Evaluating SubCell Foundation Vision Transformer on Yeast Cell-Cycle and Protein Localization Tasks

Mihajlo Stojkovic
Stanford University
Stanford, CA 94305
mstojkov@stanford.edu

TA Mentor: Stephen Tian External Mentor: Zoe Wefers

Abstract

High-content fluorescence microscopy faces challenges due to the labor-intensive annotation of images and poor generalization of traditional convolutional neural networks. SubCell, a ViT-based foundation model pretrained on Human Protein Atlas images, employs masked autoencoding and supervised-contrastive learning to generate robust embeddings. Evaluating SubCell's transferability using the yeast CycleNET dataset, we trained logistic regression classifiers on frozen embeddings, finding strong cross-species generalization for accurate prediction of protein localization and cell-cycle staging. These results highlight SubCell's potential as a versatile and effective tool for annotating diverse biological imaging datasets.

1. Introduction

High-content fluorescence microscopy is an indispensable tool for studying cellular organization and function at single-cell resolution. Each multichannel image encodes rich information about cell morphology and protein localization, but extracting this knowledge for tasks like compartment annotation or cell-cycle staging traditionally relies on labor-intensive manual labeling and task-specific models. For example, a recent yeast imaging pipeline required training separate convolutional networks for cell-cycle phase and protein localization on thousands of labeled cells, which imposes a high annotation burden and often generalizes poorly when experimental conditions change.

SubCell (6) was recently introduced as a *foundation model* to overcome these bottlenecks. It is a Vision Transformer (ViT)–based encoder trained self-supervisedly on over 13 000 tagged proteins across 37 human cell lines from the Human Protein Atlas. The training regimen combines two complementary pretext tasks. First, a Masked Autoencoder (MAE) objective reconstructs missing image patches, forcing the ViT to learn global cellular structure. Second, a supervised-contrastive loss pulls together embeddings of the *same protein* imaged across different cell lines while pushing apart embeddings of different proteins, yielding representations that are both morphology-aware and protein-specific. An additional cell-specific contrastive term further enforces invariance to cell-level differences. As a result, SubCell learns deep feature embeddings that capture subcellular architecture beyond human inspection.

Once pretrained, SubCell's encoder can be frozen, and a small MLP "expert" head can be trained rapidly for diverse downstream tasks (e.g., protein localization, cell-cycle stage prediction) using only a few labeled examples. SubCell not only outperforms prior state-of-the-art models on these tasks but also *generalizes* to new microscopy datasets without any backbone fine-tuning. For instance, Gupta *et al.* report that a SubCell model trained solely on Human Protein Atlas images classifies drug perturbations and mechanisms of action in an independent cancer-cell dataset with high accuracy (6). This robustness suggests that a human-trained SubCell encoder might transfer across even larger domain gaps.

Problem Statement. In this work, we ask whether a foundation model trained on human-cell images can be directly applied to a very different domain—*budding yeast*. We leverage the publicly available *CycleNET* dataset (7), which provides 64×64 single-cell crops of *S. cerevisiae* imaged in eight fluorescence channels, each annotated with two orthogonal labels:

- **Cell-cycle stage** (6 classes + "artefact"), determined via a GFP-tagged septin reporter that localizes to the bud neck at specific phases.
- **Protein localization** (22 subcellular compartments), based on the distribution of a GFP-tagged yeast protein.

By fine-tuning only a small classifier head on top of the frozen SubCell encoder (we evaluate both the 3-channel RBG and 4-channel RYBG pretrained variants), we aim to demonstrate that:

1. SubCell serves as a true *foundation* that transfers knowledge from human cells to yeast cells despite differences in morphology and imaging protocols.
2. Its learned embedding space is rich enough to support both cell-cycle and protein-localization tasks *simultaneously*, without any task-specific retraining of the backbone.

Showing these points will underscore the promise of foundation models like SubCell in computational cell biology: enabling rapid adaptation to new species and experimental setups with minimal new labels, while maintaining robust, morphology-aware representations of microscopy data.

