The Hair Cell Analysis Toolbox: A machine learning-based whole cochlea analysis pipeline.

Christopher J. Buswinka, David B. Rosenberg, Artur A. Indzhykulian*
Mass Eye and Ear, Harvard Medical School.
* Corresponding author: Artur Indzhykulian, inartur@hms.harvard.edu

Abstract. Auditory hair cells, the whole length of the cochlea, are routinely visualized using light microscopy techniques. It is common, therefore, for one to collect more data than is practical to analyze manually. There are currently no widely accepted tools for unsupervised, unbiased, and comprehensive analysis of cells in an entire cochlea. This represents a stark gap between image-based data and other tests of cochlear function. To close this gap, we present a machine learning-based hair cell analysis toolbox, for the analysis of whole cochleae, imaged with confocal microscopy. The software presented here allows the automation of common image analysis tasks such as counting hair cells, determining their best frequency, as well as quantifying single cell immunofluorescence intensities along the entire cochlear coil. We hope these automated tools will remove a considerable barrier in cochlear image analysis, allowing for more informative and less selective data analysis practices. **Keywords:** hair cell, stereocilia, machine-learning, cochleogram, segmentation.

Introduction

Microscopy is an essential and very common tool in investigating the histology and pathology of the inner ear. Hearing loss phenotypes are often reflected in the histology of the sensory epithelium, compromising the organ of Corti, and may be visualized with high magnification light microscopy. Collected micrographs can cover a wide spatial area, capturing data on thousands of cells, often in three spatial dimensions. Auditory hair cells of the cochlea are classified into two subtypes, inner hair cells (IHC) and outer hair cells (OHC), with varying geometric locations, sizes, shapes, and protein expression 'fingerprints', all of which may vary with age and along the tonotopic axis[1]. Each hair cell carries a bundle of actin-rich microvilli-like protrusions called stereocilia. The OHC stereocilia bundles have a characteristic V-shape and are composed of thinner stereocilia as compared to those of IHCs. Each cell is a potential datum point which must be parsed from the image into a usable format for analysis. Typically, this may have been done by hand, aided by a computer. While achievable for a small number of cells, the slow speed and tedium of this analysis poses a significant barrier when faced with large datasets, especially those generated through studies involving high-throughput screening[2]. Alternatively, (and popular in cochlear analysis) large datasets can be broken down into representative locations which can be analyzed in depth by hand. Often three small representative regions at the base, middle, and apex of the cochlear spiral[3] are chosen to reflect the tonotopic changes of biological variables. This approach may be enough to paint a general picture of the outcome of an experiment but scales poorly in comparison to other common techniques of cochlear function testing, such as recording auditory brainstem responses or distortion product otoacoustic emissions, which could be easily collected, and analyzed, at an arbitrarily large number of frequencies[4,5]. The disconnect in the frequency resolution of histopathological analysis and cochlear function testing makes image analysis on a single-cell level across the entire frequency spectrum of hearing desirable. To this end, an analysis software must *detect* each hair cell, determine its *type* (IHC vs OHC), *segment* these cells to extract geometric and fluorescence data, and assign a cell its *best frequency* based on its location along the cochlear turn.

Considerable work has been done on deep learning approaches for object detection. The predominant approach is Faster RCNN[6], a deep learning algorithm which quickly recognizes the location and position of objects in an image. While designed for use with images collected by conventional means (cell phone or camera), there has been success in applying the architecture to biomedical image analysis tasks[7-9]. We apply this algorithm to detect and classify hair cells at speeds orders of magnitude faster than manual analysis, while maintaining high accuracy. High detection accuracy is paramount to generating cochleograms – a graph of hair cell count along the length of the cochlea – a common type of manual analysis reporting hair cell survival rates along the cochlear coil most often used in human temporal bone studies.

Hair cell detection and classification alone, however, is not sufficient for the quantification of hair cell fluorescence, which has applications in gene therapy, RNA scope quantification, and various fluorescent dye

loading assays. Instead, the pixels in an image must be assigned to the high-level cell object they represent in a process known as *instance segmentation*. Cell segmentation has classically been tackled with a combination of manual thresholding and the watershed algorithm, a segmentation tool relying on intensity gradients between objects[10,11]. Often cells can be hard to detect in densely packed tissues with non-obvious delineation between cells. Poor separation negatively impacts watershed segmentation, with optimal performance heavily reliant on manual fine-tuning, which is slow and can introduce bias. More recently, machine learning approaches have been successful in supplanting the watershed algorithm for instance segmentation with increased accuracy[12-15]. Predicting the spatial embeddings of cells, a deep learning approach for generating instance segmentation masks, is highly accurate and generalizes to a wide array of cell types[16]. This approach was popularized by the *Cellpose* algorithm[17], and offers exceptional results for segmentation in two dimensions. 3D segmentation is possible by applying *Cellpose* along different coordinate planes and using the 2D masks to generate the 3D mask[17]. This is effective with isotropic and morphologically homogenous cells, however the algorithm's performance in detecting hair cells was less impressive, likely due to nonobvious spatial separation.

Leveraging these recent deep learning advances, we present here a suite of tools for cochlear hair cell image analysis, the Hair Cell Analysis Toolbox (HCAT), a consolidated software for fully unsupervised hair cell detection and segmentation.

Results

Analysis Pipeline Overview: The software utilizes deep learning algorithms which have been trained to accept input data of volumetric confocal micrographs of full cochlear turns, while also permissive to datasets containing smaller cochlear fragments. The datasets contain at least two channels of information, anti-Myo-VIIA labeling to visualize the hair cell body, and the phalloidin staining to visualize the stereocilia bundle, at a X-Y resolution of 289 nm/px (**Figure 1**). The user may either choose to run a cell detection analysis to generate a cochleogram of inner and outer hair cells (from maximal projections of confocal data) or run a segmentation analysis (from 3D volumetric confocal data), which will extract fluorescence information of all segmented hair cells along the tonotopic axis. Both algorithms scale and may be run on arbitrarily large confocal micrographs by repeatedly analyzing small local crops of the image and merging the result to form a contiguous dataset. These local crops are chosen to overlap such that each cell is guaranteed to be completely represented in at least one. Crops which do not contain Myo-VIIA fluorescence above a certain threshold are skipped, increasing speed of large image analysis and limiting false positive errors. In the case of a whole cochlear analysis, each cell is additionally assigned a best frequency via nonlinear regression and the Greenwood function as outlined below. The result of every analysis is output as an associated csv data table to enable further data analysis or downstream postprocessing.

Cochlear position detection: For both, the segmentation and detection analysis, the software must determine the place of each hair cell along the cochlea to analyze any location-based trends along the tonotopic axis. To do this, we fit a Gaussian process nonlinear regression through the Myo-VIIA fluorescence image, effectively treating each hair cell as a point in cartesian space. A line of best fit can be predicted through each hair cell and in doing so approximate the curvature of the cochlea. We can then use the length of this curve as an approximation for the length of the cochlea. For example, a cell that is 20% along the length of this curve could be interpreted as one positioned at 20% along the length of the cochlea.

To optimally perform this regression and determine cell's best frequency, multiple preprocessing steps are necessary. First, a maximum projection along the z axis of a previously predicted, whole cochlea segmentation mask, is taken, reducing a three-dimensional volume to a single plane. The image is then down sampled by a factor of ten using local averaging and converted to a binary image. Two final preprocessing steps are performed: (1) binary hole closing which closes any gaps, and (2) binary erosion which reduces the effect of external nonspecific staining. Each positive binary pixel of the resulting two-dimensional image is then treated as an X/Y pair which may be regressed against.

