for the proportional
19 data, considering statistically significant a $p \leq 0.05$ (**Table S5**). “*”, “**” and “***”
20 indicating $p \leq 0.05$, $p \leq 0.01$ and $p \leq 0.001$ respectively. Statistical analyzes and graphs
21 were performed using GraphPad Prism version 8.4.2. (GraphPad Software, La Jolla
22 California, USA, www.graphpad.com).

23

24 **CODE AVAILABILITY**

25 All the necessary materials to reproduce this study are available at the Mendeley
26 Data repository: <http://dx.doi.org/10.17632/7gbkxngpm.1>

27

28

1 BIBLIOGRAPHY

- 2 Alfonso-Pérez, T., Baonza, G., Herranz, G., & Martín-Belmonte, F. (2022).
3 Deciphering the interplay between autophagy and polarity in epithelial
4 tubulogenesis. *Seminars in Cell & Developmental Biology*, 131, 160–172.
5 <https://doi.org/https://doi.org/10.1016/j.semcd.2022.05.015>
- 6 Bailoni, A., Pape, C., Hütsch, N., Wolf, S., Beier, T., Kreshuk, A., & Hamprecht, F. A.
7 (2019). *GASP, a generalized framework for agglomerative clustering of signed*
8 *graphs and its application to Instance Segmentation.*
9 <https://doi.org/10.48550/arxiv.1906.11713>
- 10 Beck, L. E., Lee, J., Coté, C., Dunagin, M. C., Lukonin, I., Salla, N., Chang, M. K.,
11 Hughes, A. J., Mornin, J. D., Gartner, Z. J., Liberali, P., & Raj, A. (2022).
12 Systematically quantifying morphological features reveals constraints on
13 organoid phenotypes. *Cell Systems*, 13(7), 547-560.e3.
14 <https://doi.org/10.1016/j.cels.2022.05.008>
- 15 Bryant, D. M., & Mostov, K. E. (2008). From cells to organs: building polarized tissue.
16 *Nature Reviews Molecular Cell Biology*, 9(11), 887–901.
17 <https://doi.org/10.1038/nrm2523>
- 18 Carleton, A. E., Duncan, M. C., & Taniguchi, K. (2022). Human epiblast
19 lumenogenesis: From a cell aggregate to a luminal cyst. *Seminars in Cell &*
20 *Developmental Biology*, 131, 117–123.
21 <https://doi.org/10.1016/j.semcd.2022.05.009>
- 22 Cerruti, B., Puliafito, A., Shewan, A. M., Yu, W., Combes, A. N., Little, M. H.,
23 Chianale, F., Primo, L., Serini, G., Mostov, K. E., Celani, A., & Gamba, A.
24 (2013). Polarity, cell division, and out-of-equilibrium dynamics control the growth
25 of epithelial structures. *Journal of Cell Biology*, 203(2), 359–372.
26 <https://doi.org/10.1083/jcb.201305044>
- 27 Ciresan, D., Giusti, A., Gambardella, L., & Schmidhuber, J. (2012). Deep Neural
28 Networks Segment Neuronal Membranes in Electron Microscopy Images. In F.
29 Pereira, C. J. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural*
30 *Information Processing Systems* (Vol. 25). Curran Associates, Inc.
31 <https://proceedings.neurips.cc/paper/2012/file/459a4ddcb586f24efd9395aa7662bc7c-Paper.pdf>
- 33 Cousty, J., Bertrand, G., Najman, L., & Couplie, M. (2009). Watershed Cuts:
34 Minimum Spanning Forests and the Drop of Water Principle. *IEEE Transactions*
35 *on Pattern Analysis and Machine Intelligence*, 31(8), 1362–1374.
36 <https://doi.org/10.1109/TPAMI.2008.173>
- 37 Dahl-Jensen, S., & Grapin-Botton, A. (2017). The physics of organoids: a biophysical
38 approach to understanding organogenesis. *Development*, 144(6), 946–951.
39 <https://doi.org/10.1242/dev.143693>
- 40 Davidson, L. A. (2012). Epithelial machines that shape the embryo. *Trends in Cell*
41 *Biology*, 22(2), 82–87. <https://doi.org/10.1016/j.tcb.2011.10.005>

- 1 de Medeiros, G., Ortiz, R., Strnad, P., Boni, A., Moos, F., Repina, N., Challet Meylan,
2 L., Maurer, F., & Liberali, P. (2022). Multiscale light-sheet organoid imaging
3 framework. *Nature Communications*, 13(1), 4864.
4 <https://doi.org/10.1038/s41467-022-32465-z>
- 5 Drost, J., & Clevers, H. (2018). Organoids in cancer research. *Nature Reviews
6 Cancer*, 18(7), 407–418. <https://doi.org/10.1038/s41568-018-0007-6>
- 7 Dukes, J. D., Whitley, P., & Chalmers, A. D. (2011). The MDCK variety pack:
8 Choosing the right strain. In *BMC Cell Biology* (Vol. 12).
9 <https://doi.org/10.1186/1471-2121-12-43>
- 10 Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A.,
11 Deubner, J., Jäckel, Z., Seiwald, K., Dovzhenko, A., Tietz, O., Dal Bosco, C.,
12 Walsh, S., Saltukoglu, D., Tay, T. L., Prinz, M., Palme, K., Simons, M., ...
13 Ronneberger, O. (2018). U-Net: deep learning for cell counting, detection, and
14 morphometry. *Nature Methods* 2018 16:1, 16(1), 67–70.
15 <https://doi.org/10.1038/s41592-018-0261-2>
- 16 Fessenden, T. B., Beckham, Y., Perez-Neut, M., Ramirez-San Juan, G., Chourasia,
17 A. H., Macleod, K. F., Oakes, P. W., & Gardel, M. L. (2018). Dia1-dependent
18 adhesions are required by epithelial tissues to initiate invasion. *Journal of Cell
19 Biology*, 217(4), 1485–1502. <https://doi.org/10.1083/jcb.201703145>
- 20 Franco-Barranco, D., Muñoz-Barrutia, A., & Arganda-Carreras, I. (2022). Stable
21 Deep Neural Network Architectures for Mitochondria Segmentation on Electron
22 Microscopy Volumes. *Neuroinformatics*, 20(2), 437–450.
23 <https://doi.org/10.1007/s12021-021-09556-1>
- 24 Gómez, H. F., Dumond, M. S., Hodel, L., Vetter, R., & Iber, D. (2021). 3D cell
25 neighbour dynamics in growing pseudostratified epithelia. *eLife*, 10.
26 <https://doi.org/10.7554/eLife.68135>
- 27 Gómez-Gálvez, P., Anbari, S., Escudero, L. M., & Buceta, J. (2021). Mechanics and
28 self-organization in tissue development. *Seminars in Cell & Developmental
29 Biology*, 120, 147–159. <https://doi.org/10.1016/j.semcdb.2021.07.003>
- 30 Gómez-Gálvez, P., Vicente-Munuera, P., Anbari, S., Buceta, J., & Escudero, L. M.
31 (2021). The complex three-dimensional organization of epithelial tissues.
32 *Development*, 148(1), dev195669. <https://doi.org/10.1242/dev.195669>
- 33 Gómez-Gálvez, P., Vicente-Munuera, P., Anbari, S., Tagua, A., Gordillo-Vázquez,
34 C., Andrés-San Román, J. A., Franco-Barranco, D., Palacios, A. M., Velasco,
35 A., Capitán-Agudo, C., Grima, C., Annese, V., Arganda-Carreras, I., Robles, R.,
36 Márquez, A., Buceta, J., & Escudero, L. M. (2022). A quantitative biophysical
37 principle to explain the 3D cellular connectivity in curved epithelia. *Cell Systems*,
38 13(8), 631-643.e8. <https://doi.org/10.1016/J.CELLS.2022.06.003>
- 39 Gómez-Gálvez, P., Vicente-Munuera, P., Tagua, A., Forja, C., Castro, A. M., Letrán,
40 M., Valencia-Expósito, A., Grima, C., Bermúdez-Gallardo, M., Serrano-Pérez-
41 Higueras, Ó., Cavodeassi, F., Sotillos, S., Martín-Bermudo, M. D., Márquez, A.,

