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

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

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

[2]Biomedical Network Research Centre on Neurodegenerative Diseases (CIBERNED), 28029 Madrid, Spain

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

[4]Donostia International Physics Center (DIPC). 20018 San Sebastian, Spain

[5]Ikerbasque, Basque Foundation for Science. 48009 Bilbao, Spain

[6]Laboratory for Molecular Cell Biology, University College London. London, UK

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

[8] MRC Laboratory of Molecular Biology, Cambridge Biomedical Campus, Francis Crick Avenue, Cambridge, UK

[9]Department of Physiology, Development and Neuroscience, University of Cambridge, UK

ABSTRACT

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

INTRODUCTION

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

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

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

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

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

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

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

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

RESULTS

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

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

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

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

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

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

In Phase 4, a new 3D ResU-Net model ("model M2", from now on) was trained on the *"training low-resolution dataset"*, composed of the newly produced large dataset 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 performance of deep learning models is highly dependent on the number of training samples.

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

Optimization of the method

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

Performance evaluation and comparison with other methods

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

Management of results: final curation before biological analysis

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

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

Epithelial cysts adopt different shapes in 3D culture

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

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

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

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

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

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

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

The emergence of cell packing patterns in the cysts.

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

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

DISCUSSION

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

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

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

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

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

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

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

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

STAR METHODS

MDCK cyst cell culture

Type II MDCK (Madin-Darby canine kidney) cells were maintained in minimum essential medium (MEM) containing GlutaMAX (Gibco) and supplemented with 10% fetal bovine serum (FBS), 100 U/ml penicillin and 100 µg/ml streptomycin, in a 5% CO_2 humidified incubator at 37ºC. For cyst formation, MDCK cells (2500 cells/well) were suspended in complete medium containing 2% Matrigel (Corning, Life Sciences). Cell suspension was plated in a 4-well culture slide (Corning, Life Science) on a thin layer coating of 100% Matrigel. The plates were kept at 37ºC in a humidified atmosphere of 5% CO^2 for 4, 7 or 10 days and the medium was changed every 2 days.

Immunostaining and confocal imaging

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

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

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

Custom DNN architecture: 3D ResU-Net

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

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

Network optimization

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

Images preprocessing

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

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

$$[newSize_X, newSize_Y, newSize_Z] \rightarrow \left[Size_X \cdot \left(\frac{lowRes_{XY}}{highRes_{XY}}\right), Size_Y \cdot \left(\frac{lowRes_{XY}}{highRes_{XY}}\right), Size_Z \cdot \left(\frac{lowRes_Z}{highRes_Z}\right)\right]$$

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

Inference and postprocessing

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

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

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

Comparison with the state of the art

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

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

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

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

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

Training and test datasets acquisition

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

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

Proofreading of segmented cysts

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

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

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

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

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

Features extraction and cyst shape classification

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

Quantification of biological features errors in non-curated segmented cysts

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

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

Single-cell Cartography representation

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

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

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

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

Evaluation metrics

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

Precision, defined as

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

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

Recall, defined as

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

where FN is the number of false negatives.

Accuracy, defined as

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

F1 or F-score, defined as

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

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

$$panoptic = \frac{\Sigma_{(p,g)\epsilon TP} IoU(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}$$

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

Statistical analysis

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

CODE AVAILABILITY

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

BIBLIOGRAPHY

Alfonso-Pérez, T., Baonza, G., Herranz, G., & Martín-Belmonte, F. (2022). Deciphering the interplay between autophagy and polarity in epithelial tubulogenesis. *Seminars in Cell & Developmental Biology*, *131*, 160–172. https://doi.org/https://doi.org/10.1016/j.semcdb.2022.05.015

Bailoni, A., Pape, C., Hütsch, N., Wolf, S., Beier, T., Kreshuk, A., & Hamprecht, F. A. (2019). *GASP, a generalized framework for agglomerative clustering of signed graphs and its application to Instance Segmentation*. https://doi.org/10.48550/arxiv.1906.11713

Beck, L. E., Lee, J., Coté, C., Dunagin, M. C., Lukonin, I., Salla, N., Chang, M. K., Hughes, A. J., Mornin, J. D., Gartner, Z. J., Liberali, P., & Raj, A. (2022). Systematically quantifying morphological features reveals constraints on organoid phenotypes. *Cell Systems*, *13*(7), 547-560.e3. https://doi.org/10.1016/j.cels.2022.05.008

Bryant, D. M., & Mostov, K. E. (2008). From cells to organs: building polarized tissue. *Nature Reviews Molecular Cell Biology*, *9*(11), 887–901. https://doi.org/10.1038/nrm2523

Carleton, A. E., Duncan, M. C., & Taniguchi, K. (2022). Human epiblast lumenogenesis: From a cell aggregate to a lumenal cyst. *Seminars in Cell & Developmental Biology*, *131*, 117–123. https://doi.org/10.1016/j.semcdb.2022.05.009

Cerruti, B., Puliafito, A., Shewan, A. M., Yu, W., Combes, A. N., Little, M. H., Chianale, F., Primo, L., Serini, G., Mostov, K. E., Celani, A., & Gamba, A. (2013). Polarity, cell division, and out-of-equilibrium dynamics control the growth of epithelial structures. *Journal of Cell Biology*, *203*(2), 359–372. https://doi.org/10.1083/jcb.201305044

Ciresan, D., Giusti, A., Gambardella, L., & Schmidhuber, J. (2012). Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images. In F. Pereira, C. J. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 25). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2012/file/459a4ddcb586f24efd9395aa7662bc7c-Paper.pdf

Cousty, J., Bertrand, G., Najman, L., & Couprie, M. (2009). Watershed Cuts: Minimum Spanning Forests and the Drop of Water Principle. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(8), 1362–1374. https://doi.org/10.1109/TPAMI.2008.173

Dahl-Jensen, S., & Grapin-Botton, A. (2017). The physics of organoids: a biophysical approach to understanding organogenesis. *Development*, *144*(6), 946–951. https://doi.org/10.1242/dev.143693

Davidson, L. A. (2012). Epithelial machines that shape the embryo. *Trends in Cell Biology*, *22*(2), 82–87. https://doi.org/10.1016/j.tcb.2011.10.005

de Medeiros, G., Ortiz, R., Strnad, P., Boni, A., Moos, F., Repina, N., Challet Meylan, L., Maurer, F., & Liberali, P. (2022). Multiscale light-sheet organoid imaging framework. *Nature Communications*, *13*(1), 4864. https://doi.org/10.1038/s41467-022-32465-z

Drost, J., & Clevers, H. (2018). Organoids in cancer research. *Nature Reviews Cancer*, *18*(7), 407–418. https://doi.org/10.1038/s41568-018-0007-6

Dukes, J. D., Whitley, P., & Chalmers, A. D. (2011). The MDCK variety pack: Choosing the right strain. In *BMC Cell Biology* (Vol. 12). https://doi.org/10.1186/1471-2121-12-43

Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., Dovzhenko, A., Tietz, O., Dal Bosco, C., Walsh, S., Saltukoglu, D., Tay, T. L., Prinz, M., Palme, K., Simons, M., … Ronneberger, O. (2018). U-Net: deep learning for cell counting, detection, and morphometry. *Nature Methods 2018 16:1*, *16*(1), 67–70. https://doi.org/10.1038/s41592-018-0261-2

Fessenden, T. B., Beckham, Y., Perez-Neut, M., Ramirez-San Juan, G., Chourasia, A. H., Macleod, K. F., Oakes, P. W., & Gardel, M. L. (2018). Dia1-dependent adhesions are required by epithelial tissues to initiate invasion. *Journal of Cell Biology*, *217*(4), 1485–1502. https://doi.org/10.1083/jcb.201703145