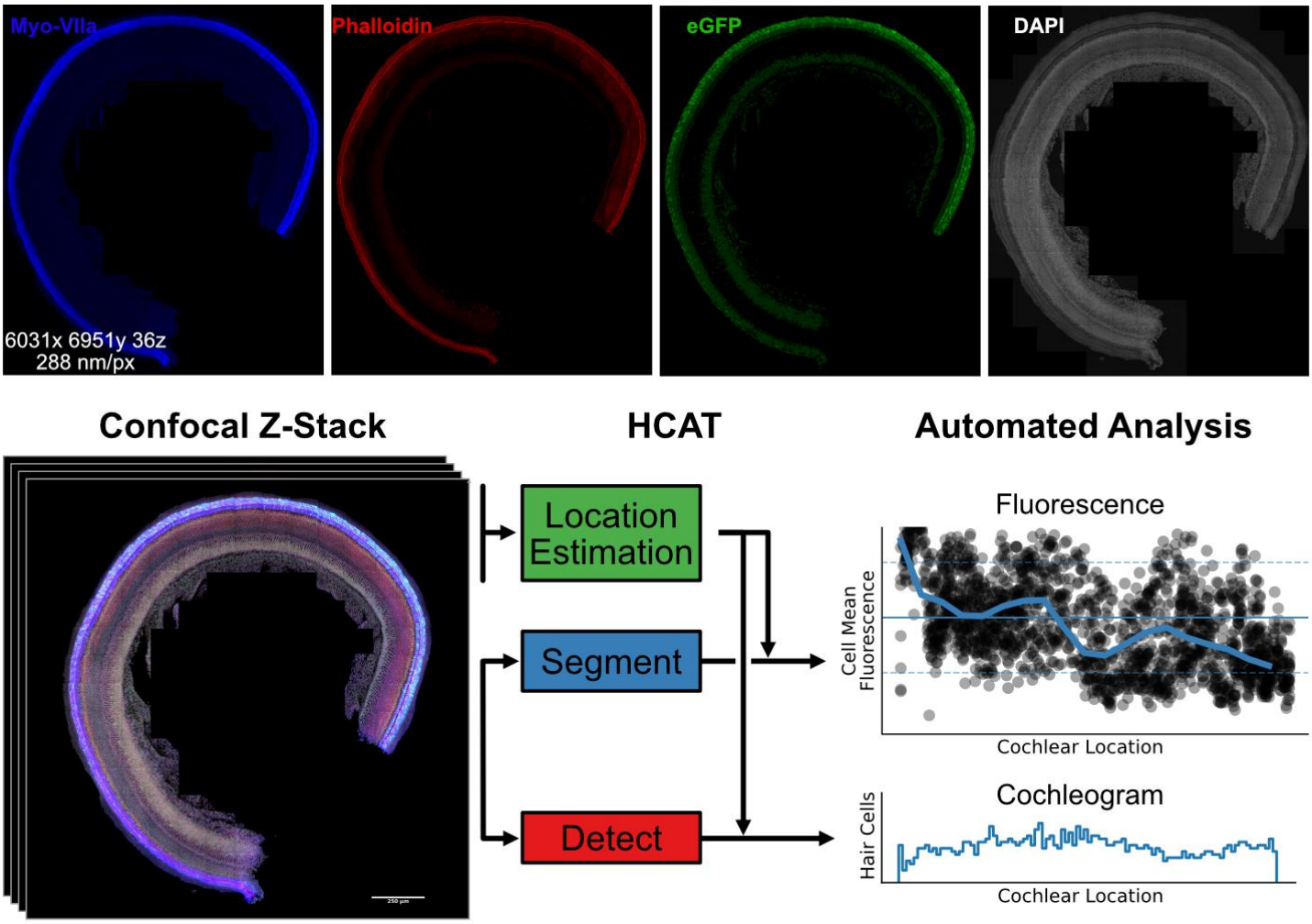


Figure 1. **Overview of Algorithmic Approach.** *Top,* Exemplar data used in this study. Cochlear coils immunolabeled against Myo-VIIA (*blue*), stained with Phalloidin (*red*) and DAPI (*grayscale*), and genetically expressing eGFP (*green*) were volumetrically imaged and input to software for analysis. *Bottom,* the user may choose to run either a 3D *segmentation* analysis (*blue*), in which cells are volumetrically delineated from each other, allowing for hair cell fluorescence intensity measurements on a single-cell level along the entire cochlear coil, or a 2D detection analysis (*red*), in which hair cells are counted and classified by type. These two analyses are paired with a cochlear location algorithm (*green*) which assigns a best frequency to each hair cell, allowing for cochleograms to be readily generated, or the fluorescence intensity measurements to be investigated as a function of frequency or cochlear location.

The resulting image of an intact cochlea, when processed as is, would likely not form a mathematical function in cartesian space, and therefore be difficult to approximate. For example, the cochlea may curve over itself such that for a single location on the X axis, there may be multiple clusters of cells at different Y values. To rectify this overlap, the data points are converted from cartesian, to polar coordinates by first shifting the points and centering the cochlear spiral around the origin. From this, each X/Y pair can be converted to a corresponding angle/radius pair. In doing so, a gap is created as a cochlea is not a closed loop. This gap is detected, and these points are shifted by one period, creating a continuous function. A Gaussian process[18], a generalized nonlinear function, is then fit to the spherical coordinates and a line of best fit is predicted. This line is then converted back to cartesian coordinates and scaled to correct for the earlier down sampling. A linear interpolation of points is performed to ensure each point is exactly 0.1% of the length of the cochlea. The apex of the cochlea is then inferred by comparing the curvature at each end of the line of best fit based on the observation that the apex has a tighter curl when mounted on a slide. Next, the location of each hair cell along the cochlea as a function of its total length (%) is determined by projecting the cell's center to the nearest point of the line of best fit. Finally, the frequency at that location is then calculated using the Greenwood function[19].

The final result is a curve of known length which tracks the hair cells of the cochlea. Any cell location may be mapped to this curve as a function of the total cochlear length (%) and a best frequency calculated using the Greenwood function. Upon our careful examination, this process proves to be highly accurate. We also manually mapped the cochlear length to cochlear frequency using a widely used *imageJ* plugin, developed by the Histology Core at the Eaton-Peabody Laboratories, Mass Eye and Ear. The plugin is available for download at https://www.masseyeandear.org/research/otolaryngology/eaton-peabody-laboratories/histology-core. Over eight manually analyzed cochleae, the *maximum* cell frequency error relative to a manually mapped best frequency result was under 10% of an octave, with the discrepancy between the two methods less than 5% for most cells (60% of a semitone, see **Supplementary Figure S2**). If the curvature estimation fails, a manually annotated text file of points following the cochlear spiral may be additionally passed to the algorithm. A smoothing spline fit through these points will be used instead for cell frequency estimation. It is also worth note, that both hair cell detection and segmentation can be carried out by the HCAT software in isolated sections of the cochlea in the absence of whole cochlear imaging. While this approach does not make use of the full functionality of the software, it does provide researchers with a valuable tool for automated and unbiased cell counting and fluorescence quantification, suitable for large datasets.

Hair Cell Detection: A maximum projection image containing cochlear hair cells can be iteratively evaluated in a deep learning model trained to detect and classify cochlear hair cells using this software. We leverage the Faster R-CNN deep learning model with a ResNet-50 backbone[20], trained on early postnatal cochlea labelled with phalloidin and antibodies against Myo-VIIA (**Figure 2**). While trained solely with these labels, the model can perform on cells labeled by other markers, provided the specimens contain both cytosolic hair cell, and stereocilia, labels.

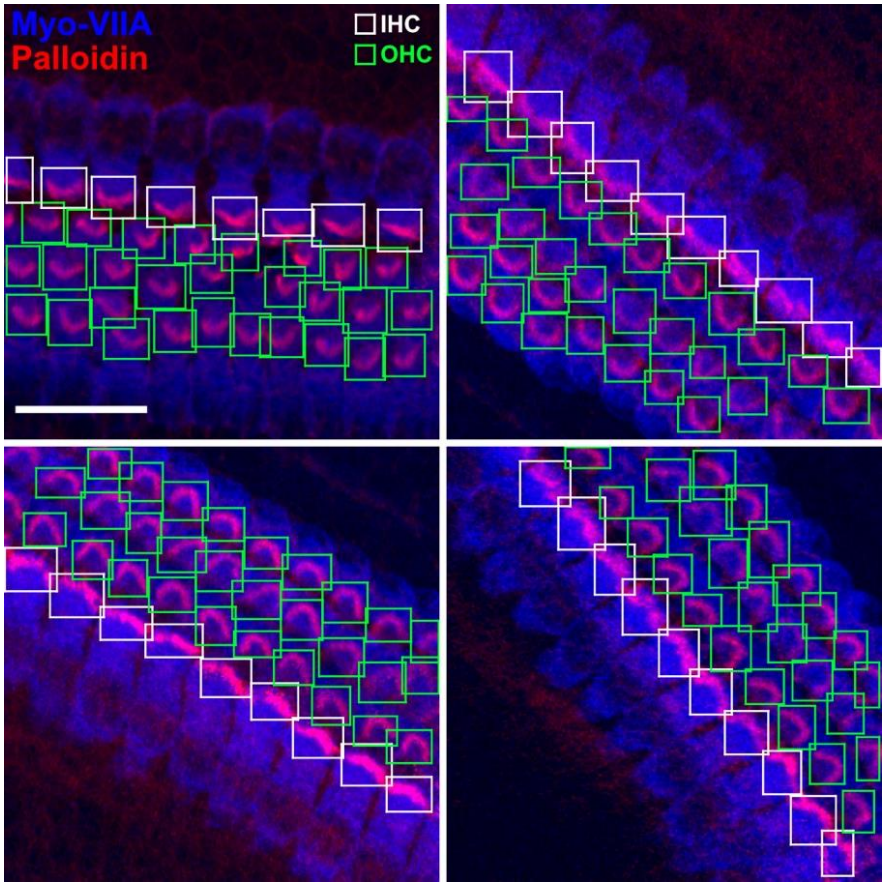


Figure 2. **Exemplar training data for the hair cell detection algorithm.** Cochlear hair cells at varied frequency location were imaged using confocal microscopy. Images were immunolabeled against Myo-VIIA (blue) and stained with phalloidin (red). Bounding boxes were manually placed around inner (*white*) and outer (*green*) hair cell stereocilia bundles. These boxes and classifications were used to train the Faster R-CNN detection algorithm. *Scale bar,* 25 µm.