- 1 Buceta, J., & Escudero, L. M. (2018). Scutoids are a geometrical solution to
2 three-dimensional packing of epithelia. *Nature Communications*, 9(1), 2960.
3 <https://doi.org/10.1038/s41467-018-05376-1>
- 4 Guo, Q., Xia, B., Moshiach, S., Xu, C., Jiang, Y., Chen, Y., Sun, Y., Lahti, J. M., &
5 Zhang, X. A. (2008). The microenvironmental determinants for kidney epithelial
6 cyst morphogenesis. *European Journal of Cell Biology*, 87(4), 251–266.
7 <https://doi.org/10.1016/j.ejcb.2007.11.004>
- 8 He, Z., Maynard, A., Jain, A., Gerber, T., Petri, R., Lin, H.-C., Santel, M., Ly, K.,
9 Dupré, J.-S., Sidow, L., Sanchis Calleja, F., Jansen, S. M. J., Riesenbergs, S.,
10 Camp, J. G., & Treutlein, B. (2022). Lineage recording in human cerebral
11 organoids. *Nature Methods*, 19(1), 90–99. <https://doi.org/10.1038/s41592-021-01344-8>
- 13 Herranz, G., & Martín-Belmonte, F. (2022). Cadherin-mediated adhesion takes
14 control. *The EMBO Journal*. <https://doi.org/10.15252/embj.2022112662>
- 15 Honda, H. (1978). Description of cellular patterns by Dirichlet domains: The two-
16 dimensional case. *Journal of Theoretical Biology*, 72(3), 523–543.
17 [https://doi.org/10.1016/0022-5193\(78\)90315-6](https://doi.org/10.1016/0022-5193(78)90315-6)
- 18 Huch, M., Knoblich, J. A., Lutolf, M. P., & Martinez-Arias, A. (2017). The hope and
19 the hype of organoid research. *Development*, 144(6), 938–941.
20 <https://doi.org/10.1242/dev.150201>
- 21 Imai, M., Furusawa, K., Mizutani, T., Kawabata, K., & Haga, H. (2015). Three-
22 dimensional morphogenesis of MDCK cells induced by cellular contractile forces
23 on a viscous substrate. *Scientific Reports*, 5(1), 14208.
24 <https://doi.org/10.1038/srep14208>
- 25 Ioannou, F., Dawi, M. A., Tetley, R. J., Mao, Y., & Muñoz, J. J. (2020). Development
26 of a New 3D Hybrid Model for Epithelia Morphogenesis. *Frontiers in
27 Bioengineering and Biotechnology*, 8. <https://doi.org/10.3389/fbioe.2020.00405>
- 28 Kappes, J. H., Speth, M., Andres, B., Reinelt, G., & Schn, C. (2011). Globally
29 Optimal Image Partitioning by Multicuts. In *Lecture Notes in Computer Science
(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in
31 Bioinformatics)*: Vol. 6819 LNCS (pp. 31–44). Springer, Berlin, Heidelberg.
32 https://doi.org/10.1007/978-3-642-23094-3_3
- 33 Kirillov, A., He, K., Girshick, R., Rother, C., & Dollar, P. (2018). Panoptic
34 Segmentation. *Proceedings of the IEEE Computer Society Conference on
35 Computer Vision and Pattern Recognition*, 2019-June, 9396–9405.
36 <https://doi.org/10.48550/arxiv.1801.00868>
- 37 Laine, R. F., Arganda-Carreras, I., Henriques, R., & Jacquemet, G. (2021). Avoiding
38 a replication crisis in deep-learning-based bioimage analysis. *Nature Methods*,
39 18(10), 1136–1144. <https://doi.org/10.1038/s41592-021-01284-3>
- 40 Laurent, J., Blin, G., Chatelain, F., Vanneaux, V., Fuchs, A., Larghero, J., & Théry,
41 M. (2017). Convergence of microengineering and cellular self-organization

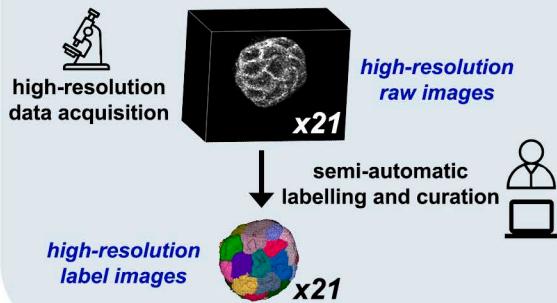
- 1 towards functional tissue manufacturing. *Nature Biomedical Engineering*, 1(12),
2 939–956. <https://doi.org/10.1038/s41551-017-0166-x>
- 3 Lecuit, T., & Lenne, P.-F. (2007). Cell surface mechanics and the control of cell
4 shape, tissue patterns and morphogenesis [Journal Article]. *Nature Reviews
5 Molecular Cell Biology*, 8(8), 633–644. <https://doi.org/10.1038/nrm2222>
- 6 Lin, Z., Wei, D., Petkova, M. D., Wu, Y., Ahmed, Z., K, K. S., Zou, S., Wendt, N.,
7 Boulanger-Weill, J., Wang, X., Dhanyasi, N., Arganda-Carreras, I., Engert, F.,
8 Lichtman, J., & Pfister, H. (2021). *NucMM Dataset: 3D Neuronal Nuclei Instance
9 Segmentation at Sub-Cubic Millimeter Scale* (pp. 164–174).
10 https://doi.org/10.1007/978-3-030-87193-2_16
- 11 Lo, Y.-H., Karlsson, K., & Kuo, C. J. (2020). Applications of organoids for cancer
12 biology and precision medicine. *Nature Cancer*, 1(8), 761–773.
13 <https://doi.org/10.1038/s43018-020-0102-y>
- 14 Lou, Y., Rupprecht, J.-F., Hiraiwa, T., & Saunders, T. E. (2022). Curvature-induced
15 cell rearrangements in biological tissues. *BioRxiv*, 2022.05.18.492428.
16 <https://doi.org/10.1101/2022.05.18.492428>
- 17 Machado, S., Mercier, V., & Chiaruttini, N. (2019). LimeSeg: a coarse-grained lipid
18 membrane simulation for 3D image segmentation. *BMC Bioinformatics*, 20(1), 2.
19 <https://doi.org/10.1186/s12859-018-2471-0>
- 20 Martín-Belmonte, F., Yu, W., Rodríguez-Fraticelli, A. E., Ewald, A., Werb, Z., Alonso,
21 M. A., & Mostov, K. (2008). Cell-Polarity Dynamics Controls the Mechanism of
22 Lumen Formation in Epithelial Morphogenesis. *Current Biology*, 18(7), 507–513.
23 <https://doi.org/10.1016/j.cub.2008.02.076>
- 24 Meijering, E. (2020). A bird's-eye view of deep learning in bioimage analysis.
25 *Computational and Structural Biotechnology Journal*, 18, 2312–2325.
26 <https://doi.org/10.1016/j.csbj.2020.08.003>
- 27 Messal, H. A., Alt, S., Ferreira, R. M. M., Gribben, C., Wang, V. M.-Y., Cotoi, C. G.,
28 Salbreux, G., & Behrens, A. (2019). Tissue curvature and apicobasal
29 mechanical tension imbalance instruct cancer morphogenesis. *Nature*,
30 566(7742), 126–130. <https://doi.org/10.1038/s41586-019-0891-2>
- 31 Metzger, J. J., Simunovic, M., & Brivanlou, A. H. (2018). Synthetic embryology:
32 controlling geometry to model early mammalian development. *Current Opinion
33 in Genetics & Development*, 52, 86–91.
34 <https://doi.org/10.1016/j.gde.2018.06.006>
- 35 Meyer, F. (1994). Topographic distance and watershed lines. *Signal Processing*,
36 38(1), 113–125. [https://doi.org/10.1016/0165-1684\(94\)90060-4](https://doi.org/10.1016/0165-1684(94)90060-4)
- 37 Moen, E., Bannon, D., Kudo, T., Graf, W., Covert, M., & van Valen, D. (2019). Deep
38 learning for cellular image analysis. *Nature Methods*, 16(12), 1233–1246.
39 <https://doi.org/10.1038/s41592-019-0403-1>

- 1 O'Brien, L. E., Zegers, M. M. P., & Mostov, K. E. (2002). Building epithelial
2 architecture: Insights from three-dimensional culture models. In *Nature Reviews
3 Molecular Cell Biology* (Vol. 3, Issue 7, pp. 531–537).
4 <https://doi.org/10.1038/nrm859>
- 5 Okuda, S., Miura, T., Inoue, Y., Adachi, T., & Eiraku, M. (2018). Combining Turing
6 and 3D vertex models reproduces autonomous multicellular morphogenesis with
7 undulation, tubulation, and branching. *Scientific Reports* 2018 8:1, 8(1), 1–15.
8 <https://doi.org/10.1038/s41598-018-20678-6>
- 9 Okuda, S., & Sato, K. (2022). Polarized interfacial tension induces collective
10 migration of cells, as a cluster, in a 3D tissue. *Biophysical Journal*, 121(10),
11 1856–1867. <https://doi.org/10.1016/J.BPJ.2022.04.018>
- 12 Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE
13 Transactions on Systems, Man, and Cybernetics*, 9(1), 62–66.
14 <https://doi.org/10.1109/TSMC.1979.4310076>
- 15 Perez, L., & Wang, J. (2017). *The Effectiveness of Data Augmentation in Image
16 Classification using Deep Learning*. <https://doi.org/10.48550/arxiv.1712.04621>
- 17 Prabhakara, C., Iyer, K. S., Rao, M., Saunders, T. E., & Mayor, S. (2022).
18 Quantitative analysis of three-dimensional cell organisation and concentration
19 profiles within curved epithelial tissues. *BioRxiv*, 2022.05.16.492131.
20 <https://doi.org/10.1101/2022.05.16.492131>
- 21 Razzak, M. I., Naz, S., & Zaib, A. (2018). *Deep Learning for Medical Image
22 Processing: Overview, Challenges and the Future* (pp. 323–350).
23 https://doi.org/10.1007/978-3-319-65981-7_12
- 24 Rupprecht, J.-F., Ong, K. H., Yin, J., Huang, A., Dinh, H.-H.-Q., Singh, A. P., Zhang,
25 S., Yu, W., & Saunders, T. E. (2017). Geometric constraints alter cell
26 arrangements within curved epithelial tissues. *Molecular Biology of the Cell*,
27 28(25), 3582–3594. <https://doi.org/10.1091/mbc.e17-01-0060>
- 28 Sakurai, A., Matsuda, M., & Kiyokawa, E. (2012). Activated Ras Protein Accelerates
29 Cell Cycle Progression to Perturb Madin-Darby Canine Kidney Cystogenesis.
30 *Journal of Biological Chemistry*, 287(38), 31703–31711.
31 <https://doi.org/10.1074/jbc.M112.377804>
- 32 Sánchez-Gutiérrez, D., Tozluoglu, M., Barry, J. D., Pascual, A., Mao, Y., &
33 Escudero, L. M. (2016). Fundamental physical cellular constraints drive self-
34 organization of tissues. *The EMBO Journal*, 35(1), 77–88.
35 <https://doi.org/10.15252/embj.201592374>
- 36 Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T.,
37 Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D. J.,
38 Hartenstein, V., Eliceiri, K., Tomancak, P., & Cardona, A. (2012). Fiji: an open-
39 source platform for biological-image analysis. *Nature Methods*, 9(7), 676–682.
40 <https://doi.org/10.1038/nmeth.2019>