Franco-Barranco, D., Muñoz-Barrutia, A., & Arganda-Carreras, I. (2022). Stable Deep Neural Network Architectures for Mitochondria Segmentation on Electron Microscopy Volumes. *Neuroinformatics*, *20*(2), 437–450. https://doi.org/10.1007/s12021-021-09556-1

Gómez, H. F., Dumond, M. S., Hodel, L., Vetter, R., & Iber, D. (2021). 3D cell neighbour dynamics in growing pseudostratified epithelia. *ELife*, *10*. https://doi.org/10.7554/eLife.68135

Gómez-Gálvez, P., Anbari, S., Escudero, L. M., & Buceta, J. (2021). Mechanics and self-organization in tissue development. *Seminars in Cell & Developmental Biology*, *120*, 147–159. https://doi.org/10.1016/j.semcdb.2021.07.003

Gómez-Gálvez, P., Vicente-Munuera, P., Anbari, S., Buceta, J., & Escudero, L. M. (2021). The complex three-dimensional organization of epithelial tissues. *Development*, *148*(1), dev195669. https://doi.org/10.1242/dev.195669

Gómez-Gálvez, P., Vicente-Munuera, P., Anbari, S., Tagua, A., Gordillo-Vázquez, C., Andrés-San Román, J. A., Franco-Barranco, D., Palacios, A. M., Velasco, A., Capitán-Agudo, C., Grima, C., Annese, V., Arganda-Carreras, I., Robles, R., Márquez, A., Buceta, J., & Escudero, L. M. (2022). A quantitative biophysical principle to explain the 3D cellular connectivity in curved epithelia. *Cell Systems*, *13*(8), 631-643.e8. https://doi.org/10.1016/J.CELS.2022.06.003

Gómez-Gálvez, P., Vicente-Munuera, P., Tagua, A., Forja, C., Castro, A. M., Letrán, M., Valencia-Expósito, A., Grima, C., Bermúdez-Gallardo, M., Serrano-Pérez-Higueras, Ó., Cavodeassi, F., Sotillos, S., Martín-Bermudo, M. D., Márquez, A.,

Buceta, J., & Escudero, L. M. (2018). Scutoids are a geometrical solution to three-dimensional packing of epithelia. *Nature Communications*, *9*(1), 2960. https://doi.org/10.1038/s41467-018-05376-1

Guo, Q., Xia, B., Moshiach, S., Xu, C., Jiang, Y., Chen, Y., Sun, Y., Lahti, J. M., & Zhang, X. A. (2008). The microenvironmental determinants for kidney epithelial cyst morphogenesis. *European Journal of Cell Biology*, *87*(4), 251–266. https://doi.org/10.1016/j.ejcb.2007.11.004

He, Z., Maynard, A., Jain, A., Gerber, T., Petri, R., Lin, H.-C., Santel, M., Ly, K., Dupré, J.-S., Sidow, L., Sanchis Calleja, F., Jansen, S. M. J., Riesenberg, S., Camp, J. G., & Treutlein, B. (2022). Lineage recording in human cerebral organoids. *Nature Methods*, *19*(1), 90–99. https://doi.org/10.1038/s41592-021-01344-8

Herranz, G., & Martín-Belmonte, F. (2022). Cadherin-mediated adhesion takes control. *The EMBO Journal*. https://doi.org/10.15252/embj.2022112662

Honda, H. (1978). Description of cellular patterns by Dirichlet domains: The two-dimensional case. *Journal of Theoretical Biology*, *72*(3), 523–543. https://doi.org/10.1016/0022-5193(78)90315-6

Huch, M., Knoblich, J. A., Lutolf, M. P., & Martinez-Arias, A. (2017). The hope and the hype of organoid research. *Development*, *144*(6), 938–941. https://doi.org/10.1242/dev.150201

Imai, M., Furusawa, K., Mizutani, T., Kawabata, K., & Haga, H. (2015). Three-dimensional morphogenesis of MDCK cells induced by cellular contractile forces on a viscous substrate. *Scientific Reports*, *5*(1), 14208. https://doi.org/10.1038/srep14208

Ioannou, F., Dawi, M. A., Tetley, R. J., Mao, Y., & Muñoz, J. J. (2020). Development of a New 3D Hybrid Model for Epithelia Morphogenesis. *Frontiers in Bioengineering and Biotechnology*, *8*. https://doi.org/10.3389/fbioe.2020.00405

Kappes, J. H., Speth, M., Andres, B., Reinelt, G., & Schn, C. (2011). Globally Optimal Image Partitioning by Multicuts. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 6819 LNCS* (pp. 31–44). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23094-3_3

Kirillov, A., He, K., Girshick, R., Rother, C., & Dollar, P. (2018). Panoptic Segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 9396–9405. https://doi.org/10.48550/arxiv.1801.00868

Laine, R. F., Arganda-Carreras, I., Henriques, R., & Jacquemet, G. (2021). Avoiding a replication crisis in deep-learning-based bioimage analysis. *Nature Methods*, *18*(10), 1136–1144. https://doi.org/10.1038/s41592-021-01284-3

Laurent, J., Blin, G., Chatelain, F., Vanneaux, V., Fuchs, A., Larghero, J., & Théry, M. (2017). Convergence of microengineering and cellular self-organization

towards functional tissue manufacturing. *Nature Biomedical Engineering*, *1*(12), 939–956. https://doi.org/10.1038/s41551-017-0166-x

Lecuit, T., & Lenne, P.-F. (2007). Cell surface mechanics and the control of cell shape, tissue patterns and morphogenesis [Journal Article]. *Nature Reviews Molecular Cell Biology*, *8*(8), 633–644. https://doi.org/10.1038/nrm2222

Lin, Z., Wei, D., Petkova, M. D., Wu, Y., Ahmed, Z., K, K. S., Zou, S., Wendt, N., Boulanger-Weill, J., Wang, X., Dhanyasi, N., Arganda-Carreras, I., Engert, F., Lichtman, J., & Pfister, H. (2021). *NucMM Dataset: 3D Neuronal Nuclei Instance Segmentation at Sub-Cubic Millimeter Scale* (pp. 164–174). https://doi.org/10.1007/978-3-030-87193-2_16

Lo, Y.-H., Karlsson, K., & Kuo, C. J. (2020). Applications of organoids for cancer biology and precision medicine. *Nature Cancer*, *1*(8), 761–773. https://doi.org/10.1038/s43018-020-0102-y

Lou, Y., Rupprecht, J.-F., Hiraiwa, T., & Saunders, T. E. (2022). Curvature-induced cell rearrangements in biological tissues. *BioRxiv*, 2022.05.18.492428. https://doi.org/10.1101/2022.05.18.492428

Machado, S., Mercier, V., & Chiaruttini, N. (2019). LimeSeg: a coarse-grained lipid membrane simulation for 3D image segmentation. *BMC Bioinformatics*, *20*(1), 2. https://doi.org/10.1186/s12859-018-2471-0

Martín-Belmonte, F., Yu, W., Rodríguez-Fraticelli, A. E., Ewald, A., Werb, Z., Alonso, M. A., & Mostov, K. (2008). Cell-Polarity Dynamics Controls the Mechanism of Lumen Formation in Epithelial Morphogenesis. *Current Biology*, *18*(7), 507–513. https://doi.org/10.1016/j.cub.2008.02.076

Meijering, E. (2020). A bird's-eye view of deep learning in bioimage analysis. *Computational and Structural Biotechnology Journal*, *18*, 2312–2325. https://doi.org/10.1016/j.csbj.2020.08.003

Messal, H. A., Alt, S., Ferreira, R. M. M., Gribben, C., Wang, V. M.-Y., Cotoi, C. G., Salbreux, G., & Behrens, A. (2019). Tissue curvature and apicobasal mechanical tension imbalance instruct cancer morphogenesis. *Nature*, *566*(7742), 126–130. https://doi.org/10.1038/s41586-019-0891-2