The deep learning model predicts three features for each predicted cell, a box encompassing the cell, a classification label (IHC vs OHC), and a confidence score (**Figure 3**). After evaluation, two post-processing steps are taken to refine the output and improve overall accuracy. First, some cells may be detected twice due to redundancies in evaluating cropped subsections of a whole image. Redundant cell detections are removed and the cell with the largest confidence score is kept. The second step is optional and relies on the estimation of the cochlear spatial path outlined earlier, with cells whose distance away from this path exceeds a threshold value are removed. This step can be enabled in cases with sub-optimal anti-MyoVIIA labeling outcomes with instances of non-specific labels away from the ribbon of hair cells along the cochlea, thus reducing the false-positive detection rate.

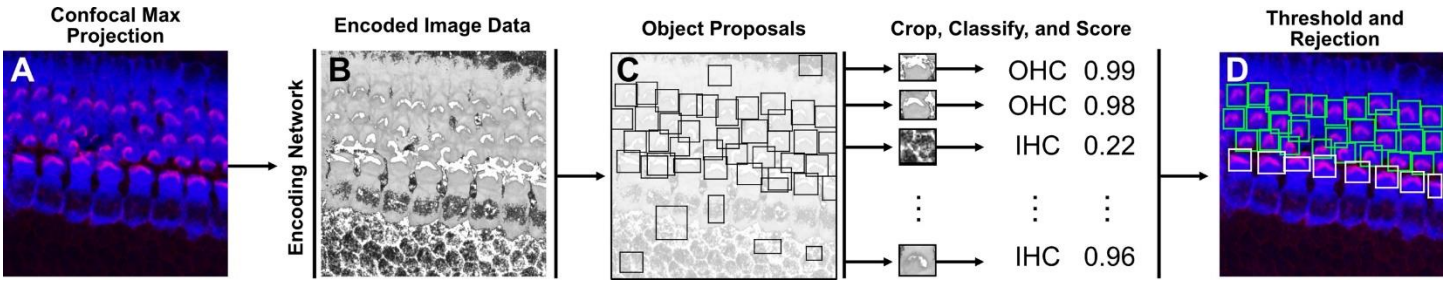


Figure 3. Simplified overview of Faster R-CNN image detection pipeline. (A) Two-dimensional maximum projection images of three-dimensional Z-stacks (Myo-VIIA in *blue* and phalloidin in *red*) are encoded into a high-level representation by a trained convolutional neural network, schematized in *(B)*. (C) The encodings are transferred to an additional object proposal network which generates bounding boxes of predicted objects. Encoded crops, based on the predicted object proposals, are classified into outer and inner hair cells, and assigned a score. (D) Based on these scores and the overlap between boxes, objects are removed, resulting in the algorithm's best guess at the location and classification of every object detected on the image.

With the deep learning model optimally trained, followed by the post-processing steps, we see highly accurate performance on whole cochlea cell detection (**Figure 4**). Compared to cochleae analyzed manually we found a mean 95.2 ± 3.6% true positive accuracy for cell identification and a 0.5± 1% classification error (8 cochlear coils, each validation shown in **Supplementary Figure S2**). This algorithm is robust against tissue idiosyncrasies such as four rows of outer hair cells (**Figure 4C**) or atypical inner hair cell locations. Furthermore, when paired with cell frequency estimation, highly accurate cochleograms may be readily generated by the software (**Figures 4F** and **4G** for IHCs and OHCs, respectively), taking just under a minute to run a full detection analysis.

Hair Cell Segmentation: To use the toolbox for the analysis of fluorescent signal in hair cells along the length of the cochlear, rather than hair cell quantification, an alternative pathway and deep learning algorithm was developed. In a similar approach to *Cellpose*[16,17,21], we elect to train a deep learning model to optimized spatial embeddings of cells for instance segmentation. The U-Net architecture has been previously utilized successfully for the performance of various biomedical segmentation tasks [14,17,22-29], and as such we employ a similar architecture. Recent work on resource constrained segmentation mask has shown that recurrently applying subsections of the U-Net architecture can improve performance without increasing model size[30]. Building on this idea, we recurrently apply each subsection of U-Net to improve performance while limiting the memory complexity of the model. To account for the anisotropy in our dataset, we employ strided convolutions[31] with different strides at each down sampling step of the architecture. A leaky ReLU[32] has been shown to be advantageous in the performance of residual neural networks[33,34] and as such is chosen as the activation function for our architecture. To maximize usability of the software and maximize flexibility in experimental design, we train the model solely on the Myo-VIIA fluorescence signal (**Figure 5**) which were manually annotated in the Amira software package[35].

An image volume with the Myo-VIIA fluorescence is first median filtered to remove noise and input into the machine learning model (**Figure 6A**) which has been trained to generate a set of spatial embedding vectors (**Figure 5B**) and a semantic segmentation mask (**Figure 6C**). From these, pixels likely belonging to a cell are

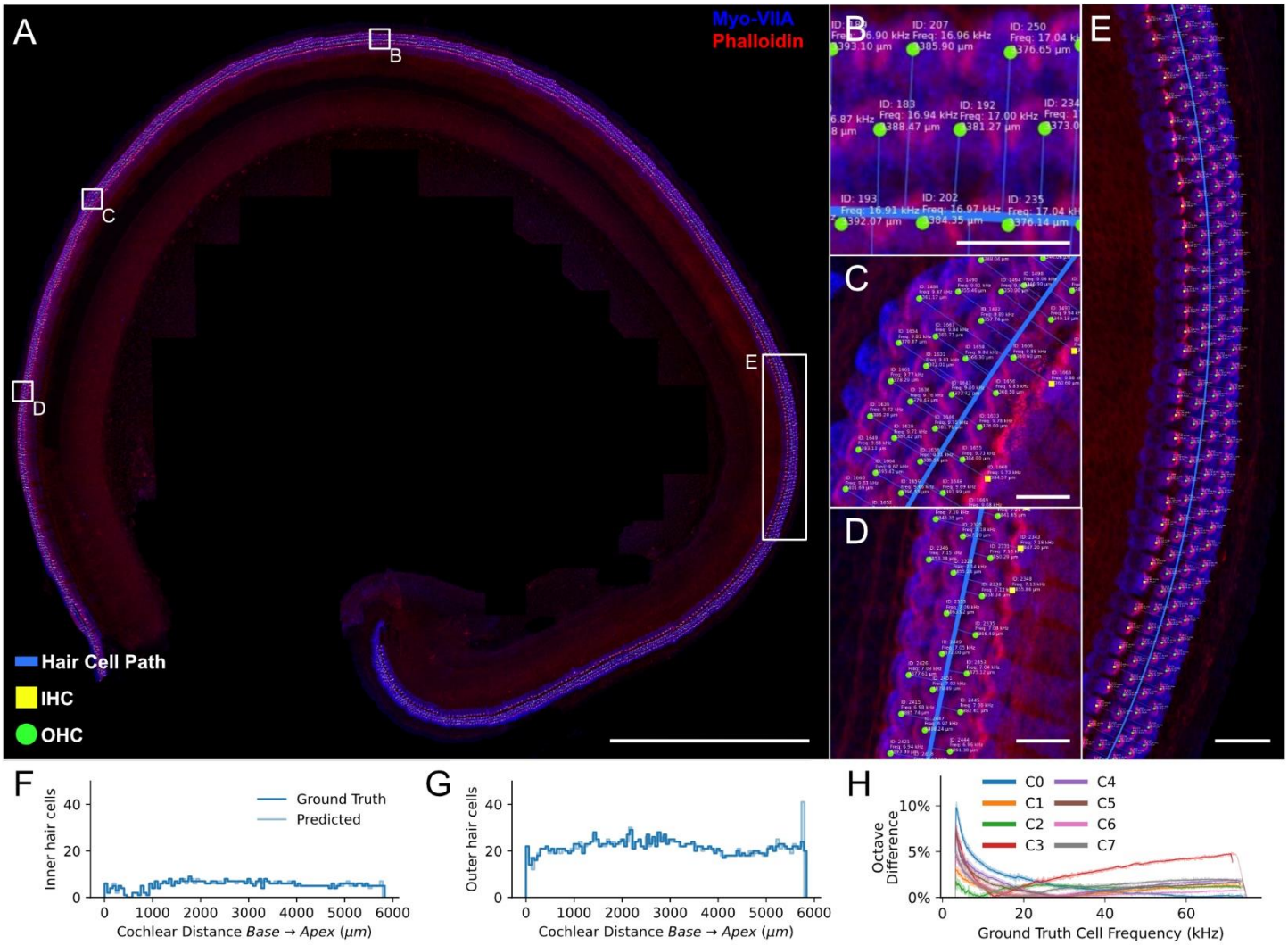


Figure 4. Validation output of the hair cell detection analysis. A validation output image is generated for each detection analysis performed by the algorithm. Each image contains visual information on cell detection locations, cell classifications and cochlear path estimation (if available). Additionally, each detected hair cell's ID, its location along the cochlear turn (distance in *μm* from the apex), and best frequency are embedded. Such an image has been automatically generated by the software here for an entire cochlea (A). While the vast majority of cells are accurately detected (B, E), regions of poor performance are also highlighted in C and D. While the algorithm is robust to four rows of outer hair cells, visual artifacts limit detection accuracy in *C* while low signal strength may explain poor detection performance in *D*. When paired with single cell frequency estimation, accurate cochleograms of inner hair cells (F) and outer hair cells (G) can be readily generated. This frequency estimation is highly accurate with a maximum error on 8 different cochleae of 10% of an octave.