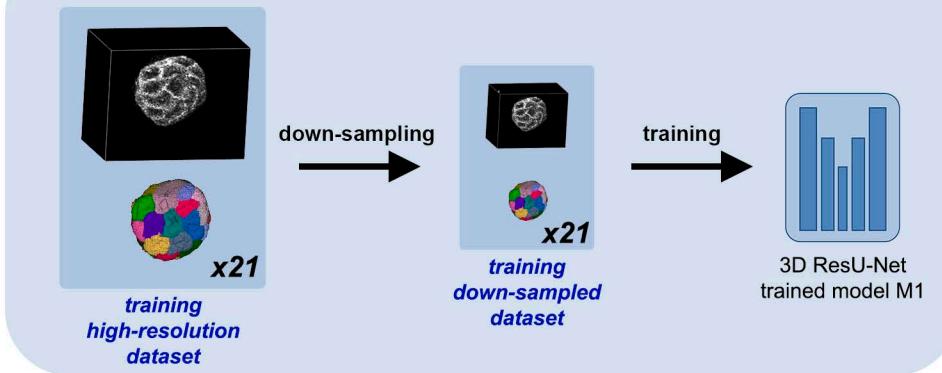
- 1 Schmeichel, K. L., & Bissell, M. J. (2003). Modeling tissue-specific signaling and
2 organ function in three dimensions. *Journal of Cell Science*, 116(12), 2377–
3 2388. <https://doi.org/10.1242/jcs.00503>
- 4 Schmidt, U., Weigert, M., Broaddus, C., & Myers, G. (2018). *Cell Detection with Star-*
5 *Convex Polygons* (pp. 265–273). https://doi.org/10.1007/978-3-030-00934-2_30
- 6 Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation
7 for Deep Learning. *Journal of Big Data*, 6(1), 1–48.
8 <https://doi.org/10.1186/S40537-019-0197-0/FIGURES/33>
- 9 Stringer, C., Wang, T., Michaelos, M., & Pachitariu, M. (2021). Cellpose: a generalist
10 algorithm for cellular segmentation. *Nature Methods*, 18(1), 100–106.
11 <https://doi.org/10.1038/s41592-020-01018-x>
- 12 Takasato, M., Er, P. X., Chiu, H. S., & Little, M. H. (2016). Generation of kidney
13 organoids from human pluripotent stem cells. *Nature Protocols*, 11(9), 1681–
14 1692. <https://doi.org/10.1038/nprot.2016.098>
- 15 Treacy, N. J., Clerkin, S., Davis, J. L., Kennedy, C., Miller, A. F., Saiani, A.,
16 Wychowaniec, J. K., Brougham, D. F., & Crean, J. (2023). Growth and
17 differentiation of human induced pluripotent stem cell (hiPSC)-derived kidney
18 organoids using fully synthetic peptide hydrogels. *Bioactive Materials*, 21, 142–
19 156. <https://doi.org/10.1016/j.bioactmat.2022.08.003>
- 20 Vidal-Quadras, M., Holst, M. R., Francis, M. K., Larsson, E., Hachimi, M., Yau, W. L.,
21 Peränen, J., Martín-Belmonte, F., & Lundmark, R. (2017). Endocytic turnover of
22 Rab8 controls cell polarization. *Journal of Cell Science*.
23 <https://doi.org/10.1242/jcs.195420>
- 24 von Chamier, L., Laine, R. F., Jukkala, J., Spahn, C., Krentzel, D., Nehme, E.,
25 Lerche, M., Hernández-Pérez, S., Mattila, P. K., Karinou, E., Holden, S., Solak,
26 A. C., Krull, A., Buchholz, T.-O., Jones, M. L., Royer, L. A., Leterrier, C.,
27 Shechtman, Y., Jug, F., ... Henriques, R. (2021). Democratising deep learning
28 for microscopy with ZeroCostDL4Mic. *Nature Communications*, 12(1), 2276.
29 <https://doi.org/10.1038/s41467-021-22518-0>
- 30 Voronoi, G. (1908). Nouvelles applications des paramètres continus à la théorie des
31 formes quadratiques. Deuxième mémoire. Recherches sur les parallélépèdres
32 primitifs. *Journal Für Die Reine Und Angewandte Mathematik*, 134, 198–287.
33 <http://eudml.org/doc/149291>
- 34 Wei, D., Lin, Z., Franco-Barranco, D., Wendt, N., Liu, X., Yin, W., Huang, X., Gupta,
35 A., Jang, W.-D., Wang, X., Arganda-Carreras, I., Lichtman, J. W., & Pfister, H.
36 (2020). *MitoEM Dataset: Large-Scale 3D Mitochondria Instance Segmentation*
37 *from EM Images* (pp. 66–76). https://doi.org/10.1007/978-3-030-59722-1_7
- 38 Weigert, M., Schmidt, U., Haase, R., Sugawara, K., & Myers, G. (2020). Star-convex
39 Polyhedra for 3D Object Detection and Segmentation in Microscopy. *2020 IEEE*
40 *Winter Conference on Applications of Computer Vision (WACV)*, 3655–3662.
41 <https://doi.org/10.1109/WACV45572.2020.9093435>

- 1 Wells, E. K., Yarborough, O., Lifton, R. P., Cantley, L. G., & Caplan, M. J. (2013).
2 Epithelial morphogenesis of MDCK cells in three-dimensional collagen culture is
3 modulated by interleukin-8. *American Journal of Physiology-Cell Physiology*,
4 304(10), C966–C975. <https://doi.org/10.1152/ajpcell.00261.2012>
- 5 Wolf, S., Pape, C., Bailoni, A., Rahaman, N., Kreshuk, A., Köthe, U., & Hamprecht,
6 F. A. (2018). *The Mutex Watershed: Efficient, Parameter-Free Image
7 Partitioning* (pp. 571–587). https://doi.org/10.1007/978-3-030-01225-0_34
- 8 Wolny, A., Cerrone, L., Vijayan, A., Tofanelli, R., Barro, A. V., Louveaux, M., Wenzl,
9 C., Strauss, S., Wilson-Sánchez, D., Lymbouridou, R., Steigleder, S. S., Pape,
10 C., Bailoni, A., Duran-Nebreda, S., Bassel, G. W., Lohmann, J. U., Tsiantis, M.,
11 Hamprecht, F. A., Schneitz, K., ... Kreshuk, A. (2020). Accurate and versatile
12 3D segmentation of plant tissues at cellular resolution. *eLife*, 9.
13 <https://doi.org/10.7554/eLife.57613>
- 14 Yan, L., Tsujita, K., Fujita, Y., & Itoh, T. (2021). PTEN is required for the migration
15 and invasion of Ras-transformed MDCK cells. *FEBS Letters*, 595(9), 1303–
16 1312. <https://doi.org/10.1002/1873-3468.14053>
- 17 Yan, Z., Yang, X., & Cheng, K.-T. T. (2018). *A Deep Model with Shape-Preserving
18 Loss for Gland Instance Segmentation* (pp. 138–146).
19 https://doi.org/10.1007/978-3-030-00934-2_16
- 20 Yonemura, S. (2014). Differential Sensitivity of Epithelial Cells to Extracellular Matrix
21 in Polarity Establishment. *PLoS ONE*, 9(11), e112922.
22 <https://doi.org/10.1371/journal.pone.0112922>
- 23 Yu, W., Fang, X., Ewald, A., Wong, K., Hunt, C. A., Werb, Z., Matthay, M. A., &
24 Mostov, K. (2007). Formation of Cysts by Alveolar Type II Cells in Three-
25 dimensional Culture Reveals a Novel Mechanism for Epithelial Morphogenesis.
26 *Molecular Biology of the Cell*, 18(5), 1693–1700.
27 <https://doi.org/10.1091/mbc.e06-11-1052>
- 28