Metzger, J. J., Simunovic, M., & Brivanlou, A. H. (2018). Synthetic embryology: controlling geometry to model early mammalian development. *Current Opinion in Genetics & Development*, *52*, 86–91. https://doi.org/10.1016/j.gde.2018.06.006

Meyer, F. (1994). Topographic distance and watershed lines. *Signal Processing*, *38*(1), 113–125. https://doi.org/10.1016/0165-1684(94)90060-4

Moen, E., Bannon, D., Kudo, T., Graf, W., Covert, M., & van Valen, D. (2019). Deep learning for cellular image analysis. *Nature Methods*, *16*(12), 1233–1246. https://doi.org/10.1038/s41592-019-0403-1

O'Brien, L. E., Zegers, M. M. P., & Mostov, K. E. (2002). Building epithelial architecture: Insights from three-dimensional culture models. In *Nature Reviews Molecular Cell Biology* (Vol. 3, Issue 7, pp. 531–537). https://doi.org/10.1038/nrm859

Okuda, S., Miura, T., Inoue, Y., Adachi, T., & Eiraku, M. (2018). Combining Turing and 3D vertex models reproduces autonomous multicellular morphogenesis with undulation, tubulation, and branching. *Scientific Reports 2018 8:1*, *8*(1), 1–15. https://doi.org/10.1038/s41598-018-20678-6

Okuda, S., & Sato, K. (2022). Polarized interfacial tension induces collective migration of cells, as a cluster, in a 3D tissue. *Biophysical Journal*, *121*(10), 1856–1867. https://doi.org/10.1016/J.BPJ.2022.04.018

Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(1), 62–66. https://doi.org/10.1109/TSMC.1979.4310076

Perez, L., & Wang, J. (2017). *The Effectiveness of Data Augmentation in Image Classification using Deep Learning*. https://doi.org/10.48550/arxiv.1712.04621

Prabhakara, C., Iyer, K. S., Rao, M., Saunders, T. E., & Mayor, S. (2022). Quantitative analysis of three-dimensional cell organisation and concentration profiles within curved epithelial tissues. *BioRxiv*, 2022.05.16.492131. https://doi.org/10.1101/2022.05.16.492131

Razzak, M. I., Naz, S., & Zaib, A. (2018). *Deep Learning for Medical Image Processing: Overview, Challenges and the Future* (pp. 323–350). https://doi.org/10.1007/978-3-319-65981-7_12

Rupprecht, J.-F., Ong, K. H., Yin, J., Huang, A., Dinh, H.-H.-Q., Singh, A. P., Zhang, S., Yu, W., & Saunders, T. E. (2017). Geometric constraints alter cell arrangements within curved epithelial tissues. *Molecular Biology of the Cell*, *28*(25), 3582–3594. https://doi.org/10.1091/mbc.e17-01-0060

Sakurai, A., Matsuda, M., & Kiyokawa, E. (2012). Activated Ras Protein Accelerates Cell Cycle Progression to Perturb Madin-Darby Canine Kidney Cystogenesis. *Journal of Biological Chemistry*, *287*(38), 31703–31711. https://doi.org/10.1074/jbc.M112.377804

Sánchez-Gutiérrez, D., Tozluoglu, M., Barry, J. D., Pascual, A., Mao, Y., & Escudero, L. M. (2016). Fundamental physical cellular constraints drive self-organization of tissues. *The EMBO Journal*, *35*(1), 77–88. https://doi.org/10.15252/embj.201592374

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., & Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nature Methods*, *9*(7), 676–682. https://doi.org/10.1038/nmeth.2019

Schmeichel, K. L., & Bissell, M. J. (2003). Modeling tissue-specific signaling and organ function in three dimensions. *Journal of Cell Science*, *116*(12), 2377–2388. https://doi.org/10.1242/jcs.00503

Schmidt, U., Weigert, M., Broaddus, C., & Myers, G. (2018). *Cell Detection with Star-Convex Polygons* (pp. 265–273). https://doi.org/10.1007/978-3-030-00934-2_30

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1), 1–48. https://doi.org/10.1186/S40537-019-0197-0/FIGURES/33

Stringer, C., Wang, T., Michaelos, M., & Pachitariu, M. (2021). Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, *18*(1), 100–106. https://doi.org/10.1038/s41592-020-01018-x

Takasato, M., Er, P. X., Chiu, H. S., & Little, M. H. (2016). Generation of kidney organoids from human pluripotent stem cells. *Nature Protocols*, *11*(9), 1681–1692. https://doi.org/10.1038/nprot.2016.098

Treacy, N. J., Clerkin, S., Davis, J. L., Kennedy, C., Miller, A. F., Saiani, A., Wychowaniec, J. K., Brougham, D. F., & Crean, J. (2023). Growth and differentiation of human induced pluripotent stem cell (hiPSC)-derived kidney organoids using fully synthetic peptide hydrogels. *Bioactive Materials*, *21*, 142–156. https://doi.org/10.1016/j.bioactmat.2022.08.003

Vidal-Quadras, M., Holst, M. R., Francis, M. K., Larsson, E., Hachimi, M., Yau, W. L., Peränen, J., Martín-Belmonte, F., & Lundmark, R. (2017). Endocytic turnover of Rab8 controls cell polarization. *Journal of Cell Science*. https://doi.org/10.1242/jcs.195420

von Chamier, L., Laine, R. F., Jukkala, J., Spahn, C., Krentzel, D., Nehme, E., Lerche, M., Hernández-Pérez, S., Mattila, P. K., Karinou, E., Holden, S., Solak, A. C., Krull, A., Buchholz, T.-O., Jones, M. L., Royer, L. A., Leterrier, C., Shechtman, Y., Jug, F., … Henriques, R. (2021). Democratising deep learning for microscopy with ZeroCostDL4Mic. *Nature Communications*, *12*(1), 2276. https://doi.org/10.1038/s41467-021-22518-0

Voronoi, G. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire. Recherches sur les parallélloèdres primitifs. *Journal Für Die Reine Und Angewandte Mathematik*, *134*, 198–287. http://eudml.org/doc/149291

Wei, D., Lin, Z., Franco-Barranco, D., Wendt, N., Liu, X., Yin, W., Huang, X., Gupta, A., Jang, W.-D., Wang, X., Arganda-Carreras, I., Lichtman, J. W., & Pfister, H. (2020). *MitoEM Dataset: Large-Scale 3D Mitochondria Instance Segmentation from EM Images* (pp. 66–76). https://doi.org/10.1007/978-3-030-59722-1_7

Weigert, M., Schmidt, U., Haase, R., Sugawara, K., & Myers, G. (2020). Star-convex Polyhedra for 3D Object Detection and Segmentation in Microscopy. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 3655–3662. https://doi.org/10.1109/WACV45572.2020.9093435

Wells, E. K., Yarborough, O., Lifton, R. P., Cantley, L. G., & Caplan, M. J. (2013). Epithelial morphogenesis of MDCK cells in three-dimensional collagen culture is modulated by interleukin-8. *American Journal of Physiology-Cell Physiology*, *304*(10), C966–C975. https://doi.org/10.1152/ajpcell.00261.2012

Wolf, S., Pape, C., Bailoni, A., Rahaman, N., Kreshuk, A., Köthe, U., & Hamprecht, F. A. (2018). *The Mutex Watershed: Efficient, Parameter-Free Image Partitioning* (pp. 571–587). https://doi.org/10.1007/978-3-030-01225-0_34

Wolny, A., Cerrone, L., Vijayan, A., Tofanelli, R., Barro, A. V., Louveaux, M., Wenzl, C., Strauss, S., Wilson-Sánchez, D., Lymbouridou, R., Steigleder, S. S., Pape, C., Bailoni, A., Duran-Nebreda, S., Bassel, G. W., Lohmann, J. U., Tsiantis, M., Hamprecht, F. A., Schneitz, K., … Kreshuk, A. (2020). Accurate and versatile 3D segmentation of plant tissues at cellular resolution. *ELife*, *9*. https://doi.org/10.7554/eLife.57613