projected from their location in space via the predicted embedding vectors, forming clusters around the centroids object instances. When predicting the segmentation masks of new cells, their centroids are not known and therefore must be inferred from the clusters of pixels (**Figure 6D**). We employ the DBSCAN clustering algorithm to predict these clusters as it is robust against noise and scales well with large datasets. To improve performance, we found that down-sampling by a factor of two increases centroid detection speed with negligible loss in detection accuracy. Pixels at the borders of objects tend to be more poorly projected to object centers, leading to sparse cluster formation. To reduce this effect, we disregard pixels near the edge of the semantic probability map via binary erosion. From the resulting centroid prediction, probability maps for each instance a re generated based on equation (1).

$$\phi_k(e_i) = exp\left(-\frac{(e_{ix}-C_{kx})^2}{2\sigma_{kx}^2} - \frac{(e_{iy}-C_{ky})^2}{2\sigma_{ky}^2} - \frac{(e_{iz}-C_{kz})^2}{2\sigma_{kz}^2}\right) \quad (1)$$

To avoid double counting cells, we perform non maximum suppression on each instance probability map[36]. To ensure each segmentation mask predicted by the algorithm is of a complete cell, we remove any segmentation masks touching the edge of an evaluated image, in addition to removing any mask whose volume is below a reasonable value of a cell chosen at 80% of the smallest volume in our training set. While the DBSCAN algorithm is fast, it relies on tuning parameters to optimally perform. We have chosen these parameters to aggressively reject loose clusters ensuring each detected cell is properly segmented, with the notable drawback of limiting the ability of the algorithm to segment every cell. Finally, we disregard cells with mean Myo-VIIA fluorescence signal intensity levels below 5% of maximum measured value, greatly reducing the number of false positive cell mask detections.

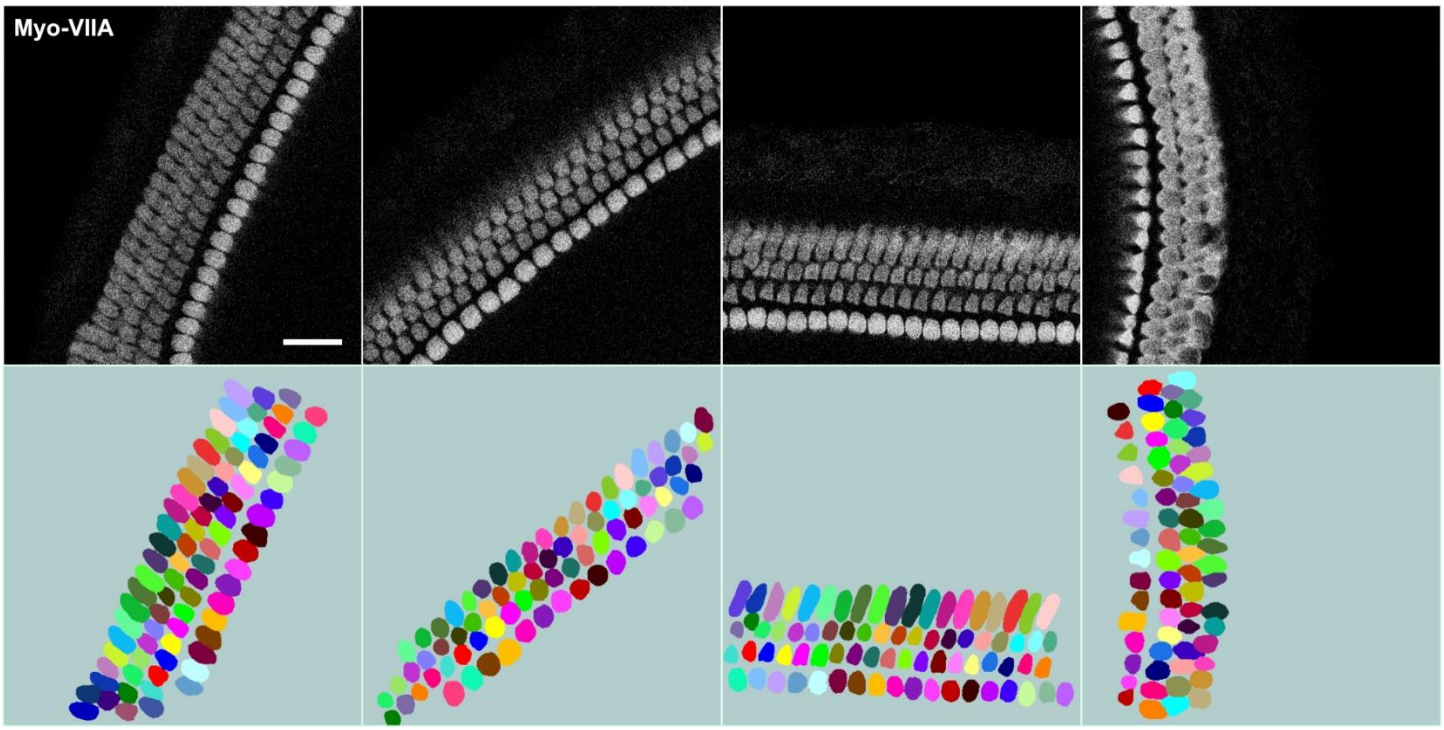


Figure 5. **Exemplar training data used to train the hair cell segmentation algorithm.** To train a deep learning network to perform an instance segmentation task, 17 confocal Z-stacks containing cochlear hair cells immunolabeled against Myo7a (*top* panels) were manually segmented in three dimensions using the Amira Software package. The resulting 3D cell segmentation masks (*bottom* panels) were subsequently used for training of the deep learning model. Presented are single frames from the z-stacks representing individual tiles of the tile scans, with the tiles taken from a number of different datasets at various cochlear locations.

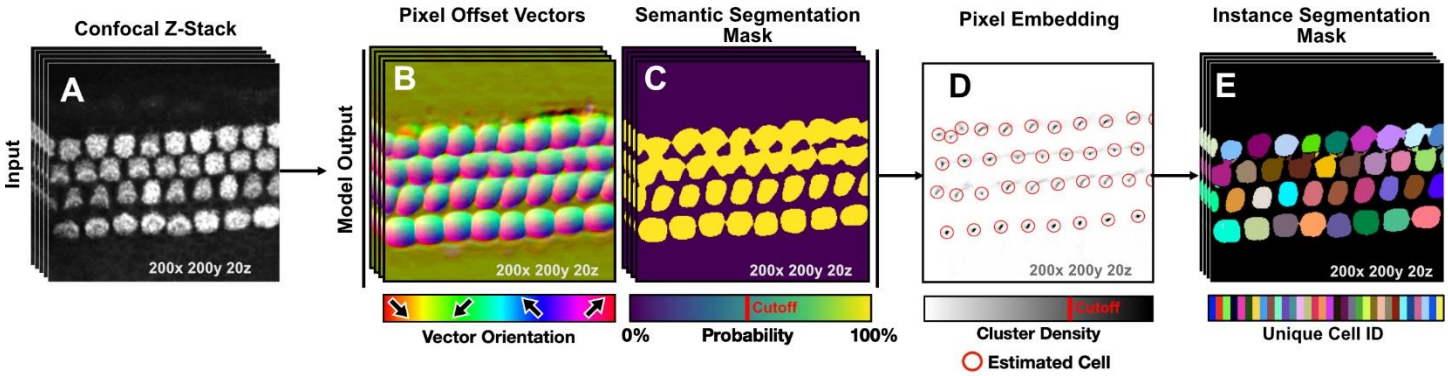


Figure 6: Overview of spatial embedding image segmentation. Volumetric confocal micrographs (A) are input to a machine learning model trained to predict spatial embedding vectors (B) in addition to a semantic segmentation mask (C). Pixels are projected using these vectors into clusters which form the "embedding space" (D) which are detected using a clustering algorithm. Each cluster gives rise to an object probability map which is used to form an instance segmentation mask of objects in the original image (E). The performance was fine-tuned by setting the thresholds in C and D (*red* "cutoff" lines). Every unique color in E represents a different hair cell.

In order to evaluate our method of instance segmentation we compared it against two other methods: 1) the generalist *Cellpose* algorithm, and 2) a semantic segmentation approach coupled with the watershed segmentation algorithm. While *Cellpose*[17] was not designed to natively segment volumes, the software provides a pseudo 3D approach in which the algorithm is applied in 2D over different axis pairs of a volume. While the segmentation performance is good, many cells were not segmented, in one case as low as only 17% of cells were segmented (**Figure 7**). As a generalist algorithm, *Cellpose* was not trained on examples of hair cells. With the same training data and deep learning architecture, we trained a deep learning model to perform instance segmentation which was paired with a volumetric watershed algorithm. The results of this approach were poor, with numerous detection and segmentation errors (**Figure 7**). We therefore find our method of instance segmentation outperform other methods of unsupervised volumetric biomedical instance segmentation.