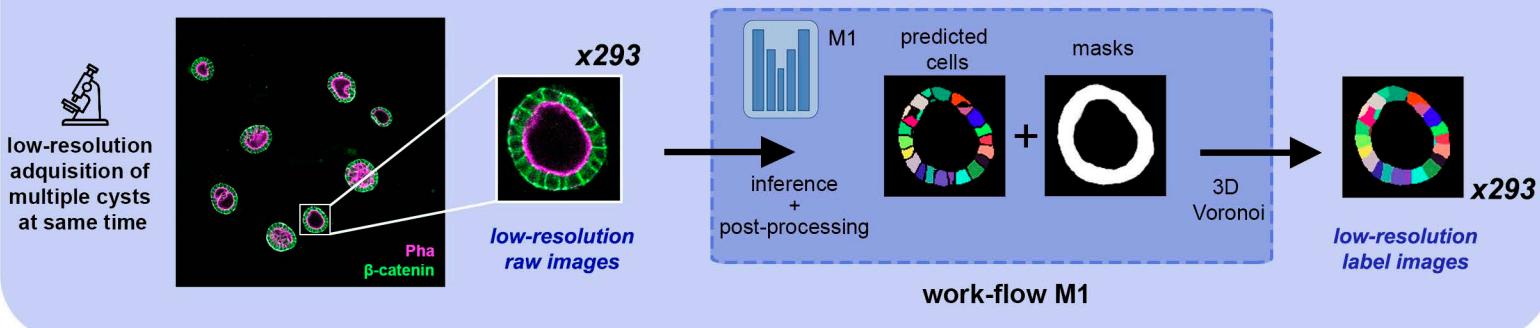
Phase 1: High-resolution cysts acquisition and curation (21 samples)



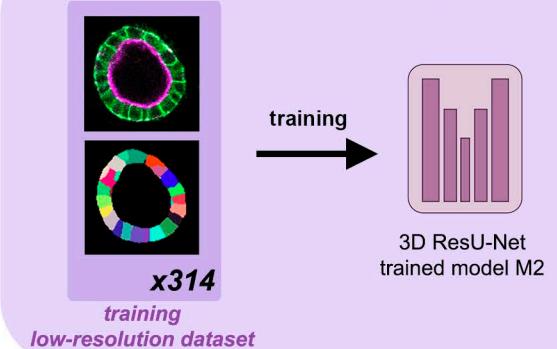
Phase 2: Down-sampling and small dataset training



Phase 3: Low-resolution cysts acquisition and training dataset extension (293 cysts)



Phase 4: Large dataset training



Phase 5: High-throughput low resolution cyst segmentation

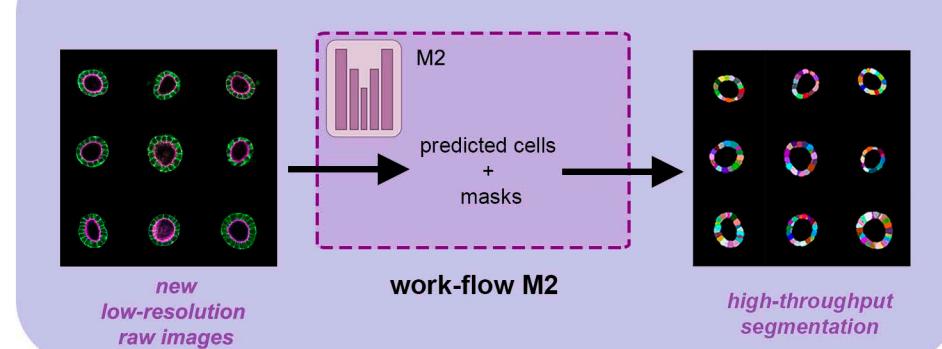


Figure 1. CartoCell pipeline for high-throughput epithelial cysts segmentation. Phase 1: The “*high-resolution raw images*” consist of confocal Z-stacks images, where the cell membrane is stained. These images are segmented and proofread using LimeSeg, and a custom Matlab code for curation to obtain the “*high-resolution label images*” (**STAR Methods**). Together, the raw and the label images encompass the “*training high-resolution dataset*”. Number of samples = 21. Phase 2: The “*training high-resolution dataset*” is down-sampled to obtain the “*training down-sampled dataset*”, which is the training set for the “model M1”. Phase 3: low-resolution images are obtained from confocal Z-stacks images, stained in a similar way to Phase 1. Number of samples = 293. Next, the “work-flow M1” is applied: Inference using “model M1” and subsequent post-processing to obtain individual cell instance predictions and cell masks, followed by the 3D Voronoi algorithm to guarantee that predicted cells remain in close contact. As a result, the “*low-resolution label images*” are generated. Phase 4: Training of the “model M2” on the large “*training low-resolution dataset*”. Number of samples = 314. Phase 5: High-throughput segmentation of new low-resolution images (unseen by the pipeline) using the “work-flow M2”, which is equivalent to the “work-flow M1” but using the “model M2”.

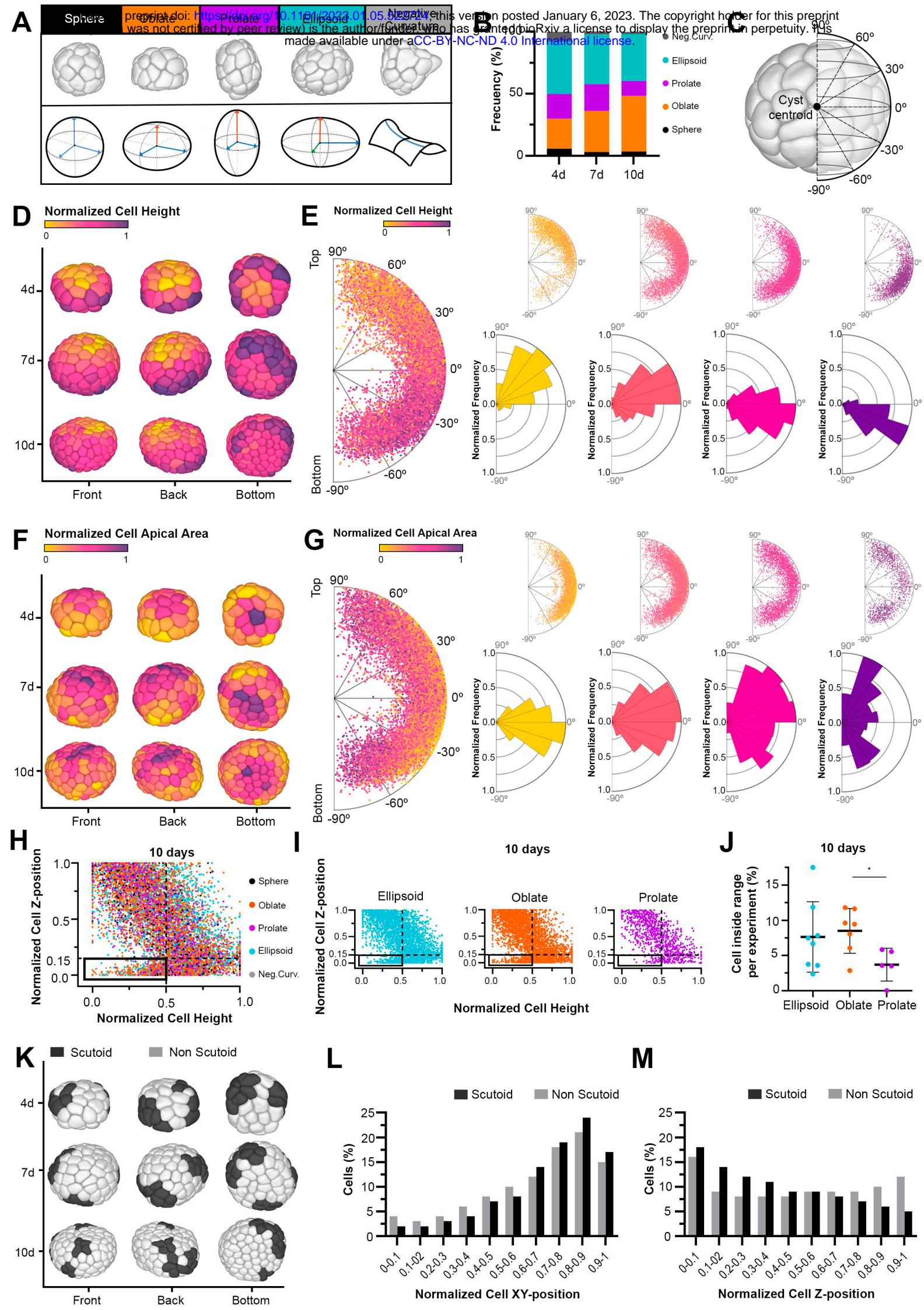


Figure 2. Realistic high-throughput 3D segmentation reveals different morphologies and cell morphology patterns in MDCK cysts.

