Leveraging foundation models and goal-dependent annotations for automated cell confluence assessment

Maximilian Josef Joas[1], Daniel Freund[2], Robert Haase[1], Erhard Rahm[1] and Jan Ewald[1]*

[1]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, Leipzig University, Leipzig, Germany, [2]Center for Regenerative Therapies, Dresden University of Technology, Dresden, Germany

Mesenchymal stem cell therapy shows promising results for difficult-to-treat diseases, but manufacturing requires robust quality control through cell confluence monitoring. While deep learning can automate confluence estimation, research on cost-effective dataset curation and the role of foundation models in this task is limited. We investigate effective strategies for AI-based confluence estimation by studying active learning, goal-dependent labeling, and foundation models that require no training or labeling effort (zero-shot). Here, we show that zero-shot inference with the Segment Anything Model (SAM) achieves excellent confluence estimation without any task-specific training, outperforming even fine-tuned and specialized models. Moreover, our findings demonstrate that active learning does not significantly enhance training and performance compared to the random selection of training samples in homogeneous cell datasets. We demonstrate that streamlined labeling approaches tailored to specific goals yield results comparable to those of exhaustive, time-consuming annotation methods. Our results challenge common assumptions about dataset curation and model training: neither active learning nor extensive fine-tuning provided significant benefits in our real-world scenarios. Instead, we found that leveraging SAM's zero-shot capabilities and goal-dependent labeling offers the most cost-effective approach for AI-based confluence monitoring. Our study provides practical guidelines for implementing automated cell quality control in MSC manufacturing, demonstrating that extensive dataset curation may be unnecessary when foundation models can effectively handle the task right out of the box.

KEYWORDS

active learning, deep learning, cell segmentation, segment anything model, computer vision

1 Introduction

Mesenchymal stem/stromal cells (MSCs) are powerful Advanced Therapy Medicinal Products (ATMPs) that can treat various conditions. Although MSCs are not yet approved for many applications, they demonstrate promising clinical outcomes in the treatment of degenerative inflammatory diseases, autoimmune disorders, tissue injuries, and chronic degenerative ailments (Strecanska et al., 2024; Galipeau and Sensébé, 2018). This is particularly significant for conditions such as rheumatic arthritis, which affects around 18 million people worldwide, imposes significant health burdens, and lacks sufficient alternative treatments (Shimizu et al., 2023; IHME, 2020).

MSC production, however, relies on limited and hard-to-access sources such as bone marrow or umbilical cord tissue, requiring *ex-vivo* expansion of these cells. During this expansion in cell cultures, the density of the cells (confluence) must be tightly controlled, as it serves as a trigger point for cell differentiation (Fernández-Santos et al., 2022; Kim et al., 2017). Therefore, scientists and laboratory technicians need to harvest the cells before they differentiate and lose their potency as stem cells. Thus, scientists and technicians monitor growth through imaging to ensure quality and optimize yield.

During MSC production, microscopic images are often analyzed manually, and scientists estimate the confluence, i.e., the fraction of area covered by cells. This erroneous and non-standardized process leads to increased labor efforts and fails to optimize yield by missing the optimal harvest point. For AI-based confluence estimation, we need: (a) a method for cell segmentation in live-cell images and (b) data with ground truth, particularly images with known confluence and segmented cells as labels. Traditional image processing techniques include thresholding methods (Shen et al., 2018), such as edge detection, and region-growing approaches (Panagiotakis and Argyros, 2018). However, advances in artificial intelligence (AI) have shown that AI models often outperform traditional methods for cell segmentation (Chen and Murphy, 2023). A typical approach to training AI models involves collecting data, labeling it, and training models from scratch, such as a U-Net model for segmentation (Ronneberger et al., 2015). Recently, pre-trained and large generalist foundation models have gained popularity due to their good performance and broad applicability (Han et al., 2021; Chen and Murphy, 2023). In computer vision, several such models have been developed, including generalists for image segmentation [SAM (Kirillov et al., 2023), Detectron2 (Wu et al., 2019)] and specialists for cell segmentation [Cellpose (Stringer et al., 2020), LiveCell (Edlund et al., 2021)]. Although training custom models has become more accessible to end-users without large computing resources (von Chamier et al., 2021), the main advantage of pre-trained models is that they can be used with little to no labeled images. This is especially important since human labeling is costly and time-consuming. Moreover, MSCs, like other underrepresented cells in live-cell datasets, are non-round and irregularly shaped, making the segmentation task challenging. Additionally, ATMP manufacturing processes prohibit staining for higher contrast. Consequently, generating a sufficiently detailed and diverse training dataset for custom model training would require significant effort, highlighting the utility of pre-trained models.

With both foundation models and untrained models, existing approaches naturally raise the question of how to estimate confluence with minimal labeled data while maintaining sufficient performance. With a large number of images, or data, available, there are three strategies for labeling and applying AI models: (a) zero-shot, which involves no labeling or model training, (b) using all images to label and train models, and (c) active learning, where the n most informative samples are selected for labeling and training. AL is a data-centric approach that reduces the effort of data labeling by selecting the next datapoint(s) not randomly, but based on either uncertainty, diversity, or clustering (Monarch, 2021). Interestingly, using AL to choose only a core set of the entire

dataset can yield results similar to or even better than labeling all the data (Jafari et al., 2024). For our model-driven approach, we focus on uncertainty-based methods due to their broad applicability across various models.

Since current research on cell segmentation (Chen and Murphy, 2023) and AL (Sayin et al., 2021; Monarch, 2021) focus on large datasets in theoretical contexts; we aim to apply AL to real-world small datasets using state-of-the-art models for confluence estimation. In our study, we are interested in four insights:

1. Impact and effectiveness of active learning.
2. Applicability of simplified goal-dependent, i.e., "lazy", labeling.
3. Active learning selection patterns in a time-resolved cell culture, "movie context."
4. Performance of zero-shot inference.

To gain these insights, we describe in the following section (*cf.* Section 2) our experimental setup, which includes three datasets for comparing learning strategies across four models. The selected models range from U-Net (Ronneberger et al., 2015), developed from scratch, to large generalist foundation models such as Detectron2 (Wu et al., 2019) and Meta's Segment Anything Model (SAM) (Kirillov et al., 2023), as well as the specialist pre-trained model Cellpose (Stringer et al., 2020). By leveraging our experiments, which are graphically summarized in Figure 1, we elaborate on the listed insights point by point in Section 3 and describe their impact, along with their limitations, in the final Section 4.

2 Material and methods

In the following data, we describe models and experiments. To ensure full transparency and reproducibility, we provide all relevant code and scripts in our GitLab repository (https://git.informatik. uni-leipzig.de/joas/confluence-unet).