While this algorithm relies on a strong Myo-VIIA fluorescence signal to perform, the resulting cell segmentation mask can be further used to measure cell-specific fluorescence intensity data on *any* addition channels of imaging information accompanying the dataset. As such, the main purpose of this segmentation process is to delineate individual hair cells within their borders by assigning all pixels within a volume (i.e., voxels) representing each hair cell to a single segmentation mask with a unique ID and illustrated with a unique color on **Figure 8**. These masks are then used to measure fluorescence intensity levels across different channels of imaging data on a single-cell level along the cochlear coil. As each cell has a best frequency automatically inferred as described above, not only do measurements of whole cell fluorescence intensity levels in 3-dimensional data become automated, simple, and unbiased, they also allow to present these fluorescence measurements as a function of location along the tonotopic axis of the cochlea (**Figure 8**).

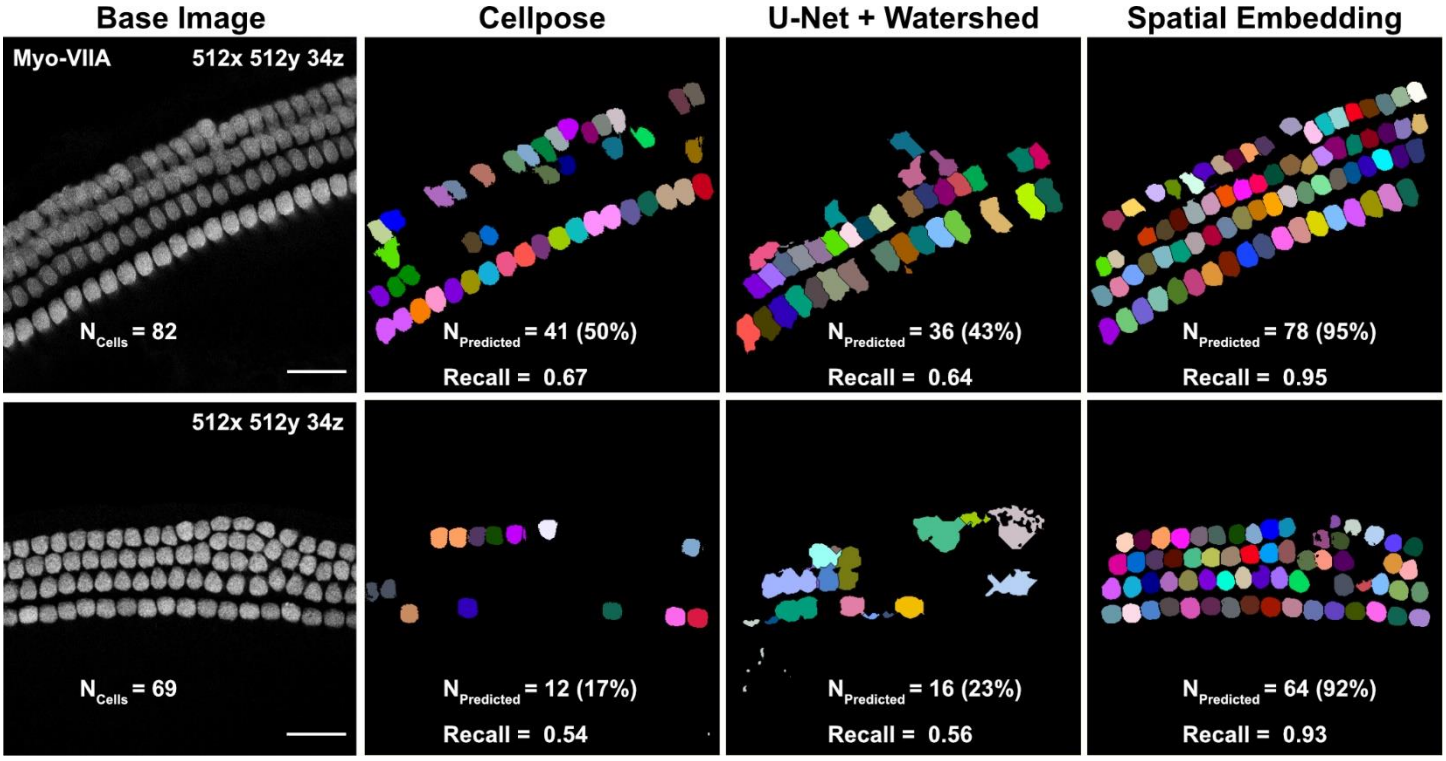


Figure 7. **Example segmentation outcomes of three different segmentation methods evaluated in this study.** Comparison of multiple approaches at volumetric instance segmentation of confocal z-stack of cochlear hair cells. The spatial embedding approach more accurately detected hair cells compared to *U-Net + watershed* approach or the generalist *Cellpose* tool.

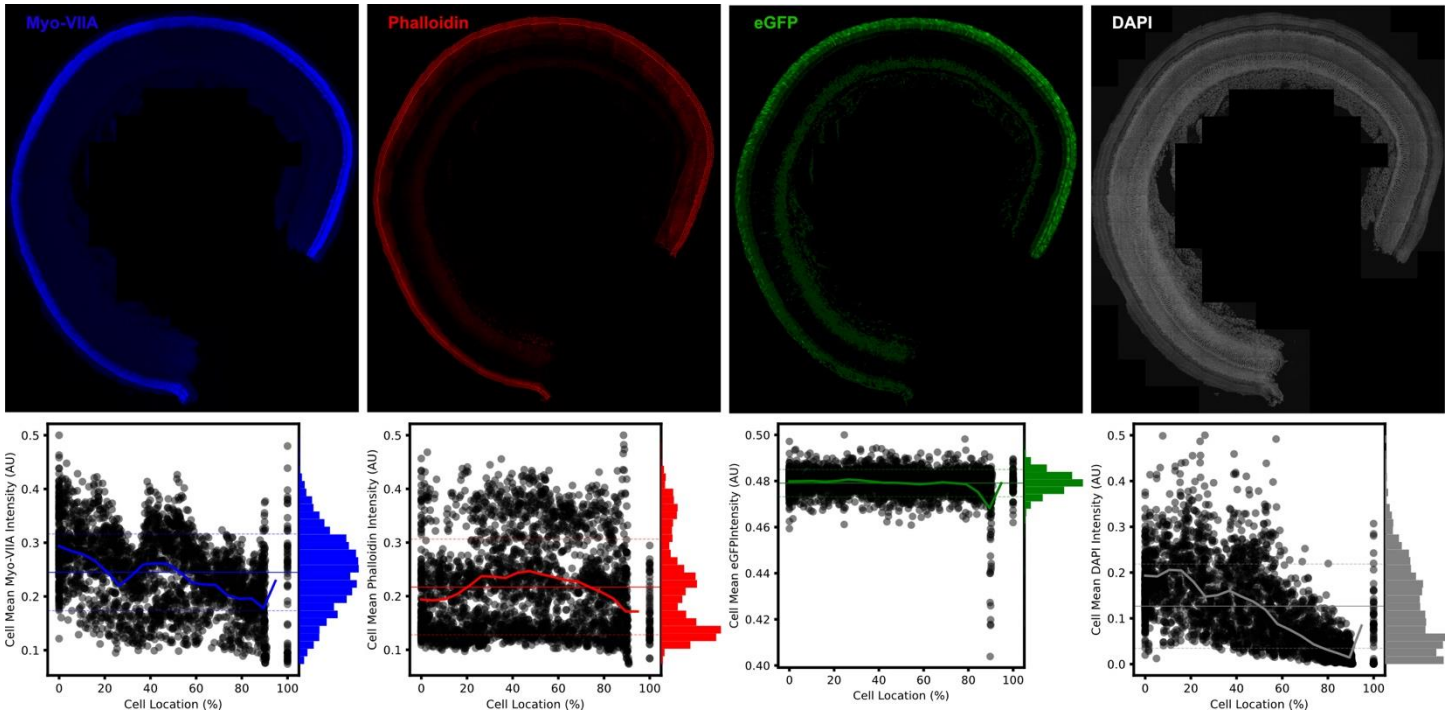


Figure 8: Single-cell fluorescence intensity analysis of hair cells along the cochlear coil. *Top*, exemplar maximum projection image of the Z-stack of a mouse cochlear coil analyzed with the hair cell segmentation algorithm. *Bottom*, Single-cell volumetric fluorescence intensity measurements of four fluorescence signals obtained from predicted instance segmentation masks. Paired with cell frequency estimation, the fluorescence intensity measurements are presented as a function of cochlear location from apex (0%) to base (100%) of the cochlea. Black dots represent single cell measurements, with an average fluorescence intensity value along the cochlear turn shown as a color line, with the population histogram presented on the right of each panel.

