SubCell: Vision foundation models for microscopy capture single-cell biology

Authors: Ankit Gupta[1], Zoe Wefers[2,3], Konstantin Kahnert[2], Jan N. Hansen[1,2], Will Leineweber[2], Anthony Cesnik[2], Dan Lu,[4] Ulrika Axelsson [1], Frederic Ballllosera Navarro[2], Theofanis Karaletsos[4], Emma Lundberg[1,2,5, 6,*]

Affiliations:

1 Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Stockholm, Sweden

2 Bioengineering Department, Stanford University, Stanford, California, USA

3 Computer Science Department, Stanford University, Stanford, California, USA

4 Chan Zuckerberg Initiative, Redwood City, California, USA

5 Pathology Department, Stanford University, Stanford, California, USA

6 Chan Zuckerberg Biohub, San Francisco, California, USA

*Corresponding author. Email: emmalu@stanford.edu

Abstract

Cells are the functional units of life, and the wide range of biological functions they perform are orchestrated by myriad molecular interactions within an intricate subcellular architecture. This cellular organization and functionality can be studied with microscopy at scale, and machine learning has become a powerful tool for interpreting the rich information in these images. Here, we introduce SubCell, a suite of self-supervised deep learning models for fluorescence microscopy that are designed to accurately capture cellular morphology, protein localization, cellular organization, and biological function beyond what humans can readily perceive. These models were trained using the metadata-rich, proteome-wide image collection from the Human Protein Atlas. SubCell outperforms state-of-the-art methods across a variety of tasks relevant to single-cell biology. Remarkably, SubCell generalizes to other fluorescence microscopy datasets without any finetuning, including dataset of drug-perturbed cells, where SubCell accurately predicts drug perturbations of cancer cells and mechanisms of action. Finally, we construct the first proteome-wide hierarchical map of proteome organization that is directly learned from image data. This vision-based multiscale cell map defines cellular subsystems with large protein-complex resolution, reveals proteins with similar functions, and distinguishes dynamic and stable behaviors within cellular compartments. In conclusion, SubCell enables deep image-driven representations of cellular architecture applicable across diverse biological contexts and datasets.

Introduction

As a fundamental unit of life, cells have a remarkable capacity to perform a wide variety of functions, adopting diverse morphologies and carefully orchestrating a multitude of biological processes simultaneously. Proteins and other biomolecules localize to specific cellular substructures, undergo shifts in localization, and are subject to local modifications, enabling the cell to adapt its function and identity. Therefore, understanding cell behavior requires comprehensive spatial mapping of cell morphology, subcellular architecture and dynamics that currently only microscopy can capture at scale.

High-content imaging has become an indispensable tool in genetic perturbation (*1*, *2*) and drug screening (*1*, *3*), serving as the standard single-cell resolved readout. The Human Protein Atlas (HPA) has generated an extensive, near proteome-wide image collection with associated expert annotations on subcellular protein localization for over 65% of the protein-coding genes in humans - in other words, the majority of all intracellular proteins - across 37 cell lines representing cells of different developmental origins and tissue of origin with vastly different phenotypes (e.g., suspension cells, epithelial cells, fibroblasts) (*4*). The HPA dataset contains high-resolution immunofluorescence images stained with reference markers for the nucleus (DAPI), microtubules (MT), endoplasmic reticulum (ER), and an antibody targeting a protein of interest (*Figure 1*A), thus making it an ideal resource for modeling the subcellular organization in cells to understand cell behavior.

Deep learning has emerged as a transformative approach for modeling and analyzing large-scale image datasets, offering significant advantages over traditional methods that rely on handcrafted texture or intensity features (*5*, *6*). However, developing machine learning models that deliver generalizable cell embeddings remains challenging due to the extraordinary plasticity of cells and the spatial distribution diversity of the human proteome. Approximately 63% of proteins are found in multiple subcellular locations (*4*) and around 23% exhibit variable localization or expression between individual cells, reflecting dynamic cell states (*7*, *8*). Additionally, the spatial distribution of proteins is highly skewed: about 46% are localized to the nucleoplasm, 38% to the cytosol, and 8% to mitochondria, while fewer proteins are located in smaller, specialized compartments such as the centrosome (3.5%) or cell junctions (2.4%). Previous deep-learning approaches have used weakly-supervised paradigms to encode the images and predict protein localization patterns (*9*, *10*). Although such models perform well in their specified tasks, the cell embeddings often fail to generalize to other tasks. Annotations of the images are often performed on the image- or field-of-view (FOV)- level, which also limits supervised learning of single-cell heterogeneity.