2.1 Data

We utilized three datasets, one of which is labeled with two strategies, resulting in four datasets for model training. Additionally, one dataset was derived from a larger published live-cell imaging dataset ["lc-external" (Edlund et al., 2021)], while three originated from our lab ("internal"). One of these internal datasets contains live-cell imaging data obtained using a CytoSmart Lux microscope (10x magnification; 5-megapixel camera). This dataset was labeled in a standard manner ("lc-internal") and with goal-dependent labeling ("lc-internal-lazy") for a direct comparison. In the original LC-internal dataset, each cell was labeled individually, whereas in the "lazy" labeled set, cohesive clusters of cells were labeled as single objects. All images had dimensions of 1280x960 pixels (see Table 1). An additional internal dataset contains standard microscopy ("sc-internal") images acquired with a ZEISS Axiovert 40 CFL microscope [10x objective Ph1 (phase contrast); Axiocam ERc 5s, 5-megapixel camera]. The images in this dataset measured 512x512 pixels. We filtered the external dataset for the A172 cell line (which has morphological similarities to MSC cells
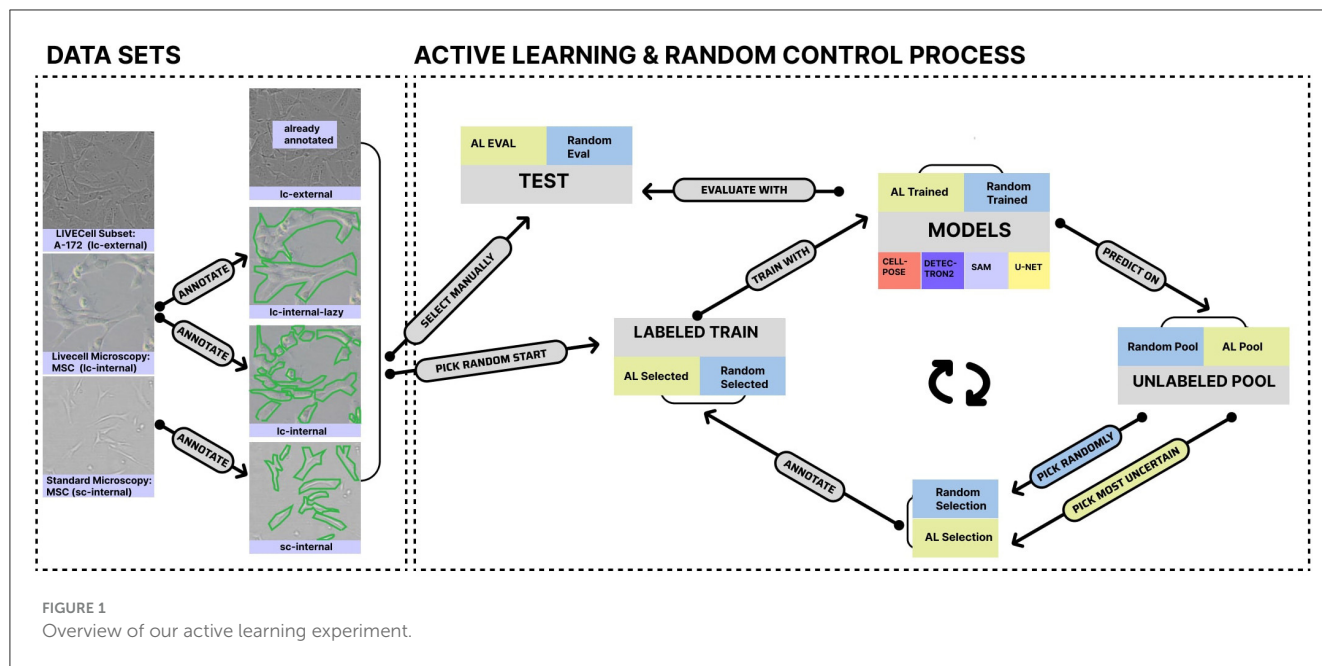
FIGURE 1
Overview of our active learning experiment.

TABLE 1 Dataset characteristics showing the number of regions of interest (ROIs), number of images, and image dimensions for each subset.

Dataset	#ROIs	#Images	Image size
lc-internal-train	2,243	16	$1,280 \times 960$
lc-internal-test	400	2	$1,280 \times 960$
lc-internal-train-lazy	897	16	$1,280 \times 960$
lc-internal-test-lazy	224	4	$1,280 \times 960$
lc-external-train	70,758	333	702×520
lc-external-test	30,005	138	702×520
sc-train	2,467	12	512×512
sc-test	446	3	512×512

despite its origin from glioblastomas) to roughly match the cell shapes of the other datasets. Table 1 provides an overview of our dataset, including the number of regions of interest (ROIs), the number of images, and image dimensions. An ROI is defined as a single labeled cell in an image.

We annotated the internal datasets using ImgLab[1] following instructions from wet lab scientists and obtained the annotations in the COCO JSON format (Lin et al., 2014). Since U-Net and Cellpose require instance masks instead of COCO JSON files, we converted the files to masks with custom scripts (https://git.informatik.uni-leipzig.de/joas/confluence/-/blob/main/utils/coco_to_mask.py?ref_type=heads). We divided each dataset into a training set and a testing set to evaluate model performance. To capture a variety of characteristics, we used a combination of datasets, as summarized in Table 1. The LC-external dataset, drawn from the LIVECell paper (Edlund et al., 2021), is the largest, containing the most ROIs. The LC-internal dataset includes full

and "lazy" subsets and offers the highest resolution. Finally, the sc-internal dataset provides additional data, but with significantly lower resolution and fewer images than the other datasets. Figure 1 shows sample images and annotations for each dataset.

2.2 Active learning for dataset curation

Upon processing the dataset, we manually selected the test set for the internal datasets to ensure that images with approximately 50% confluence were included. This confluence range is the most critical in production, making it essential for evaluating model performance. For AL, we defined a pool set, which can be selected for training, and an initial training set, chosen randomly with a size ranging from two to ten, depending on the total number of images. Selected images in each round were transferred from the pool to the overall train set (either physically on our machine or by filtering the COCO JSON file). We subsequently trained our four models using these initial training sets and in all other rounds for 100 epochs. Figure 1 visually represents this process, and all models were trained on NVIDIA A100 GPUs. To evaluate the performance of our model, we compared the predicted masks with the actual ground-truth masks. We did this by calculating the IoU (Intersection over Union) for each image. IoU measures the overlap between the predicted and actual masks in relation to their combined area. We calculated the mean IoU, standard deviation, and interquartile range across all images in the test set. Additionally, as a second criterion inspired by use cases, we calculated the absolute differences in the confluence between ground truth and predictions.

For our uncertainty-based active learning approach, we created probability maps for all images in the dataset (excluding the Detectron2 model). We then calculated the Shannon entropy on these probability maps to quantify uncertainty. Images in the pool set were ranked according to their entropy values, with those having

---

1 https://github.com/NaturalIntelligence/imglab

the highest entropy (indicating the greatest uncertainty) selected for the next training iteration. To accommodate varying dataset sizes, we chose one image per round from internal datasets and ten images per round from the larger external dataset.

This approach, known as entropy-based sampling (Monarch, 2021; Yin et al., 2023), is a widely used technique in uncertainty-based AL (Zhu et al., 2008). We iteratively retrain the model in each round until all images are transferred from the pool set to the training set. After every training iteration, we assess the model on the predefined test set. To make a direct comparison, we trained another set of models using randomly chosen images. These models adhered to the same training protocol as our uncertainty-based approach, but without utilizing entropy for image selection. This results in ten rounds for the standard microscopy dataset, 14 rounds for the LC-internal datasets, and 34 rounds for the LC-external dataset.

To account for randomness in the image selection, we repeated the previously described process ten times. We aggregated the evaluation metrics from each round across all ten experiments by calculating the mean and the interquartile range of these metrics. To assess whether AL significantly improves training, we compared the random selections with the AL selections and conducted the Mann-Whitney U-test (Mann and Whitney, 1947) for the means of the evaluation metrics in each round across all ten experiments.

2.3 Goal-dependent labeling

Most segmentation tasks are sensitive to the shape of the object. In adherent cell cultures, cells often grow in close proximity to one another or overlap, forming complex shapes that resemble blobs or clusters. Labeling such blobs of cells requires much less effort ("lazy") than labeling each cell individually. Since confluence estimation necessitates full cell segmentation, we hypothesized that it is sufficient to label blobs of cells instead of labeling them separately. Therefore, we compared the model performance between the lc-internal and lc-internal-lazy datasets.

2.4 Active learning in a microscopy movie context

The images from the "sc-internal" dataset originated from a time-resolved microscopy movie. This means that the later the images were captured in the movie, the more the cells have grown (higher confluence). Therefore, we expect that AL may select images with varying levels of confluence, and subsequently, the movie positions influence the selection process. Based on the results from the AL experiment, we calculated the number of positions at which the selected image differs from the image chosen in the previous step. Again, we aggregated these differences by calculating the mean and determining the interquartile range across all ten random runs. As a control, we compared the image selected by AL with the randomly selected image. We tested for significance using the Mann-Whitney U Test (Mann and Whitney, 1947).