Discussion

Here we present the first fully automated cochlear hair cell analysis pipeline where users can enter multiple confocal micrographs of cochleae and quickly detect hair cells to generate cochleograms or segment hair cells enabling fluorescence intensity measurements on a single-cell level. We consider this to represent a considerable advancement in the unbiased analysis and quantification of hair cell imaging datasets.

Previous methods of extracting quantitative data from images of hair cells has been varied, and often adapted for a particular experimental outcome. For example, in hair cell survival studies it is often fastest to manually count hair cells in order to generate cochleograms[37]. While the accuracy that can be achieved in this task by a trained individual is the target of any algorithmic approach, the significant time cost of performing such analysis makes an automated solution desirable even with rare errors. Automated hair cell counting algorithms already exist but are less feature rich, limiting adoption. One relies on the homogeneity of structure in the organ of Corti and fails when irregularities, such as four rows of outer hair cells, are present[38]. Another may count hair cells but cannot differentiate between inner and outer hair cells[39].

The presence of a hair cell represents the most basic of quantification and therefore the quickest to do manually, although still represents a significant burden for large datasets. Image analysis of increasing complexity, such as counting cells above a fluorescent intensity threshold, as is common in the study of protein expression levels for example, take considerably longer to the point where they are no longer routinely performed over the entire cochlea[40-45]. Analysis of fluorescence levels within hair cells represents even greater complexity. While some of these analyses are being enabled more efficiently via general use software, such as ImageJ[46], the highly specialized geometry of hair cells limits the effectiveness of these nonspecific tools for use in an automated, unsupervised manner. Furthermore, to carry out the highly favorable, more sophisticated quantification of fluorescence in 3D cell volumes, discussed further below, requires volumetric segmentation of cells. *Cellpose*

represents the cutting edge of generalist cell segmentation, however, fails to achieve acceptable performance levels on our datasets.

The HCAT segmentation pipeline, presented here, offers fully automated volumetric segmentation of hair cells along the entire length of the cochlea. This task would be prohibitively labor intensive to carry out manually, however is carried out here at only marginally poorer segmentation performance when compared to manual segmentation, removing the most significant barrier to analyzing cochlear-wide fluorescence within large datasets. The most accurate measurement of the fluorescence intensity signal within the cell are, three-dimensional, volume-normalized measurements. It is common for manual analysis methodology to be performed on a two-dimensional projection of a three-dimensional micrograph, yet these projections can bias results. For example, a maximum projection, where any pixel is the maximum value of the entire z column of pixels, is increasingly biased with poorer signal-to-noise ratios. Alternatively, analysis on a summed projection, where any pixel is the summed value of all underlying z planes, may be particularly sensitive to hair cell orientation or morphology. The HCAT segmentation pipeline is a tool for the automated quantification of hair cell fluorescence from three-dimensional imaging data, which can be carried out on full or partial cochlear samples. Furthermore, in combination with the automated calculation of each hair cell's predicted best frequency in full cochleae, determined to be in good agreement with manual estimates of hair cell best frequency using the widely accepted EPL method, this enables the creation of hair cell fluorescence cochleograms.

While our models were trained solely using the Myo-VIIa and phalloidin labels, the model can perform on specimens labeled with other markers, provided they contain both cytosolic hair cell, and stereocilia, labels. An example of cytosolic hair cell labels might include (i) hair-cell-specific expression of a genetically encoded fluorescence marker, such as *Atoh1-eGFP;* (ii) cochlear samples collected from animals following AAV injection resulting in strong expression of fluorescence markers in hair cells; or (iii) cochlear samples treated with FM1-43 styryl dye known to selectively permeate into hair cells with functional mechanotransduction channel when briefly applied to live cochlea. An example of stereocilia-specific label is an immunolabeling against a highly enriched stereociliary protein espin, widely used to label hair cell stereocilia in adult cochlear preparations.

While our segmentation accuracy offers superior performance compared to other methods, this increase in segmentation accuracy comes at a cost of cell detection accuracy, as seen by the variability in the cochleogram seen in **Figure 5F**, likely due to the limitation in deep learning architecture complexity for volumetric analysis. Thus, in the HCAT software we offer two distinct pipelines for cell detection and segmentation analysis.

While our study shows that these algorithms perform well on data collected in-house, their generalizability across datasets collected by other research groups is pending further validation. Deep learning models tend to generalize poorly; when evaluated on community-supplied data, HCAT was no exception. While there were several instances of highly accurate performance on community-provided datasets, the deep learning models performed sub-optimally in cases where hair cells on community-supplied datasets were visually different than those in our training data. To our surprise, while the hair cell detection pipeline was trained exclusively with confocal microscopy data, it performed well on a set of community-provided images of hair cells collected using widefield fluorescence. All community-provided datasets that resulted in poor HCAT performance will be used to retrain the model with a prior permission from the owner(s) of the data set.

Overall, we are planning to expand the training data employing a wider variety of hair cells, such as those from adult mice and examples of cochlear coils with hair cells following a treatment with ototoxic drugs. Furthermore, we will endeavor to continually update and maintain the software as well as periodically update the machine learning model as the state of 3D cell segmentation and cell detection advances. Further development of this software in the future will provide support for automatic segmentation of adult cochlear tissue, often imaged in pieces rather than as a contiguous piece of tissue, due to dissection limitations. Additionally, we will continue to re-train the algorithm as more training data become available, improving its performance over time to a wider array of tissue preparations and ages.

To our knowledge, this is the first whole cochlear analysis pipeline capable of accurately and quickly detecting or segmenting hair cells. Offering state of the art performance, this hair cell analysis toolbox (HCAT) enables expedited cochlear image data analysis while maintaining high accuracy. This accurate and unsupervised data analysis approach will both facilitate ease of research and improve experimental rigor.

Materials and Methods

Some of the procedures described below, including the staining authors used to prepare cochlear samples labeled with four different fluorescent markers as shown in **Figure 1** are not required for the use of the presented hair cell analysis toolbox. They do, however provide an example of a dataset for which the single-cell fluorescence intensity quantification feature following the application of the hair cell segmentation pipeline can be utilized.

Sample preparation, confocal microscopy. Postnatal day (P) 0 *C57Bl6* mice of either sex were cryoanesthetized and injected with AAVs encoding eGFP through the round window membrane of the cochlea as described previously. The animals were then returned to the dam for recovery. Organs of Corti were dissected at P5 in Leibovitz's L-15 culture medium (21083-027, Thermo Fisher Scientific) and fixed in 4% formaldehyde for 1 hour. The samples were permeabilized with 0.2% Triton-X for 30 minutes and blocked with 10% goat serum in calcium-free HBSS for two hours. To visualize the hair cells, samples were labeled with an anti-Myosin VIIA antibody (#25-6790 Proteus Biosciences, 1:400) and goat anti-rabbit CF568 (Biotium). Additionally, samples were labeled with Phalloidin to visualize actin filaments (Biotium CF640R Phalloidin) and with DAPI to visualize cell nuclei (Molecular Probes DAPI, #D1306). Samples were then mounted on slides using ProLong® Gold Antifade Mounting kit (P36931, Thermo Fisher Scientific,) and imaged with a Leica SP8 confocal microscope (Leica Microsystems) using a 63x/1.3 NA objective. Confocal Z-stacks of 512x512 pixel images with an effective pixel size of 288 nm were collected using the tiling functionality of the Leica LASX acquisition software. All experiments were carried out in compliance with ethical regulations and approved by the Animal Care Committees of Massachusetts Eye and Ear.

Computational Environment: All scripts were run on a custom-built analysis computer running Ubuntu 20.04.1 LTS, an open-source Linux distribution from Canonical based on Debian. The workstation was equipped with an AMD Ryzen 7 3700X 8-Core processor with 64 GB of RAM. Deep learning computations were accelerated by a Nvidia 2080Ti graphics card with 11GB of DDR6 video memory. Many scripts were custom written in python 3.8 using open source scientific computation libraries including numpy[47], matplotlib[48], scikit-learn[49]. All deep learning architectures, training logic, and much of the data transformation pipeline was written in pytorch[50]. All code has been hosted on github and is available at: https://github.com/buswinka/hcat.

Training Data Annotation: Two deep learning models were trained on distinct datasets.

For the *hair cell segmentation task*, 17 training and 2 validation Z-stacks of cochlear hair cells were selected from different datasets and ensuring equal representation along the entire length of the cochlea, each containing ~50-100 hair cells. Each image was manually segmented using the Amira Software package (Thermo Fisher Scientific) running on a Lenovo ThinkPad X1 yoga touchscreen laptop. Cell outlines were delineated by hand using a stylus and saved with a unique cell ID to create cell masks.

For the *hair cell detection task* training data, bounding boxes for hair cells seen in maximum projected z-stacks were manually annotated in the labelImg software[51] and saved as an xml file. For whole cochlear cell annotation, a "human in the loop" approach was taken, first evaluating the deep learning model on the entire cochlea, then manually correcting errors.