As cell morphology carries a lot of information about cell state, morphological profiling has emerged as a powerful tool in phenotypic drug discovery (*11*). There is growing interest in building robust morphological feature extractors that can be used in large image-based perturbation screens to determine the effects and associations between genetic and chemical perturbations. While the most widely used profiling tools, such as CellProfiler, (*12*) use hand-crafted features, many works have applied machine-learning architectures to learn important morphological features from images (*5*, *6*, *13*). However, all such machine learning methods were trained on perturbation data, specifically with perturbation profiling in mind. Self-supervised learning has emerged as a powerful paradigm for general feature representation in biology (*5*, *6*, *14*, *15*). Recently, the set of DINO4Cells models was developed to obtain unbiased representations of cellular morphology (*5*). The results highlighted various aspects of a general-purpose feature extractor; however, it failed to achieve the same performance as supervised models in protein localization. Thus, a clear and

unfulfilled need remains for a vision model with a generalized understanding of both subcellular organization and cellular morphology.

In this work, we seek to train such a foundational model capable of performing a variety of tasks relevant to single-cell biology. We present SubCell, a suite of self-supervised vision-based foundation models to analyze fluorescence microscopy images of cells. SubCell's multi-task learning framework efficiently encodes both the protein localization patterns and cellular morphology in the HPA dataset on the single-cell level and learns biologically relevant information. Importantly, we demonstrate that SubCell is highly generalizable, performing effectively across various independent datasets - such as the OpenCell (*16*), Bridge2AI-CM4AI (*17*), and JUMP (*1*) datasets - without further parameter tuning. SubCell consistently matches or outperforms other models specifically trained on these datasets in various tasks such as protein localization, drug treatment, and mechanism of action prediction, highlighting its capability to extract meaningful biological insight from images.

Building foundational self-supervised models for single-cell images

In this study, we sought to develop a generalizable deep-learning model from fluorescent microscope images that can represent broad cellular characteristics, including protein localization patterns and morphology. We examined various state-of-the-art unsupervised learning methods, including masked autoencoders (MAEs) (*18*) and self-supervised learning techniques such as contrastive learning (*19*) and self-distillation (*20*). In addition to the standard reconstruction objective, we introduced the concepts of "cell-specific" and "protein-specific" objectives (Figure 1B), which were aimed at learning different aspects of biological information present in the images (See Methods). We combined these three objectives in a multi-task learning framework (Figure 1B) to train a vision transformer (ViT) model (*21*) to achieve a generalized representation of the single-cell images. We added an attention-pooling module (*22*) to the output of the encoder to prevent the model from learning spurious background features. The pooling modules weighted each patch token based on its importance to the final feature output, allowing the model to focus on the most indicative features of the image. We also developed a novel data-augmentation technique that preferentially masks tokens inside cell masks, encouraging models to focus on a single cell at the center of an image (Supplementary Figure S1).

We optimized the multi-task learning framework by training with various combinations of the three learning objectives, and the performance of the models was evaluated on two tasks: cell line and protein localization prediction. We started the experimentation with the MAEs and found that although MAEs learned cell morphology comparably to the DINO4Cells-HPA model (best configuration's macro average precision (AP) of 0.94 vs 0.957), they performed markedly worse in protein localization (best configuration's macro-AP of 0.528 vs. 0.705) (Supplementary Figure S2). We further explored contrastive loss, termed "cell-specific" loss here, aimed at minimizing the difference between feature representations of augmented views of a cell. We combined the cell-specific objective with MAEs, similar to (*19*, *23*) where the model was jointly trained on the reconstruction and the contrastive objectives. Combining MAE and self-supervision on the cell image level improved protein localization performance (best config. macro-AP of 0.666 vs. 0.705), although still lagging behind the DINO4Cells-HPA model's performance (Supplementary Figure S3).

We further introduced a "protein-specific" objective to specifically push the model to learn the protein localization patterns across cell types. The "protein-specific" loss minimizes the

differences between representations of cells from different cell lines and experiments which were stained with the same antibody. A similar but supervised approach was used in (*15*), where protein identity prediction from images was used as an auxiliary task to train a protein localization prediction model. Our self-supervised approach aims to learn more general features than on the individual protein level and improves performance on the protein localization prediction task significantly (best config. macro-AP of 0.742 vs. 0.705) and is comparable to the supervised "bestfitting" model's performance (macro-AP of 0.77) (Supplementary Figure S5). Finally, to prevent the model from learning uninformative background features, we added an attention-pooling module (*22*). The attention module further improved the models' performance, matching the supervised baseline on protein localization prediction (best config. macro-AP of 0.76 vs. 0.77) and surpassing the baseline in cell-line classification compared to the DINO4Cells-HPA (best config. macro-AP of 0.976 vs. 0.957) (Supplementary Figure S6).

