BEN-GURION UNIVERSITY OF THE NEGEV
FACULTY OF COMPUTER AND INFORMATION SCIENCE

PROGRAM OF INFORMATION SYSTEMS ENGINEERING

Detecting Failures in Image-to-Image Translation for In Silico Labeling

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE M.Sc. DEGREE

By: Gad M. Miller

December 2025

BEN-GURION UNIVERSITY OF THE NEGEV
FACULTY OF COMPUTER AND INFORMATION SCIENCE

PROGRAM OF INFORMATION SYSTEMS ENGINEERING

Detecting Failures in Image-to-Image Translation for In Silico Labeling

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE M.Sc. DEGREE

By: Gad M. Miller

Supervised by: Professor Assaf Zaritsky

Author:  Date: 28.12.2025

Supervisor:  Date: 28.12.2025

Chairman of Graduate Studies Committee:…… Date:……..

December 2025

Abstract

Image-to-image translation learns a mapping from one image type to another, so the computer can infer a target view from an input view. In silico labeling applies this to microscopy: it translates easy to acquire, uninformative microscopy images into detailed microscopy images, producing “virtual” labeled structures in the cell without additional acquisition. It has the potential to revolutionize live-cell imaging by avoiding the phototoxicity, perturbation, and spectral overlap of conventional staining. Yet adoption is limited as it is not clear how much each prediction can be trusted for quantitative analysis. Even when predictions look plausible, scientists need a principled, per-cell estimate of how much to trust them.

Our contribution is a model-agnostic pipeline that gives the outputs of in silico labeling a measure of confidence at a single-cell resolution. We design a 3D ResNet based model that ingests two channels: the in silico labeling prediction and an explanation-derived importance mask that encodes which regions in the input image the translator relies on, and predicts the expected error for each organelle in the cell. This reframes explanations from “why the model looked there” into “how much this particular prediction should count,” producing a confidence metric rather than post-hoc visual explanations. The approach is plug-in: the in silico labeling model is treated as a black box, and the importance masks are used as signals, not objectives, so no changes to in silico labeling training are required.

Our confidence scores track true discrepancies between in silico labeling outputs and ground truth across eight different organelles, and outperform baselines that use the in silico labeling image alone or hand-crafted heuristics. We are able to assess prediction quality in single-cell resolution, and show that the scores align with common applicative metrics measured against the ground truth, enabling simple operating points: accept high-confidence cells for quantitative readouts, flag low-confidence cells for review, and prioritize ambiguous regions for data curation.

In summary, we provide a pipeline to make in silico labeling measurable at the most important unit: the cell. By coupling translation outputs with explanation-derived signals in a learned confidence estimator, we deliver per-cell reliability maps that support reviewer triage, reliability-aware quantification, and data selection, moving in silico labeling from visually convincing images to quantitative measures biologists can use, in hopes to make in silico labeling a practical tool in biological research.

Keywords

- In silico labeling
- Image-to-image translation
- Uncertainty quantification
- Interpretability
- Explainable AI (XAI)
- Confidence estimation
- Error prediction
- Biological image analysis
- Fluorescence microscopy
- Single-cell analysis

Acknowledgments

I would like to express my deepest gratitude to my advisor, Prof. Assaf Zaritsky, for his guidance, support, and encouragement throughout this research. His insights and critical thinking have shaped both this thesis and my academic growth.

I am also thankful to the members of the Computational Cell Dynamics Lab for their collaboration, stimulating discussions, and friendship, which created a supportive and inspiring environment.

Finally, I am grateful to the university and the research community whose resources, knowledge, and inspiration have made this work possible.

Table of Contents

Figures and Tables

1) Introduction

As biological research advances, the need for trustworthy, high-quality microscopic data grows. Cellular functions involve dynamic and coordinated rearrangements of subcellular organelles and cytoskeletal elements that define the cell's architecture. Uncovering the complex mechanisms of cellular processes requires continuous measurements of a cell's subcellular organization over time. Understanding these processes can lead to breakthroughs in the way we look at diseases and develop treatments.

One imaging technique for obtaining cellular data over time is **Fluorescence Microscopy**. This method involves binding fluorophores to specific cellular components, producing bright, high-contrast images of the structures of interest against a dark background. It allows the observation of dynamic interactions within cells in real-time, tracking the movement of molecules, and understanding cellular functions in detail. However, the number of fluorophores that can be simultaneously tagged and distinguished in a live cell microscopy image are severely limited, and increasing the number of subcellular structures, experiment time, imaging depth or fluorophore density leads to a reduction in signal intensity over time (photobleaching) and induce phototoxic effects in living cells (phototoxicity) that alter the cell physiology.

In 2018, two seminal studies demonstrated computationally translating label-free transmitted-light (brightfield) microscopy images to synthetic organelle-specific fluorescent images, a method coined as 'in silico labeling' or 'virtual staining'. These in silico labeling proof of principle studies have been replicated, and applied to new organelles, cell types, and imaging techniques. However, the practical routine use of in silico labeling remains limited. The main caveats these Image-to-image translation models face are the **hallucination problem**, where the model creates artifacts due to inadequate training data, and the **generalization problem**, where models fail to generalize to unseen data. Additionally, the difficulty in interpreting the models' results poses a significant challenge, as biologists are reluctant to use frameworks that operate as complete black boxes without explanation.

Interpretability is crucial for the reliable application of in silico labeling models. Qualitative (visual) and quantitative explanations can reveal when models fail, for example due to batch effects, rare cell states, or experimental perturbations. This enables researchers to filter out unreliable predictions and focus downstream analyses on cell subpopulations and organelles

where reconstructions are trustworthy. Such advances can transform in silico labeling from a proof-of-principle into a practical method for systematic study of cellular organization.

One approach to address this need is the **Mask Interpreter model**, which was developed to enhance in silico labeling interpretability. Mask Interpreter learns to generate an importance mask over the input image, highlighting regions that are most critical for the in silico labeling model to produce accurate outputs. These masks provide insight into which morphological features the in silico labeling model relies on and can help identify conditions under which the model may produce erroneous or biased predictions. However, while Mask Interpreter offers valuable explanations, it does not directly quantify how reliable a given prediction is.

This thesis builds on the idea of interpretability but introduces a complementary direction: **confidence estimation**. Instead of only visualizing important regions, we train a model to map importance masks and in silico labeling predictions into a quantitative measure of prediction reliability. This framework enables biologists to not only interpret *how* an in silico labeling model makes predictions, but also assess *how much trust* to place in each prediction. By combining interpretability with confidence estimation, the method developed here moves in silico labeling closer to routine use, bridging the gap between complex computational models and practical applications in biology.

2) Literature Review

This review surveys recent progress in in silico labeling alongside uncertainty and interpretability methods for image-to-image translation models in microscopy, with the goal of enabling reliable, routine biological use. We begin with experimental motivation: practical limits of fluorescence imaging motivate predicting fluorescent readouts directly from label-free images. In silico labeling realizes this idea using deep models that translate brightfield to synthetic organelle-specific signals, yet adoption is hindered by hallucinations, generalization gaps, metric mismatches, and limited model transparency. In silico labeling training was traditionally done using U-Net models, while alternative approaches explore global context, adversarial objectives, and multi-task or context-aware training.

Uncertainty methods (stochastic inference, model ensembles) estimate how much to trust a given prediction, whereas interpretability methods aim to explain which input structures support reconstructed signals. Interpretability of Image-to-image translation remains comparatively underdeveloped, motivating approaches tailored to translation tasks and their relation to quantitative reliability.

Section 2.1 reviews in silico labeling advances and limitations. Section 2.2 covers uncertainty and interpretability, distinguishing explanation tools from confidence tools and positioning this thesis's spatially informed confidence framework within that landscape.

2.1) In Silico Labeling

2.1.1) Motivation for in silico labeling

Fluorescence microscopy allows targeted visualization of subcellular structures, but the number of fluorophores that can be tagged and distinguished in live cells is severely limited[1]. Increasing the number of labeled structures, experiment time, imaging depth, or fluorophore density reduces signal over time (photobleaching) and can induce phototoxic effects that alter cell physiology[2]. Even attaching a fluorescent label may disrupt protein folding or structure, alter interactions, or impede movement of the molecules being studied, and thus interfere with their natural function. These constraints motivate approaches that recover fluorescent readouts from label-free images.

2.1.2) In silico labeling emergence

In 2018, two studies introduced in silico labeling: predicting organelle-specific fluorescence from brightfield images[3,4]. In silico labeling learns a supervised mapping on paired acquisitions and produces channel-specific synthetic images that approximate experimentally acquired fluorescence. Since then, in silico labeling has been replicated and extended to additional organelles, cell types, and imaging modalities[5,6,7]. These works place in silico labeling within the broader area of augmented microscopy and show that virtual labels can support analysis while reducing labeling burden. In silico labeling has also been used within downstream pipelines, for example to study dynamics or to supply inputs for other models. Routine use still depends on reliability and interpretability, which are the focus of later sections.

2.1.3) Baseline architectures and training setup

Most in silico labeling implementations use a U-Net encoder–decoder with skip connections because it preserves fine structures while capturing context[8]. Models are trained in a supervised way on paired label-free and fluorescence images, often with patch-based training in 2D or 3D to fit memory and organelle scales. Common losses include L1 or L2 on intensities, sometimes SSIM[40] or perceptual losses, and multi-channel predictions sum losses across channels. Evaluation typically reports PSNR, SSIM, and correlation, but these image similarity metrics can miss biologically important errors such as wrong localization or missing structures[9]. U-Net remains a strong baseline, yet it can hallucinate or fail to generalize across microscopes and batches, which motivates the alternative architectures.

2.1.4) Global context models

Several works replace or augment U-Net with architectures that capture global context, such as vision transformers and voxel transformers, to model long-range morphology that may drive organelle patterns[10]. These models report small improvements on label-free to fluorescence prediction compared to strong convolutional baselines, especially when structures span large spatial scales. The trade-off is higher memory and runtime, which becomes a bottleneck for full fields-of-view and 3D stacks. Training can also be less stable and may require larger datasets or stronger regularization than standard CNNs. As a result,

global-context models are promising but have not yet displaced U-Net as the default backbone for in silico labeling, while diffusion models are also strong candidates to inherit U-Net's position[32].

2.1.5) Multi-task, context-aware training, and downstream use

Several works extend the basic in silico labeling setup with auxiliary objectives or context-aware training to improve robustness[11]. Multi-task learning ties the translation to related tasks such as segmentation or morphology prediction and can stabilize features that matter for biology. In silico labeling is also used for downstream pipelines. For tracking, predicting nuclei and membrane channels via robust virtual staining supports contour extraction and more reliable cell linking in time-lapse data[12]. In drug screening, models predict a small set of reliable channels to recover common biomarkers at scale[11]. These directions position in silico labeling as a component in broader analysis workflows rather than a stand-alone endpoint.

2.1.6) Practical limitations

Despite steady progress, in silico labeling is not yet a routine tool in the lab. Models often **fail to generalize** across batches, microscopes, labs, or cell states, which leads to shifts in brightness, texture, or morphology that the network did not see during training[13]. **Hallucinations and missed structures** remain a risk, producing plausible but incorrect signal where there is little evidence in the input[9]. Standard image similarity metrics (PSNR, SSIM, correlation) can look good while **biologically important errors** persist, such as wrong localization or missing organelles, so reported scores may not track what biologists care about[33,34]. Data issues also matter: imperfect registration of label-free and fluorescent pairs, inconsistent staining quality, and class imbalance for rare structures can bias training and evaluation[35,36,37]. Cross-site benchmarking is limited, which makes it hard to compare methods and to estimate robustness in realistic settings. Finally, most in silico labeling systems provide **no built-in reliability mechanism** and offer limited interpretability, so users cannot easily filter low-trust predictions before downstream analysis. These gaps motivate methods that combine stronger generalization with **uncertainty** and **explanations**.

2.2) Uncertainty & Interpretability for Image-to-Image Translation

2.2.1) Uncertainty basics for imaging

Uncertainty estimation tells us how much to trust each prediction and where to focus review or repeat imaging. In imaging we usually distinguish aleatoric uncertainty (noise in the data, ambiguous labels, imperfect registration) from epistemic uncertainty (model uncertainty due to limited training data or distribution shift)[38]. Both are relevant to in silico labeling: aleatoric arises from optical noise and staining quality in the ground truth, while epistemic appears when batches, microscopes, or cell states differ from training[14]. Uncertainty can be reported as maps (per pixel or voxel) or as scalars at the level of a cell, field of view, or experiment. Useful uncertainty should be calibrated, meaning higher predicted uncertainty aligns with higher actual error and flags out-of-distribution inputs. In the next paragraphs we review common tools for this goal in biomedical imaging and translation tasks: SoftMax-based confidence, MC Dropout, and Deep Ensembles.

2.2.2) SoftMax confidence and entropy

SoftMax-based confidence is a simple baseline used widely in classification and semantic segmentation. The maximum SoftMax probability or the entropy of the SoftMax distribution can be taken as a pixelwise or image-level uncertainty signal[15]. These signals are cheap to compute and easy to visualize, and they can help triage predictions or guide manual review. However, SoftMax scores are often miscalibrated and can be overconfident on out-of-distribution inputs; temperature scaling and related fixes help but do not fully solve this[16]. For in silico labeling, which is a regression-style translation, SoftMax is not directly applicable, though some works adapt it by discretizing intensities into bins or by predicting class-conditional outputs to obtain probability maps. In practice, SoftMax-based maps provide a useful but weak baseline compared to stochastic or ensemble methods, which better capture model uncertainty.