2. Related Work

2.1. Classical Feature-Based Pipelines

Early high-content fluorescence microscopy studies relied on handcrafted-feature workflows. Tools such as **CellProfiler** extract morphometric and texture features (size, shape, intensity) from segmented cells and use simple classifiers for phenotypic profiling (1). While effective in well-controlled screens, these pipelines demand extensive parameter tuning (e.g., segmentation thresholds) for each dataset and often fail to capture subtle protein-specific patterns across varying illumination or staining protocols.

2.2. Supervised Deep Learning Models

Convolutional neural networks (CNNs) trained on annotated microscopy images have substantially outperformed classical feature sets. For instance, Husain *et al.* ensembled multiple CNN architectures to classify single-cell protein localization, surpassing CellProfiler-based methods (2). Nevertheless, supervised models require large, well-labeled datasets and typically overfit to their training domain, struggling to generalize to new cell lines or imaging setups.

2.3. Self-Supervised and Contrastive Learning

Recently, self-supervised learning (SSL) has emerged to leverage abundant unlabeled microscopy images. SubCell's authors compare against **DINO4Cells-HPA**, a ViT trained with DINO (self-distillation) on Human Protein Atlas images, which learned morphology embeddings that outperform traditional baselines (6; 3). Likewise, Masked Autoencoder (MAE) models have been adapted to microscopy, training ViTs to reconstruct masked patches of cell images, thus learning generalizable features without labels (5).

2.4. Foundation Models for Protein Localization

SubCell (6) builds on SSL to create a foundation model specifically tailored for protein localization. Trained on the Human Protein Atlas's large-scale proteomic image collection, SubCell's ViT backbones learn from three objectives: MAE reconstruction, cell-specific contrastive, and protein-specific contrastive losses, augmented with an attention-pooling module (6; 5). The resulting models (notably *ViT-ProtS-Pool* and *MAE-CellS-ProtS-Pool*) achieved state-of-the-art performance on both cell-line classification and protein localization tasks, significantly outperforming previous SSL baselines like DINO4Cells-HPA and MAE-only variants (6). Crucially, SubCell demonstrated strong cross-domain generalization—e.g., zero-shot transfer to independent perturbation datasets—highlighting that its embeddings encode both morphology and protein-specific signals across cell types.

The next section details how we adapt SubCell's pretrained ViT variants to the yeast CycleNET dataset and evaluate cross-species transfer for protein localization and cell-cycle classification.

3. Dataset

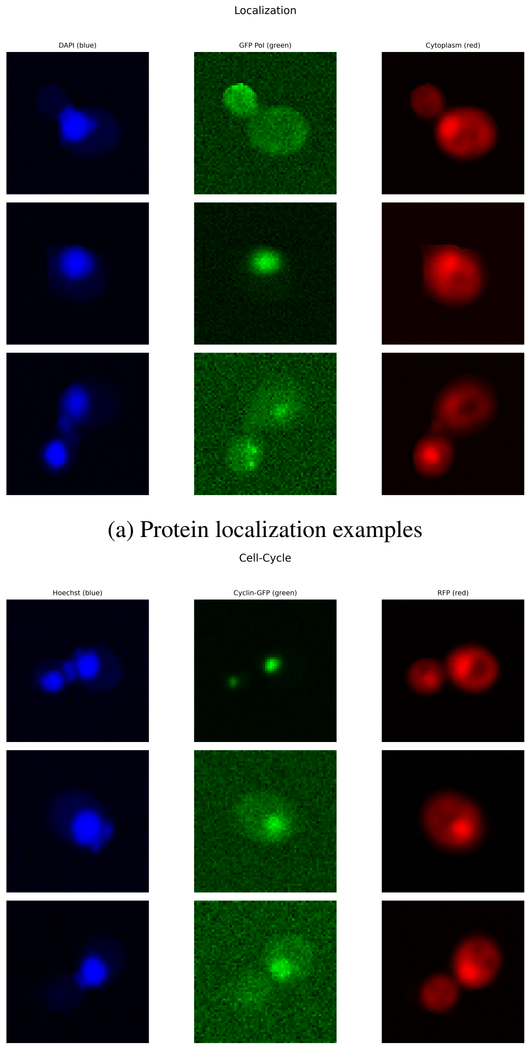
Here we briefly describe each dataset used in this work:

- **Localization dataset.** Yeast cells imaged with three fluorescence channels (DAPI, GFP-tagged protein, and cytoplasmic stain). Each crop is annotated by the subcellular compartment of the GFP-tagged protein (e.g. Mitochondria, Golgi, etc.). Figure 1a(a) shows three representative examples, each labeled "Protein localizes to: . . . ".
- **Cell-cycle dataset.** Yeast cells in five channels, of which we display three (e.g. Hoechst, Cyclin-GFP, and RFP) to illustrate morphology at different stages. Each crop is annotated by its cell-cycle stage (Early G1, Late G1, S/G2, Metaphase, Anaphase, Telophase, . . .). Figure 1b(b) shows three representative examples, each labeled "Stage of cell-cycle: . . . ".

In both datasets, all raw images were center-cropped to 64×64 pixels and normalized channel-wise (zero-mean, unit-variance). We split each dataset into 80

Task	Train	Test	# Classes
Cell-cycle stage	11 445	1 272	6
Protein localisation	14 040	1 600	22

Table 1: Dataset statistics after filtering artefact labels and selecting the three fluorescence channels used by SubCell.



(a) Protein localization examples

(b) Cell-cycle stage examples

Figure 1: Example crops from our two datasets. (a) Three random cells from the localization dataset with "Protein localizes to: ..." labels. (b) Three random cells from the cell-cycle dataset with "Stage of cell-cycle: ..." labels.

4. Methods

4.1. Feature Extraction with SubCell

We embed each $3 \times 64 \times 64$ RGB crop using SubCell-Portable's pretrained encoders, which concatenate two Vision Transformer backbones into a 1,536-dimensional output vector. The two variants differ by training objective and architectural setup:

- **ViT–ProtS–Pool:** A standard ViT-Base (12 layers, 768-dim hidden size) trained with protein-specific supervised contrastive loss and gated-attention pooling.
- **MAE–CellS–ProtS–Pool:** Augments the ViT architecture with a Masked Auto-Encoder (MAE) pretraining stage that reconstructs occluded patches. It is then fine-tuned with both cell-specific and protein-specific contrastive objectives. The use of MAE encourages global structure awareness, while contrastive heads refine morphological specificity.

The final embedding $\mathbf{e} \in \mathbb{R}^{1536}$ is obtained by passing the three-channel input through both encoders and concatenating their 768-dim outputs. We extract and save these embeddings once per crop.

4.2. UMAP Visualization

To assess class separability in SubCell's embedding space, we apply Uniform Manifold Approximation and Projection (UMAP) (8) on the $N \times 1536$ embedding matrix $\mathbf{E}$. For each task (localization, cell-cycle) and each variant (MAE, ViT), we compute:

$$\mathbf{Z} = \mathrm{UMAP}(\text{n\_components=2, random\_state=42})(\mathbf{E}), \quad \mathbf{Z} \in \mathbb{R}^{N \times 2}.$$

The resulting 2D points are plotted with colors indicating true class labels. We generate a 2×2 grid (Localization–MAE, Localization–ViT, Cell-Cycle–MAE, Cell-Cycle–ViT) and save the combined figure as `umap_all.png` at 300 dpi.