Training: Each deep learning model was trained to accurately perform its designed task. The procedure to train the Faster R-CNN detection algorithm has been standardized and extensively documented in previous reports[6], while the training procedure for instance segmentation is more involved. A spatial embedding approach has been shown to be a versatile method to perform biomedical instance segmentation[16,17,21,52] , however the optimal way to train a network for this task is less well defined. We optimize spatial embeddings based on the distance of each pixel (i) in embedding space from the known centroid (C) of the underlying object. From this distance, each pixel (i) can be assigned a probability (ϕ) of belonging to an object (k) by equation (1), regularized by the parameter sigma (**σ**). Due to the anisotropy of the dataset, we elect to choose a proportionally smaller sigma when computing the loss in Z compared to X and Y, thereby improving embedding accuracy along that dimension. For the spatial embedding instance segmentation, the architecture outputs spatial embedding vectors in addition to a probability map. In the post processing of the spatial embedding, object detection heavily relies on tight clusters of pixels, therefore it is advantageous to penalize false negative results more heavily than false positive ones. This contrasts with the probability map, where false negative results should be more heavily penalized to ensure the detection of cell borders. To rectify these opposing needs, we chose to optimize the Tversky loss[53] of the spatial embedding and semantic probability map with differing penalties for each.

When using a probability map to aid in watershed segmentation, it is critical the borders of cells are well defined. To this end, we optimize binary cross entropy loss with a pixel-by-pixel penalty based on the distance from the border of an object, with pixels closer to the object penalized more heavily. Due to the class imbalance between background and foreground, optimizing over binary cross entropy alone leads the model to a local minimum where the entire output is predicted to be background. Therefore, we additionally include the dice coefficient in loss calculation.

For both tasks, the deep learning architectures were trained with the Adam optimizer with a learning rate starting at 1e-4 and decaying based on cosine annealing with warm restarts with a period of 100 epochs. When training the model for the spatial embedding task, we initialize sigma to be large values, and progressively decay by half at set epochs. Due to the labor-intensive nature of the manual segmentation workflow, we are limited in the total number of training volumes we can generate. In cases with a small number of training images, deep learning models tend to fail to generalize and instead “memorize” the training data. To avoid this, we make heavy use of image transformations which randomly add variability to our images and synthetically increase the variety of our training data sets[54] (**Supplementary Figure S1**).

Acknowledgements. We would like to thank Dr. Marcelo Cicconet (Image and Data Analysis Core at Harvard Medical School) and Haobing Wang, MS (Mass Eye and Ear Light Microscopy Imaging Core Facility) for their assistance in this project. We thank Dr. Guy Richardson, Dr. Corne Kros, Dr. Bradley Walters, Dr. Albert Edge, Dr. Yushi Hayashi, Dr. Lisa Cunningham, Dr. Mark Rutherford, Dr. Tejbeer Kaur, Dr. Vijayprakash Manickam, Dr. Ksenia Gnedeva, the members of their teams and all other research groups, for providing their datasets to evaluate the HCAT. We also thank Dr. Richard Osgood and Evan Hale, MS for critical reading of the manuscript.

Code availability.

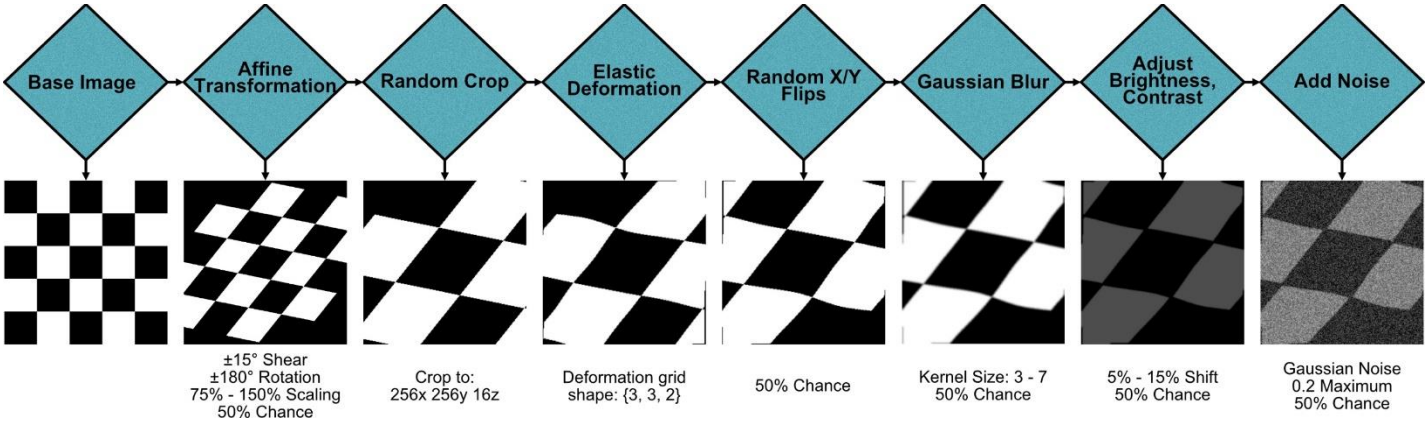
All code has been hosted on github and is available at: https://github.com/buswinka/hcat.

References

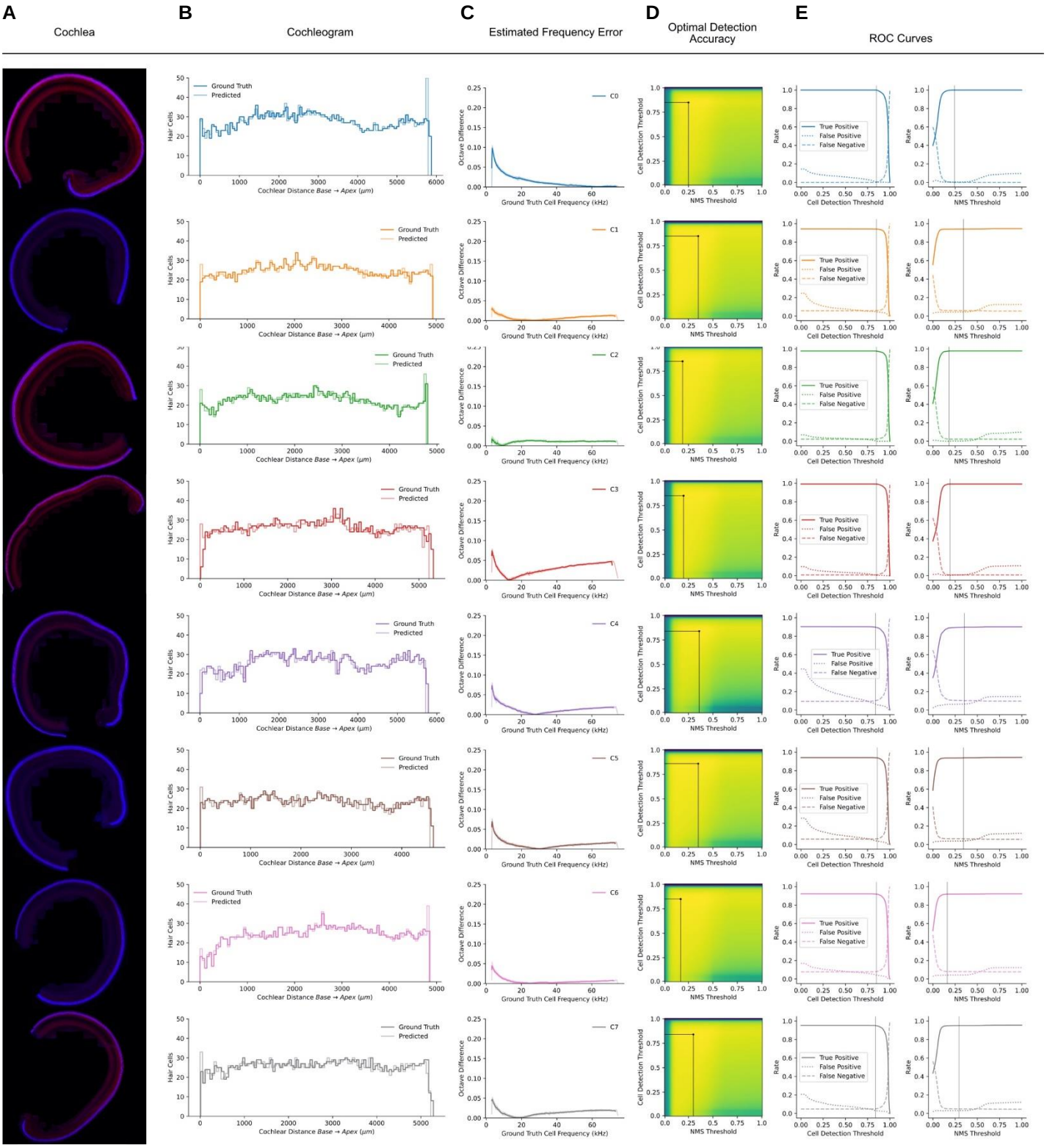
1. Ashmore J. Tonotopy of cochlear hair cell biophysics (excl. mechanotransduction). *Current opinion in physiology.* 2020;18:1-6.
2. Potter PK, Bowl MR, Jeyarajan P, et al. Novel gene function revealed by mouse mutagenesis screens for models of age-related disease. *Nature communications.* 2016;7(1):1-13.
3. Gray SJ, Foti SB, Schwartz JW, et al. Optimizing promoters for recombinant adeno-associated virus-mediated gene expression in the peripheral and central nervous system using self-complementary vectors. *Human gene therapy.* 2011;22(9):1143-1153.
4. Kalluri R, Shera CA. Distortion-product source unmixing: A test of the two-mechanism model for DPOAE generation. *The Journal of the Acoustical Society of America.* 2001;109(2):622-637.
5. Glasscock ME. *The ABR handbook : auditory brainstem response.* 2nd ed. ed. New York: Thieme Medical Publishers; 1987.
6. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence.* 2016;39(6):1137-1149.
7. Ezhilarasi R, Varalakshmi P. Tumor detection in the brain using faster R-CNN. Paper presented at: 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2018 2nd International Conference on2018.
8. Yang S, Fang B, Tang W, Wu X, Qian J, Yang W. Faster R-CNN based microscopic cell detection. Paper presented at: 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)2017.
9. Zhang J, Hu H, Chen S, Huang Y, Guan Q. Cancer cells detection in phase-contrast microscopy images based on Faster R-CNN. Paper presented at: 2016 9th international symposium on computational intelligence and design (ISCID)2016.
10. Strahler AN. Quantitative analysis of watershed geomorphology. *Eos, Transactions American Geophysical Union.* 1957;38(6):913-920.
11. Atta-Fosu T, Guo W, Jeter D, Mizutani CM, Stopczynski N, Sousa-Neves R. 3D Clumped Cell Segmentation Using Curvature Based Seeded Watershed. *Journal of imaging.* 2016;2(4):31.
12. Jones TR, Kang IH, Wheeler DB, et al. CellProfiler Analyst: data exploration and analysis software for complex image-based screens. *BMC bioinformatics.* 2008;9(1):482-482.
13. Beliën JAM, van Ginkel HAHM, Tekola P, et al. Confocal DNA cytometry: a contour-based segmentation algorithm for automated three-dimensional image segmentation. *Cytometry (New York, NY).* 2002;49(1):12-21.
14. Xiaowei C, Xiaobo Z, Wong STC. Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy. *IEEE transactions on biomedical engineering.* 2006;53(4):762-766.
15. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. Paper presented at: Proceedings of the IEEE international conference on computer vision2017.
16. Neven D, Brabandere BD, Proesmans M, Gool LV. Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition2019.
17. Stringer C, Wang T, Michaelos M, Pachitariu M. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods.* 2021;18(1):100-106.
18. Rasmussen CE. Gaussian processes in machine learning. Paper presented at: Summer school on machine learning2003.
19. Greenwood DD. A cochlear frequency-position function for several species—29 years later. *The Journal of the Acoustical Society of America.* 1990;87(6):2592-2605.
20. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition2016.
21. Zhang D, Chun J, Cha SK, Kim YM. Spatial semantic embedding network: fast 3D instance segmentation with deep metric learning. *arXiv preprint arXiv:200703169.* 2020.