2.2.3) MC Dropout

MC Dropout estimates uncertainty by keeping dropout active at test time and running the model multiple times on the same input. The spread of the resulting predictions gives a pixelwise or cell-level variance map, and the average prediction is used as the output image[17]. This variance can also be looked at as an uncertainty map that often correlates with absolute error and helps flag regions likely to be wrong or out of distribution[18]. It is lighter than training full ensembles, but it still requires several forward passes per image and its quality depends on where dropout is placed and what rate is used. Calibration may require post-processing, and MC Dropout can miss systematic biases that affect all passes in the same way. Even so, it is a practical first step for adding spatial uncertainty to translation models.

2.2.4) Deep ensembles

Deep ensembles train several independent models and use their disagreement as an uncertainty signal[19]. In imaging, ensembles often give better calibration than single models and their variance maps correlate with reconstruction error, including in image-to-image translation tasks[20]. For in silico labeling, ensemble disagreement can highlight regions where the synthetic fluorescence is unreliable and help flag out-of-distribution inputs[21]. The main cost is computation and memory, since multiple models must be trained and stored, and several predictions are needed at inference. Overall, ensembles are a strong baseline for in silico labeling uncertainty.

2.2.5) Interpretability in imaging: saliency limits

Saliency and Grad-CAM were developed mainly for classification, where a single score is explained by highlighting influential pixels[22]. In image-to-image translation the output is an image, not a class score, so it is unclear which target to explain and how to summarize many pixelwise influences into a clear map. Backpropagation through many output pixels often produces diffuse or noisy highlights and can react to edges or contrast rather than the structures that truly drive the reconstruction. These methods are also sensitive to preprocessing and scaling choices, which can change the visual explanation without changing the underlying model behavior[23]. As a result, classification-oriented saliency can be informative but is not a faithful tool for explaining what evidence an in silico labeling model

used to generate a specific synthetic channel. This motivates translation-specific, task-faithful approaches.

2.2.6) Perturbation and sensitivity for translation models

Perturbation methods explain image-to-image translation models by changing the input and measuring how the output image changes[24]. Common choices include occluding patches, adding controlled noise, shuffling regions, or erasing specific structures, then scoring each region by the drop in output similarity to the unperturbed prediction. This ties the explanation to the actual translation objective and is often more faithful than classification-style saliency for in silico labeling. These methods can be expensive because many perturbed runs are needed, and naive occlusion gives only local views unless the search is optimized.

2.2.7) Mask Interpreter

Mask Interpreter is a perturbation-based method, developed at the Lab of Computational Cell Dynamics of Ben-Gurion University, that explains image-to-image translation models by learning an importance mask over the input that must be preserved for accurate translation. During training it applies controlled noise to candidate regions and optimizes a mask so that noising the unimportant areas leaves the in silico labeling output unchanged, while noising important areas degrades it. The result highlights the minimal evidence the model relies on and shows how required spatial context differs across organelles and conditions. These maps help reveal failure modes, such as sensitivity in mitotic cells or batch-specific artifacts, and support filtering of low-trust cases in analysis pipelines. Mask Interpreter provides explanations, but it does not output a numeric reliability score for each prediction. In other words, it improves transparency but not confidence estimation.

2.2.8) Gap: spatially informed confidence

Current uncertainty tools give useful signals but are generic and often expensive, while explanation tools show where the model looked but do not say how much to trust a prediction[9]. This leaves a gap between **why** the model produced a result and **whether** that result is reliable for analysis. In this thesis we use the information in explanation maps to

predict reliability at the single-cell level. We are inspired by a similar approach used for predicting quality of segmentation results[25]. here we combine the in silico labeling prediction with its importance mask and train a model to output a quantitative confidence score per cell. In practice, the score supports filtering unreliable cells, prioritizing review, and focusing downstream analyses on subpopulations where in silico labeling is trustworthy. This bridges interpretability and uncertainty and moves in silico labeling closer to routine use in biology.

3) Research Goal

Our goal is to make simple, non-invasive microscopy useful for studying how organelles interact over time, without harming the cells. We focus on in silico labeling, which converts label-free images into fluorescent-like images where subcellular structures are visible. To be used routinely, these predictions need to be understandable and trustworthy. We address both needs by pairing explanations of what the model relies on with a way to estimate how reliable each prediction is.

Concretely, we study an image-to-image in silico labeling model and develop a framework that explains which regions in the input support each predicted signal and that assigns a confidence score to each cell. The explanations come from importance masks that highlight the minimal evidence needed to keep the prediction accurate. The confidence model uses these masks together with the in silico labeling output to predict how much the result can be trusted at the single-cell level. This combination is meant to help biologists filter low-trust cells, focus on reliable subpopulations, and detect conditions where in silico labeling is likely to fail.

The guiding question is whether we can quantify the quality of in silico labeling predictions in a way that promotes regular use by biologists. Our working hypothesis is that importance-based explanations, when coupled with a learned confidence score, will track real error and reveal batch effects, cell states, or perturbations that degrade performance. Success means showing that confidence correlates with error across organelles and datasets, and that filtering by confidence improves downstream analyses without adding complexity to the imaging process.

4) Methodology

4.1) Overview

We develop a three-stage pipeline (Figure 1) that (i) predicts organelle-specific fluorescence from label-free microscopy, (ii) explains which input regions support each prediction (Mask Interpreter), and (iii) estimates reliability of those predictions (Confidence Model).
Given a 3D transmitted-light volume, the in silico labeling model produces organelle channels; the Mask Interpreter yields a compact importance mask indicating the minimal evidence required to preserve each in silico labeling output; the Confidence Model consumes the in silico labeling output together with the importance mask and outputs a continuous confidence (predicted error).
Our experiments probe three axes: (1) organelle type (multi-organelle evaluation), (2) data regime (train/val/test splits without leakage) and (3) different modalities (with vs. without importance masks; alternative baselines). All models are trained on training sets only; validation selects hyperparameters; reporting uses held-out tests.
We deem the approach successful if (i) confidence (or predicted error) correlates with true per-cell error across organelles, (ii) filtering by confidence improves downstream in silico labeling quality without altering acquisition, and (iii) calibration and robustness remain acceptable on held-out data (more information in the following sections).

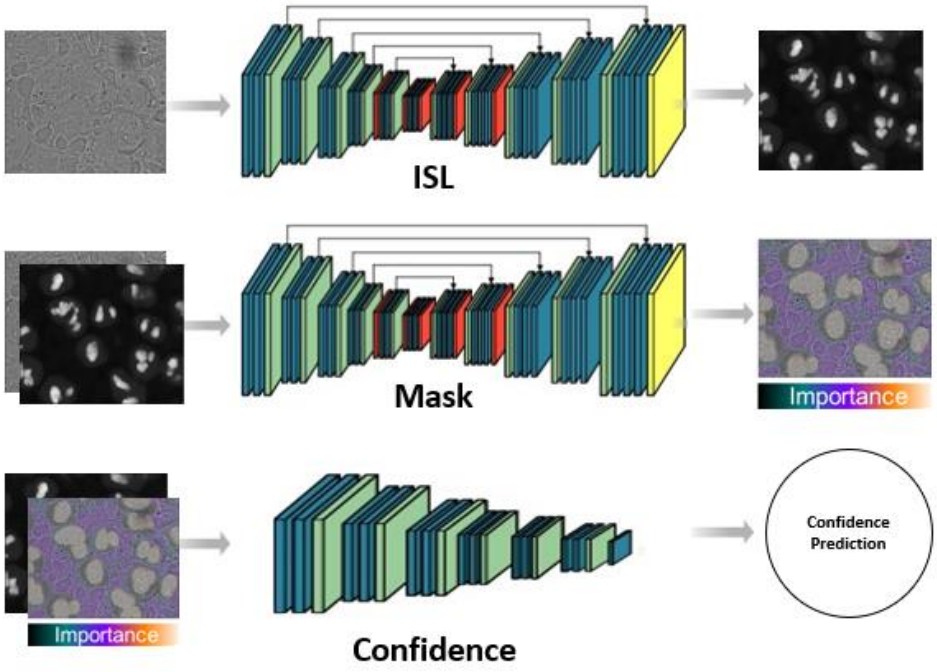

Figure 1. Suggested pipeline. *The fluorescence label is predicted from a bright-field image (top row), the most important areas used for prediction are extracted using Mask Interpreter (middle row), the two outputs are then fed to our confidence model (bottom row).*

4.2) Data

4.2.1) Datasets

The data used for this research are sourced from the Allen Institute for Cell Science's microscopy pipeline and was used for the development of the original in silico labeling model[3]. It consists of three-dimensional light microscopy z-stacks of various cell types, including human embryonic kidney cells (HEK293), human fibrosarcoma cells (HT-1080), genome-edited human induced pluripotent stem cells (hiPSCs) expressing proteins tagged with monomeric enhanced green fluorescent protein (mEGFP) or red fluorescent protein (mTagRFP), and hiPSC-derived cardiomyocytes. The tagged proteins localize to specific subcellular structures such as microtubules, actin filaments, desmosomes, the nuclear envelope, nucleoli, actomyosin bundles, the endoplasmic reticulum, the Golgi apparatus, mitochondria, and tight junctions. Figure 2 shows an example of a bright-field image of a tissue, the ground truth fluorescent image and the labeled image for the DNA channel. Imaging was performed on a Zeiss spinning disk microscope with various settings for

different channels, capturing high-resolution z-slice images with specific pixel scales. The data include multi-channel z-stacks paired with transmitted-light and EGFP/RFP channels to train models for predicting the localization of tagged subcellular structures.

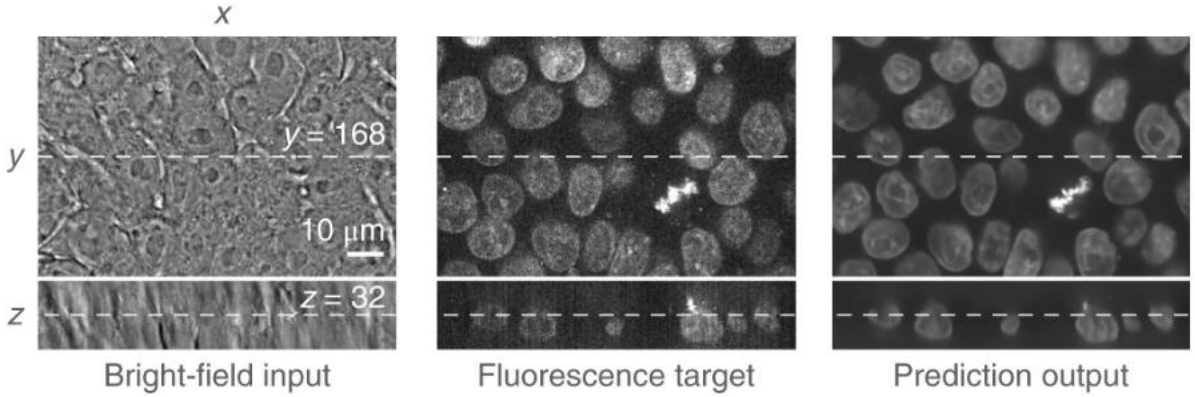

Figure 2. Data example. *The non-informative bright-field image is used as an input for the in silico labeling model, which predicts the localization of the DNA in the tissue. Image was adapted from Ounkomol et al.[3].*

4.2.2) Inclusion criteria and splits

We include volumes that contain both transmitted-light and the corresponding fluorescence channel for the organelle of interest. Data are split into train/validation/test at the acquisition (FOV) level to prevent leakage across splits.

4.2.3) Data Preprocessing

To ensure consistent intensity scaling across the entire image volume, pixel intensities were normalized to their z-scores. The normalization for each pixel (x_i') was performed using the equation:

1. $x_i' = \frac{x_i - \mu}{\sigma}$

where x_i is the original pixel intensity, μ is the mean pixel intensity of the entire volume, σ is the standard deviation of the pixel intensities, and x_i' is the normalized pixel intensity. Due to the large size of the input volumes, training was conducted on randomly cropped patches from the input volumes, each of size 32×128×128 voxels, corresponding to the z, x and y axes, respectively. To obtain the in silico predictions required as inputs for the model, we

utilized the designated pre-trained in silico labelling and mask interpreter models specific to each organelle.

4.2.4) Postprocessing

During inference, given the large volume of the input data, predictions were generated on overlapping patches. These patch-wise predictions were then merged using a weighted triangle function, assigning more weight to the predictions at the center of each patch and less weight to those at the peripheral areas. This technique, also employed in Ounkomol et al.[3], was used in generating the full importance mask and confidence prediction that were necessary for the single-cell analysis.

The triangle weight function w(u) for a single dimension is defined as:

2. $w(u) = 1 - \frac{2\left|u - \frac{L-1}{2}\right|}{L-1}$

where u is the coordinate within the patch along a given axis, and L is the size of the patch along that axis. The weights peak at the center of the patch and decrease linearly towards the edges. For three-dimensional patches, the overall weight W(x,y,z) at voxel (x,y,z) is calculated as the product of the weights along each dimension:

3. $W(x, y, z) = w(x) * w(y) * w(z)$

The final prediction at each voxel is obtained by averaging the weighted predictions from all overlapping patches.

4.2.5) Single-cell ROI generation

For single-cell analysis, segmentation masks are used to extract per-cell ROIs from the predicted in silico labeling channel, the full importance mask and the full confidence prediction. All downstream cell-level metrics (IoU/Dice/SSIM[40]) are computed within these ROIs.