Yan, L., Tsujita, K., Fujita, Y., & Itoh, T. (2021). PTEN is required for the migration and invasion of Ras-transformed MDCK cells. *FEBS Letters*, *595*(9), 1303–1312. https://doi.org/10.1002/1873-3468.14053

Yan, Z., Yang, X., & Cheng, K.-T. T. (2018). *A Deep Model with Shape-Preserving Loss for Gland Instance Segmentation* (pp. 138–146). https://doi.org/10.1007/978-3-030-00934-2_16

Yonemura, S. (2014). Differential Sensitivity of Epithelial Cells to Extracellular Matrix in Polarity Establishment. *PLoS ONE*, *9*(11), e112922. https://doi.org/10.1371/journal.pone.0112922

Yu, W., Fang, X., Ewald, A., Wong, K., Hunt, C. A., Werb, Z., Matthay, M. A., & Mostov, K. (2007). Formation of Cysts by Alveolar Type II Cells in Three-dimensional Culture Reveals a Novel Mechanism for Epithelial Morphogenesis. *Molecular Biology of the Cell*, *18*(5), 1693–1700. https://doi.org/10.1091/mbc.e06-11-1052

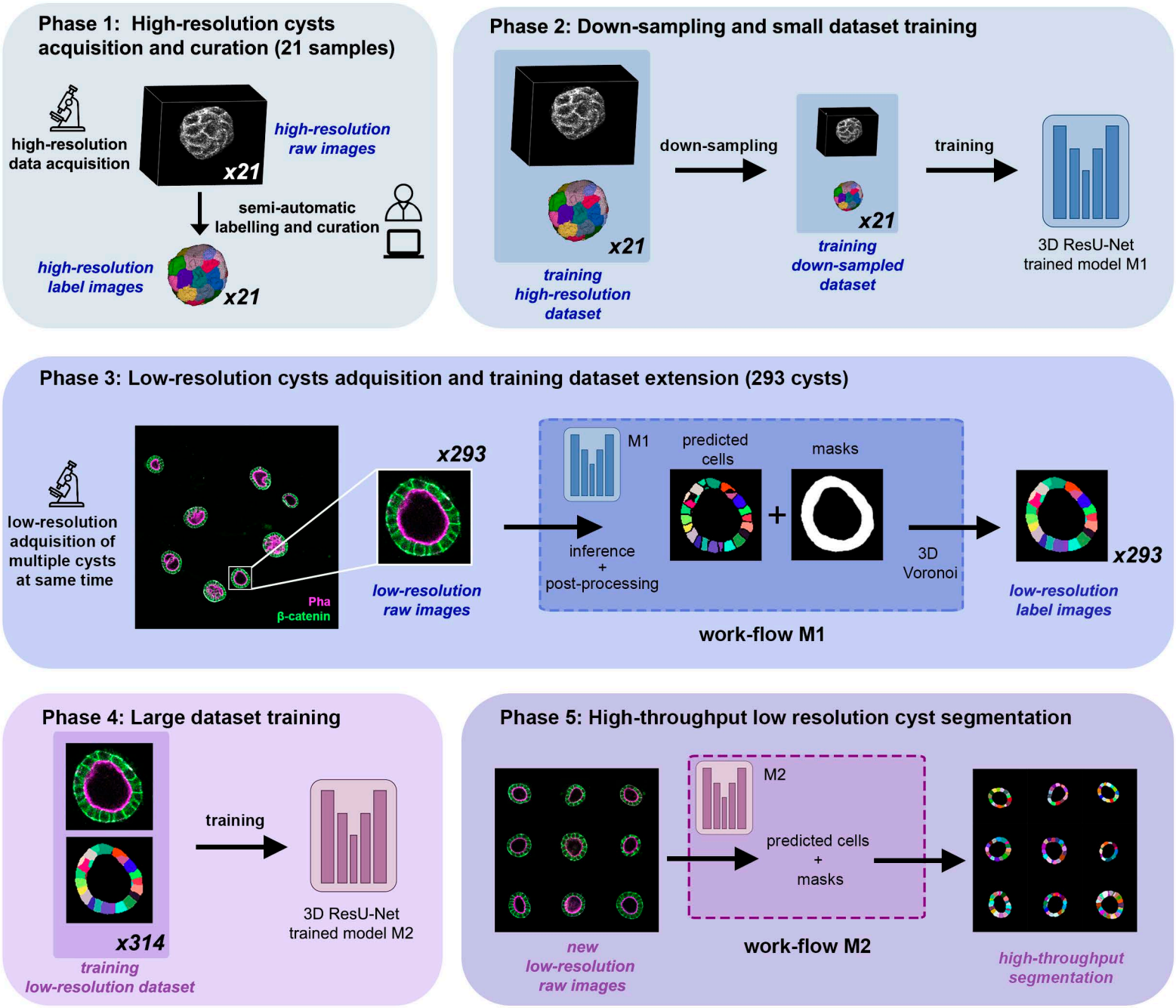

Phase 1: High-resolution cysts acquisition and curation (21 samples)
high-resolution data acquisition
x21
high-resolution raw images
semi-automatic labelling and curation
high-resolution label images
x21
Phase 2: Down-sampling and small dataset training
x21
training high-resolution dataset
down-sampling
x21
training down-sampled dataset
training
3D ResU-Net trained model M1
Phase 3: Low-resolution cysts adquisition and training dataset extension (293 cysts)
low-resolution adquisition of multiple cysts at same time
Pha
β-catenin
x293
low-resolution raw images
M1
inference + post-processing
predicted cells
masks
3D Voronoi
x293
low-resolution label images
work-flow M1
Phase 4: Large dataset training
x314
training low-resolution dataset
training
3D ResU-Net trained model M2
Phase 5: High-throughput low resolution cyst segmentation
new low-resolution raw images
M2
predicted cells + masks
work-flow M2
high-throughput 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".