2.5 Fine-tuning compared to zero-shot learning

We evaluated the necessity of fine-tuning by first using all models without additional training and directly feeding the test set into them for inference (zero-shot). In contrast, we fine-tuned the models with all available labeled data, training them for 500 epochs (without early stopping). We performed inference on datasets and their test sets. As before, we used IoU and the absolute delta in confluence as evaluation metrics. To provide better context for the models' performance, we included a baseline confluence detector that does not incorporate modern deep learning. This baseline algorithm processes grayscale images and detects edges with the Canny edge detector (Canny, 1986). It then fills gaps in the detected edges using binary hole-filling. Subsequently, it removes small objects and detects contours in the processed image with the "marching squares" algorithm, finding iso-valued contours at a specific level. After detecting the contours, any open contours are closed and simplified using polygonal approximation. This baseline algorithm then draws these contours onto a blank mask and interpolates between them to create a filled mask, which can be used to calculate metrics similarly to the other models.

2.6 Models

2.6.1 Cellpose

We utilized the Cellpose model from Stringer et al. (2020), based on version 3.0.0, and added functionality to enable custom names for standard Cellpose log files. This modification was reviewed and incorporated into Cellpose's code by the authors. We utilized the train function from the Cellpose model with the model type "cyto" for fine-tuning. Cellpose relies on the mean cell diameter as input. Therefore, we calculated the mean cell diameter with a custom script (https://git.informatik.uni-leipzig.de/joas/confluence/-/blob/main/cellpose_main.py?ref_type=heads) using our data. For inference, Cellpose requires two key thresholds: the cell probability threshold and the flow threshold. The cell probability threshold determines the minimum probability for pixels to be classified as part of a cell, while the flow threshold controls the tolerance for errors in detecting cells (Stringer et al., 2020). We observed that the model's performance is sensitive to these thresholds; thus, we implemented an automatic tuning process to optimize them for our training set. We developed a method that systematically explores different threshold combinations. The function evaluates the model's performance on the training set using ground-truth masks as a reference. It iteratively tests a range of flow thresholds (from 0 to 3) and cell probability thresholds (from -6 to 6) in 0.5 steps to identify the optimal combination (IoU score). Additionally, a penalty is imposed if no cells are detected in the predicted masks. This ensures that the model does not optimize for precision and only outputs the background without any masks.

Besides the thresholds, we did not alter any other hyperparameters from the default values of the Cellpose model

class and the Cellpose train method. For the AL part, we obtained the cell probabilities directly from the Cellpose model's `eval` method and calculated the Shannon entropy for each pool image. Furthermore, we trained all models in the AL experiments for 100 epochs.

2.6.2 Detectron2

We employed the Detectron2 framework to perform instance segmentation as published by Meta Research (Wu et al., 2019). The model was configured using a Mask R-CNN architecture with a ResNet-50 backbone and Feature Pyramid Network (FPN) as specified in the `mask_rcnn_R_50_FPN_3x.yaml` configuration file. The model was designed to detect a single class, corresponding to the cells in our images. We configured the model with a base learning rate of 0.00025 and a batch size of 128 images for the region of interest (ROI) heads. To ensure reproducibility, we established a fixed random seed. The model was trained for 10 iterations, with a 5-iteration warmup period. For systems without GPU acceleration, the model defaults to CPU computation automatically.

To determine image selection in the AL process, we cannot obtain the probability masks directly from Detectron2. Instead, the model provides pre-calculated confidence scores for each mask. We use the average of these scores as a measure of uncertainty. In the inference step, we select the image(s) with the lowest score to label next.

2.6.3 Segment anything model

We extended SAM, which does not natively support fine-tuning, by constructing a custom module to enable this functionality, as described in detail in a dedicated blog post (see footnote[2]) that makes fine-tuning SAM publicly accessible. Our results are based on SAM version 1.0.0 (Kirillov et al., 2023).

We developed a custom wrapper class (`ModelSimple`) around SAM's architecture to enable supervised training on our datasets. The key innovation in our approach was selectively freezing specific components of the network while allowing others to be updated during backpropagation. Specifically, we froze the image encoder and prompt encoder parameters to preserve the model's pre-trained feature extraction capabilities while making only the mask decoder trainable. This strategy significantly reduced computational requirements, allowing the model to adapt to our specific segmentation task.

Since cell segmentation lacks predefined spatial locations, we adjusted the standard SAM inference pipeline to function without explicit prompts or bounding boxes. Our implementation directly processes the input images to generate segmentation masks, eliminating the need for user interaction or a predefined ROI. We maintained SAM's native input resolution of 1,024 × 1,024 pixels to leverage the generation of probability mask maps that were upsampled to match the original image dimensions. To optimize our approach, we

employed a combined loss function as specified in the original SAM paper:

$$\mathcal{L}_{total} = 20 \cdot \mathcal{L}_{focal} + \mathcal{L}_{dice} \qquad (1)$$

The model was trained using the Adam optimizer, and to ensure reproducibility, we set the random seed to 100 for all random operations (Kirillov et al., 2023). For AL, we calculated the Shannon entropy of the predicted probability maps and selected the image(s) with the highest entropy.

2.6.4 U-Net

We implemented the U-Net architecture according to the original design proposed by Ronneberger et al. (2015). Our implementation includes a symmetric encoder-decoder structure with skip connections to maintain spatial information throughout the network. The encoder pathway consists of four down-sampling blocks, each containing two 3 × 3 convolutional layers with ReLU activation, followed by a 2 × 2 max pooling operation. The number of feature channels doubles at each down-sampling step, starting with 64 channels after the initial convolution and expanding to 128, 256, 512, and finally 1,024 channels at the bottleneck (or 512 when using bilinear up-sampling). The decoder pathway mirrors the encoder with four up-sampling blocks. The final layer consists of a 1×1 convolution that translates the 64-channel feature map into the desired number of output classes, resulting in pixel-wise classification for the segmentation mask. In our cell segmentation task, the network produces a single-channel probability map that indicates the likelihood of each pixel belonging to a cell.