4.3) In Silico Labeling

In silico labeling[3,4] was originally developed in 2018 as a U-Net[8] based method to facilitate the prediction of subcellular structures from label-free microscopy images. The U-Net architecture leverages convolutional layers to learn complex nonlinear relationships across multiple spatial areas without the need for data-specific feature extraction pipelines. In silico labeling models transform transmitted-light microscopy images into fluorescent subcellular structures, thus overcoming the invasive and phototoxic limitations of traditional fluorescence microscopy. An illustration of in silico labeling's pipeline is presented in Figure 3. The training of the model involves optimizing parameters through stochastic gradient descent to minimize the mean squared error between output and target images. We use in silico labeling here as the 'black box' model we aim to interpret.

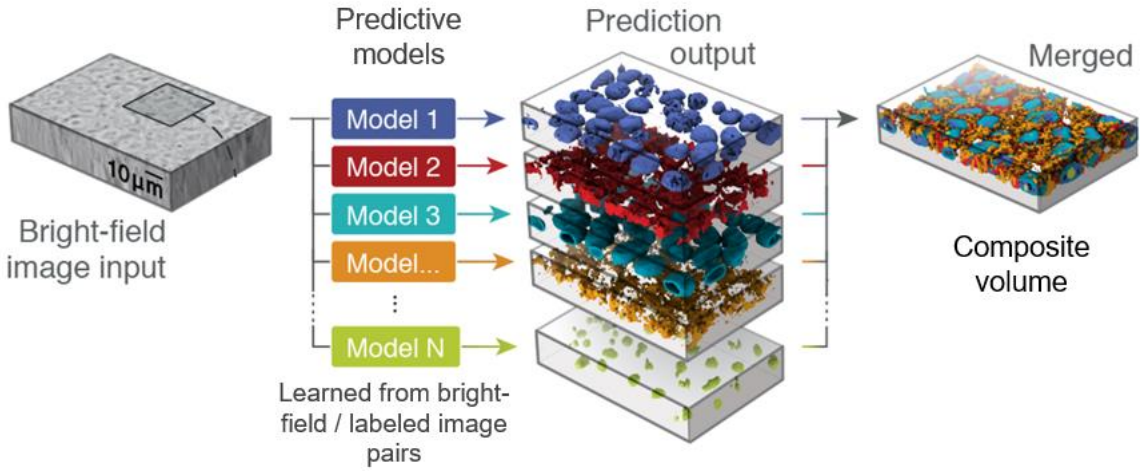

Figure 3. In silico labeling pipeline. *A bright-field image is taken as an input; For each organelle a specific model is trained to predict the fluorescence channel of this organelle. These predictions can later be used separately or combined for a holistic view. This image was adapted from Ounkomol et al.[3].*

4.4) Mask Interpreter

Mask Interpreter is a novel model developed to enhance the interpretability of image-to-image translation models, particularly in silico labeling. This model is designed to address the inherent "black box" nature of deep learning models by providing clear, visual explanations of the predictions made by these models. Mask Interpreter generates importance masks that identify the critical regions in an input image necessary for maintaining accurate predictions.

These masks highlight which parts of the image are essential for the model's decision-making process, allowing users to understand and trust the predictions better.

Given an Image-to-Image translation model and an input image, Mask Interpreter outputs an importance mask with values ranging from 0 to 1, where 0 indicates an irrelevant pixel and 1 indicates a highly important pixel. This importance mask highlights the critical regions of the input that are essential for the model's prediction. In the training phase, the Mask Interpreter model follows a unique pipeline to determine these critical regions (Figure 4). First, it creates the importance mask. Then, noise is added to the input image according to the importance of each pixel, to create a noisy version of the image. The model then generates two predictions using the Image-to-Image translation model: one from the original input and one from the noisy input. Several terms are calculated for both images, and three loss components are optimized: the Mean Squared Error (MSE) loss, which minimizes the difference between the in silico labeling predictions of the noisy and original inputs, ensuring high-quality predictions from the noisy input; the Pearson Correlation Coefficient (PCC) loss, which maximizes the correlation between the predictions from the noisy and original inputs; and the Mask loss, which minimizes the importance mask to highlight only the most critical regions in the input image. Finally, a threshold can be chosen and only pixels above this threshold are used for creating the explanation. The rationale behind this method is that by learning what can and cannot be noised, the model identifies what is important for its predictions.

The novelty of Mask Interpreter lies in its ability to provide intuitive and detailed visual explanations of model decisions, surpassing traditional interpretability methods like GradCAM and guided backpropagation. This superiority was demonstrated through qualitative and quantitative evaluations, where Mask Interpreter maintained higher prediction accuracy despite the introduction of more noise compared to other methods.

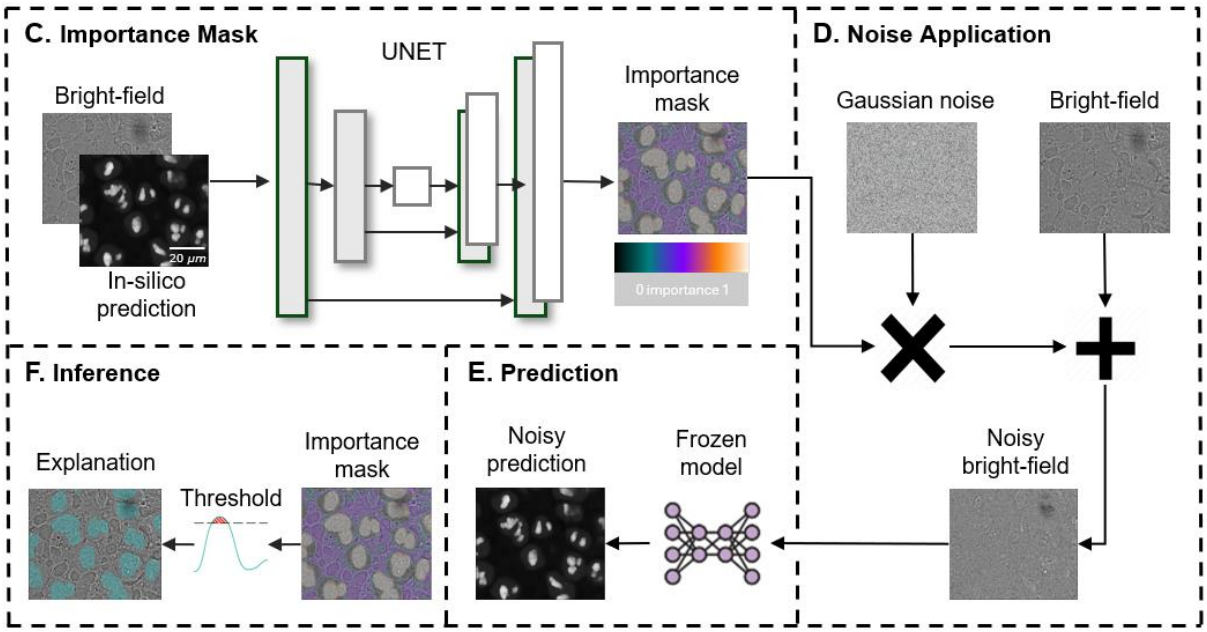

Figure 4. Mask Interpreter pipeline. *This figure was adapted from a similar one by Lion Ben Nedava, who developed Mask Interpreter.*

Mask interpreter can also be used in inference to try and assess in silico labeling quality. A method termed ‘mask efficacy score’ which measures the Pearson Correlation Coefficient between the in silico labeling prediction from noisy input and the in silico labeling prediction from clean input was introduced as part of Mask Interpreter development and helps identify cases where the mask failed to recognize important areas in the input, assuming missed areas are under-represented in the data and would lead to poor in silico labeling predictions. We use this method as one of our comparison baselines as specified in the ablation study section.

One interesting phenomenon observed using Mask Interpreter is the occurrence of typical explanation masks. For instance, Figure 5 illustrates a case where a deteriorated explanation mask suggested an erroneous prediction of the nuclear envelope. Upon further examination using the DAPI channel, it was discovered that the cell in question was undergoing mitosis, which explained the model's failure. This insight highlights the ability of Mask Interpreter to not only identify critical regions for accurate predictions but also to reveal instances where the model may struggle due to uncommon cellular states. This observation has **motivated the development of our method**, which we will present next, aiming to enhance the robustness and reliability of in silico labeling predictions by addressing such atypical cases.

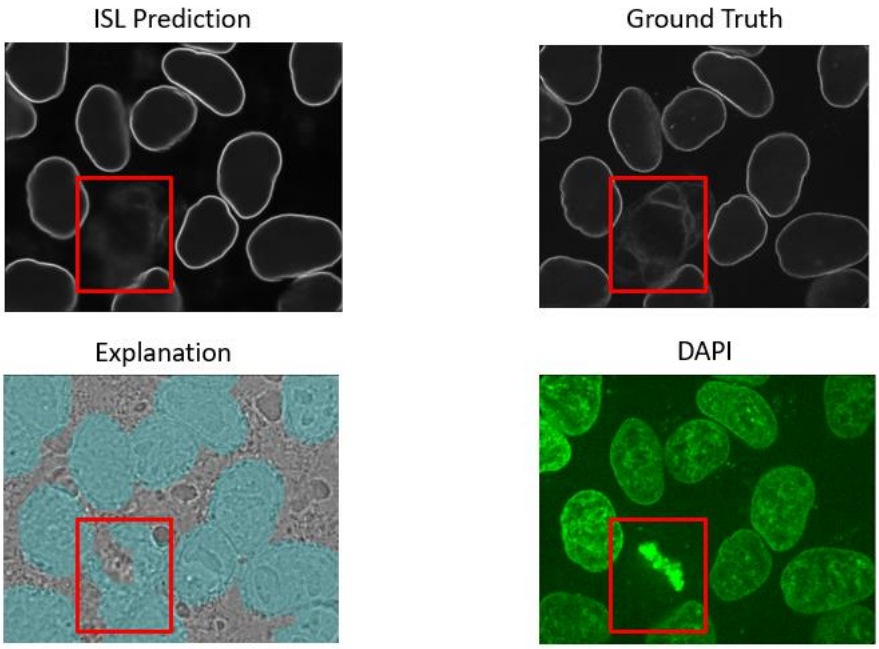

Figure 5. Mask Interpreter results provide information that can be used to find in silico labeling errors. *This figure was adapted from a similar one by Lion Ben Nedava.*

4.5) Confidence Model

This is the main contribution of my thesis.

4.5.1) Architecture

To estimate the reliability of in silico labeling predictions, we developed a confidence model based on a modified 3D ResNet-18 architecture[41] (r3d_18, pretrained on Kinetics[42]). This architecture was chosen as it's suitable for 3D data (originally for spatiotemporal data, but also works well for volumetric data), fairly light in comparison to other suitable models, it has shown good results on medical data[43], and similar approaches succeeded in related tasks such as assessing segmentation quality[44]. The model was adapted for two-channel input which are the in silico labeling prediction and its corresponding importance mask, by replacing the initial convolutional layer to accept 2-channel volumetric data. The final classification head was replaced with a regression module comprising two fully connected layers (512→128→1), with ReLU activations. The model outputs a scalar between 0 and 1 that is the estimated error of the in silico labeling prediction. This is the Pearson correlation

error as explained in Figure 6, between the in silico labeling prediction and fluorescent ground truth. We use the terms prediction error and confidence score interchangeably.

Three fine-tuning modes were evaluated: full model training, partial tuning of deeper layers, and training only the regression head. Partial fine-tuning (layers in layer4 and the FC head) yielded the best tradeoff between performance and generalization.

4.5.2) Pipeline and Model Description

Using in silico labeling, a transmitted-light image is converted into a prediction of subcellular structures. Mask Interpreter then generates an importance mask for this prediction, highlighting critical regions necessary for accurate labeling. The importance mask is used to assess the quality of the prediction. The model employs a convolutional neural network (CNN) to capture patterns within the importance mask, providing insights into the prediction's reliability.

4.5.3) Supervised Learning and Output

This is a supervised model where the output is a confidence score ranging from 0 to 1. This score is the predicted error is the performance of the base in silico labeling model. The ground truth for the training process is 1 minus the Pearson Correlation Coefficient (PCC) between the predicted patch and the original fluorescence image. In training, the Mean Squared Error (MSE) between these two values is used for optimizing the model.

4.5.4) Evaluation Metrics

The evaluation of the confidence model utilizes several metrics that are measured between model’s predictions and ground truth results: Mean Squared Error, Pearson Correlation Coefficient, Spearman Coefficient, R Squared, Mutual Information and DREMI[39] with Kernel-Density Estimation, Mutual Information and DREMI with K-Nearest Neighbours. These metrics help quantify the accuracy of the confidence scores in reflecting the true quality of the predictions.