A) Shape classification of single-layer, single-lumen cysts. (Top) The 3D rendering of representative segmented cysts. (Bottom) Schematic representation of the morphological cyst classification. **B)** Frequency distribution of the different cyst shapes at 4, 7 and 10 days. **C)** Schematic representation of the relation between 3D reconstructions of the cyst and the 2D plots. The drawing shows how the position and angle of the cell centroids is represented. **D-E)** Single-cell Cartography of “cell height” feature. **D)** Computer rendering of three representative segmented cysts: front, back and bottom views. The cell color scale symbolizes the value of the cell height (normalized by cyst). **E)** (Left) Polar scatter showing the normalized distance and angle of cell centroids regarding the cyst centroid (considering 0 degrees any vector contained in the XY plane of the cyst centroid; positive angles correlate with cell centroids placed over the cyst centroid, and cells located below the cyst centroid XY plane are represented with negative angles, as indicated in panel **C**), with a heatmap coloring the normalized value of cell height. 20391 cells from 353 cysts were represented. (Right) Cell sorting by the normalized cell height values, from left to right: 0-0.25, 0.25-0.5, 0.5-0.75. 0.75-1. On top, a scatter polar diagram showing the angle and distance of cell centroids regarding the cyst centroid, with a heat map coloring based on the quantification of the normalized cell height; on bottom, a polar histogram accounting for the frequency of cell positions. **F-G)** Single-cell Cartography of the “apical area” feature. The same color code and plotting properties as panels **D-E**. **H-J)** The cyst morphology affects the cartography of cell height at 10 days. **H)** Cell Z-position versus cell height (normalized per cyst). The color of the dots is determined by the kind of cyst the cells belong to. 8528 cells from 116 cysts of 10 days were analyzed. **I)** Scatter plots showing the comparison in cell Z-position versus cell height in 10-day ellipsoid, oblate and prolate cysts. **J)** Cell proportion in the highlighted quadrant of normalized cell Z-position < 0.15 and normalized cell height <0.5 in 10-day ellipsoid, oblate and prolate cysts. The percentage is computed with respect to the total number of cells for each morphology and experiment. Data are represented as mean ± SD. * p value <0.05 (t-Student test). **K-M)** Single-cell Cartography of scutoids. **K)** Computer rendering of three

representative segmented cysts in 3 perspectives, showing scutoids in black and non-scutoidal cells in light gray. **L)** Histogram representing the cell proportion at different intervals of the normalized XY-position distance with respect to the XY coordinate of the cyst centroid. Percentage of all cells from the 353 segmented cysts. **M)** Histogram representing the cell proportion of scutoidal and non-scutoidal cells at different intervals of the Z-position distance with respect to the Z coordinate of the cyst centroid. Percentage of all cells from the 353 segmented cysts.

3D RESU-NET

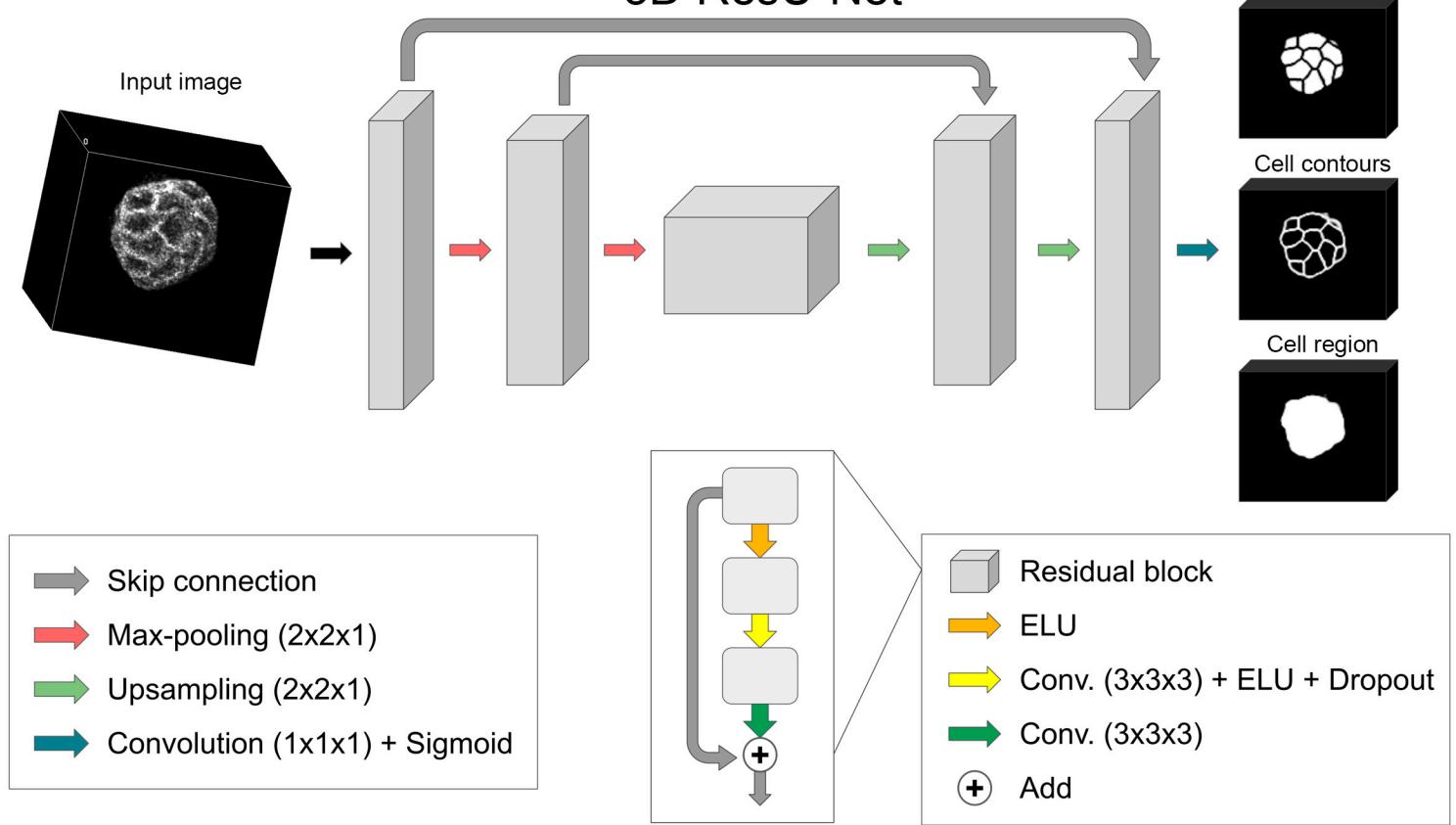


Figure S1. Architecture of the 3D ResU-Net segmentation network.

Our proposed 3D ResU-Net is a U-Net-like network with three depth levels and residual blocks. It receives 3D single-channel images as input and produces three different outputs representing the probabilities of the individual cell masks, contours and whole cell region. Convolutional filters are applied in 3D, while the down-sampling and up-sampling operations are only performed in 2D due to the anisotropy of image data.

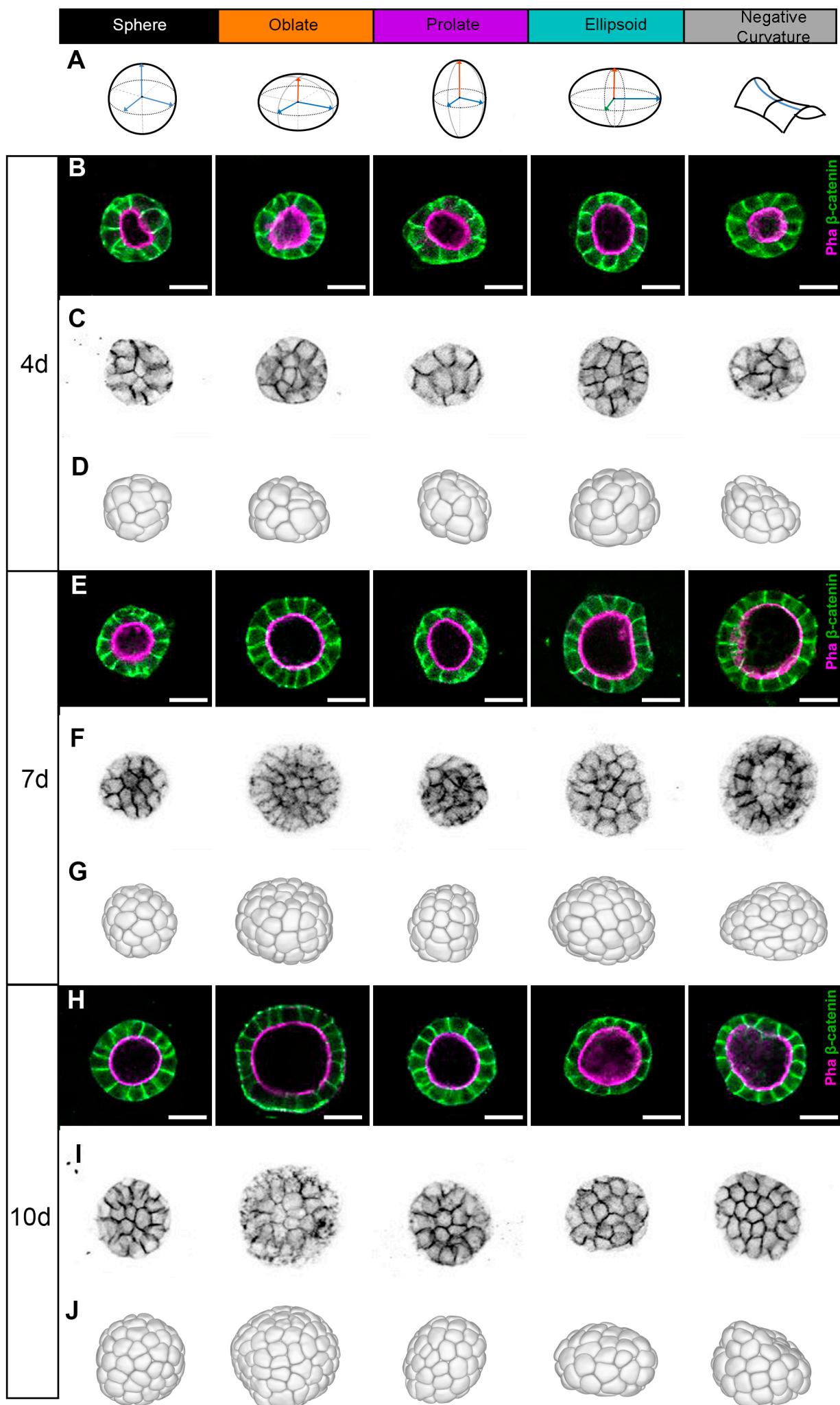
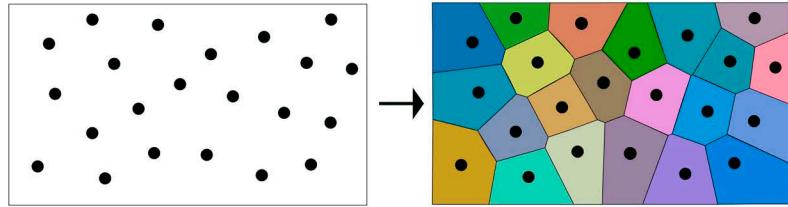