The best-performing models were the ViT-ProtS-Pool, trained with only protein-specific objective, and the MAE-CellS-ProtS-Pool (Supplementary Figure S4), trained with the multi-task objective of reconstruction, cell-specific and protein-specific objectives, and were explored further for various tasks in the paper.

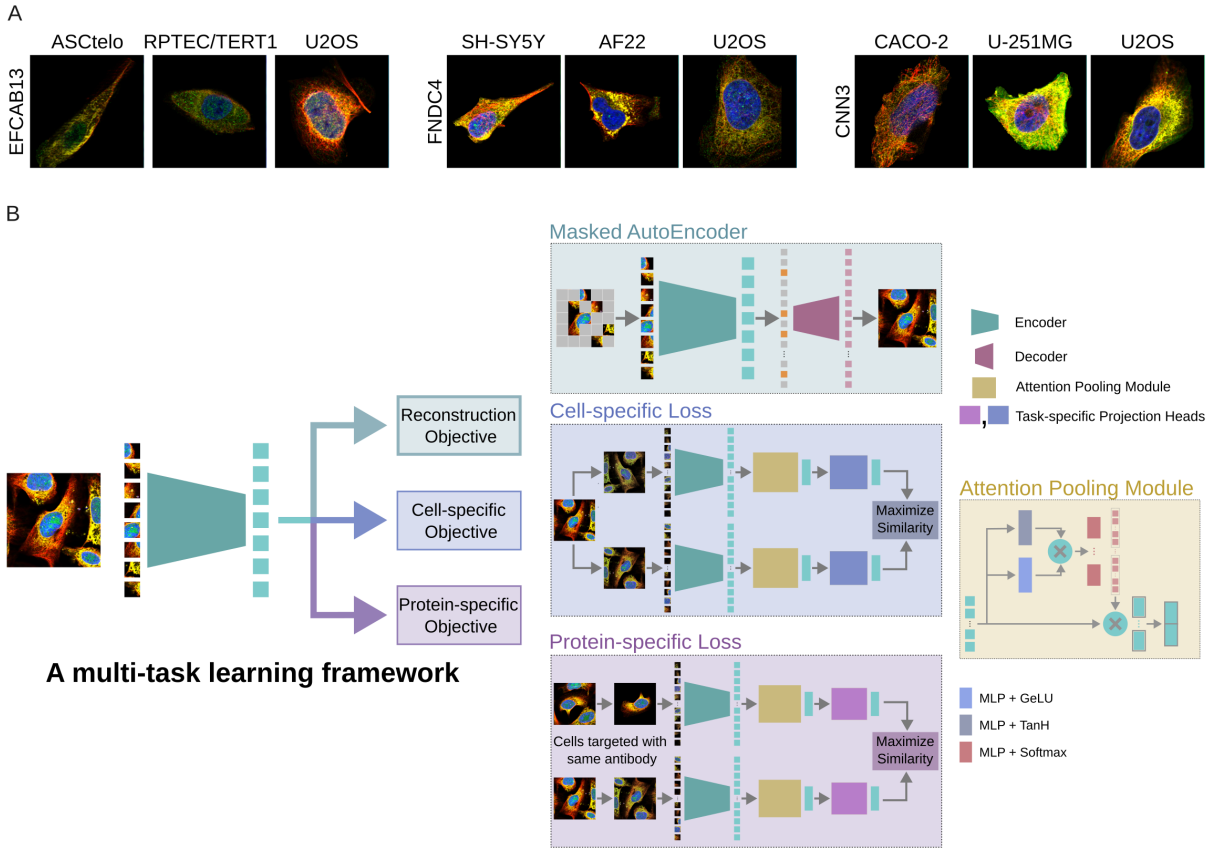


Figure 1. ***Overview of the SubCell self-supervised learning framework****. (A) Representative images of single-cell crops from the Human Protein Atlas dataset demonstrating the diverse cell morphologies and protein subcellular localizations present in a variety of human cells (Blue: Nucleus; Red: Microtubules; Yellow: endoplasmic reticulum; Green: Protein of interest). The images show three different proteins (EFCAB13, FNDC4 and CNN3) in a variety of human, primary and cancer cell lines derived from different organs of origin in donors of different age and sex (ASC52telo: mesenchyme, adipose tissue; RPTEC/TERT1: epithelium, kidney; U-2 OS: osteosarcoma; SH-SY5Y: brain; AF22: neuroepithelial stem cells; CACO-2: adenocarcinoma; U-251MG: glioblastoma (more details see Materials and Methods)) (B) Illustration depicting our multi-task learning approach to train a vision transformer model. We use three objectives to train our model: reconstruction, cell-specific, and protein-specific objectives.*