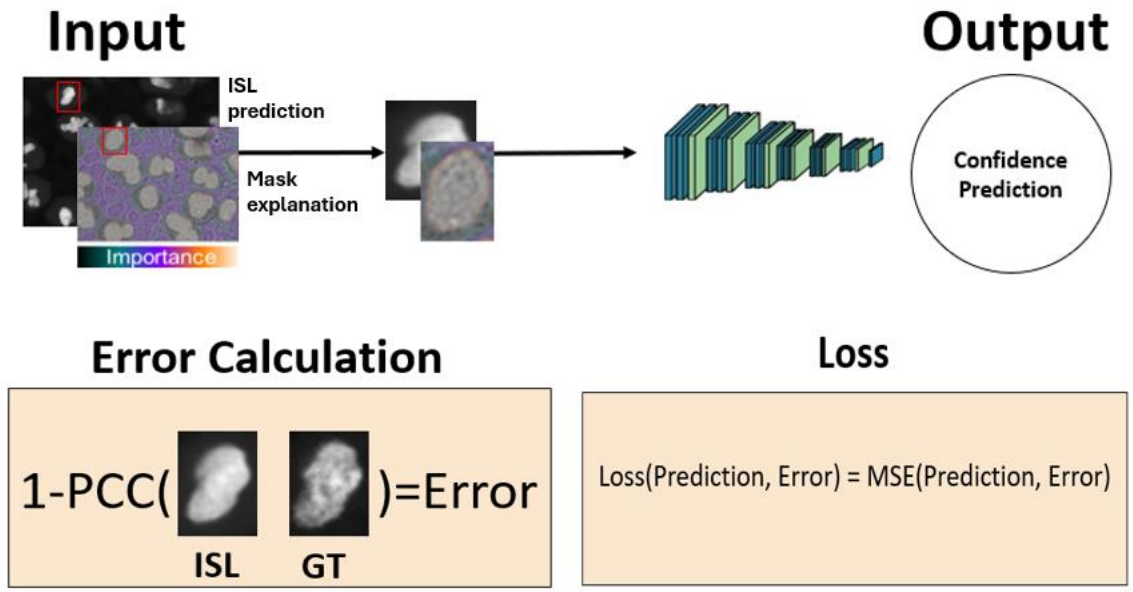

Figure 6. Confidence network pipeline. *The model takes in the in silico labeling patch and importance patch and outputs a confidence score.*

4.6) Permutation Test

We evaluated reliability using a one-sided, non-parametric permutation test on mean-squared error (MSE; lower is better). Under the null H_0, predicted errors are unrelated to true errors. As the test statistic T, we used the MSE between predicted and true errors. We permuted the true-error labels $B = 1000$ times, recomputed T for each permutation to form the null distribution, and computed the p-value with the add-one rule $\hat{p} = (1 + \#\{T_{\text{null}} \leq T_{\text{obs}}\})/(B + 1)$ (ties counted as 'as-extreme').

4.7) Ablation Study

We evaluate the confidence model in a controlled setting that isolates the contribution of explanations and probes robustness. The base condition uses the frozen in silico labeling predictor and the frozen Mask Interpreter to produce, per cell, an in silico labeling prediction and its aligned importance mask; these two channels form the input to our confidence model. As **baselines**, we compare against (i) an in silico labeling-only variant that receives the prediction channel without any mask, (ii) the mask efficacy score.

We also tested multiple inputs to the confidence model to assess which contributes the most information and improves predictions. Full results are reported in the tables that are in the supplementary materials.

4.8) Single Cell Analysis

We binned cells into confidence quartiles (Q1=highest…Q4=lowest) and tested for performance differences using a non-parametric Kruskal-Wallis omnibus test. Upon significance, we ran pairwise Mann-Whitney U tests with Benjamini-Hochberg FDR correction and report Cliff's Δ as effect size. Because quartiles are ordered, we also report a Spearman trend between quartile index and the metric.

5) Results

Sections 5.1.1, 5.1.2 and 5.2 discuss results at the patch resolution. Sections 5.1.3 and 5.3 discuss results at the full FOV or single-cell resolution. Moving from the patch resolution to other resolutions is feasible using the pipeline described in Methods: Postprocessing.

5.1) Confidence Model Identifies Faulty Predictions

5.1.1) Main Results

To ensure the method is robust, we chose to look at 8 organelles: Actin Filaments, DNA (not an organelle but can be predicted), Endoplasmic Reticulum, Microtubules, Mitochondria, Nuclear Envelope, Nucleolus Granular Components and the Plasma Membrane. Across all evaluated organelles, the confidence model's predicted error closely tracked the true error in a held-out test set (see Methods: Evaluation metrics). A reminder, the error we predict in our confidence model is the difference between the Pearson correlation of the in silico labeling and the ground truth image from a perfect score of 1 (specified in Figure 6).
Predicted errors and true errors for all organelles are plotted in Figure 7. Full statistical analysis between true and predicted errors is reported in supplementary tables 1-7. Results show promising potential in the method's ability to detect unreliable predictions. True and predicted errors align on the red diagonal in the graph (X=Y), signifying that although in silico labeling quality varies across organelles, our model is consistently able to detect prediction errors and quantify them closely to the true error. This opens up the possibility for practical use of in silico labeling in biological pipelines. We expand this idea in the discussion.

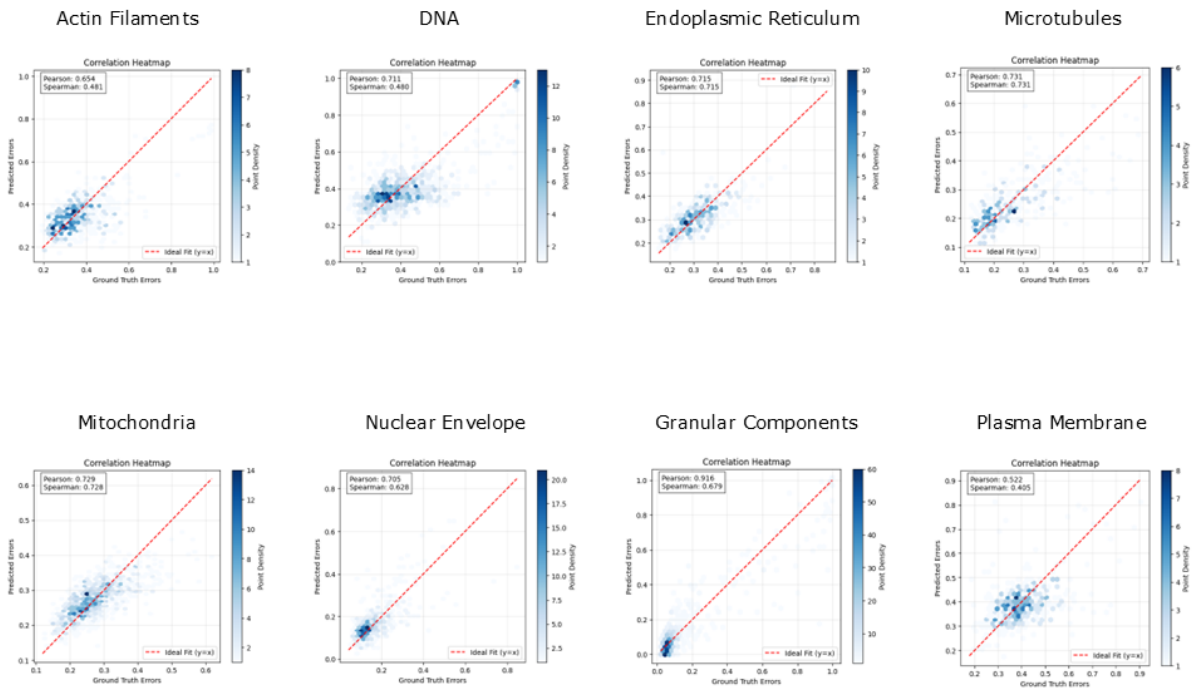

Figure 7. Confidence model accurately predicts in silico labeling errors across organelles.
Correlation heatmaps of predicted (y-axis) vs. ground truth (x-axis) errors for all eight organelles in the test set. Each subplot shows the Pearson and Spearman correlation coefficients between the model's confidence prediction and the true in silico labeling error (1 – PCC). The red dashed line indicates ideal fit. This analysis is in the patch-level, colorbar show the density by number of patches. High correlations demonstrate the generalizability of our explanation-driven confidence model across diverse structural targets.

5.1.2) Inspection of Results at the Patch Level

Motivated by the model's performance, we wanted to understand how patches who receive different confidence scores look. We chose to focus on the Nucleolus Granular Components for this analysis as they're fairly easy to predict and we expect mistakes to be meaningful (have a biological significance).

We divided the graph into four groups: low error and low prediction, high error and high prediction, mild error and mild prediction, high error and low prediction.

Each figure in this subsection shows the correlation heatmap from the main results next to the full graph, in which every point is a patch. The black circle shows which group of patches we analyze. Under the graphs, an example representing the group is shown. Top row contains the bright field, in silico labeling prediction and Mask Interpreter explanation. Bottom row contains the fluorescence ground truth next to the in silico labeling prediction (repeated for

clear visualization) and the DNA channel. When looking at the low error and low prediction group (Figure 8), in silico labeling model works as expected. The Granular Components have a clear shape and are accurately predicted from the bright-field image. The explanation from Mask Interpreter is complete and has a clear round shape which is typical for this organelle. The DNA channel helps us ensure that we are in fact looking at a nucleus. Most patches belong to this group as clearly seen in the heatmap and we conclude that these patches are reliable.

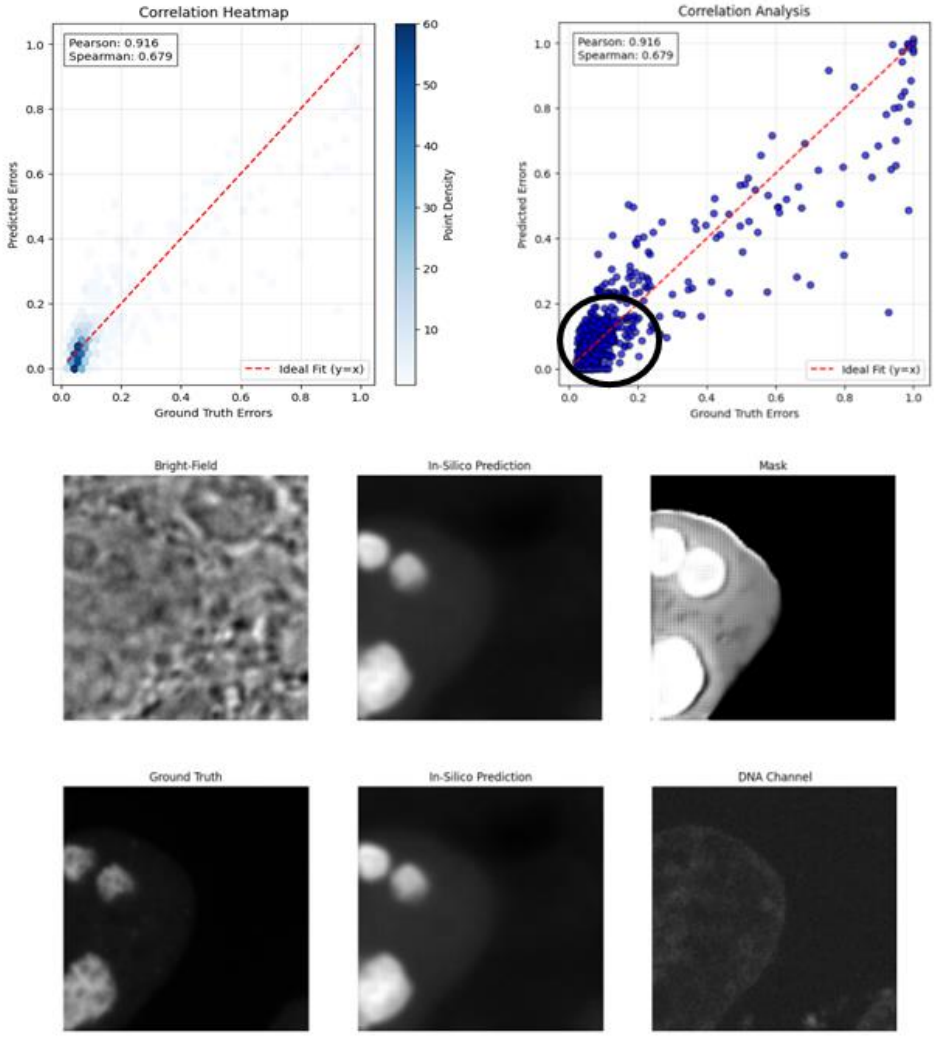

Figure 8. Low error and low prediction region in the graph.

Moving forward to the next group, containing high error and high predictions (Figure 9), we mostly see noise or blur. This happens most likely because the patch does not contain cells or because of experimental artifacts. Accordingly, Mask Interpreter prediction shows no important pixels and DNA channel shows no sign of cellular components. These patches are not reliable but very easily discovered.

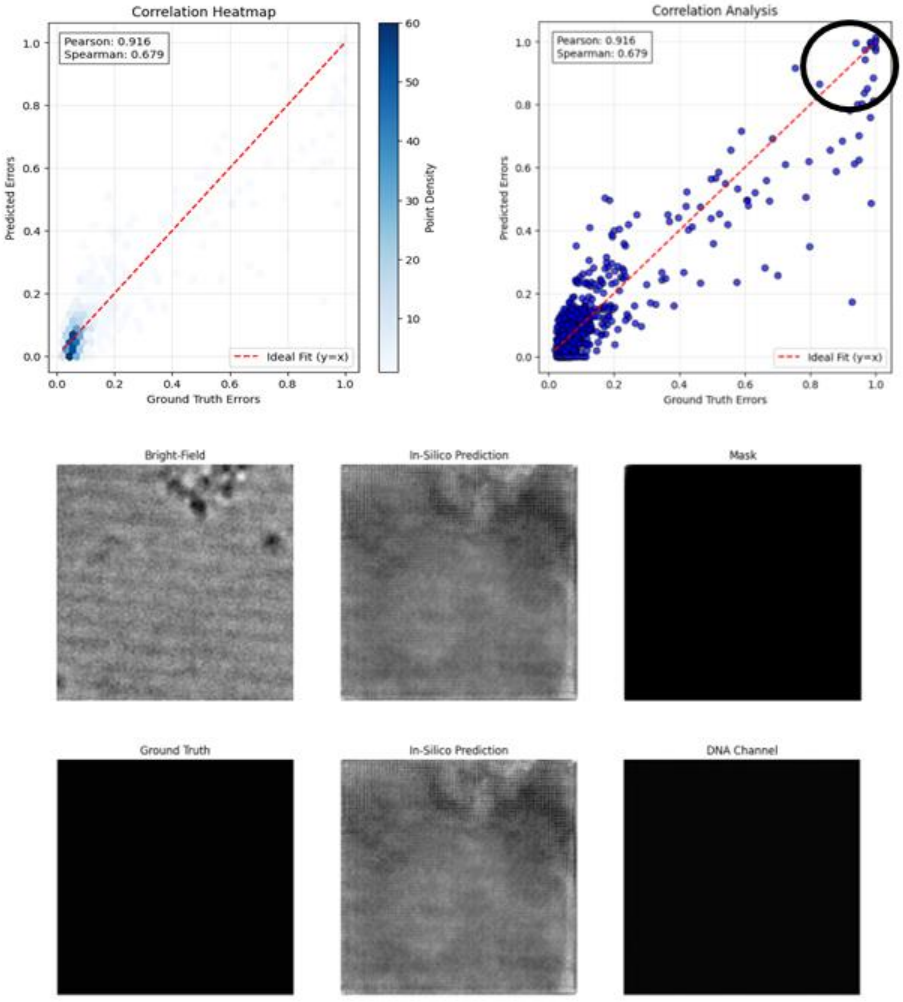

Figure 9. High error and high prediction region in the graph.