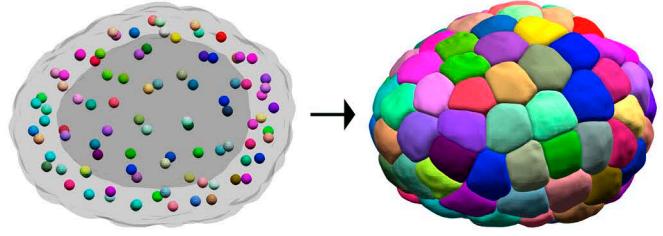
Figure S2. MDCK cysts adopt different shapes in 3D cell cultures.

A) Schematic representation of the morphological cyst classification. **B, E, and H)** Middle sections of top-to-bottom confocal microscopy Z-stack images of representative cysts belonging to each morphological category at 4, 7 and 10 days. Cell contours were stained with Alexa Fluor 647 phalloidin (magenta) and anti- β -catenin antibody (green) (**STAR Methods**). Scale bars, 20 μm . **C, F and I)** Half projections of representative cysts belonging to each morphological category at 4, 7 and 10 days (cell contours stained by Alexa Fluor 647 phalloidin and anti- β -catenin antibody (**STAR Methods**). **D, G and J)** Computer rendering (front or back view) of the segmented cysts belonging to each morphological category at 4, 7 and 10 days.

A



B



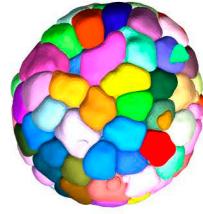
C



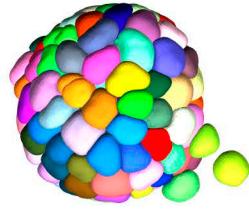
ground-truth



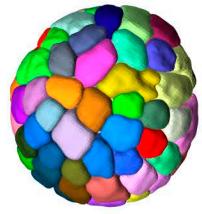
CartoCell



PlantSeg



StarDist



Cellpose

Figure S3. Voronoi tessellations and comparison of custom three-dimensional Voronoi based model against state-of-the-art methods

A) Voronoi tessellation in two dimensions occupying the space of a rectangle using randomly distributed seeds. **B)** Three-dimensional Voronoi tessellation of a cyst filling the space determined by the cell tissue using the centroids of each cell as seeds. **C)** Comparison of a manually labeled cyst (ground-truth) against our proposed method (CartoCell) and the three state-of-the-art methods chosen for comparison: PlantSeg, StarDist and Cellpose.

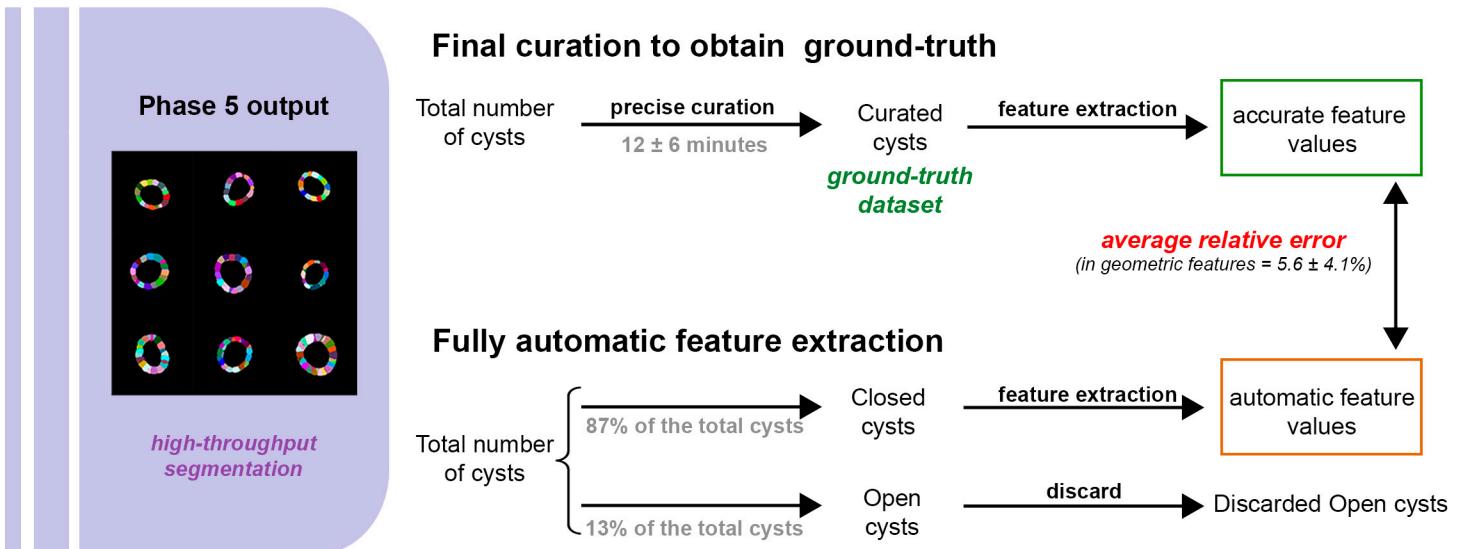
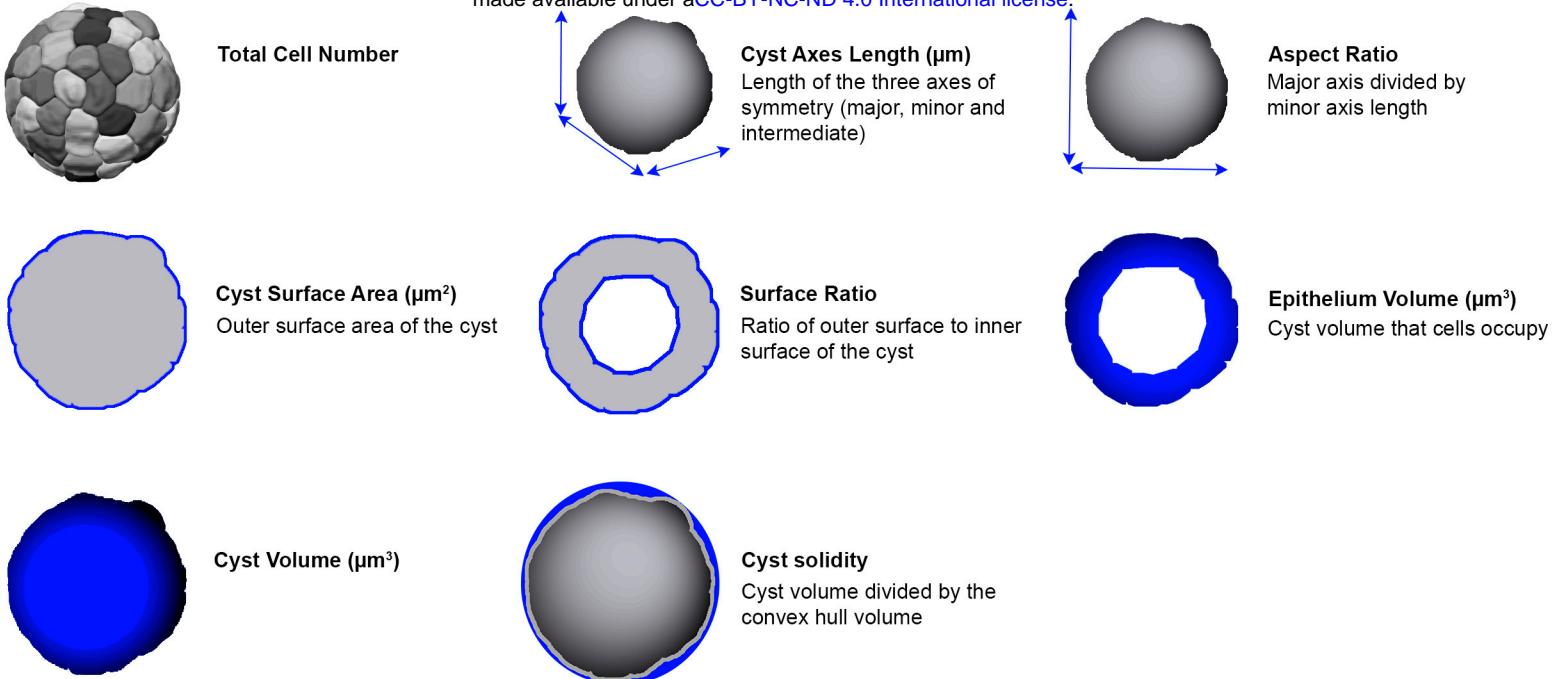
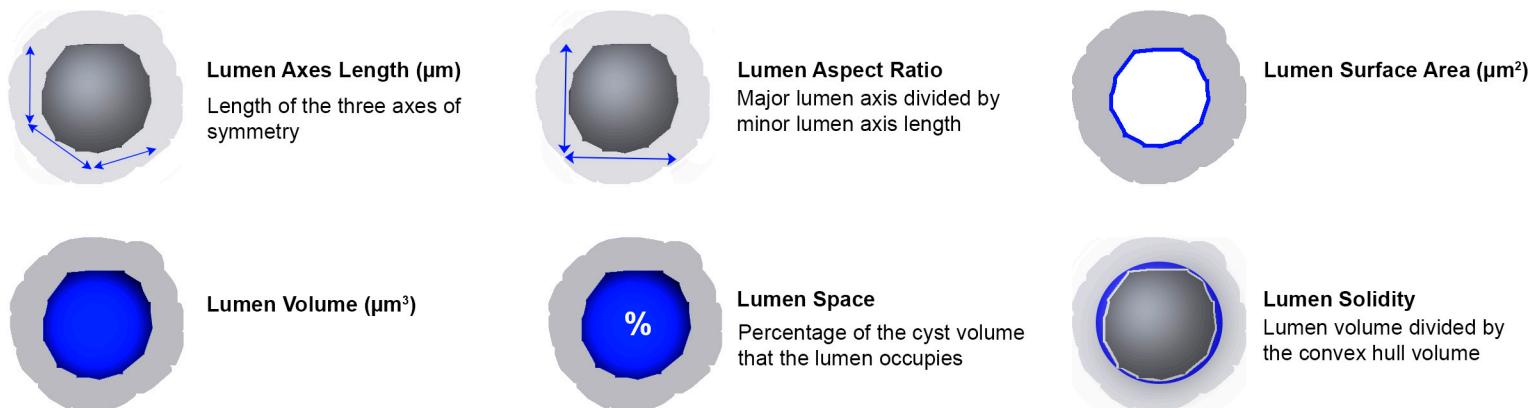