For AL, we used the output probability maps directly from the U-Net model. We calculated the Shannon entropy of these predicted probability maps to select the images with the highest entropy, or uncertainty.
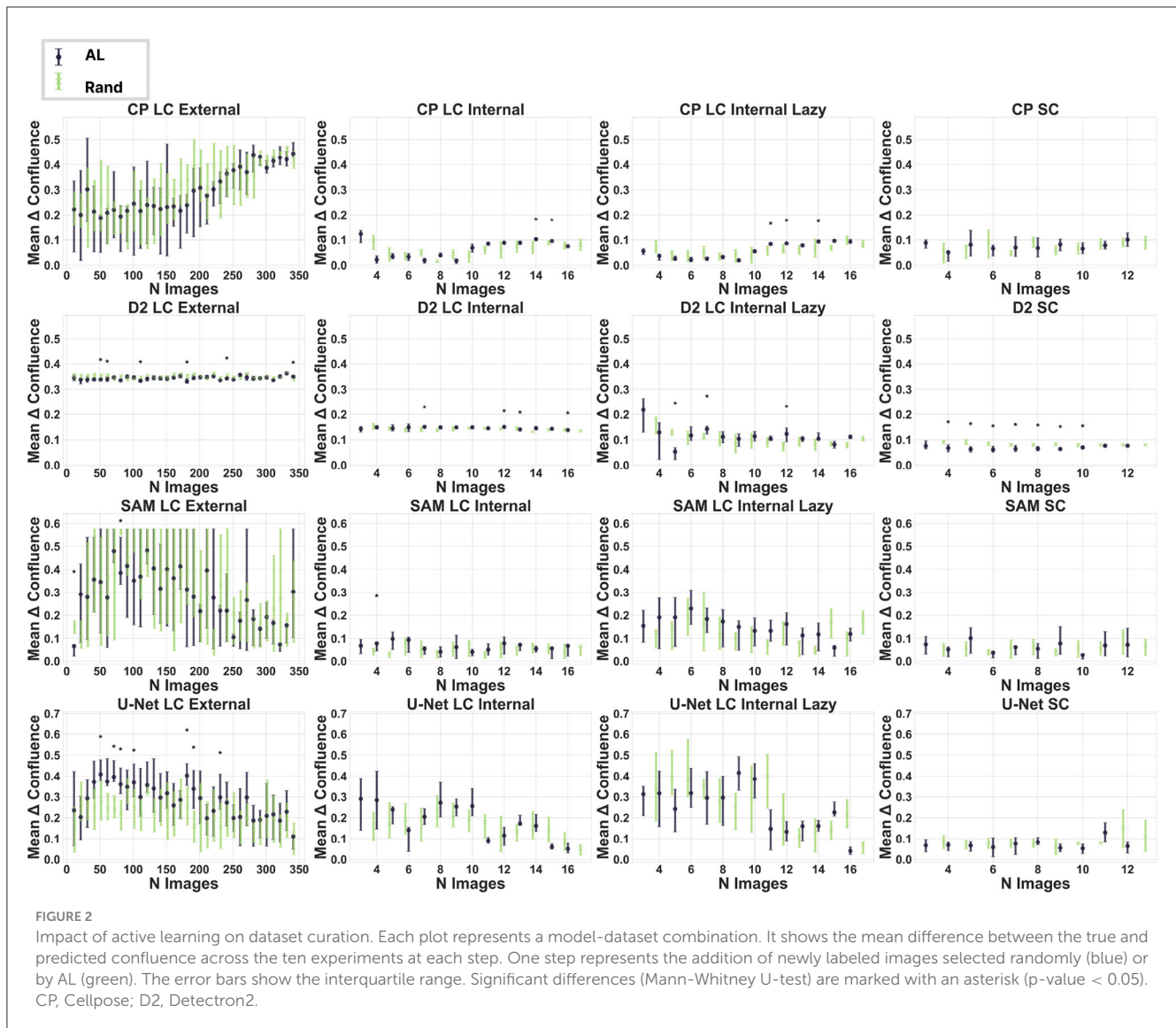
3 Results

3.1 Active learning performance

We analyzed whether AL is an effective approach for improving cell segmentation and reducing labeling effort when using four commonly used models for segmentation. We demonstrate that uncertainty-based AL provides no improvement in confluence prediction performance in our chosen datasets. The experiments with Cellpose and SAM exhibited the smallest differences between random and AL image selection, with only four (Cellpose) and three (SAM) statistically significant steps out of 72 total steps (*cf.* Figure 2, Supplementary Figure S2) We observe significantly better performance from the U-Net model when using a random dataset curation approach on the external dataset. Detectron2 is the only model where AL improved confluence prediction in the standard microscopy dataset and, to some extent, in the external dataset.

We hypothesized that both specialized and generalist pretrained segmentation models benefit from fine-tuning, given the unique and complex shape of MSC cells. Surprisingly, the impact of fine-tuning is limited and largely depends on the dataset. In seven experiments, we observed the best scores during the early stages of fine-tuning in AL, indicating that more data does not

---

2   https://maxjoas.medium.com/finetune-segment-anything-sam-for-images-with-multiple-masks-34514ee811bb

FIGURE 2
Impact of active learning on dataset curation. Each plot represents a model-dataset combination. It shows the mean difference between the true and predicted confluence across the ten experiments at each step. One step represents the addition of newly labeled images selected randomly (blue) or by AL (green). The error bars show the interquartile range. Significant differences (Mann-Whitney U-test) are marked with an asterisk (p-value < 0.05). CP, Cellpose; D2, Detectron2.

always lead to improved results. Our control experiments, in which random picks were used as the next image for labeling and training, yielded similar results, with six of the best performances occurring in the first half of the fine-tuning. From a model perspective, we observe that Detectron2 does not benefit from fine-tuning, Cellpose gets even worse, and for SAM and U-Net, we do not see a clear trend.

Specifically, nine out of 16 experiments (model and dataset combinations) achieved a minimal absolute delta in confluence of no more than 0.05, while three experiments achieved a minimal absolute delta in confluence of no more than 0.10. When comparing the mean performances across all datasets, SAM predicts the confluence most accurately, with an absolute delta value of 0.05 ± 0.02, while Detectron2 shows an absolute delta value of 0.15 ± 0.13. Performance analysis across datasets indicates that the goal-dependent labeled dataset ("lazy") achieves the best results (mean 0.04 ± 0.02), whereas models perform worst on the external dataset (mean 0.18 ± 0.11). Table 2 provides a detailed aggregation of the best results across models and datasets. Furthermore,

non-active learning (randomized) exhibits similar trends (*cf.* Supplementary Table S1). While cf. Supplementary Table S2 shows the absolute delta in confluence for all experiments, we observe similar trends when using IoU as a performance metric (*cf.* Supplementary Tables S4, S5).

3.2 Goal-dependent labeling

For Confluence prediction, accurately segmenting individual cells is unnecessary because a foreground/background classification would suffice. Therefore, we investigated whether faster goal-dependent labeling of cell clusters as single clusters ("lazy labeling") affects model performance. The results achieved through goal-dependent labeling do not universally enhance the performance of every model. We observe a clear trend with the Detectron2 model, where ten out of 14 steps show significant improvement with the lazy labeling method. In contrast, for U-Net and Cellpose, there is no evident difference. Interestingly, the SAM model demonstrates

a) Model aggregation		
Model	Min	Convergence point
Cellpose	0.07 ± 0.08	0.29 ± 0.20
Detectron2	0.15 ± 0.13	0.50 ± 0.37
SAM	0.05 ± 0.02	0.56 ± 0.41
U-Net	0.06 ± 0.03	0.95 ± 0.11
b) Dataset aggregation		
Dataset	Min	Convergence point
lc-external	0.18 ± 0.11	0.41 ± 0.45
lc-internal	0.06 ± 0.05	0.75 ± 0.29
lc-internallazy	0.04 ± 0.02	0.63 ± 0.40
sc-internal	0.05 ± 0.01	0.50 ± 0.34

The column *Convergence Point* represents the fraction of Active Learning iterations completed, where 0 indicates no additional data added, and 1 signifies that all available data has been included.

even better results for the precisely labeled images in terms of confluence estimation.

The IoU is significantly higher by a large margin (see Figure 3) when data are labeled lazily for all models, indicating that this metric is not invariant to labeling strategy, cell size, and cluster shapes.

3.3 Active learning in a microscopy movie context

We expected that uncertainty-based AL would ideally choose images with varying confluence, which sequentially increases in our standard microscopy dataset recorded as a movie of a cell culture. On the contrary, we observe no consistent differences or tendencies in the movie positions of the selected images between the AL and random selection processes across our four segmentation models (see Figure 4). This suggests that cell density is not a significant factor in the model's uncertainty, likely because cell shape remains relatively unchanged over time. Furthermore, the results align with the observation that AL does not significantly improve model performance for cell segmentation and confluence prediction, as shown in Figure 2.

3.4 Zero-shot inference or full fine-tuning

Since we observed that models exhibited mixed behavior during fine-tuning, such as Cellpose experiencing a decrease in performance, we analyzed the zero-shot capability of the models. The performance of SAM has an absolute delta of 0.05 $\pm$ 0.036 in confluence estimation, which is nearly perfect across all datasets, even without fine-tuning. As expected, deep learning-based approaches significantly outperform the algorithmic image segmentation baseline (see Figure 5, Supplementary Figure S2). We

observe a significant performance improvement for Detectron2 when fine-tuning on our internal datasets. However, we observe a decline in performance when fine-tuning on the external dataset. Additionally, we see small dataset-dependent fluctuations for the Cellpose model.