The third group is more interesting. It contains patches with a mild error and mild prediction (Figure 10). The in silico labeling model fails in these cases, however our confidence model recognizes the failure and the confidence score aligns with true performance. The prediction on the patch level looks real but looking at the ground truth we don't see an organelle. The explanation from Mask Interpreter does have a normal structure, however the values of the pixels are lower in comparison to group 1 which could signify lack of ability to predict on

this patch. The DNA channel shows that there is some structure in the patch, but unlike in group 1, this structure is not well defined and is most likely either a cell during mitosis or faulty labeling of the fluorescence marker. This group is quite rare and these patches are unreliable, and our model captures it.

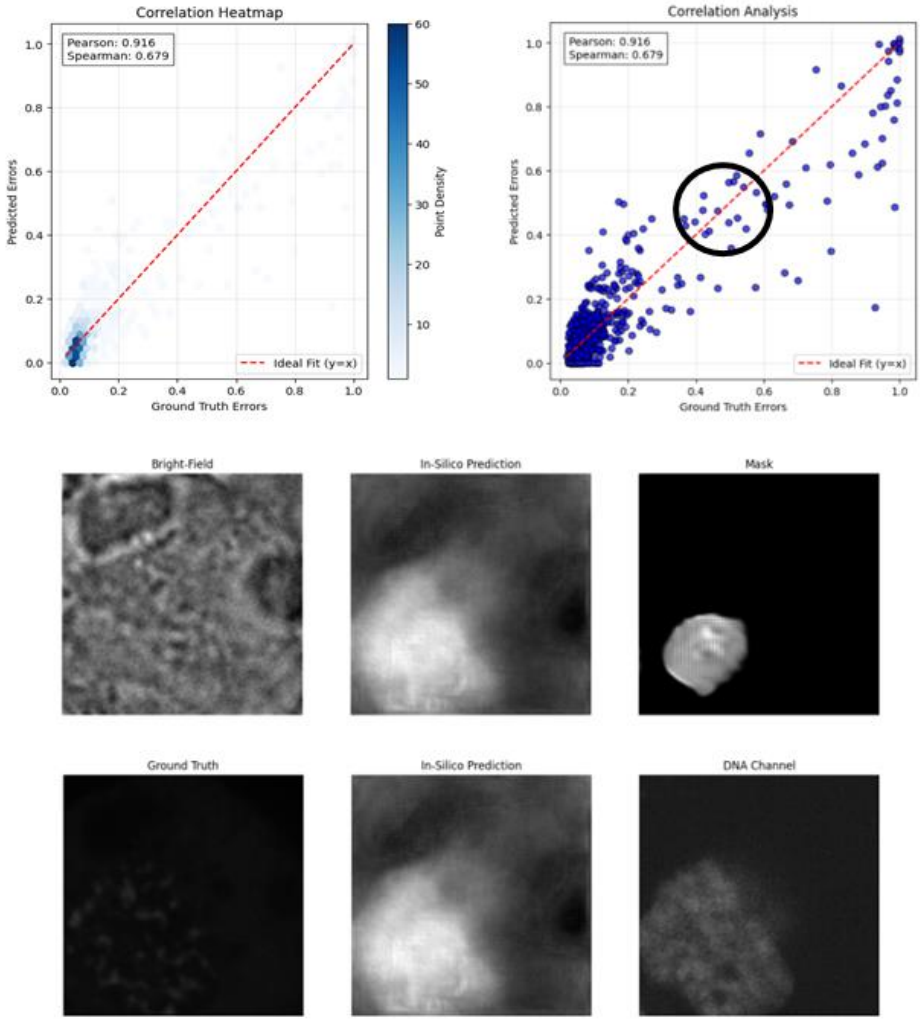

Figure 10. Mild error and mild prediction region in the graph.

Lastly, the fourth group contains patches with a high error and low prediction (Figure 11). This is where our model misses the true quality of in silico labeling predictions. The prediction looks real but again, there's no organelle. The explanation from Mask Interpreter looks convincing which could be why our model tends to trust this prediction. The DNA channel helps us verify there is no biological process happening and that this is a full hallucination that isn't captured. This group is very rare and these patches are unreliable. We

analyzed these patches further and found that this hallucination happens mostly in the edges of patches and they are not consistent when shifting the location of the patch by a few pixels. As we wanted to avoid these mistakes and to get a more holistic view of confidence at the cell level and not at the patch level, the natural thing to do was to apply the same pipeline of the original in silico labelling model (Methods: Postprocessing). This pipeline solves these inconsistent hallucinations by averaging confidence from overlapping patches while applying more weight to the middle of the patch. It also let's us move from the patch resolution to the FOV resolution, and ultimately to the single-cell resolution. The following section presents the results of this process.

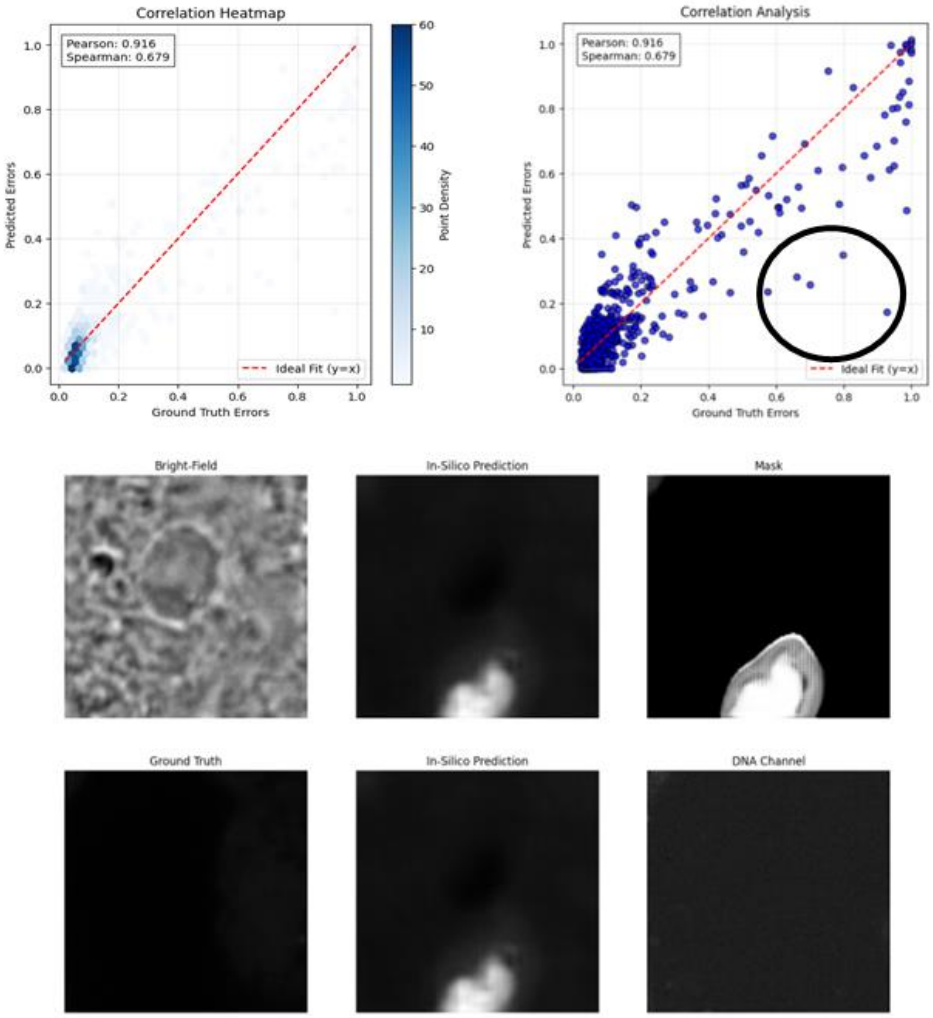

Figure 11. High error and low prediction region in the graph.

5.1.3) Confidence Heatmap Generation

This section deals with the move from patch resolution to single-cell resolution. After applying the triangle weight function at inference, we were able to get a FOV-level confidence. We decided to visualize this confidence map over the in silico labeling prediction, to see which parts of the prediction are reliable (Figure 12, left). We also wanted to compare this results with the true errors in prediction and to do so we plotted them over the in silico labeling prediction as well. (Figure 12, right). This gives us a strong tool to see which cells or organelles cannot be trusted, and how this compares to the true performance of the in silico labeling model. Examples for additional organelles can be found in Figure 15 in the supplementary materials.

Now that we have the FOV-level confidence, we can segment cells and observe how a practical biological task is done on cells at different confidence levels. We present this analysis in section 5.3.

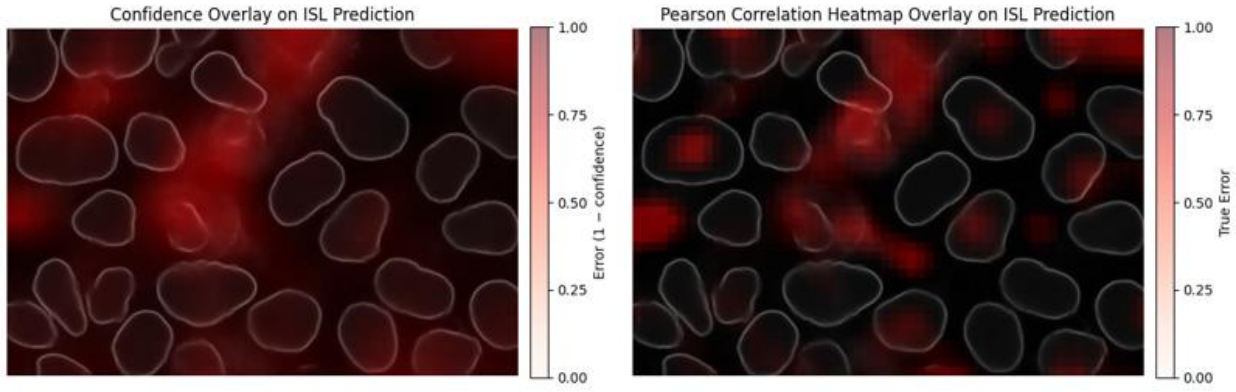

Figure 12. Confidence model enhanced with mask interpreter outperforms baselines and predicts unreliable regions.

Visualization of confidence and actual error over the in silico labeling prediction. Left: Confidence overlay highlighting low-confidence regions (in red) overlaid on the in silico labeling output. Right: Pearson correlation heatmap (1 - PCC) overlaid on the in silico labeling output, marking true error regions. The spatial correspondence between low confidence and high error regions confirms the model's ability to identify underperforming predictions without access to ground truth.

5.2) Incorporating Importance Masks Improves Model's Performance

The ablation and permutation tests described here are done at the patch level. To validate the reliability of our model's confidence predictions, we conducted two control analyses. First, we performed a permutation test as described in "Methods: Permutation Test". Our model consistently outperformed even the best-scoring permutation, confirming that it captures a meaningful relationship between inputs and true in silico labeling error. For all organelles, we received a p-value of 0.001 ($B = 1000$).

Second, we compared the model's predictions to a naïve baseline that always outputs the global mean error prediction. Our confidence model outperformed this baseline as well, demonstrating that it does not merely learn the error distribution but instead tailors its predictions to the content of each input patch.

Beyond validation against randomized and naïve baselines, we further evaluated our model by comparing it to alternative error prediction strategies and trained baselines. This comparison was done at the patch level as all models were trained and tested on this level. Specifically, we benchmarked against the mask efficacy score (see Methods: Ablation Study), which measures the performance drop when perturbing high-importance regions. We also compared against a learned baseline model trained to predict error using only the in silico labeling output, without access to the corresponding explanation mask. While this in silico labeling-only model does capture some predictive signal, it lacks the contextual information provided by importance masks.

Across all organelles tested, our confidence model consistently outperformed both baselines. It was always **significantly better** than the mask efficacy score, and either **matched or surpassed** the performance of the in silico labeling-only model. Full comparison of dual-input models and single-input models are reported in supplementary tables 1-7.