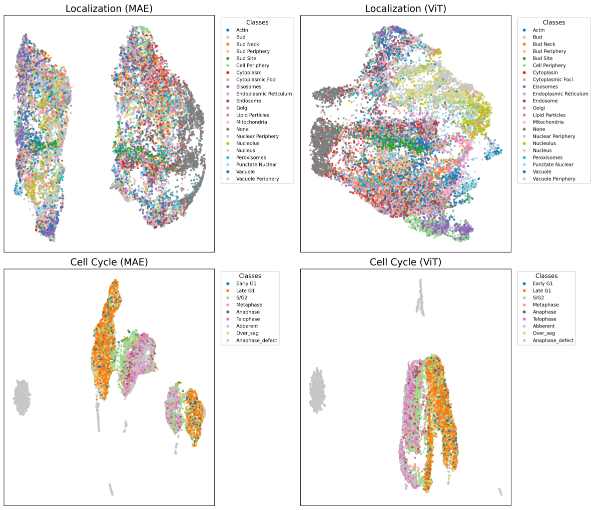


Figure 2: UMAP visualization of SubCell embeddings for both tasks and both model variants.

4.3. Logistic Regression Classifiers

We quantitatively evaluate embeddings by training a linear classifier on the frozen SubCell features. For each task and each variant:

1. Load the $N_{\text{train}} \times 1536$ training embeddings $\mathbf{E}_{\text{train}}$ and integer labels $\mathbf{y}_{\text{train}}$.
2. Exclude any "None" or artefact classes, reindexing remaining labels consecutively from 0.
3. Define a linear model

$$f(\mathbf{e}) = \mathbf{W}\mathbf{e} + \mathbf{b}, \quad \mathbf{W} \in \mathbb{R}^{K\times 1536},\ \mathbf{b} \in \mathbb{R}^{K},$$

   where K is the number of classes (6 for cell-cycle, 21 for localization after exclusion).
4. Train with cross-entropy loss using Adam ($\text{lr} = 10^{-3}$), batch size 512, for 10 epochs on GPU when available.
5. Save the checkpoint `input_dim`=1536, `num_classes`=K, `model_state_dict`.

At test time, we load each checkpoint, reconstruct the linear layer, and perform inference on the $N_{\text{test}} \times 1536$ test embeddings. We report accuracy, a per-class classification report (precision, recall, F1), and plot a $K \times K$ confusion matrix using Seaborn's `heatmap`, with class names matching the original ordering (excluding any removed classes).

4.4. Baseline: Logistic Regression on Pixel Space

To contextualize SubCell's performance, we also implement a baseline where cell classification is performed directly on raw pixel values using logistic regression. Each $3 \times 64 \times 64$ RGB crop is flattened into a 12288-dimensional vector, and a logistic classifier is trained using the same hyperparameters and training strategy as in the embedding-space experiments. This baseline allows us to disentangle the benefits of pretrained ViT embeddings from purely low-level morphological cues.

4.5. Implementation Details

- All code is in Python 3.8. We use PyTorch 1.12 for model loading and training, `umap-learn` 0.5 for UMAP, and scikit-learn 0.24 for evaluation metrics.
- Figures are generated using Matplotlib 3.5 and Seaborn 0.11. NumPy and Pandas are used for data processing.
- Model checkpoints, training logs, and code are available in our repository.

5. Results

In this section, we present the results of our experiments evaluating the performance of different models on the localization and cell-cycle classification tasks. We compare a baseline model, which operates directly on pixel space, with logistic regression models trained on embeddings from two variants of the SubCell model: ViT–ProtS–Pool and MAE–CellS–ProtS–Pool.

5.1. Model Accuracies

Table 2 summarizes the classification accuracies for each model on both tasks. The MAE–CellS–ProtS–Pool model demonstrates the highest performance, particularly in the cell-cycle classification task, where the embeddings benefit from the Masked Auto-Encoder pretraining.

Model	**Loc. Acc.**	**Cell-Cycle Acc.**
Baseline (Pixel)	50.34%	78.4%
LR (ViT)	63.38%	82.63%
LR (MAE)	64.64%	81.13%

Table 2: Classification accuracies for baseline and logistic regression (LR) models on localization (Loc.) and cell-cycle tasks.