In summary, our results indicate a strong indication that, for the confluence prediction of MSC-like cells, generalist foundation models, such as SAM, outperform specialized models, such as Cellpose, and fine-tuning is unnecessary. Furthermore, in the case of Cellpose, results indicate that fine-tuning with irregular cell shapes (MSC) may result in decreased performance rather than the expected improvements.

3.5 Qualitative analysis and usability

When deciding how to label a dataset for confluence prediction and which model to choose, many practical considerations arise beyond just performance. We will provide insights on (a) the difficulty of fine-tuning the models, (b) implementing an uncertainty-based AL approach, and (c) overall computational considerations.

Detectron2 (Wu et al., 2019) is the simplest model to fine-tune, as fine-tuning is a built-in feature. The documentation for Detectron2 is clear and easy to follow. Detectron2 supports input data in COCO annotations, which is a widely used format. While it is easy to use, it offers less customizability and control. Additionally, Detectron2 is not the most cutting-edge model and does not specialize in cell segmentation.

In contrast, Cellpose (Stringer et al., 2020) specializes in cell segmentation and provides robust fine-tuning options. However, Cellpose predictions are highly sensitive to parameter settings in cell probability and flow thresholds. These thresholds can be adjusted when the ground truth is known, but for automatic segmentation on unknown data, this is not feasible and would require manual intervention to identify the optimal thresholds. Cellpose requires ground-truth masks as input for fine-tuning, which is also a standard practice. With releases in February 2024, Cellpose offers updated models and some customizability regarding cell types.

SAM (Kirillov et al., 2023) was the most challenging model to fine-tune because this option is not supported. We needed to write custom wrapper classes to enable fine-tuning, which is not possible without significant technical expertise in deep learning. On the other hand, SAM is easy to use for zero-shot learning and is the most powerful of all the models used. SAM supports annotations in the common COCO JSON format.

We trained U-Net from scratch without any fine-tuning. U-Net requires implementation knowledge, such as PyTorch or Keras, and needs to be trained from scratch. Due to its limited performance, it does not provide a good trade-off for efficient confluence prediction.

To combine a dataset with AL, we need to obtain uncertainty measures from the models. Detectron2 was the only model that returned confidence scores for the mask predictions. However, there was no built-in functionality to obtain probability masks. In our custom implementations of U-Net and SAM, we made
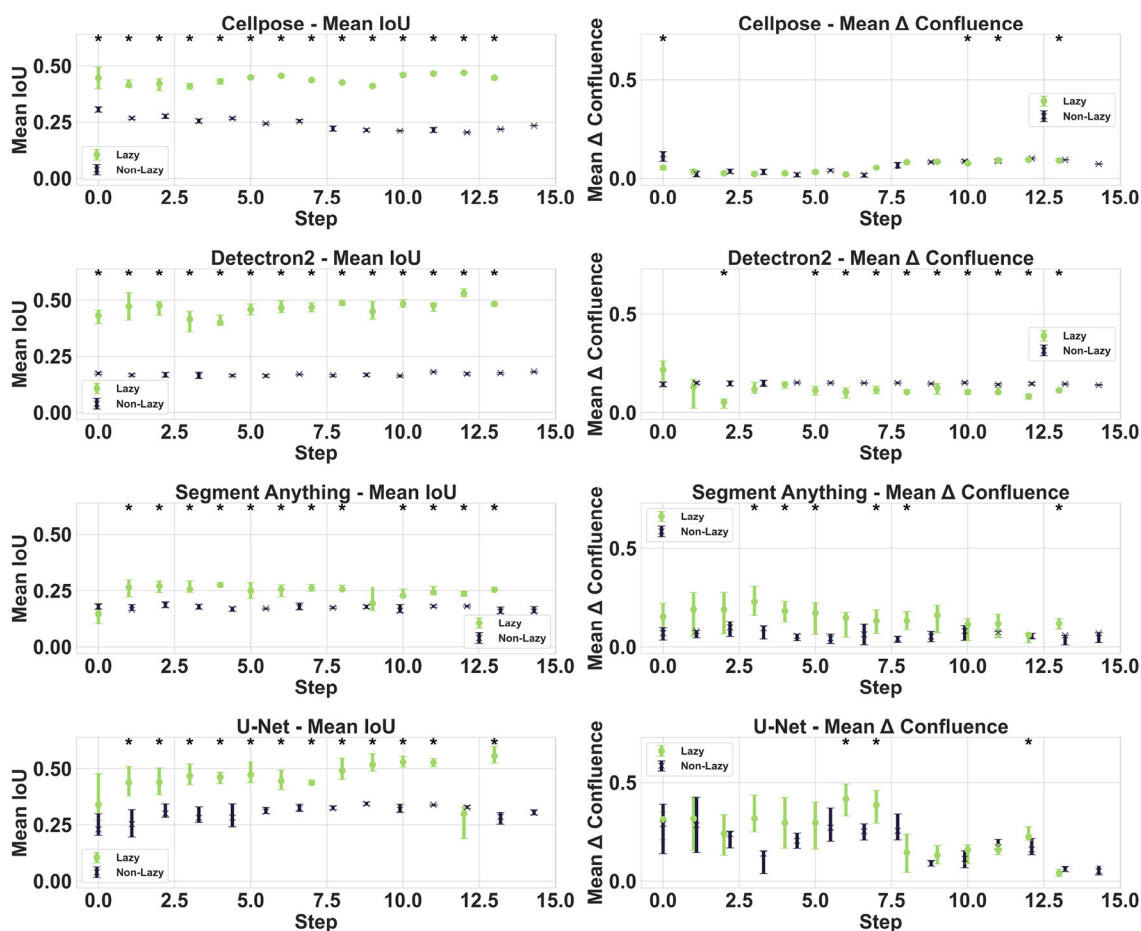
FIGURE 3
Impact of lazy labeling. The plot compares the performance of lazy and exact labeling methods during the dataset curation process for all models. The first column displays the IoU metric, while the second column illustrates the differences in confluence.

it easy to obtain the probability maps directly. Nevertheless, this comes before custom implementation, which requires expertise in deep learning. Cellpose returns the cell probabilities directly, making an uncertainty-based AL approach more straightforward to implement.