5.3) Applicative Use of the Confidence Model for Single Cell Analysis

We wanted to show a practical application in which the confidence model could be useful for a biologist. As mentioned before, we made the transition from the patch-level to the FOV-level. With this holistic view of confidence across regions, we were now able to segment

cells from the full FOV and look at the single-cell level to determine which organelles can be trusted. The application we chose is organelle segmentation. Organelle segmentation transforms microscopy from pictures into quantitative, reproducible, and mechanistically interpretable measurements. We hypothesized that where the confidence is high, the segmentation of an organelle from the in silico labeling prediction will be close to that of the ground truth fluorescence image. Organelles chosen for this analysis were the Granular Components, Nuclear Envelope, Microtubules and Mitochondria. We followed Allan Institute's segmentation pipeline[45] for each organelle and compared the segmentations from in silico labeling to those from the ground truth using several metrics (IoU, Dice Coefficient, SSIM). We plotted the results against the confidence score of each organelle and divided the results onto quartiles based on the level of confidence (see Methods: Single Cell Analysis). Our hypothesis was correct as shown in figure 13.

For the Nuclear Envelope, performance increased with confidence: quartiles differed (Kruskal–Wallis $H = 24.14$, $p = 2.34 \times 10^{-5}$, $\varepsilon^2 = 0.164$); medians fell from Q1/Q2 ≈0.56/0.72 to Q4 ≈0.34/0.50 (IoU/Dice). A monotonic decline was confirmed (Spearman $\rho \approx -0.41$ to -0.44, all $p < 1.5 \times 10^{-6}$). Pairwise contrasts were largest for Q1 vs Q4 (FDR $p < 2 \times 10^{-4}$; Cliff's $\Delta \approx 0.60$–0.63, large).

For the Granular Components, the separation was even clearer (Kruskal–Wallis $H = 45.91$, $p = 5.93 \times 10^{-10}$, $\varepsilon^2 = 0.254$); medians decreased from 0.854→0.769 (IoU) and 0.921→0.870 (Dice) from Q1→Q4, with a strong monotonic trend ($\rho = -0.477$, $p = 3.34 \times 10^{-11}$). Q1/Q2 vs Q4 remained highly significant after FDR (all $p_{\text{adj}} \leq 3.6 \times 10^{-8}$); effects were large (Cliff's $\Delta \approx 0.71$–0.73).

For the Microtubules, confidence ranks patches in the right direction, but evidence is limited by power; Our key takeaway is that effects should be validated with larger n.

For Mitochondria, confidence again ranked predictions (Kruskal–Wallis $H = 28.96$, $p = 2.28 \times 10^{-6}$, $\varepsilon^2 = 0.114$); medians declined 0.523→0.434 (IoU) and 0.687→0.605 (Dice), with a clear trend ($\rho = -0.337$, $p = 1.47 \times 10^{-7}$). Q1/Q2/Q3 each exceeded Q4 after FDR (all $p_{\text{adj}} \leq 0.0019$); Q1 vs Q4 showed a **large** effect ($\Delta \approx 0.53$). SSIM showed the strongest separation (Kruskal–Wallis $H = 56.44$, $p = 3.38 \times 10^{-12}$, $\varepsilon^2 = 0.235$; Q1 vs Q4 $\Delta \approx 0.71$).

The fact that the confidence score was a reliable estimator of segmentation accuracy across all metrics, and both in high segmentation performance organelles (Granular Components, Nuclear Envelop) and in low segmentation performance organelles which harder to predict (Microtubules, Mitochondria), underscores its generality and robustness, supporting its use

for trustworthy quality control, reviewer triage, and active data selection across diverse morphologies and imaging conditions.

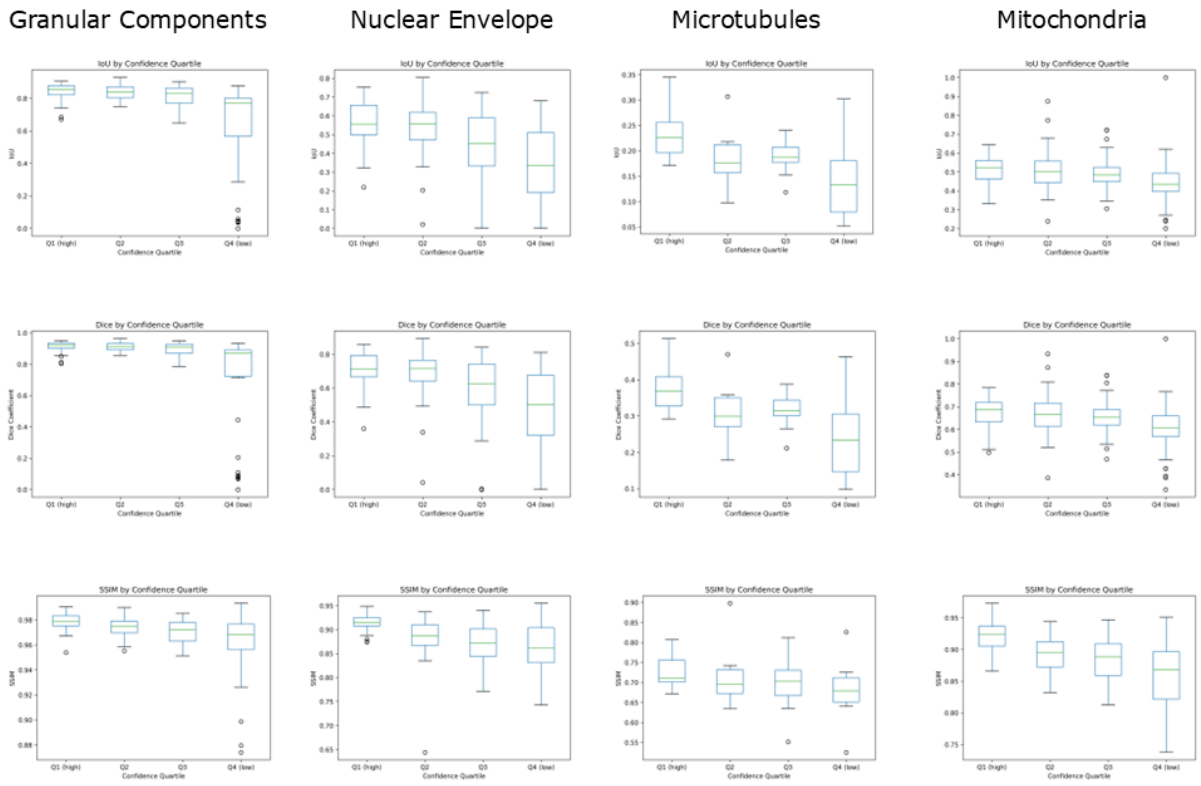

Figure 13. Confidence model predictions align with other known metrics.

Box plots show the distribution of Intersection over Union (IoU, top row), Dice coefficient (middle row), and Structural Similarity Index Measure (SSIM, bottom row) between in silico labeling predictions and fluorescence ground truth, across four organelles (columns): Granular Components (173 cells), Nuclear Envelope (133 cells), Microtubules (33 cells), and Mitochondria (231 cells). Cells were grouped into quartiles (Q1–Q4) based on predicted confidence scores, with Q1 corresponding to the highest confidence and Q4 to the lowest. Metrics were computed per cell after segmentation and thresholding of in silico labeling predictions and ground truth images.

5.4) Applying Mask Interpreter to Single Cell Modality

Independent of the confidence model, we sought to show that Mask Interpreter adapts to the predictive context of the model it explains. Demonstrating this adaptation strengthens our motivation of incorporating Mask Interpreter predictions in our pipeline as it explains why adding importance masks improves calibration and robustness. We chose to focus on the nuclear envelope for this analysis, as it had an interesting Mask Interpreter explanation: the explanation showed that apart from the envelope itself, the in silico model also uses the nucleus as context for prediction (Figure 5). This is a non-trivial explanation as the envelope is quite clear and should be sufficient for prediction. We hypothesize that there are many elliptical envelope-like patterns in the input patches and that leads the model to use the nucleus to verify it deals with a cell. We also hypothesized that if the model was trained on single cells, it would yield a different explanation mask. We decided to train a new in silico labeling model to test our hypothesis. We used a custom dataset of images that contained only a single cell. This dataset was created from the original Allen Institute's dataset so apart from the single-cell feature the datasets were identical. When we trained an in silico labeling model and the according Mask Interpreter model on this single-cell dataset, we discovered that Mask Interpreter produces importance masks that no longer emphasize the nucleus as necessary context, unlike the explanations from the original patch-based model where the nucleus is prominently highlighted. We sought to verify our findings: we chose a single cell from a specific FOV and tested how perturbing its nucleus affects the in silico prediction in the original model vs in the single-cell model. We injected noise around the nucleus in the input, bright-field image, and expected the performance to deteriorate in the original patch-based in silico model, but not in the new single-cell model. Figure 14 Shows this validation, where noise is added gradually and our expectations are met. This result illustrates the power of Mask Interpreter.

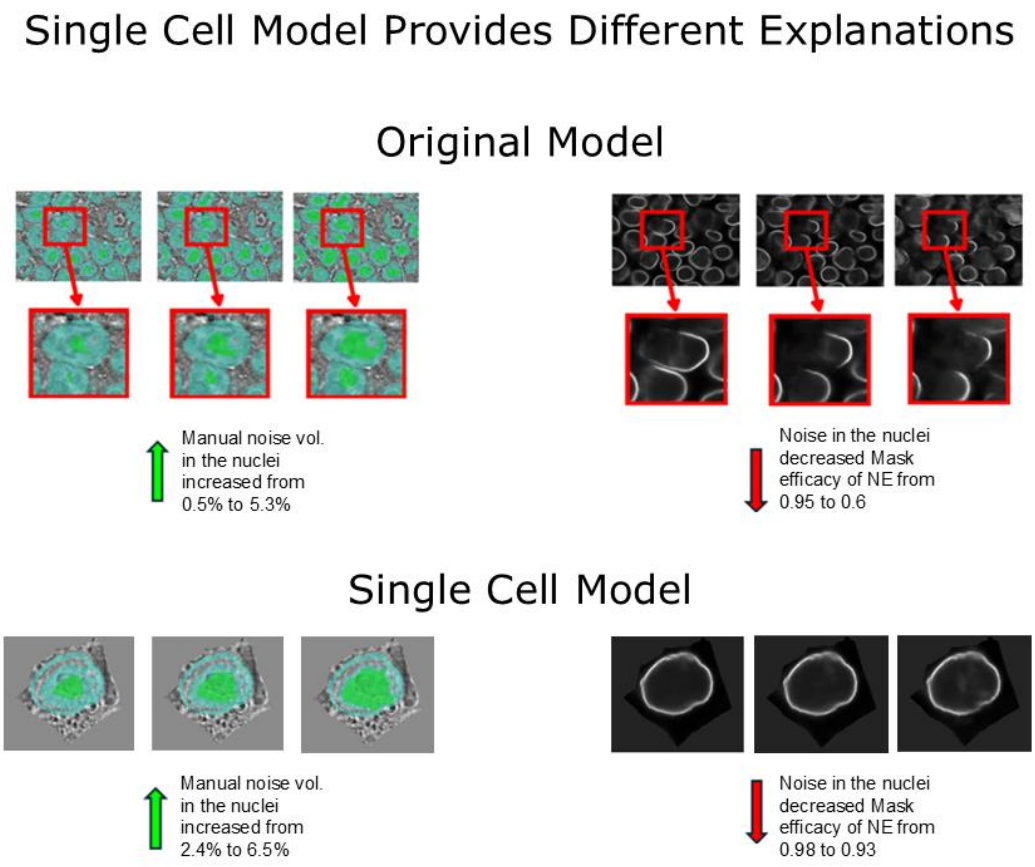

Figure 14-Single cell nuclear envelope prediction doesn't require inner context.

Top images show original in silico labeling model's explanations with different levels of manual noise (left) and the original model's predictions and performance in comparison to clean predictions (right). Bottom images show the same analysis but this time for a model that trains on a single cell dataset. More noise is introduced and performance is still high.

6) Conclusions & Discussion

In silico labeling is emerging as a powerful alternative to fluorescence microscopy, enabling multi-organelle inference from label-free images while mitigating multiplexing limits and phototoxicity, which is essential for long-term live-cell studies[1]. Yet, broader adoption has been slowed by concerns about interpretability and generalizability across laboratories, microscopes, and cellular states[9]. Our study addresses these barriers by turning model behavior into human-readable evidence. Rather than asking users to trust an opaque image-to-image translation model, we expose what visual context the model uses and quantify when its outputs are likely to be reliable.

Using Mask Interpreter, we probe our in silico labeling pipeline and relate explanations to reliability. Across our data, different organelles show distinct and reproducible bright-field context requirements, indicating reliance on structured cues rather than incidental correlations. Explanation signatures also act as sensitive indicators of out-of-distribution inputs: in mitotic cells they become non-stereotypical and coincide with degraded predictions, consistent with altered intracellular organization. The behavior is model-aware in practice as well; when used on a single-cell nuclear-envelope model, the masks no longer emphasize nuclear interior context seen with models trained on confluent scenes, likely because the single-cell formulation reduces reliance on neighboring nuclei. Finally, explanation-mask efficacy provides a practical, label-free quality control signal for batch effects and other acquisition-specific issues by flagging cases where the model appears to rely on weak, inconsistent, or spatially misplaced evidence.

We leverage Mask Interpreter explanations in the development of our model: a confidence model that predicts in silico labeling quality from the in silico labeling output and its explanation mask. The resulting confidence scores correlate with real errors across metrics and organelles, and are especially discriminative in edge cases such as mitosis and apoptosis. This enables triage of high-confidence cells for downstream quantification and cautious handling or review of low-confidence cases. The approach does require reliable supervision during training and inherits any biases in explanation quality, which motivates continued work on explanation calibration and label-efficient training strategies.