SubCell efficiently predicts both cellular morphology and protein localization

To date, most models for microscopy images are either focused on the analysis of subcellular patterns (*9*, *15*) or cellular morphology (*24*). We evaluated SubCell's performance on both these tasks based on an expert-annotated HPA dataset for the protein localization task (HPAv23) and a single-cell annotated reference dataset from an HPA Kaggle competition (*9*) . We used cell line classification as an indicator of morphological performance. We compared the models to recently published models: a self-supervised model, DINO4Cells-HPA (*5*), and a supervised model, "bestfitting," which emerged from a Kaggle challenge focused on predicting protein localization patterns in the HPA dataset (*9*). To compare SubCell's performance against a weakly-supervised baseline with the same architecture, we trained a supervised ViT model on the image-level annotations. We evaluated the performance of the models at the FOV level on the HPAv23 test set and on the single-cell level on the hidden test set of the Kaggle Single-Cell classification challenge (Figure 2A, B).

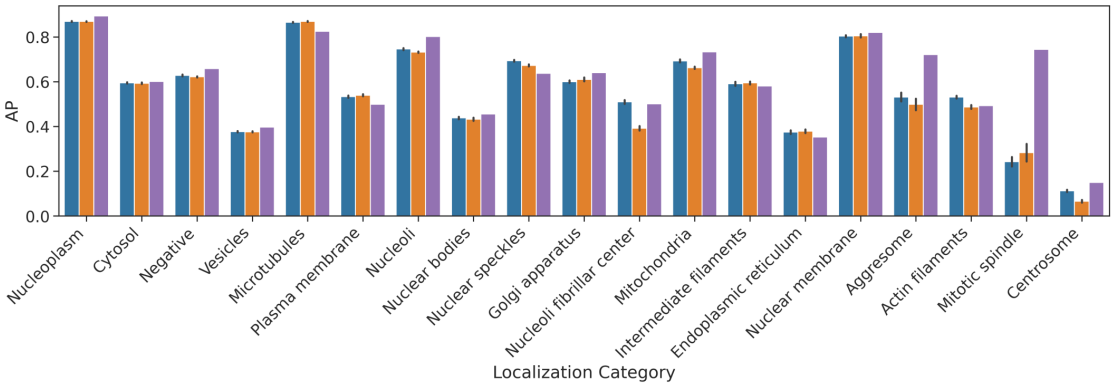
For protein localization prediction, SubCell's best models (ViT-ProtS-Pool and MAE-CellS-ProtS-Pool) outperformed the DINO4Cells-HPA model on both datasets (micro-AP of 0.881 vs. 0.863 on HPAv23 and 0.651 vs. 0.631 on the Kaggle test set, Figure 2A). SubCell's best model, the ViT-ProtS-Pool model, outperformed the bestfitting model on the HPAv23 test set (0.881 vs 0.83); however, it fell behind on the Kaggle test set (0.651 and 0.663). This difference was mainly reflected by performance on rare classes, such as mitotic spindle and centrosome, for which the bestfitting model employed different techniques for boosting the performance on these specific classes (*9*). However, our models consistently outperformed the weakly-supervised ViT baseline (0.876 on HPAv23 and 0.647 on the Kaggle test set). For morphology evaluation, our models outperformed the DINO4Cells-HPA model for cell-line classification (Figure 2A) and achieved superior performance in both test sets. In conclusion, these results show that our models learn more versatile features than the supervised models, which are trained for a particular task and tend not to generalize over different tasks.