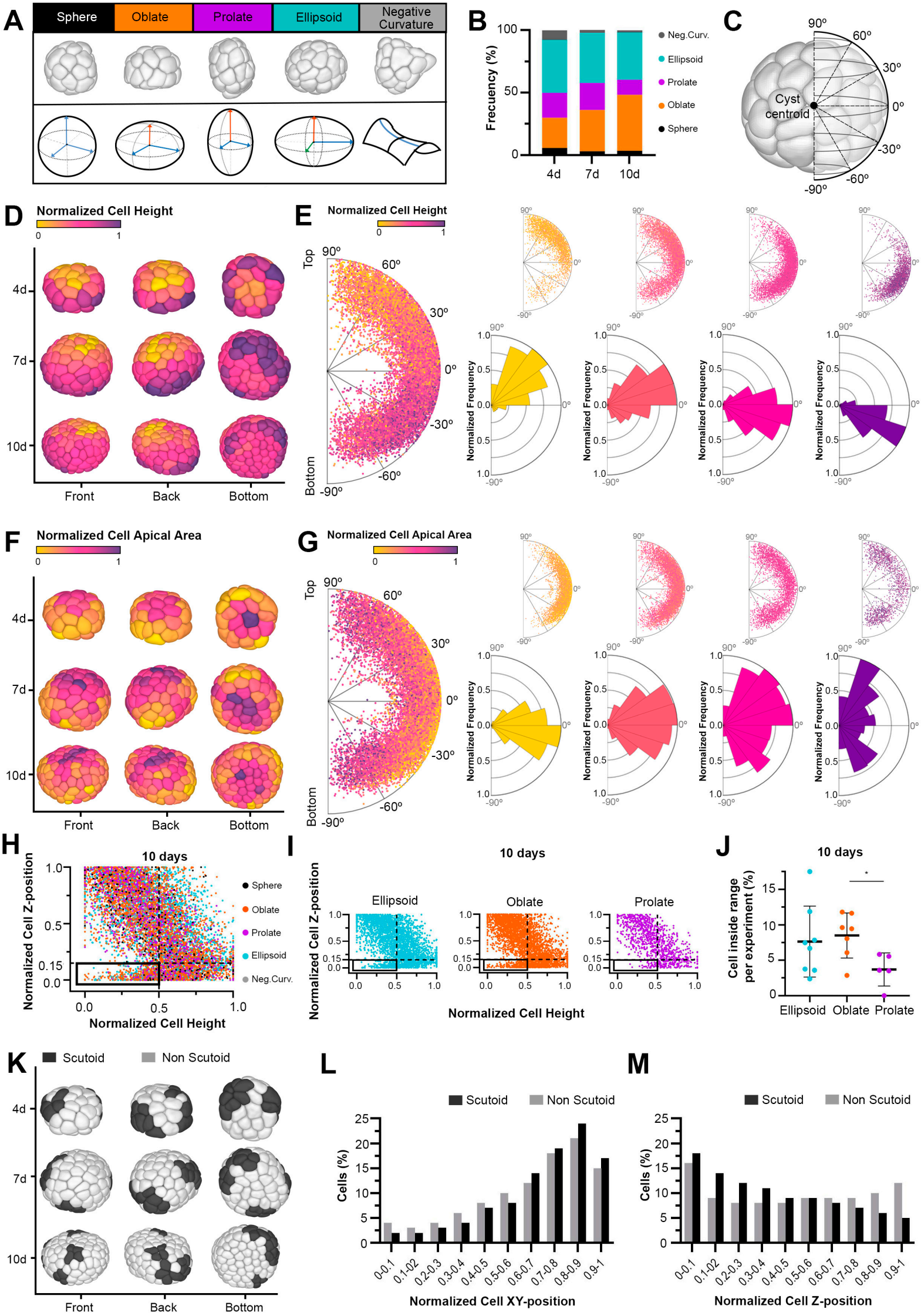

A
Sphere
Oblate
Prolate
Ellipsoid
Negative Curvature
B
Frecuency (%)
100
50
0
4d
7d
10d
Neg.Curv.
Ellipsoid
Prolate
Oblate
Sphere
C
90º
60º
30º
0º
-30º
-60º
-90º
Cyst centroid
D
Normalized Cell Height
0
1
4d
7d
10d
Front
Back
Bottom
E
Normalized Cell Height
Top
Bottom
Normalized Frequency
F
Normalized Cell Apical Area
G
Normalized Cell Apical Area
H
10 days
Normalized Cell Z-position
Normalized Cell Height
Sphere
Oblate
Prolate
Ellipsoid
Neg.Curv.
I
10 days
Ellipsoid
Oblate
Prolate
Normalized Cell Z-position
Normalized Cell Height
J
10 days
Cell inside range per experiment (%)
Ellipsoid
Oblate
Prolate
K
Scutoid
Non Scutoid
4d
7d
10d
Front
Back
Bottom
L
Scutoid
Non Scutoid
Cells (%)
0-0.1
0.1-02
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0.9-1
Normalized Cell XY-position
M
Scutoid
Non Scutoid
Cells (%)
Normalized Cell Z-position

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.

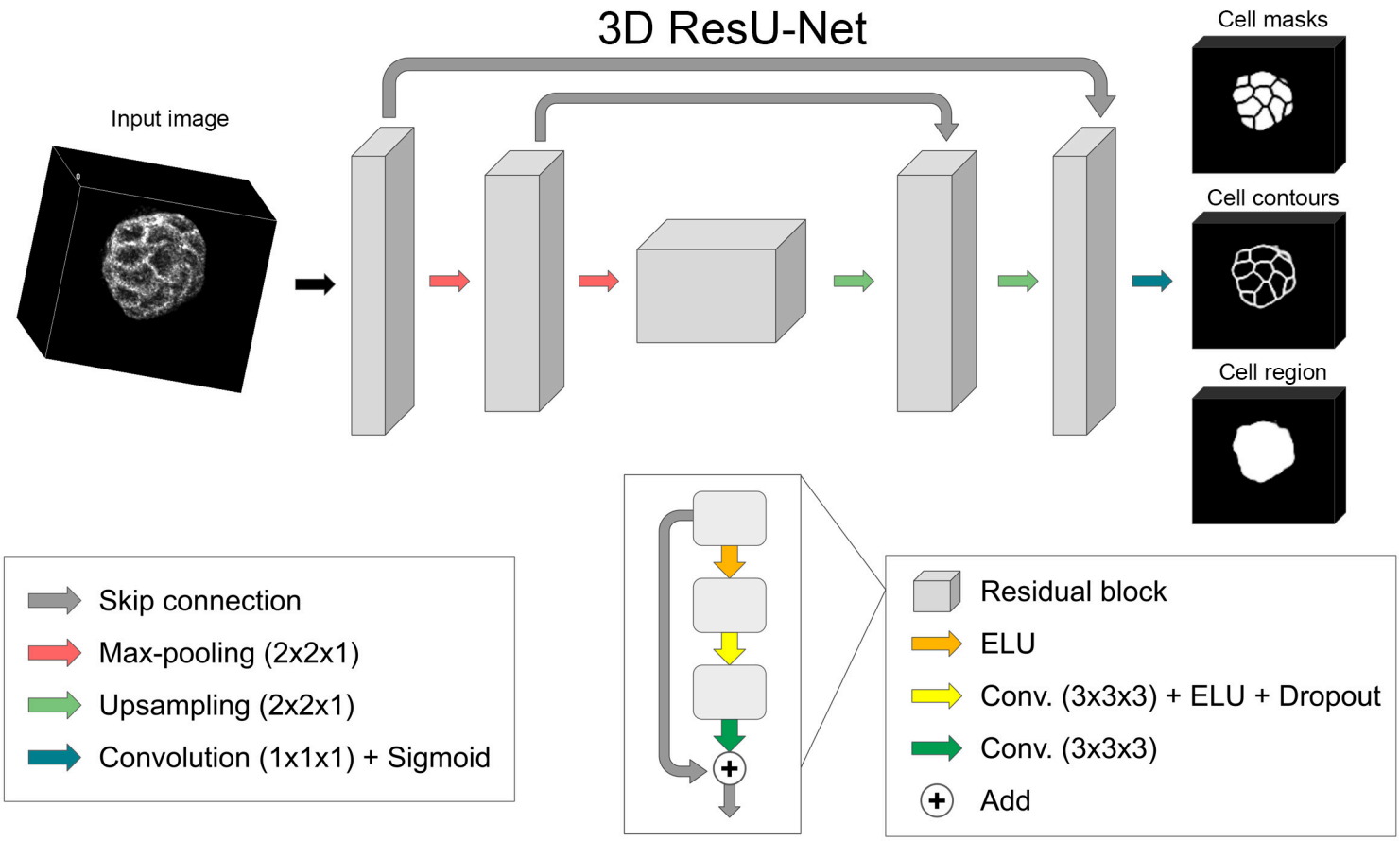

3D ResU-Net
Input image
Cell masks
Cell contours
Cell region
Skip connection
Max-pooling (2x2x1)
Upsampling (2x2x1)
Convolution (1x1x1) + Sigmoid
Residual block
ELU
Conv. (3x3x3) + ELU + Dropout
Conv. (3x3x3)
Add

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.