Our results also help situate where future performance gains are likely to come from. Architectural changes alone have produced only marginal improvements for conventional transmitted-light inputs, hinting that we are approaching an information ceiling set by the

physical specificity of the modality[3,10]. Complementary imaging such as quantitative phase imaging can raise that ceiling by injecting richer, organelle-specific contrast and has already shown promise for virtual staining and live-cell phenotyping[7,26]. Temporal information is another underused axis. Enforcing consistency across frames and exploiting motion patterns can disambiguate morphologically similar organelles and stabilize predictions, as shown in related segmentation studies and proposed in recent spatiotemporal in silico labeling designs[27]. These directions are compatible with our framework, since Mask Interpreter and confidence estimation can be extended to quantify how temporal or phase cues are utilized.

Equally important is how we evaluate success. Pixel-wise scores such as correlation, MAE, MSE, SSIM, or PSNR are convenient but abstract for biological interpretation and can propagate errors into downstream analyses[28]. We advocate pairing them with application-specific metrics that reflect the intended use, for example per-organelle localization, counts, shape descriptors, and contact statistics[5]. Public benchmarks that include label-free inputs, fluorescent ground truth, and validated per-organelle segmentation, such as the Allen Institute resources, can accelerate this shift by standardizing practical evaluations and enabling automated selection of imaging conditions for a given assay.

Finally, our findings point to a closed-loop path for robust deployment. Explanations diagnose where and why a model fails, confidence scores quantify when to trust predictions, and both can drive interventions that improve the pipeline. Looking ahead, semantic confidence can trigger event-driven microscopy that acquires fluorescence only when and where predicted reliability dips, reducing phototoxicity while maintaining accuracy[29,30,31]. In parallel, combining in silico labeling of well-localized structures with sparse ground truth for harder targets can make high-dimensional live-cell assays feasible at scale.

In summary, our Mask Interpreter based confidence model turn in silico labeling from a black box into a measured, inspectable system. By revealing the spatial evidence behind predictions, surfacing OOD states and batch effects, and enabling reliability-aware post-processing, our framework improves the interpretability, trustworthiness, and practical value of in silico labeling. The same principles are general and can extend to other image-to-image translation problems in biomedical imaging, particularly when paired with richer physics-based contrast, temporal modeling, and application-grounded benchmarks.

7) Supplementary Materials

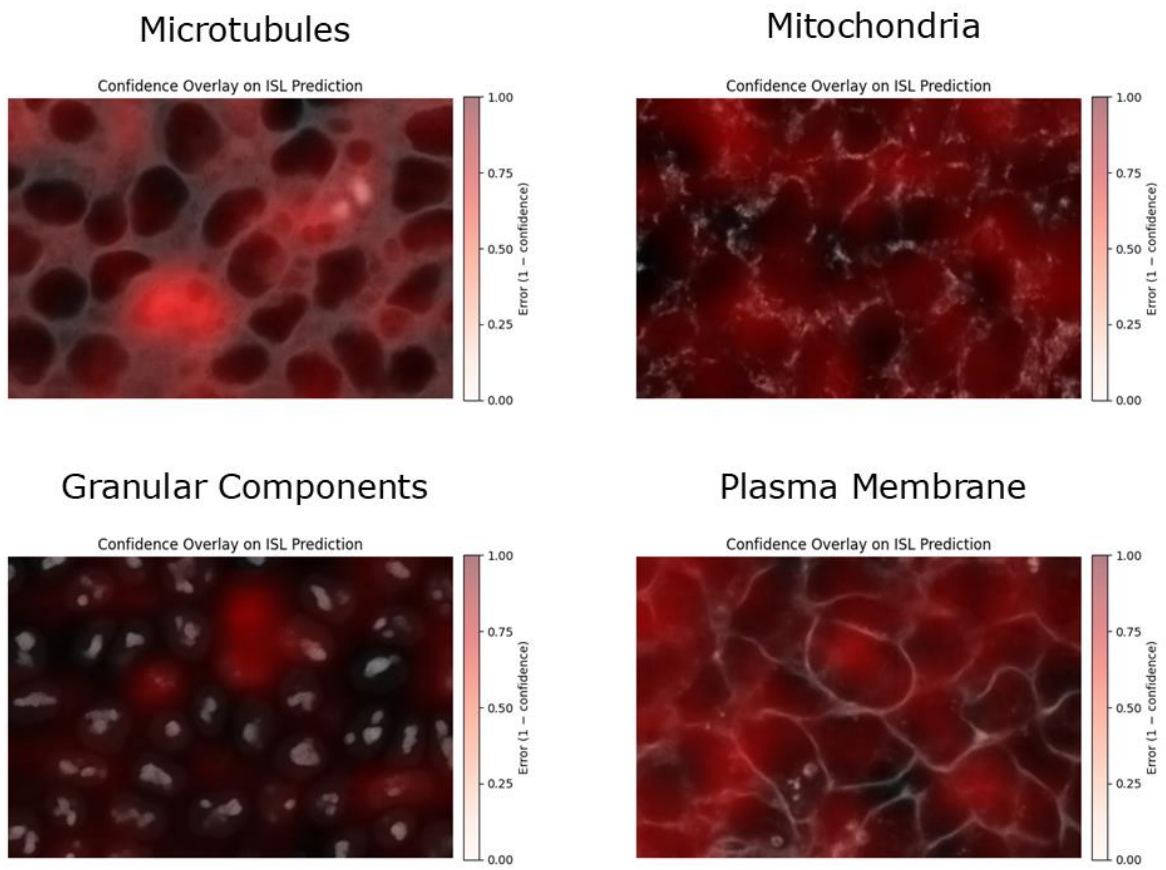

Figure 15. Confidence overlays identify low-quality in silico labeling regions across organelles.

Overlay visualizations showing confidence on top of in silico labeling predictions for microtubules, mitochondria, granular components, and plasma membrane. Brighter red indicates lower confidence, often corresponding to regions of signal dropout, morphological complexity, or ambiguous structure. These overlays highlight the model's ability to detect spatially localized prediction failures without access to ground truth, making them useful for quality control in downstream phenotypic analysis.

Nucleolus Granular Components

Model Input	MSE	PCC	Spearman	R-Squared	KDE MI	KDE DREMI	KNN MI	KNN DREMI
ISL Prediction	0.0093	0.8690	0.6260	0.7504	0.3100	0.6200	0.4200	0.5500
ISL Prediction+Noisy ISL Prediction	0.0065	0.9110	0.6150	0.8258	0.3600	0.7100	0.4500	0.6600
ISL Prediction+Mask Prediction	0.0061	0.9160	0.6790	0.8364	0.3700	0.7200	0.5200	0.6700
ISL Prediction+Difference in Predictions	0.0067	0.9080	0.6880	0.8221	0.3400	0.7100	0.4600	0.6700
Noisy ISL Prediction	0.0068	0.9050	0.6590	0.8172	0.3500	0.7000	0.4800	0.6700
Noisy ISL Prediction+Mask Prediction	0.0073	0.9030	0.6800	0.8058	0.3300	0.6900	0.4800	0.6600
Noisy ISL Prediction+Difference in Predictions	0.0058	0.9200	0.5870	0.8446	0.3400	0.7200	0.4400	0.7200
Mask Prediction	0.0089	0.8890	0.5250	0.7626	0.2900	0.6100	0.4500	0.6800
Mask Prediction+Difference in Prediction	0.0079	0.9030	0.5800	0.7890	0.3300	0.6600	0.4600	0.7000
Difference in Prediction	0.0077	0.8940	0.5060	0.7947	0.3100	0.6800	0.4000	0.6600
Bright Field	0.0184	0.7240	0.2670	0.5067	0.1300	0.3800	0.2100	0.4200

Table 1. Nucleolus Granular Components results.

Nuclear Envelope

Model Input	MSE	PCC	Spearman	R-Squared	KDE MI	KDE DREMI	KNN MI	KNN DREMI
ISL Prediction	0.0048	0.6660	0.5770	0.3839	0.2100	0.5400	0.2800	0.3800
ISL Prediction+Noisy ISL Prediction	0.0050	0.6190	0.5350	0.3561	0.1700	0.4900	0.2300	0.3200
ISL Prediction+Mask Prediction	0.0043	0.7050	0.6280	0.4381	0.2100	0.5000	0.3100	0.3800
ISL Prediction+Difference in Predictions	0.0051	0.6160	0.5320	0.3412	0.1800	0.4900	0.2400	0.3300
Noisy ISL Prediction	0.0076	0.3970	0.3110	0.0159	0.0800	0.2700	0.1300	0.1900
Noisy ISL Prediction+Mask Prediction	0.0059	0.5500	0.4400	0.2394	0.1300	0.5100	0.1600	0.2700
Noisy ISL Prediction+Difference in Predictions	0.0049	0.6520	0.5720	0.3646	0.2000	0.5600	0.2700	0.3700
Mask Prediction	0.0058	0.6020	0.4900	0.2513	0.1500	0.5500	0.1900	0.3000
Mask Prediction+Difference in Prediction	0.0056	0.5490	0.4280	0.2780	0.1500	0.4800	0.1700	0.3000
Difference in Prediction	0.0054	0.5930	0.5560	0.2954	0.1500	0.4500	0.2300	0.3300
Bright Field	0.0073	0.3670	0.3190	0.0497	0.0700	0.1400	0.1100	0.1300

Table 2. Nuclear Envelope results.

Mitochondria

Model Input	MSE	PCC	Spearman	R-Squared	KDE MI	KDE DREMI	KNN MI	KNN DREMI
ISL Prediction	0.0026	0.7100	0.6830	0.4939	0.2500	0.6000	0.3400	0.4800
ISL Prediction+Noisy ISL Prediction	0.0028	0.6940	0.6670	0.4596	0.2400	0.6000	0.3300	0.4500
ISL Prediction+Mask Prediction	0.0024	0.7290	0.7280	0.5288	0.2700	0.6200	0.4100	0.5000
ISL Prediction+Difference in Predictions	0.0022	0.7640	0.7370	0.5738	0.3000	0.6800	0.4500	0.5600
Noisy ISL Prediction	0.0031	0.6910	0.6590	0.4110	0.2300	0.5900	0.3300	0.4600
Noisy ISL Prediction+Mask Prediction	0.0029	0.6860	0.6600	0.4499	0.2400	0.6700	0.3200	0.4900
Noisy ISL Prediction+Difference in Predictions	0.0035	0.5820	0.5710	0.3198	0.1600	0.4100	0.2400	0.3400
Mask Prediction	0.0036	0.5880	0.5410	0.3145	0.1600	0.5100	0.2200	0.3500
Mask Prediction+Difference in Prediction	0.0028	0.6730	0.6520	0.4515	0.2200	0.6100	0.3200	0.4300
Difference in Prediction	0.0038	0.5600	0.5300	0.2708	0.1500	0.4600	0.2100	0.3300
Bright Field	0.0050	0.3310	0.2740	0.0310	0.0400	0.1800	0.0800	0.1500

Table 3. Mitochondria results.

Actin Filaments

Model Input	MSE	PCC	Spearman	R-Squared	KDE MI	KDE DREMI	KNN MI	KNN DREMI
ISL Prediction	0.0089	0.6220	0.4120	0.3398	0.0900	0.5600	0.1300	0.2600
Mask Prediction	0.0097	0.5650	0.2660	0.2822	0.0800	0.6700	0.0800	0.2300
ISL Prediction+Mask Prediction	0.0081	0.6540	0.4810	0.4003	0.1300	0.6100	0.1800	0.2600
ISL Prediction+Noisy ISL Prediction	0.0073	0.6880	0.5100	0.4543	0.1000	0.5500	0.1700	0.2900
ISL Prediction+Difference in Predictions	0.0087	0.6100	0.3960	0.3499	0.0900	0.4800	0.1200	0.2500
Noisy ISL Prediction+Difference in Predictions	0.0080	0.6390	0.4280	0.4036	0.1000	0.6500	0.1400	0.2800

Table 4. Actin Filaments results.

Endoplasmic Reticulum

Model Input	MSE	PCC	Spearman	R-Squared	KDE MI	KDE DREMI	KNN MI	KNN DREMI
ISL Prediction	0.0049	0.7020	0.6740	0.4281	0.2300	0.8400	0.2800	0.3600
Mask Prediction	0.0057	0.6250	0.5820	0.3266	0.1800	0.7400	0.2500	0.3400
ISL Prediction+Mask Prediction	0.0044	0.7150	0.7150	0.4774	0.2400	0.8400	0.3400	0.4100
ISL Prediction+Noisy ISL Prediction	0.0054	0.6630	0.6270	0.3596	0.1900	0.7200	0.2700	0.3800
ISL Prediction+Difference in Predictions	0.0043	0.7460	0.6610	0.4972	0.2300	0.8700	0.2800	0.3500
Noisy ISL Prediction+Difference in Predictions	0.0050	0.6610	0.6170	0.4130	0.1800	0.7600	0.2600	0.3600

Table 5. Endoplasmic Reticulum results.

Microtubules

Model Input	MSE	PCC	Spearman	R-Squared	KDE MI	KDE DREMI	KNN MI	KNN DREMI
ISL Prediction	0.0062	0.7300	0.7030	0.5042	0.2500	0.4000	0.3100	0.3400
Mask Prediction	0.0083	0.6060	0.6040	0.3433	0.1600	0.2900	0.2100	0.2200
ISL Prediction+Mask Prediction	0.0060	0.7310	0.7310	0.5225	0.2500	0.4300	0.3200	0.3200
ISL Prediction+Noisy ISL Prediction	0.0064	0.7020	0.6970	0.4917	0.2500	0.4000	0.3000	0.3300
ISL Prediction+Difference in Predictions	0.0076	0.6770	0.6590	0.3958	0.2100	0.3600	0.2600	0.2600
Noisy ISL Prediction+Difference in Predictions	0.0088	0.6350	0.6180	0.2967	0.1800	0.3400	0.2300	0.2800

Table 6. Microtubules results.