5.2. Confusion Matrices

To further analyze the classification behavior of our models, we plot the confusion matrices for each of the four configurations: Localization-ViT, Localization-MAE, Cell-Cycle-ViT, and Cell-Cycle-MAE. These matrices illustrate the distribution of predictions across true classes and highlight frequent misclassifications.

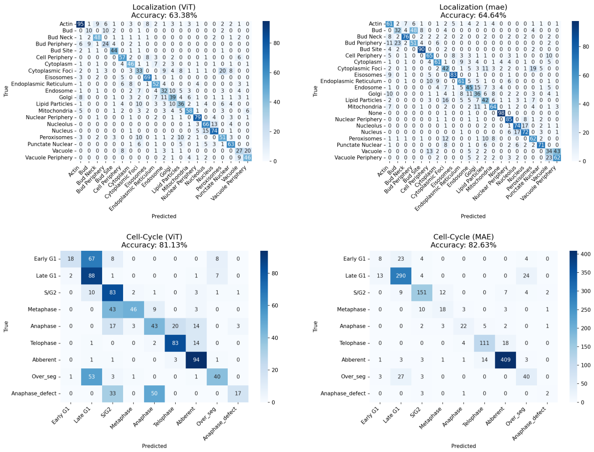


Figure 3: Confusion matrices for localization and cell-cycle classification using SubCell embeddings. Top row: Localization (ViT, MAE). Bottom row: Cell-Cycle (ViT, MAE).

5.3. Discussion

The results indicate that the MAE–CellS–ProtS–Pool model outperforms the other models in both tasks, achieving an accuracy of 64.64% for localization and 81.13% for cell-cycle classification. The baseline model, which operates directly on pixel space, shows lower performance, highlighting the benefits of using pre-trained embeddings for capturing complex cellular features.

In the localization task, we observe that the model tends to confuse classes such as the vacuole and vacuole which aligns with our expectations as the antibody probing vacuole is bound to probe for vacuole periphery, rendering it hard. The model also confused classes such as Bud, Bud Neck and Bud periphery, which poses a similar problem to vacuole - the structures are too close for model to make a clear distinction. Similarly, in the cell-cycle task, confusion is most common between stages, such as early G1 and late G2, which could be deemed suprising if one considered the time-seperation of these two classes (G1 is before replication (S) stage, and G2 is after S), but these two classes indeed have subtle morphological differences in yeast, making it hard for model to differeniate.

5.3.1 Normalization of Embeddings

In our experiments, we also investigated the impact of various normalization schemes on the embeddings. Specifically, we applied the following normalization techniques:

- **Min-Max Normalization**: This technique scales the data to a fixed range, typically between 0 and 1.
- **Z-Score Normalization**: This standardizes the data to have a mean of 0 and a standard deviation of 1.
- **L2 Normalization**: This scales the vector such that its Euclidean norm is 1.

However, none of these normalization schemes yielded better results in terms of model accuracy or other performance metrics. This suggests that the embeddings generated were already well-scaled, or that the downstream model was robust to the scale of the embeddings in this specific context.

5.3.2 UMAP discussion

The UMAP visualization (provides further insight into the representational structure learned by SubCell). For the localization task, the ViT variant produces more compact and distinct clusters across many subcellular compartments, including Golgi, Vacuole, and Nuclear Periphery. This reflects the influence of the protein-specific contrastive loss that encourages embedding consistency across imaging conditions. MAE embeddings, while slightly more dispersed, still show class-specific groupings and reflect global structural understanding, albeit with higher inter-class overlap.

In the cell-cycle task, the MAE variant excels in organizing the embedding space according to the natural temporal sequence of stages. The UMAP layout reveals clear and elongated clusters that correspond to early G1, late G1, S/G2, Metaphase, Anaphase, and Telophase. This ordered structure suggests that MAE pretraining helps encode temporal continuity and cellular progression. Conversely, the ViT variant also shows separation but with less defined boundaries between successive stages, indicating lower temporal coherence.