A GPU for model training or fine-tuning is almost essential for all four models. For inference, a CPU is adequate for U-Net, Detectron2, and Cellpose. This is especially important for integrating confluence prediction into real-world automation systems in cell production, where high-performance GPUs may not be available. However, inference with SAM on a CPU is impractical due to the model's size and slow performance.
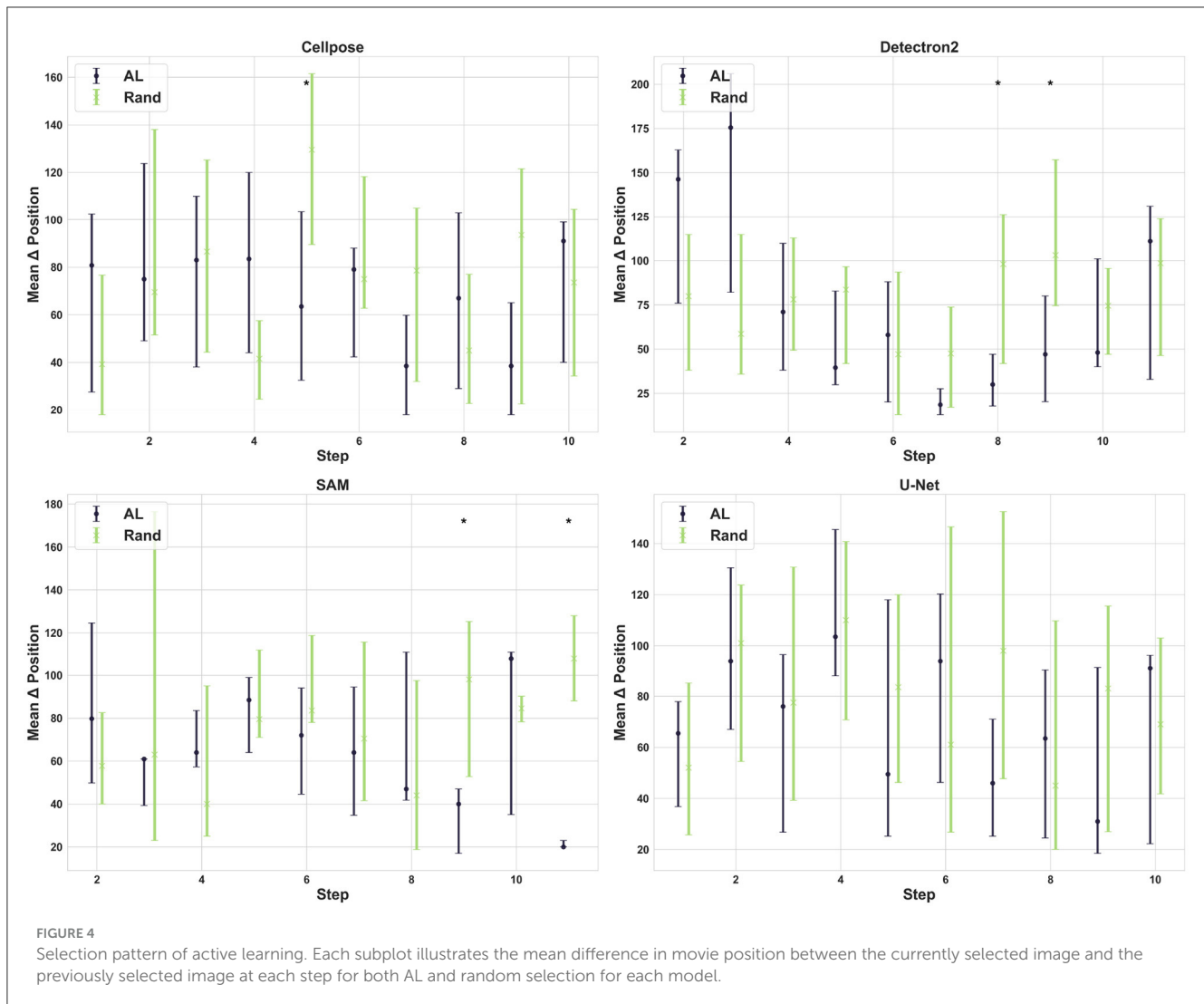
The fine-tuning process requires a GPU for all models to ensure completion within a reasonable timeframe. Fine-tuning on larger datasets with a CPU is impractical, as it could take weeks and provide minimal benefit compared to zero-shot training. In contrast, fine-tuning with only a few images can be completed within hours for Detectron2 and Cellpose, offering significant performance improvements for these models. Given the high computational cost of fine-tuning and the lack of substantial performance gains, we do not recommend fine-tuning SAM in contexts such as our use case.

4 Discussion

In our study, we compared four models for cell segmentation across various datasets inspired by real-world MSC manufacturing. This comparison aims to gain insights on how to leverage AI-based confluence estimation most efficiently. The results provide actionable strategies applicable to similar contexts. First, we demonstrated that zero-shot inference with SAM achieved near-perfect confluence estimation. Second, we observed that goal-dependent labeling outperformed traditional labeling methods in terms of IoU. Finally, we demonstrate that AL is suboptimal for MSC microscopy images. The limited benefit of AL can be explained by: (a) a lack of diversity within the dataset, (b) the use of large pre-trained models, (c) a basic AL approach, and (d) the simple binary classification (foreground/background) task.

Since our datasets contain only one cell type, the primary variation lies in the growth state or cell density. This does not fundamentally alter the characteristics of the objects to be segmented (i.e., the cells). Consequently, the timing of when a given example is presented during training has a limited impact on model performance. Our analysis revealed that AL did not select images based on their temporal position in the growth

FIGURE 4
Selection pattern of active learning. Each subplot illustrates the mean difference in movie position between the currently selected image and the previously selected image at each step for both AL and random selection for each model.
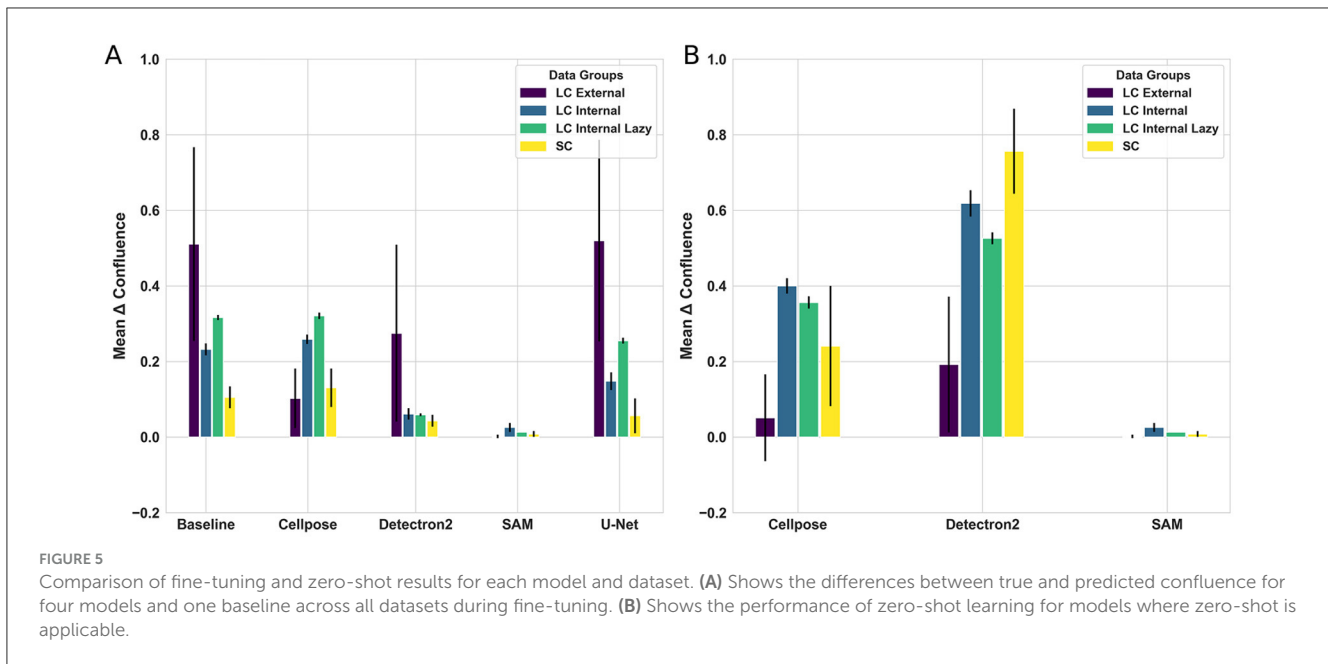
sequence, showing no significant difference from random selection. This finding supports our explanation that cell density alone does not create enough variation that AL strategies typically exploit (Monarch, 2021). The model's uncertainty, which drives the AL selection process, appears to be independent of the growth state. This suggests that once the model learns to segment cells at one density, it can readily generalize to other densities.

We also hypothesize that the benefit of AL in pre-trained models is minimal due to the extensive data exposure these models have already experienced, which diminishes the impact of new data points. Additionally, our observations indicate that U-Nets underperform when faced with a small number of highly diverse instances through AL, while random selection retains a distribution that is more representative of the entire dataset. Furthermore, for usability reasons, we adopt a straightforward maximum entropy approach for uncertainty-based AL. However, capturing the complexities of this data and model may require more complex AL strategies involving combinations of techniques to manage variations more effectively. Although prior research on biomedical images demonstrates that AL techniques can achieve

comparable performance with a reduced sample size (Nath et al., 2021; Kim et al., 2024; Li et al., 2023; Huang et al., 2024), these studies did not incorporate pre-trained models, used diverse data, and employed more complex AL methods. Considering our findings and previous research, we conclude that AL is best utilized when pre-trained models are not appropriate for a given use case and when complex AL algorithms are available for specific problems, ideally in a diverse multi-classification task.