Plasma Membrane

Model Input	MSE	PCC	Spearman	R-Squared	KDE MI	KDE DREMI	KNN MI	KNN DREMI
ISL Prediction	0.0105	0.4490	0.3850	0.1345	0.1000	0.3500	0.1200	0.2100
Mask Prediction	0.0097	0.5230	0.4110	0.1949	0.1300	0.3300	0.1400	0.2300
ISL Prediction+Mask Prediction	0.0100	0.5220	0.4050	0.1738	0.1200	0.4000	0.1400	0.2500
ISL Prediction+Noisy ISL Prediction	0.0095	0.4850	0.4310	0.2137	0.1000	0.2900	0.1300	0.2000
ISL Prediction+Difference in Predictions	0.0105	0.4960	0.3760	0.1326	0.1000	0.4000	0.1400	0.2500
Noisy ISL Prediction+Difference in Predictions	0.0104	0.5210	0.3940	0.1383	0.1000	0.4000	0.1200	0.2200

Table 7. Plasma Membrane results.

8) Bibliography

[1] Garini Y, Young IT, McNamara G. Spectral imaging: principles and applications. *Cytometry A*. 2006;69(8):735-747. doi:10.1002/cyto.a.20311

[2] Icha J, Weber M, Waters JC, Norden C. Phototoxicity in live fluorescence microscopy, and how to avoid it. *Bioessays*. 2017;39(8):10.1002/bies.201700003. doi:10.1002/bies.201700003

[3] Ounkomol C, Seshamani S, Maleckar MM, Collman F, Johnson GR. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nat Methods*. 2018;15(11):917-920. doi:10.1038/s41592-018-0111-2

[4] Christiansen EM, Yang SJ, Ando DM, et al. In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images. *Cell*. 2018;173(3):792-803.e19. doi:10.1016/j.cell.2018.03.040

[5] LaChance J, Cohen DJ. Practical fluorescence reconstruction microscopy for large samples and low-magnification imaging. *PLoS Comput Biol*. 2020;16(12):e1008443. Published 2020 Dec 23. doi:10.1371/journal.pcbi.1008443

[6] Guo SM, Yeh LH, Folkesson J, et al. Revealing architectural order with quantitative label-free imaging and deep learning. *Elife*. 2020;9:e55502. Published 2020 Jul 27. doi:10.7554/eLife.55502

[7] Cheng S, Fu S, Kim YM, et al. Single-cell cytometry via multiplexed fluorescence prediction by label-free reflectance microscopy. *Sci Adv*. 2021;7(3):eabe0431. Published 2021 Jan 15. doi:10.1126/sciadv.abe0431

[8] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

[9] Elmalam N, Ben Nedava L, Zaritsky A. In silico labeling in cell biology: Potential and limitations. *Curr Opin Cell Biol*. 2024;89:102378. doi:10.1016/j.ceb.2024.102378

[10] Wang, Z., Xie, Y. & Ji, S. Global voxel transformer networks for augmented microscopy. *Nat Mach Intell* **3**, 161–171 (2021). https://doi.org/10.1038/s42256-020-00283-x

[11] Lee G, Oh JW, Her NG, Jeong WK. DeepHCS^{++}: Bright-field to fluorescence microscopy image conversion using multi-task learning with adversarial losses for label-free high-content screening. *Med Image Anal*. 2021;70:101995. doi:10.1016/j.media.2021.101995

[12] Liu, Z., Hirata-Miyasaki, E., Pradeep, S. *et al.* Robust virtual staining of landmark organelles with Cytoland. *Nat Mach Intell* **7**, 901–915 (2025). https://doi.org/10.1038/s42256-025-01046-2

[13] Belthangady C, Royer LA. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nat Methods*. 2019;16(12):1215-1225. doi:10.1038/s41592-019-0458-z

[14] Abdar, Moloud, et al. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges." *Information fusion* 76 (2021): 243-297.

[15] Hendrycks, Dan, and Kevin Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks." *arXiv preprint arXiv:1610.02136* (2016).

[16] Guo, Chuan, et al. "On calibration of modern neural networks." *International conference on machine learning*. PMLR, 2017.

[17] Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." *international conference on machine learning*. PMLR, 2016.

[18] Kendall, Alex, and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?." *Advances in neural information processing systems* 30 (2017).

[19] Ganaie, Mudasir A., et al. "Ensemble deep learning: A review." *Engineering Applications of Artificial Intelligence* 115 (2022): 105151.

[20] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." *Advances in neural information processing systems* 30 (2017).

[21] Imboden S, Liu X, Payne MC, Hsieh CJ, Lin NYC. Trustworthy in silico cell labeling via ensemble-based image translation. *Biophys Rep (N Y)*. 2023;3(4):100133. Published 2023 Oct 18. doi:10.1016/j.bpr.2023.100133

[22] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

[23] Adebayo, Julius, et al. "Sanity checks for saliency maps." *Advances in neural information processing systems* 31 (2018).

[24] Fong, Ruth C., and Andrea Vedaldi. "Interpretable explanations of black boxes by meaningful perturbation." *Proceedings of the IEEE international conference on computer vision*. 2017.

[25] DeVries, Terrance, and Graham W. Taylor. "Leveraging uncertainty estimates for predicting segmentation quality." *arXiv preprint arXiv:1807.00502* (2018).

[26] Bai, Bijie, et al. "Deep learning-enabled virtual histological staining of biological samples." *Light: Science & Applications* 12.1 (2023): 57.

[27] Arbelle, Assaf, Shaked Cohen, and Tammy Riklin Raviv. "Dual-task ConvLSTM-UNet for instance segmentation of weakly annotated microscopy videos." *IEEE Transactions on Medical Imaging* 41.8 (2022): 1948-1960.

[28] Tonks, Samuel, et al. "Evaluating virtual staining for high-throughput screening." *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023.

[29] Balzarotti F, Eilers Y, Gwosch KC, et al. Nanometer resolution imaging and tracking of fluorescent molecules with minimal photon fluxes. *Science*. 2017;355(6325):606-612. doi:10.1126/science.aak9913

[30] Royer LA, Lemon WC, Chhetri RK, et al. Adaptive light-sheet microscopy for long-term, high-resolution imaging in living organisms. *Nat Biotechnol*. 2016;34(12):1267-1278. doi:10.1038/nbt.3708

[31] Mahecic D, Stepp WL, Zhang C, Griffié J, Weigert M, Manley S. Event-driven acquisition for content-enriched microscopy. *Nat Methods*. 2022;19(10):1262-1267. doi:10.1038/s41592-022-01589-x

[32] Cross-Zamirski, Jan Oscar, et al. "Class-guided image-to-image diffusion: Cell painting from brightfield images with class labels." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

[33] Dohmen, Melanie, et al. "Similarity and quality metrics for MR image-to-image translation." *Scientific Reports* 15.1 (2025): 3853.

[34] Chen, Jianxu, Matheus P. Viana, and Susanne M. Rafelski. "When seeing is not believing: application-appropriate validation matters for quantitative bioimage analysis." *nature methods* 20.7 (2023): 968-970.

[35] Martell, Matthew T., et al. "Deep learning-enabled realistic virtual histology with ultraviolet photoacoustic remote sensing microscopy." *Nature Communications* 14.1 (2023): 5967.

[36] Cimini, Beth A. "Creating and troubleshooting microscopy analysis workflows: common challenges and common solutions." *Journal of microscopy* 295.2 (2024): 93-101.

[37] Johnson, Justin M., and Taghi M. Khoshgoftaar. "Survey on deep learning with class imbalance." *Journal of big data* 6.1 (2019): 1-54.

[38] Hüllermeier, Eyke, and Willem Waegeman. "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods." *Machine learning* 110.3 (2021): 457-506.

[39] Krishnaswamy, Smita, et al. "Conditional density-based analysis of T cell signaling in single-cell data." *Science* 346.6213 (2014): 1250689.

[40] Wang, Zhou, et al. "Image quality assessment: from error visibility to structural similarity." *IEEE transactions on image processing* 13.4 (2004): 600-612.

[41] Hara, Kensho, Hirokatsu Kataoka, and Yutaka Satoh. "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018.

[42] Kay, Will, et al. "The kinetics human action video dataset." *arXiv preprint arXiv:1705.06950* (2017).

[43] Chen, Sihong, Kai Ma, and Yefeng Zheng. "Med3D: transfer learning for 3D medical image analysis. arXiv." *arXiv preprint arXiv:1904.00625* (2019).

[44] Robinson, Robert, et al. "Real-time prediction of segmentation quality." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer International Publishing, 2018.

[45] Chen, Jianxu, et al. "The Allen Cell and Structure Segmenter: a new open source toolkit for segmenting 3D intracellular structures in fluorescence microscopy images." *BioRxiv* (2018): 491035.

תקציר

המרת תמונה לתמונה היא שיטה בה ניתן לחזות תמונת פלט מתמונת קלט באמצעות המחשב. שיטת תיוג אברוני התא באמצעים ממוחשבים (in silico labeling) מיישמת את רעיון זה על ידי המרה של תמונות מיקרוסקופ אור-נראה אשר אינן אינפורמטיביות וניתן לרכושן יחסית בפשטות לתמונה עם תיוג פלורוסנטי ממוחשב של האברונים בתא, שהיא אינפורמטיבית מאוד. שיטה זו עשויה להוות התקדמות משמעותית בדימות תאים חיים כיוון שאינה דורשת תיוג פלורוסנטי, ובכך נמנעת מהוספת רעש לתמונה, פגיעה בהתנהלות התקינה של התא וחפיפה ספקטרלית המאפיינות שיטות פלורוסנטיות מסורתיות.
למרות זאת, השיטה אינה נמצאת בשימוש שוטף כיוון שלא ברור עד כמה ניתן לסמוך על כל חיזוי לצורך ביצוע אנליזות. גם כאשר התוצרים נראים משכנעים, מדענים זקוקים להוכחה כמותית, ברמת תא בודד, של מידת האמינות של כל חיזוי.
התרומה שלנו היא מודל שמקנה לחיזויים האלו מדד ביטחון כמותי ברמת תא בודד. המודל הינו תלת-ממדי, מבוסס ResNet וקולט שני ערוצים: החיזוי הממוחשב שהתבצע ומסכת אזורי חשיבות שמופקת בהליך נפרד ומקודדת אילו אזורים בתמונת הקלט הם אלה שעליהם מודל התיוג מסתמך. מודל הביטחון שלנו מנבא את השגיאה הצפויה עבור כל תא. כך אנו ממירים הסברים מסוג "למה המודל הסתכל כאן" ל"כמה משקל כדאי לתת לחיזוי המסוים הזה", ומספקים ביטחון מעשי במקום ויזואליזציה שנעשית פוסט-הוק. מודל התיוג מטופל כקופסה שחורה, ומסכות החשיבות משמשות כאותות קלט ולא כמטרות אימון ולכן אין צורך בשינויים באימון המודל המקורי.
במדידות על פני אברונים רבים, ציוני הביטחון שלנו עוקבים אחר הפערים האמיתיים בין פלטי מודל התיוג לבין התוצאות האמיתיות ועולים בביצועיהם על מודלי בסיס המשתמשים רק בתמונה הפלט של מודל התיוג או בהיוריסטיקות אחרות. כאשר מעבירים את התחזיות מרמת פאטץ' לרמת שדה ראייה שלם ואז לרמת תא בודד, הציונים מתואמים באופן מונוטוני עם מטריקות פרקטיות. בפועל, הדבר מאפשר קבלת החלטות פשוטה: לקבל תאים בעלי ביטחון גבוה למדידות כמותיות, לדחות או לבקר תאים בעלי ביטחון נמוך, ולתעדף אזורים אמביוולנטיים במידה ונרצה לחזור אליהם ולתקן את מאגר הנתונים.
בנוסף, אנחנו בוחנים קונטקסטים ביולוגיים שונים ומראים כי מדד הביטחון נשאר אמין גם במקרים שבהם המראה הוויזואלי עלול להטעות.
לסיכום, אנו מציעים כלי מעשי ההופך את תהליך תיוג האברונים בצורה ממוחשבת למדיד ברמה החשובה ביותר לביולוגים: התא. אנו מספקים מדד ביטחון העוזר בכימות אמין של תהליכים ביולוגיים ומקדם את שיטת התיוג הממוחשבת, מתמונות משכנעות למדדים כמותיים שמדענים יכולים להשתמש בהם, בתקווה להפוך אותו לכלי יישומי במחקר ביולוגי.

אוניברסיטת בן-גוריון בנגב
הפקולטה למדעי המחשב והמידע

התוכנית להנדסת מערכות מידע

זיהוי טעויות במודלי תמונה-תמונה ביולוגיים

חיבור זה מהווה חלק מהדרישות לקבלת תואר מגיסטר בהנדסה

מאת: גד מ. מילר

מנחה: פרופסור אסף זריצקי

חתימת המחבר  תאריך 28.12.2025

אישור המנחה/ים  תאריך 28.12.2025

אישור יו"ר ועדת תואר שני מחלקתית..... תאריך.....

טבת תשע"ו דצמבר 2025

אוניברסיטת בן-גוריון בנגב
הפקולטה למדעי המחשב והמידע

התוכנית להנדסת מערכות מידע

זיהוי טעויות במודלי תמונה-תמונה ביולוגיים

חיבור זה מהווה חלק מהדרישות לקבלת תואר מגיסטר בהנדסה

מאת: גד מ. מילר

טבת תשע״ו דצמבר 2025