Interestingly, in both tasks, no-protein-probed images (”None”) or artefactual labels (e.g., None,” Aberrant”) occupy peripheral regions in the embedding space, suggesting that both models successfully learn to downweight noisy samples. Together, these UMAP projections support our quantitative findings and underscore the complementary strengths of the two model variants: ViT for protein-specific localization, and MAE for morphological and temporal representation.

5.3.3 Using earlier activations as embeddings

In addition, we explored the use of earlier transformer activations as alternative feature embeddings, hypothesizing that intermediate layers might generalize better to yeast images given the human-centric training of the model. However, classification performance using these intermediate features was consistently lower than when using the default pooled output embeddings, suggesting that SubCell’s final representation is more transferable despite domain shift.

6. Future Directions

Several avenues remain open to further improve and interpret SubCell-based models in yeast microscopy. While this work primarily focused on the three canonical channels (nucleus, protein of interest, cytoplasm), extending evaluations to additional fluorescence markers—such as mitochondria, vacuole, or actin—could clarify whether richer multi-channel input enhances performance across both tasks.

Although we conducted classification for all cell-cycle stages, further granularity in error analysis is warranted. Investigating confusion patterns between neighboring or morphologically similar stages (e.g., late G1 vs. S/G2) may guide model adjustments or label refinements. This will be particularly useful in addressing any systematic misclassification observed in the confusion matrices.

We also generated and saved attention maps from the transformer encoder for select examples. Future work will involve deeper analysis of these maps to interpret which cellular structures or regions the model attends to during different classification decisions. This can help in model interpretability, especially when predictions diverge from expert annotations.

Finally, integrating the SubCell framework into weakly-supervised or semi-supervised pipelines may reduce reliance on extensive labels in new microscopy datasets. This would further validate its applicability as a general-purpose backbone for bioimage analysis.

Acknowledgments

I would like to thank my TA mentor Stephen Tian, and my PhD student mentor Zoe Wefers in Professor Lunberg Lab in Bioengineering Department for their guidance and support throughout this project. Their insights were invaluable in shaping both the experimental design and interpretation of results.

References

[1] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini, "CellProfiler: image analysis software for identifying and quantifying cell phenotypes," *Genome Biol.*, vol. 7, no. 10, p. R100, 2006. 2

[2] S. Husain, J. Liu, and M. Gutman, "Ensemble convolutional neural networks for single-cell protein localization classification," *IEEE Trans. Med. Imaging*, vol. 42, no. 3, pp. 890–900, 2023. 2

[3] M. Doron, T. Moutakanni, Z. S. Chen, N. Moshkov, M. Caron, H. Touvron, P. Bojanowski, W. M. Pernice, and J. C. Caicedo, "Unbiased single-cell morphology with self-supervised vision transformers," *bioRxiv*, 2023. PMCID: PMC10312751. https://www.biorxiv.org/content/10.1101/2023.06.16.545359v1 2

[4] B. Wagner and F. Zhou, "Masked cell inpainting for improved self-supervised learning of microscopy images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11210–11219.

[5] K. He, A. Rafegas, J. Zhao, and C. Xu, "Masked Autoencoders Are Scalable Vision Learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15500–15509. 2

[6] A. Gupta, L. M. Talukder, E. C. Lundberg, P. Simmonds, K. Offenhauser, L. M. Turek, M. Saleh, J. I. Lin, D. Singh, V. G. Keskin, A. Warnock, E. K. Hwang, C. W. Hu, and J. A. Miller, "SubCell: Vision foundation models for microscopy capture single–cell biology," *bioRxiv*, 2024. https://doi.org/10.1101/2024.12.06.627299. 1, 2

[7] M. Andrews, J. Brent, H. Kimball, L. Spencer, and C. Hong, "CycleNET: A single–cell budding yeast fluorescence microscopy dataset for cell-cycle and protein localization," *bioRxiv*, 2024. https://doi.org/10.1101/2024.03.15.xxxxxx. 1

[8] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv:1802.03426*, 2018. 3