Beyond AL, we explored simplified goal-dependent labeling directly linked to the desired outcome, namely, confluence estimation. Interestingly, for the IoU metric, goal-dependent labeled data significantly outperformed traditional labeling approaches. We attribute this success to simpler shapes that are easier for models to learn. Even when examining the Confluence task directly by calculating the difference from the ground-truth Confluence, we observe no drop in performance when utilizing goal-dependent labeling, or "lazily". Notably, when employing the SAM model, traditionally labeled data performed slightly better, which we attribute to SAM's extensive pre-training on precisely annotated datasets. Additionally, lazy labeling introduces

FIGURE 5
Comparison of fine-tuning and zero-shot results for each model and dataset. **(A)** Shows the differences between true and predicted confluence for four models and one baseline across all datasets during fine-tuning. **(B)** Shows the performance of zero-shot learning for models where zero-shot is applicable.

irreducible error, acting as a source of noise that is more pronounced when model performance is overall very high, as seen with SAM. Importantly, the substantial reduction in labeling effort makes the lazy labeling approach attractive for confluence estimation since detailed cell segmentation is not required (e.g., to derive individual cell characteristics).

While this specific labeling strategy may not be applicable to all problems, it demonstrates the value of developing task-specific labeling approaches that strike a balance between annotation effort and model performance. To our knowledge, no existing research directly addresses this confluence of estimation-specific annotation. However, recent studies in biomedical imaging have highlighted the use of time-efficient annotation techniques combined with self- and human-supervised learning (human-in-the-loop) to reduce labeling demands. For example, in nucleus segmentation, some approaches focus on selectively annotating only a small subset of critical image patches, utilizing human-supervised methods and data augmentation to match the performance of fully supervised models while minimizing the requirements for labeled data (Lou et al., 2023). Similarly, in cell segmentation, weakly supervised methods use single-point annotations per cell, combined with self- and co-training strategies, to achieve segmentation accuracy close to that of fully supervised methods (Zhao and Yin, 2021). Krishnan et al. (2022)'s review further highlights how efficient annotation methods, when combined with self-supervised learning, enable models to leverage large volumes of unannotated data, thereby enhancing model development while reducing expert annotation time.

Furthermore, tools like LABKIT (Arzt et al., 2022) provide interfaces for efficient annotation and human-in-the-loop processes, combined with supervised deep learning. Although we implemented only a time-efficient annotation strategy without any form of weak human supervision, the relative simplicity of our task suggests that these findings still support our results.

They indicate that efficient annotation methods, particularly when paired with self- and human-supervised learning, can significantly reduce labeling efforts in active learning while maintaining model performance.

While efficient labeling strategies may reduce annotation effort, completely eliminating the need for labeling would be even more desirable. We demonstrate that SAM achieves nearly perfect confluence estimation with zero-shot inference, indicating that the cost-benefit ratio of fine-tuning decreases with larger foundation models. Although other models exhibited marginal improvements with fine-tuning, and even SAM showed slight gains, these advantages were minimal compared to SAM's zero-shot performance. Interestingly, across all models, we discovered that the best performance was achieved using only a subset of the available training data, suggesting that more is not always better. The demands of fine-tuning, including (a) computational resources, (b) technical expertise for model adaptation, and (c) time invested in data annotation, far outweigh modest performance gains. This cost-benefit analysis strongly favors the use of large foundation models like SAM in its zero-shot configuration. Additionally, the recently released SAM 2 (Ravi et al., 2024) may demonstrate even higher accuracy for such tasks and focuses on object tracking in video contexts, which is relevant in time-series data from cellular production systems.

When comparing our results with existing research, we see that zero-shot inference with SAM is powerful even for more complex tasks. However, the optimal approach appears task-dependent, and some amount of fine-tuning or combining it with other models still seems to have a benefit: Baral and Paing (2024) used a fine-tuned object detection model (YOLOv9-E) to generate prompts for zero-shot SAM inference, followed by traditional image processing refinements. This hybrid approach achieved highly accurate performance (94% mAP50) for cell segmentation across varying difficulty levels without fine-tuning SAM directly.

In contrast, more specialized applications require more complex architectures. CryoSegNet (Gyawali et al., 2024) demonstrated that the combination of SAM with a task-specific U-Net significantly improved protein particle detection in cryo-EM images compared to using SAM alone. Similarly, the Segment Anything for Microscopy project (Archit et al., 2023) demonstrated that specialized training for multi-dimensional microscopy data significantly improved segmentation quality across various imaging conditions. In this context, the additional effort is likely warranted due to the complexity of volumetric segmentation and tracking tasks compared to ours. These varying approaches highlight the importance of considering task complexity and resource constraints when choosing between zero-shot applications, hybrid solutions, and full fine-tuning of foundation models.

While our findings provide insights into confluence estimation for MSC production standardization, we acknowledge several limitations of our study. First, our results specifically focus on confluence estimation, reflecting the practical needs of our applied domain. Consequently, our methods are not directly applicable to other tasks that may require tracking, such as spatial colony growth monitoring in bacteria (Kindler et al., 2019). Additional research is necessary to compare specialized tracking tools, such as TRACKASTRA (Gallusser and Weigert, 2025), with generalist foundation models, such as SAM 2 (Ravi et al., 2024).

While not exhaustive, our selected models strategically covered a representative spectrum of approaches: from models trained from scratch to cell-specific models and powerful general-purpose foundation models. This selection allowed us to compare different paradigms in model development while maintaining practical feasibility. Similarly, while our datasets were limited to one cell type, they enabled us to draw important conclusions about the impact of data diversity on active learning effectiveness in real-world scenarios.

Despite these limitations, our study provides several valuable contributions: We demonstrated that for homogeneous cell cultures, (a) SAM delivers excellent results without requiring resource-intensive methods such as active learning or fine-tuning. We provide (b) technical guidelines for implementing active learning in cell imaging applications, demonstrate (c) the importance of goal-specific labeling strategies, and highlight (d) how data homogeneity influences active learning performance. These insights can guide future research in automated cell culture monitoring and quality control. These practical insights significantly lower the barrier to implementing automated quality control in cell manufacturing. A prototype of our confluence detection software is publicly available (https://livinglab. scadsai.uni-leipzig.de/cell-confluence/), enabling immediate community adoption.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/MaxJoas/confluence-paper.

Author contributions

MJ: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. DF: Data curation, Resources, Writing – original draft, Writing – review & editing. RH: Conceptualization, Investigation, Resources, Supervision, Writing – original draft, Writing – review & editing. ER: Funding acquisition, Project administration, Resources, Writing – original draft, Writing – review & editing. JE: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

Acknowledgments

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. Generative AI was used to improve grammar and language style.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp. 2025.1562358/full#supplementary-material

References

Archit, A., Nair, S., Khalid, N., Hilt, P., Rajashekar, V., Freitag, M., et al. (2023). *Segment Anything for* microscopy. *bioRxiv*. doi: 10.1101/2023.08.21.554208

Arzt, M., Deschamps, J., Schmied, C., Pietzsch, T., Schmidt, D., Tomancak, P., et al. (2022). LABKIT: Labeling and segmentation toolkit for big image data. *Front. Comp. Sci.* 4:777728. doi: 10.3389/fcomp.2022.777728

Baral, S., and Paing, M. P. (2024). Instance segmentation of cells and nuclei from multi-organ cross-protocol microscopic images. *Quant. Imaging Med. Surg.* 14, 6204–6221. doi: 10.21037/qims-24-801

Canny, J. (1986). A computational approach to edge detection. *IEEE Trans. Pattern Analys. Mach. Intellig.* 8, 679–698. doi: 10.1109/TPAMI.1986.4767851