Figure S4. Appendix to Fig.1 showing two possible paths for feature extraction. After the Phase 5, a ground-truth dataset can be obtained by final manual curation of the results (12.0 ± 6 minutes per cyst) to extract the accurate feature values. On the other hand, there is an alternative option, opting for the automatic feature extraction of geometric data assuming an average relative error: Based on the data of our study, 46 of 353 cysts (13%) presented under-segmentation defects which induced open cysts and impaired the feature extraction. The extraction of the features of the remaining cyst (87%, closed cysts) provided results with an average error of $5.6 \pm 4.1\%$ in geometric characteristics compared with the accurate feature values (**Table S4** and **STAR Methods**).

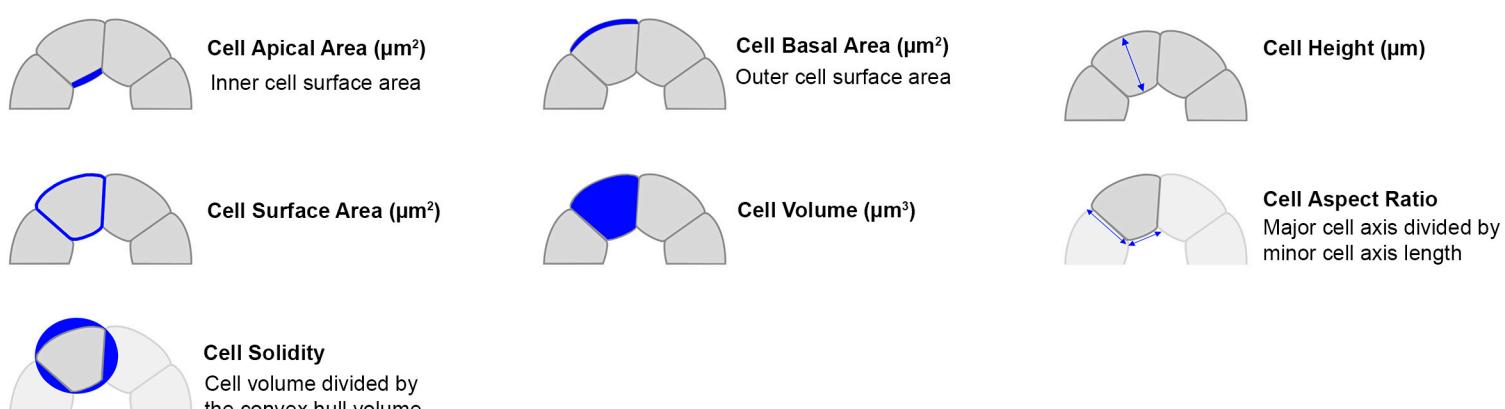
Cyst geometric features



Lumen geometric features



Cell geometric features



Cell packing features

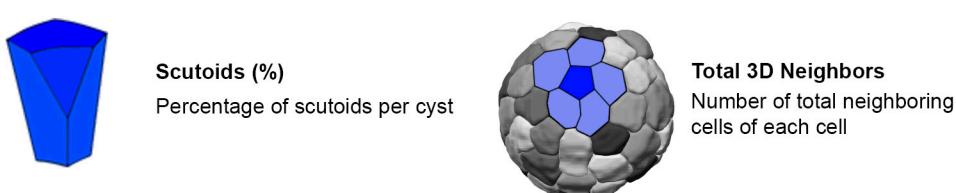
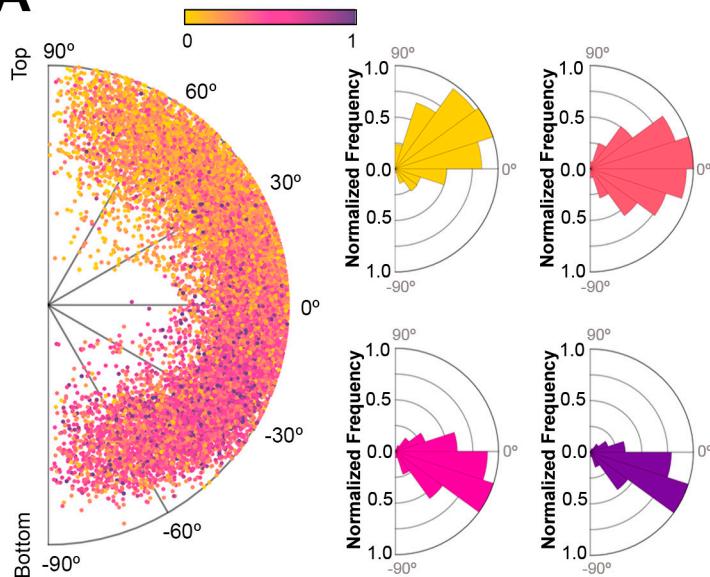


Figure S5. Qualitative definition of the geometric and packing features describing the cyst architecture.