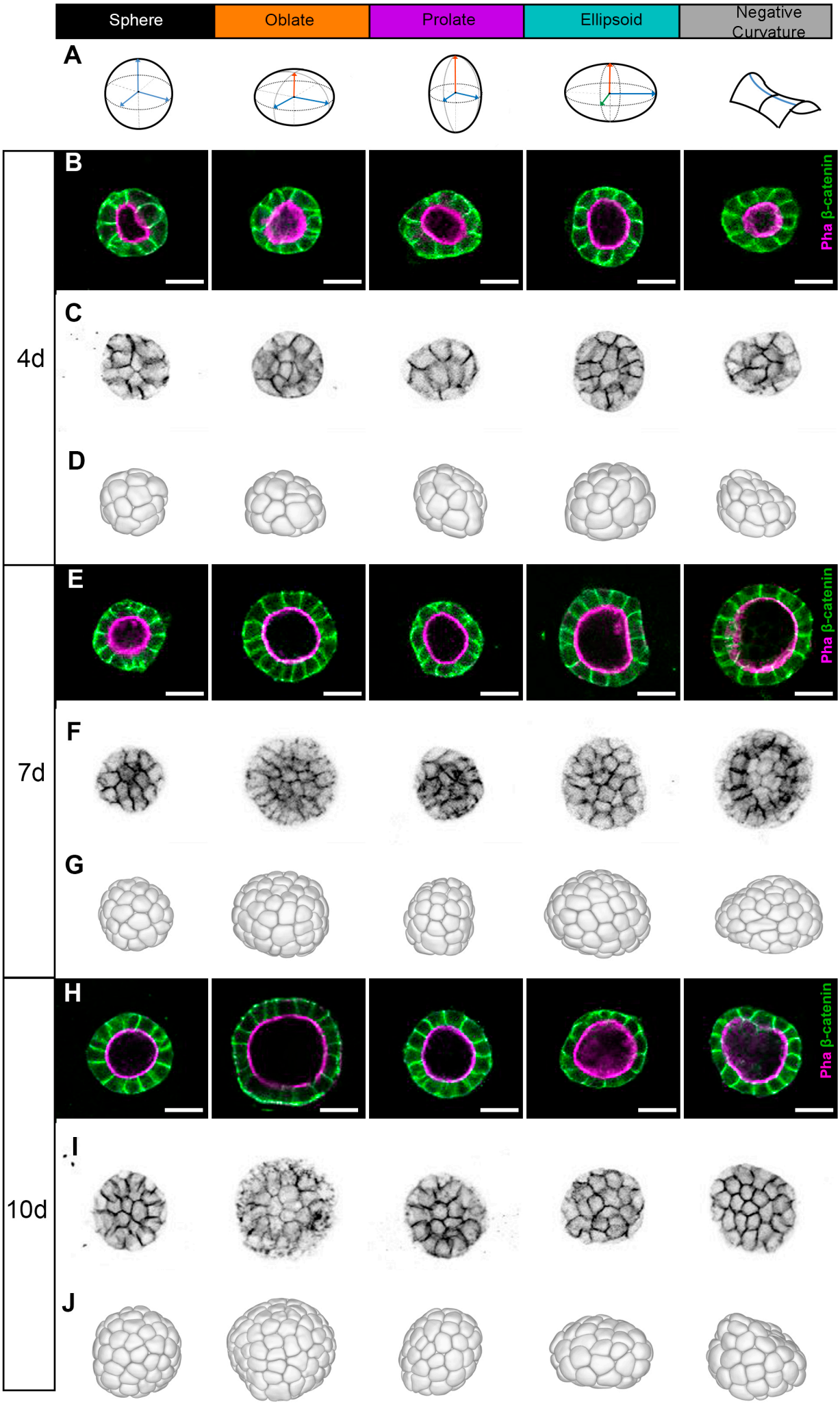

Sphere
Oblate
Prolate
Ellipsoid
Negative Curvature
A
B
C
D
E
F
G
H
I
J
4d
7d
10d
Pha β-catenin
Pha β-catenin
Pha β-catenin

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.

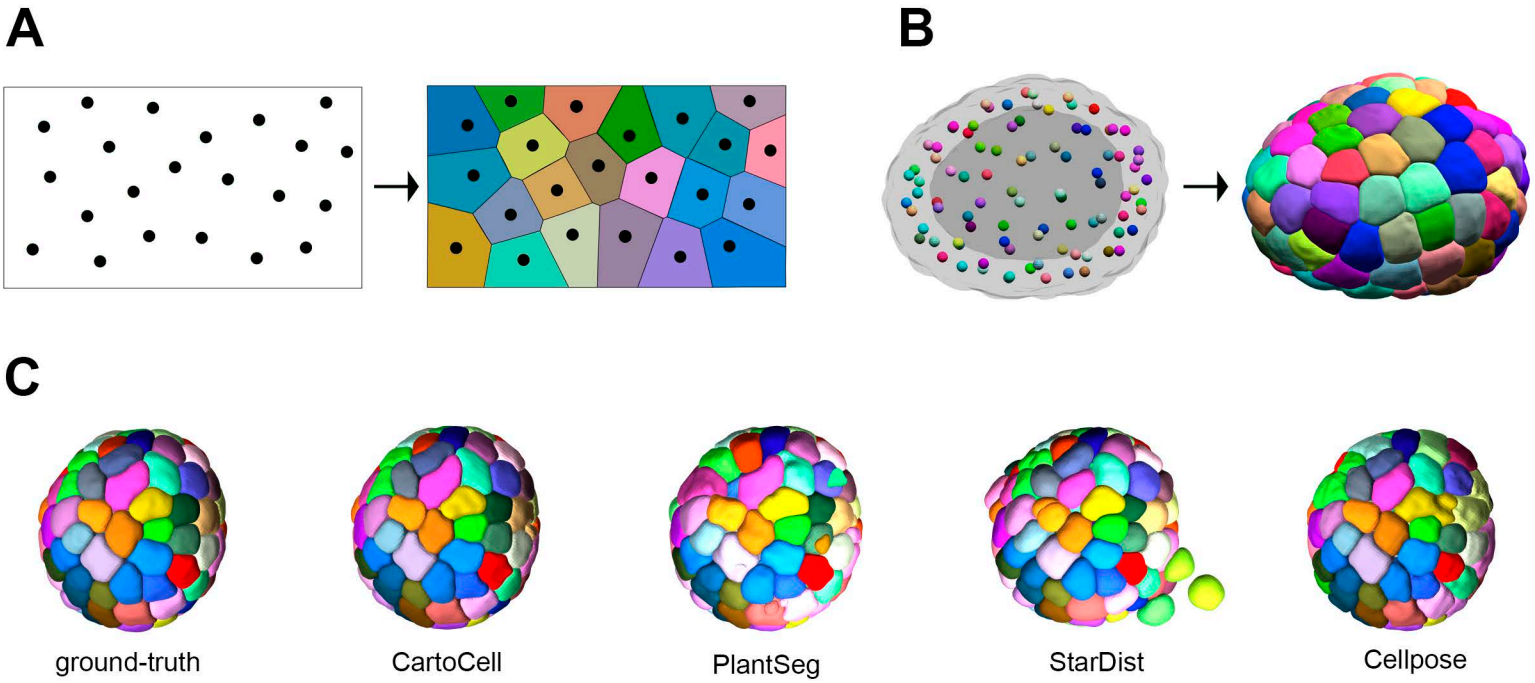

A
B
C
ground-truth
CartoCell
PlantSeg
StarDist
Cellpose

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.