Chen, H., and Murphy, R. F. (2023). Evaluation of cell segmentation methods without reference segmentations. *Mol. Biol. Cell* 34:6. doi: 10.1091/mbc.E22-08-0364

Edlund, C., Jackson, T. R., Khalid, N., Bevan, N., Dale, T., Dengel, A., et al. (2021). LIVECell—a large-scale dataset for label-free live cell segmentation. *Nat. Methods* 18, 1038–1045. doi: 10.1038/s41592-021-01249-6

Fernández-Santos, M. E., Garcia-Arranz, M., Andreu, E. J., García-Hernández, A. M., López-Parra, M., Villarón, E., et al. (2022). Optimization of mesenchymal stromal cell (MSC) manufacturing processes for a better therapeutic outcome. *Front. Immunol.* 13:918565. doi: 10.3389/fimmu.2022.918565

Galipeau, J., and Sensébé, L. (2018). Mesenchymal stromal cells: Clinical challenges and therapeutic opportunities. *Cell Stem Cell* 22, 824–833. doi: 10.1016/j.stem.2018.05.004

Gallusser, B., and Weigert, M. (2025). "TRACKASTRA: transformer-based cell tracking for live-cell microscopy," in *Computer Vision-ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol (Cham: Springer Nature Switzerland), 467–484.

Gyawali, R., Dhakal, A., Wang, L., and Cheng, J. (2024). CryoSegNet: accurate cryo-EM protein particle picking by integrating the foundational AI image segmentation model and attention-gated U-Net. *Brief. Bioinform.* 25:bbae282. doi: 10.1093/bib/bbae282

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., et al. (2021). Pre-trained models: Past, present and future. *AI Open* 2, 225–250. doi: 10.1016/j.aiopen.2021.08.002

Huang, J., Farpour, N., Yang, B. J., Mupparapu, M., Lure, F., Li, J., et al. (2024). Uncertainty-based active learning by bayesian U-Net for multi-label cone-beam CT segmentation. *J. Endod.* 50, 220–228. doi: 10.1016/j.joen.2023.11.002

IHME (2020). "Global burden of disease study 2019 (GBD 2019) disease and injury burden 1990-2019," in *Technical report, Institute for Health Metrics and Evaluation (IHME)* (Seattle: IHME).

Jafari, M., Zhang, Y., Zhang, Y., and Liu, S. (2024). "The power of few: Accelerating and enhancing data reweighting with coreset selection," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Seoul: IEEE), 7100–7104.

Kim, D. D., Chandra, R. S., Yang, L., Wu, J., Feng, X., Atalay, M., et al. (2024). Active learning in brain tumor segmentation with uncertainty sampling and annotation redundancy restriction. *J. Imag. Inform. Med.* 37, 2099–2107. doi: 10.1007/s10278-024-01037-6

Kim, D. S., Lee, M. W., Lee, T.-H., Sung, K. W., Koo, H. H., and Yoo, K. H. (2017). Cell culture density affects the stemness gene expression of adipose tissue-derived mesenchymal stem cells. *Biomed. Reports* 6, 300–306. doi: 10.3892/br.2017.845

Kindler, O., Pulkkinen, O., Cherstvy, A. G., and Metzler, R. (2019). Burst statistics in an early biofilm quorum sensing model: the role of spatial colony-growth heterogeneity. *Sci. Rep.* 9:1. doi: 10.1038/s41598-019-48525-2

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). "Segment Anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (Paris: IEEE), 3992–4003.

Krishnan, R., Rajpurkar, P., and Topol, E. J. (2022). Self-supervised learning in medicine and healthcare. *Nat. Biomed. Eng.* 6, 1346–1352. doi: 10.1038/s41551-022-00914-1

Li, X., Xia, M., Jiao, J., Zhou, S., Chang, C., Wang, Y., et al. (2023). HAL-IA: A hybrid active learning framework using interactive annotation for medical image segmentation. *Med. Image Anal.* 88:102862. doi: 10.1016/j.media.2023.102862

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., et al. (2014). "Microsoft COCO: Common objects in context," in *Computer Vision-ECCV 2014* (Cham: Springer International Publishing), 740–755.

Lou, W., Li, H., Li, G., Han, X., and Wan, X. (2023). Which pixel to annotate: A label-efficient nuclei segmentation framework. *IEEE Trans. Med. Imaging* 42, 947–958. doi: 10.1109/TMI.2022.3221666

Mann, H. B., and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.* 18, 50–60. doi: 10.1214/aoms/1177730491

Monarch, R. (2021). *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-centered AI*. Shelter Island, NY: Manning.

Nath, V., Yang, D., Landman, B. A., Xu, D., and Roth, H. R. (2021). Diminishing uncertainty within the training pool: active learning for medical image segmentation. *IEEE Trans. Med. Imaging* 40, 2534–2547. doi: 10.1109/TMI.2020.3048055

Panagiotakis, C., and Argyros, A. A. (2018). "Cell segmentation via region-based ellipse fitting," in *2018 25th IEEE International Conference on Image Processing (ICIP)* (Athens: IEEE), 2426–2430. doi: 10.1109/ICIP.2018.8451852

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., et al. (2024). SAM 2: Segment anything in images and videos. *arXiv* [preprint] arXiv:2408.00714. doi: 10.48550/arXiv.2408.00714

Ronneberger, O., Fischer, P., and Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Cham: Springer International Publishing, 234–241.

Sayin, B., Krivosheev, E., Yang, J., Passerini, A., and Casati, F. (2021). A review and experimental analysis of active learning over crowdsourced data. *Artif. Intellig. Rev.* 54, 5283–5305. doi: 10.1007/s10462-021-10021-3

Shen, S. P., Tseng, H.-a., Hansen, K. R., Wu, R., Gritton, H. J., Si, J., et al. (2018). Automatic cell segmentation by adaptive thresholding (ACSAT) for large-scale calcium imaging datasets. *eNeuro* 5:ENEURO.0056-18.2018. doi: 10.1523/ENEURO.0056-18.2018

Shimizu, Y., Ntege, E. H., Azuma, C., Uehara, F., Toma, T., Higa, K., et al. (2023). Management of rheumatoid arthritis: Possibilities and challenges of mesenchymal stromal/stem cell-based therapies. *Cells* 12:1905. doi: 10.3390/cells12141905

Strecanska, M., Sekelova, T., Csobonyeiova, M., Danisovic, L., and Cehakova, M. (2024). Therapeutic applications of mesenchymal/medicinal stem/signaling cells preconditioned with external factors: are there more efficient approaches to utilize their regenerative potential? *Life Sci.* 346:122647. doi: 10.1016/j.lfs.2024.122647

Stringer, C., Wang, T., Michaelos, M., and Pachitariu, M. (2020). Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* 18, 100–106. doi: 10.1038/s41592-020-01018-x

von Chamier, L., Laine, R. F., Jukkala, J., Spahn, C., Krentzel, D., Nehme, E., et al. (2021). Democratising deep learning for microscopy with ZeroCostDL4Mic. *Nat. Commun.* 12:2276. doi: 10.1038/s41467-021-22518-0

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). *Detectron2*. Available online at: https://github.com/facebookresearch/detectron2 (accessed September 20, 2024).

Yin, T., Panapitiya, G., Coda, E. D., and Saldanha, E. G. (2023). Evaluating uncertainty-based active learning for accelerating the generalization of molecular property prediction. *J. Cheminform.* 15:1. doi: 10.1186/s13321-023-00753-5

Zhao, T., and Yin, Z. (2021). Weakly supervised cell segmentation by point annotation. *IEEE Trans. Med. Imaging* 40, 2736–2747. doi: 10.1109/TMI.2020.3046292

Zhu, J., Wang, H., Yao, T., and Tsou, B. K. (2008). "Active learning with sampling by uncertainty and density for word sense disambiguation and text classification," in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, D. Scott, and H. Uszkoreit (Manchester, UK:Coling 2008 Organizing Committee), 1137–1144.