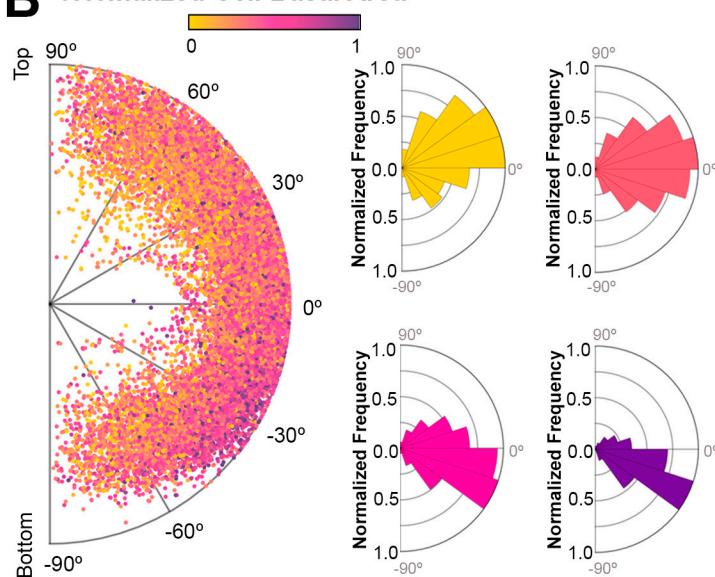
A

Normalized Cell Volume



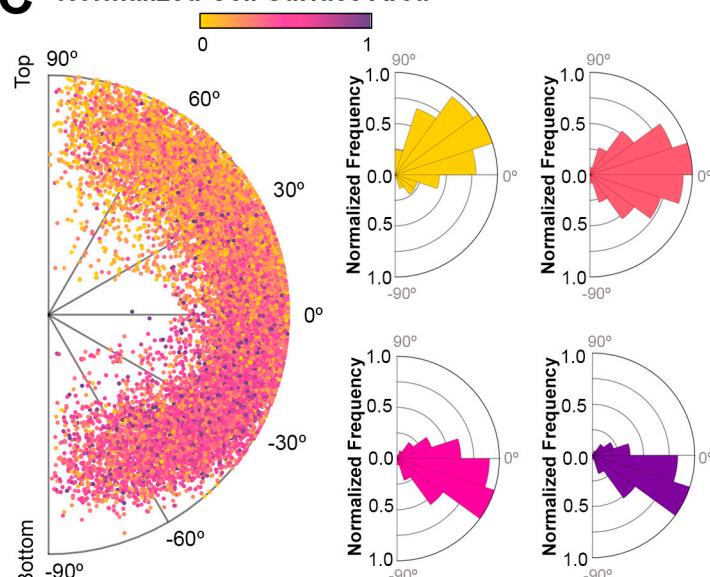
B

Normalized Cell Basal Area



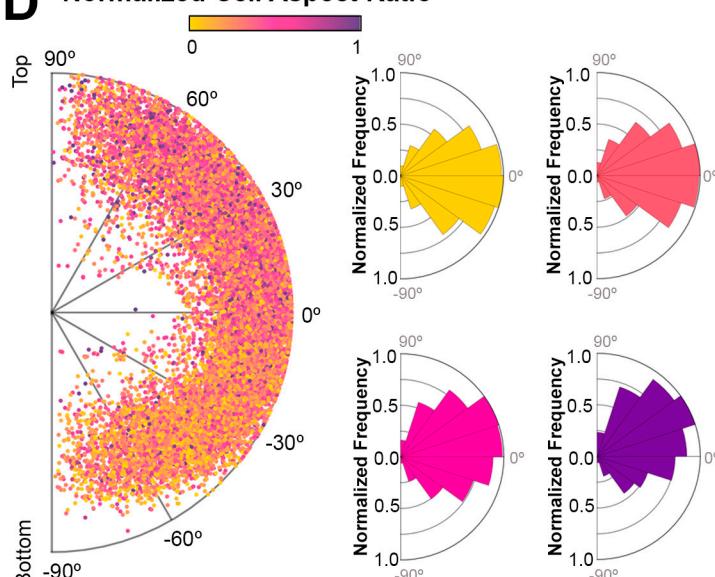
C

Normalized Cell Surface Area



D

Normalized Cell Aspect Ratio



E

Normalized Cell Solidity

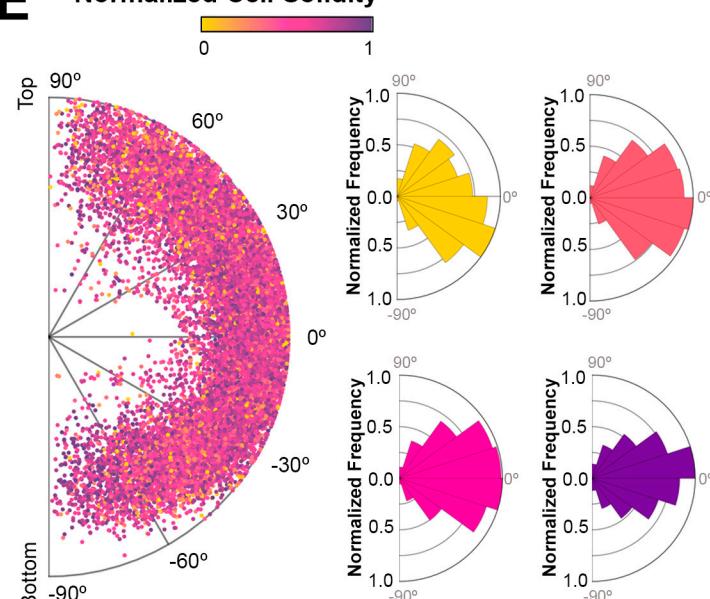


Figure S6. Examples of spatial patterns in the distribution of geometric feature values. Same representations as **Fig. 2E, G** to map the features: cell volume (**A**), cell basal area (**B**), cell surface area (**C**), cell aspect ratio (**D**) and cell solidity (**E**).

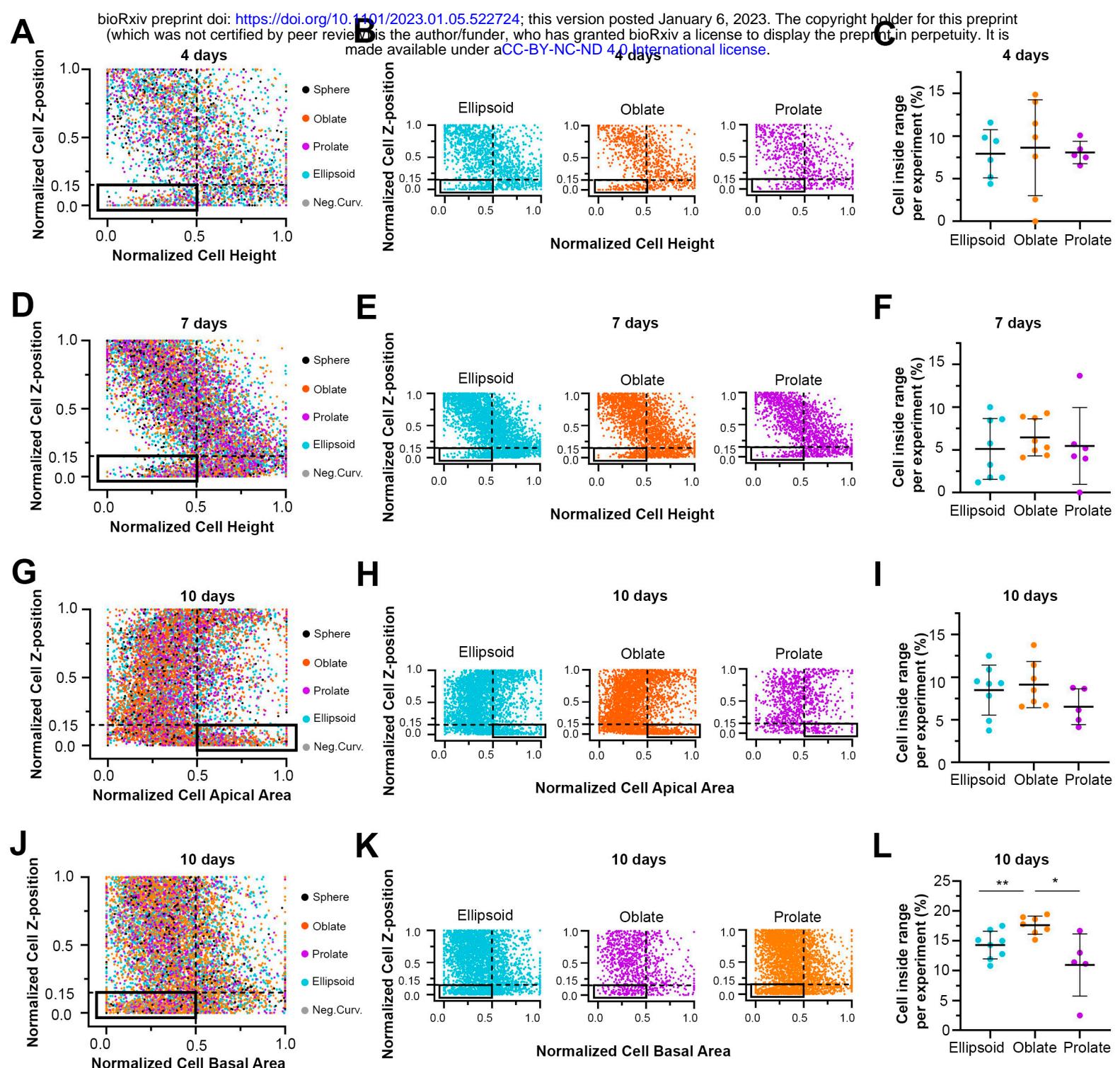


Figure S7. The cyst morphology affects the cartography of cell geometric features. Same representations as **Fig. 2H-J** to explore the cell Z-position versus cell height at 4-day (**A-C**) and 7-day cysts (**D-F**), cell apical area at 10-day cysts (**G-I**) and cell basal area at 10-day cysts (**J-L**).

TABLES

Table S1. Extracted features from 353 curated cysts (104 cyst at 4 days, 133 cysts at 7 days, 116 cysts at 10 days). Tab1 (GlobalFeatures) shows the mean of geometric and packing features describing the whole cyst. Tab 2 (CellParameters) shows the mean and standard deviation of cellular geometric and packing features.

Table S2. Hyperparameter search space for our proposed 3D ResU-Net.

Table S3. Comparison of state-of-the-art segmentation methods against our proposed pipeline (CartoCell) using the evaluation metrics described in **STAR Methods**.

Table S4. Relative error between features extracted using automatically segmented cysts and manually curated cysts (**STAR Methods**).

Table S5. Tab 1 (Cyst morphology properties) shows the statistical differences in cell proportions after comparing different cyst morphologies. Tab 2 (Stats for position of scutoids) shows the statistical comparison between the spatial position of scutoids and non scutoids cells.