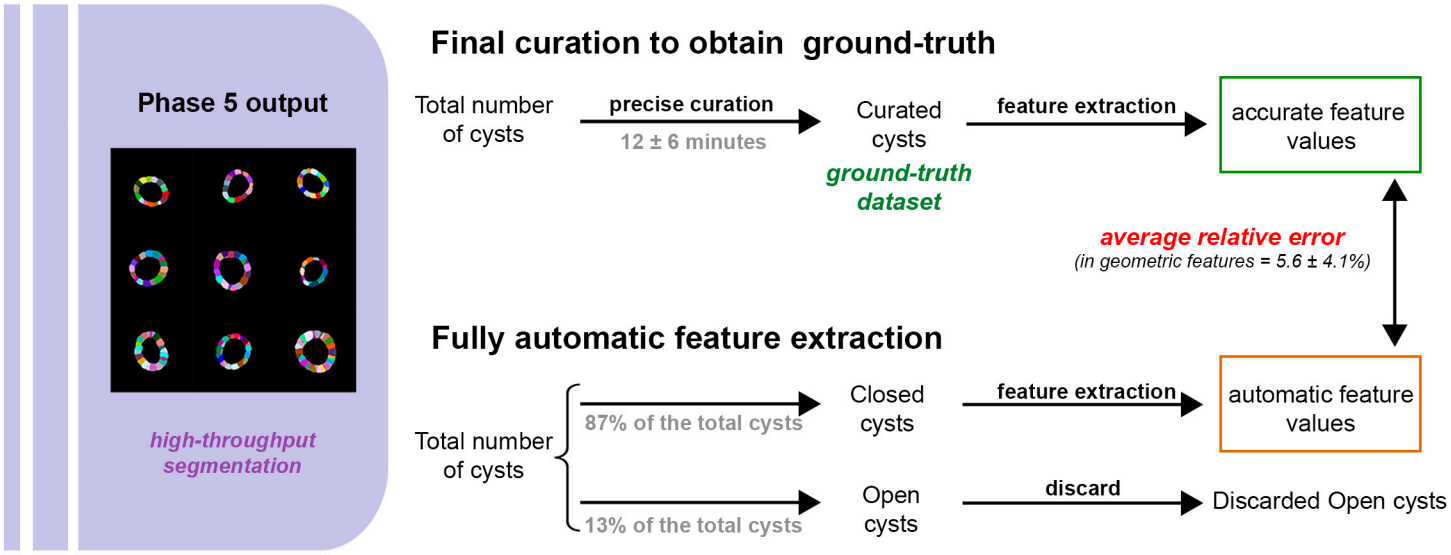
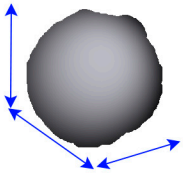

Phase 5 output
high-throughput segmentation
Final curation to obtain ground-truth
Total number of cysts
precise curation
12 ± 6 minutes
Curated cysts
ground-truth dataset
feature extraction
accurate feature values
average relative error
(in geometric features = 5.6 ± 4.1%)
Fully automatic feature extraction
Total number of cysts
87% of the total cysts
Closed cysts
feature extraction
automatic feature values
13% of the total cysts
Open cysts
discard
Discarded Open cysts

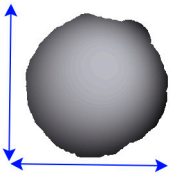
Figure S4. Appendix to Fig.1 showing two possible paths for feature extraction. After the Phase 5, a ground-truth dataset can be obtained by final manual curation of the results (12.0 ± 6 minutes per cyst) to extract the accurate feature values. On the other hand, there is an alternative option, opting 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 ± 4.1% in geometric characteristics compared with the accurate feature values (**Table S4** and **STAR Methods**).

Cyst geometric features

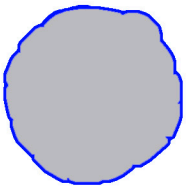
Total Cell Number



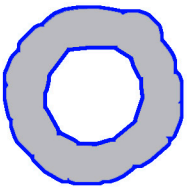
Cyst Axes Length (µm)
Length of the three axes of symmetry (major, minor and intermediate)



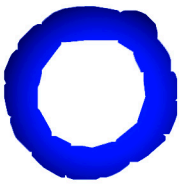
Aspect Ratio
Major axis divided by minor axis length



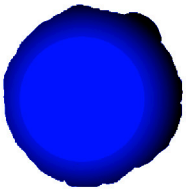
Cyst Surface Area (µm²)
Outer surface area of the cyst



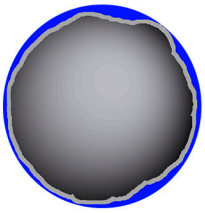
Surface Ratio
Ratio of outer surface to inner surface of the cyst



Epithelium Volume (µm³)
Cyst volume that cells occupy

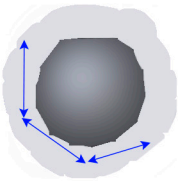


Cyst Volume (µm³)

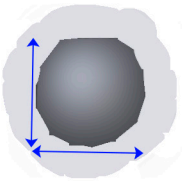


Cyst solidity
Cyst volume divided by the convex hull volume

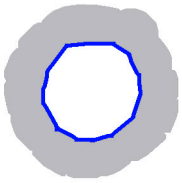
Lumen geometric features



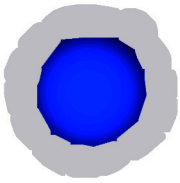
Lumen Axes Length (µm)
Length of the three axes of symmetry



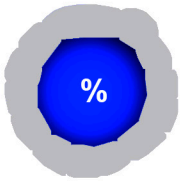
Lumen Aspect Ratio
Major lumen axis divided by minor lumen axis length



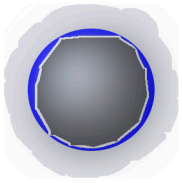
Lumen Surface Area (µm²)



Lumen Volume (µm³)

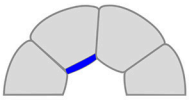


Lumen Space
Percentage of the cyst volume that the lumen occupies

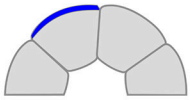


Lumen Solidity
Lumen volume divided by the convex hull volume

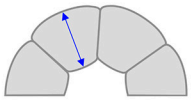
Cell geometric features



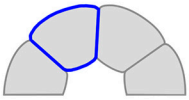
Cell Apical Area (µm²)
Inner cell surface area



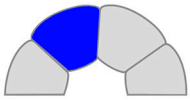
Cell Basal Area (µm²)
Outer cell surface area



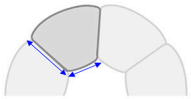
Cell Height (µm)



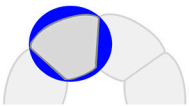
Cell Surface Area (µm²)



Cell Volume (µm³)



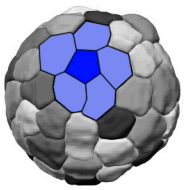
Cell Aspect Ratio
Major cell axis divided by minor cell axis length