22. Yao C, Tang J, Hu M, et al. Claw U-Net: A Unet-based Network with Deep Feature Concatenation for Scleral Blood Vessel Segmentation. *arXiv preprint arXiv:201010163.* 2020.
23. Qamar S, Jin H, Zheng R, Ahmad P, Usama M. A variant form of 3D-UNet for infant brain segmentation. *Future Generation Computer Systems.* 2020;108:613-623.
24. Wang C, MacGillivray T, Macnaught G, Yang G, Newby D. A two-stage 3D Unet framework for multi-class segmentation on full resolution image. *arXiv preprint arXiv:180404341.* 2018.
25. Weigert M, Schmidt U, Haase R, Sugawara K, Myers G. Star-convex polyhedra for 3d object detection and segmentation in microscopy. Paper presented at: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision2020.
26. Schmidt U, Weigert M, Broaddus C, Myers G. Cell detection with star-convex polygons. Paper presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention2018.
27. Sheridan A, Nguyen T, Deb D, et al. Local shape descriptors for neuron segmentation. *bioRxiv.* 2021.
28. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Paper presented at: International Conference on Medical image computing and computer-assisted intervention2015.
29. Lucchi A, Smith K, Achanta R, Knott G, Fua P. Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. *IEEE transactions on medical imaging.* 2011;31(2):474-486.
30. Pohlen T, Hermans A, Mathias M, Leibe B. Full-resolution residual networks for semantic segmentation in street scenes. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition2017.
31. Dumoulin V, Visin F. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:160307285.* 2016.
32. Xu B, Wang N, Chen T, Li M. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:150500853.* 2015.
33. Sun T, Tsang WM, Park WT, Cheng KJ, Merugu S. Modeling in vitro neural electrode interface in neural cell culture medium. *Microsyst Technol.* 2015;21(8):1739-1747.
34. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. Paper presented at: European conference on computer vision2016.
35. Stalling D, Westerhoff M, Hege H-C. Amira: A highly interactive system for visual data analysis. *The visualization handbook.* 2005;38:749-767.
36. Neubeck A, Van Gool L. Efficient non-maximum suppression. Paper presented at: 18th International Conference on Pattern Recognition (ICPR'06)2006.
37. Viberg A, Canlon B. The guide to plotting a cochleogram. *Hearing Res.* 2004;197(1-2):1-10.
38. Urata S, Iida T, Yamamoto M, et al. Cellular cartography of the organ of Corti based on optical tissue clearing and machine learning. *eLife.* 2019;8.
39. Cortada M, Sauteur L, Lanz M, Levano S, Bodmer D. A deep learning approach to quantify auditory hair cells. *Hearing Res.* 2021;409:108317.
40. Li H, Liu H, Corrales CE, et al. Differentiation of neurons from neural precursors generated in floating spheres from embryonic stem cells. *BMC neuroscience.* 2009;10(1):122-122.
41. Garza LA, Yang C-C, Zhao T, et al. Bald scalp in men with androgenetic alopecia retains hair follicle stem cells but lacks CD200-rich and CD34-positive hair follicle progenitor cells. *The Journal of clinical investigation.* 2011;121(2):613-622.
42. György B, Sage C, Indzhykulian AA, et al. Rescue of Hearing by Gene Delivery to Inner-Ear Hair Cells Using Exosome-Associated AAV. *Mol Ther.* 2017;25(2):379-391.
43. Rhee C-K, He P, Jung JY, et al. Effect of low-level laser treatment on cochlea hair-cell recovery after ototoxic hearing loss. *Journal of biomedical optics.* 2013;18(12):128003-128003.
44. Gersten BK, Fitzgerald TS, Fernandez KA, Cunningham LL. Ototoxicity and Platinum Uptake Following Cyclic Administration of Platinum-Based Chemotherapeutic Agents. *Journal of the Association for Research in Otolaryngology.* 2020;21(4):303-321.

45. Liu Z, Owen T, Fang J, Srinivasan RS, Zuo J. In vivo notch reactivation in differentiating cochlear hair cells induces sox2 and prox1 expression but does not disrupt hair cell maturation. *Developmental Dynamics.* 2012;192(4):684-696.
46. Rasband WS. ImageJ. Bethesda, MD; 1997.
47. Van Der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. *Computing in science & engineering.* 2011;13(2):22-30.
48. Hunter JD. Matplotlib: A 2D graphics environment. *IEEE Annals of the History of Computing.* 2007;9(03):90-95.
49. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research.* 2011;12:2825-2830.
50. Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:191201703.* 2019.
51. Lin T. LabelImg. Github; 2015.
52. Sheridan A, Nguyen T, Deb D, et al. Local Shape Descriptors for Neuron Segmentation. *bioRxiv.* 2021:2021.2001.2018.427039.
53. Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. Paper presented at: International workshop on machine learning in medical imaging2017.
54. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *Journal of Big Data.* 2019;6(1):1-48.



Supplementary Figure S1: Training data augmentation pipeline. Training images for each deep learning approach underwent identical data augmentation steps, increasing the variability of our dataset and improving performance. Each of these augmentation steps were probabilistically applied sequentially (left to right).



Supplementary Figure S2: Validation of hair cell detection analysis and location estimation pipeline. Whole cochlear turns (*A*) were manually annotated and evaluated with the HCAT detection analysis pipeline. Each analysis generated highly accurate cochleograms (*B*), reporting the 'ground truth' result obtained from manual segmentation (*dark lines*) superimposed onto the cochleogram generated from hair cells detected by the HCAT detection analysis (*light lines*), reporting highly accurate results. The frequency estimation error was then calculates as an octave difference to quantify the accuracy of predicted frequency estimation for every hair cell vs their ground truth frequency (*C*). Optimal cell detection and non-maximum suppression thresholds were discerned via a grid search by maximizing the true positive rate penalized by the false positive and false negative rates (*D*). Black lines on the ROC curves (*E*) denote the optimal hyperparameter value.