Cell Solidity
Cell volume divided by the convex hull volume

Cell packing features

Scutoids (%)
Percentage of scutoids per cyst



Total 3D Neighbors
Number of total neighboring cells of each cell

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

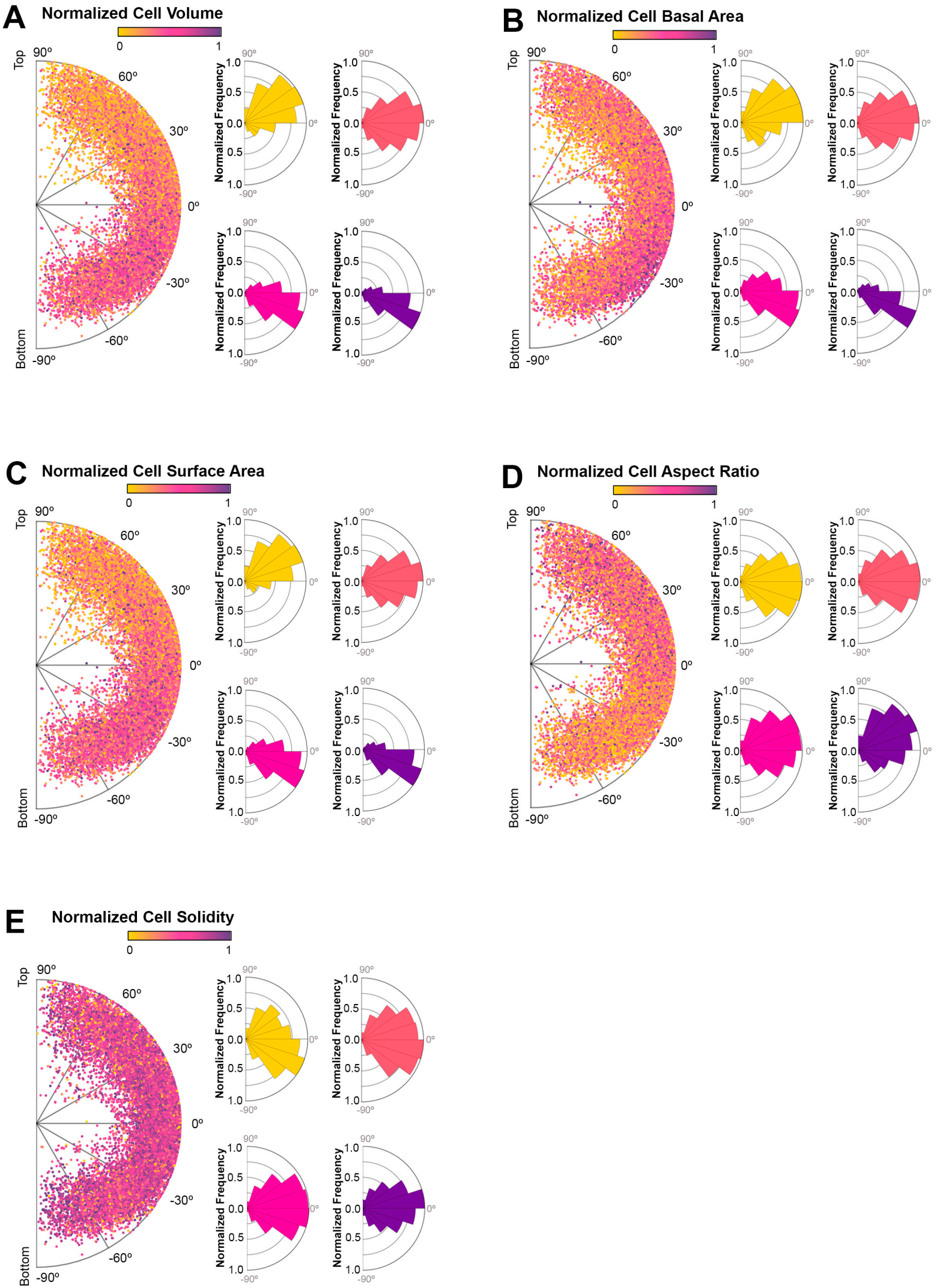

A
Normalized Cell Volume
B
Normalized Cell Basal Area
C
Normalized Cell Surface Area
D
Normalized Cell Aspect Ratio
E
Normalized Cell Solidity
0
1
Top
Bottom
90°
60°
30°
0°
-30°
-60°
-90°
Normalized Frequency
1.0
0.5
0.0
0.5
1.0

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**).

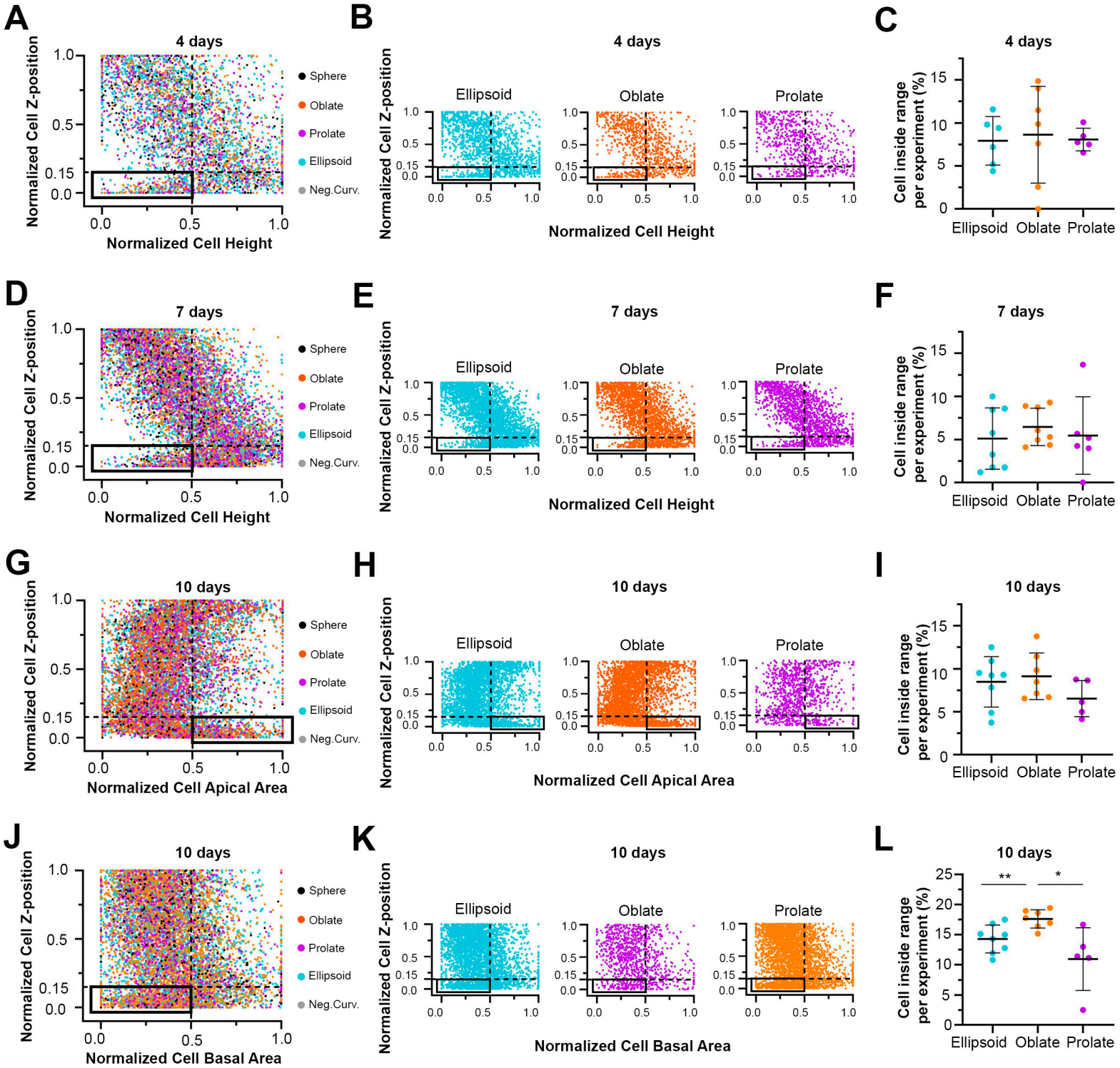

A
4 days
Normalized Cell Z-position
Normalized Cell Height
Sphere
Oblate
Prolate
Ellipsoid
Neg.Curv.
B
4 days
Ellipsoid
Oblate
Prolate
Normalized Cell Z-position
Normalized Cell Height
C
4 days
Cell inside range per experiment (%)
Ellipsoid
Oblate
Prolate
D
7 days
Normalized Cell Z-position
Normalized Cell Height
E
7 days
Ellipsoid
Oblate
Prolate
Normalized Cell Z-position
Normalized Cell Height
F
7 days
Cell inside range per experiment (%)
Ellipsoid
Oblate
Prolate
G
10 days
Normalized Cell Z-position
Normalized Cell Apical Area
H
10 days
Ellipsoid
Oblate
Prolate
Normalized Cell Z-position
Normalized Cell Apical Area
I
10 days
Cell inside range per experiment (%)
Ellipsoid
Oblate
Prolate
J
10 days
Normalized Cell Z-position
Normalized Cell Basal Area
K
10 days
Ellipsoid
Oblate
Prolate
Normalized Cell Z-position
Normalized Cell Basal Area
L
10 days
**
*
Cell inside range per experiment (%)
Ellipsoid
Oblate
Prolate

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.