To evaluate the importance of the three reference channels for model performance, we trained models with different input channel combinations, specifically DNA-Protein, ER-DNA-Protein, and MT-DNA-Protein (Figure 2D). Protein localization slightly improved when the ER channel was removed (0.651 vs. 0.654 and 0.647 vs. 0.652 for ViT-ProtS-Pool and MAE-CellS-ProtS-Pool, respectively), but the cell-line classification task performed worse (0.98 vs. 0.837 and 0.977 vs. 0.941). Removing the MT channel greatly reduced performance for the protein localization task (0.651 vs. 0.626 and 0.647 vs. 0.624) but not as much in cell line classification (0.98 vs. 0.89 and 0.977 vs. 0.963). Model performance was the lowest for both tasks when using only DNA as a reference channel, indicating the importance of other marker channels for broader generalization on these tasks. These results also hint that a broad exploration of reference channels beyond what is commonly used in HPA and Cell Paint assays (*3*, *25*) should be systematically assessed, which might lead to the design of better image-based assays with paired models.

A

Dataset	Experiment	Protein Localization		Cell Line Prediction	
		Macro AP	Micro AP	Macro AP	Micro AP
HPAv23 (FOV)	ViT-ProtS-Pool	0.760 ± 0.0020	**0.881 ± 0.0008**	**0.976 ± 0.0011**	0.993 ± 0.0001
	MAE-CellS-ProtS-Pool	0.739 ± 0.0024	0.876 ± 0.0004	**0.976 ± 0.0009**	**0.995 ± 0.0001**
	DINO4Cells-HPA	0.705 ± 0.0022	0.863 ± 0.0010	0.957 ± 0.0014	0.990 ± 0.0002
	ViT-Weak-Supervised	0.734 ± 0.0008	0.869 ± 0.0004	0.552 ± 0.0103	0.813 ± 0.0034
	bestfitting	**0.770**	0.830	0.342 ± 0.0038	0.682 ± 0.0013
Kaggle (Cell)	ViT-ProtS-Pool	0.565 ± 0.0020	0.651 ± 0.0013	**0.968 ± 0.0009**	**0.980 ± 0.0005**
	MAE-CellS-ProtS-Pool	0.552 ± 0.0018	0.647 ± 0.0027	0.965 ± 0.0013	0.977 ± 0.0006
	DINO4Cells-HPA	0.519 ± 0.0040	0.631 ± 0.0033	0.950 ± 0.0014	0.964 ± 0.0014
	ViT-Weak-Supervised	0.544 ± 0.0005	0.644 ± 0.0010	0.531 ± 0.0058	0.613 ± 0.0046
	bestfitting	**0.606**	**0.663**	0.316 ± 0.0037	0.420 ± 0.0029

B



C

Channels	Experiment	Protein Localization		Cell Line Prediction	
		Macro AP	Micro AP	Macro AP	Micro AP
DNA-Protein	ViT-ProtS-Pool	0.497 ± 0.0030	0.610 ± 0.0009	0.641 ± 0.0081	0.727 ± 0.0058
	MAE-CellS-ProtS-Pool	**0.532 ± 0.0056**	**0.618 ± 0.0015**	**0.874 ± 0.0021**	**0.910 ± 0.0013**
ER-DNA-Protein	ViT-ProtS-Pool	0.521 ± 0.0013	**0.626 ± 0.0021**	0.855 ± 0.0043	0.898 ± 0.0030
	MAE-CellS-ProtS-Pool	**0.539 ± 0.0027**	0.624 ± 0.0022	**0.941 ± 0.0014**	**0.963 ± 0.0006**
MT-DNA-Protein	ViT-ProtS-Pool	0.534 ± 0.0034	**0.654 ± 0.0007**	0.770 ± 0.0059	0.837 ± 0.0047
	MAE-CellS-ProtS-Pool	**0.561 ± 0.0035**	0.652 ± 0.0012	**0.916 ± 0.0015**	**0.941 ± 0.0010**
All Channels	ViT-ProtS-Pool	**0.565 ± 0.0020**	**0.651 ± 0.0013**	**0.968 ± 0.0009**	**0.980 ± 0.0005**
	MAE-CellS-ProtS-Pool	0.552 ± 0.0018	0.647 ± 0.0027	0.965 ± 0.0013	0.977 ± 0.0006

Figure 2. ***Quantitative evaluation of the features learned by the models in predicting protein localization patterns and cell-based morphology.*** *(**A**) The results of the protein localization prediction task and cell line classification tasks on the HPAv23 test set (FOV-level) and hidden Kaggle test set (single-cell level), showing the overall macro and micro average precision. (**B**) The protein localization results for each category in the Kaggle test set. The bar plots show the average precision of each category for the bestfitting (violet), ViT-ProtS-Pool (blue), and MAE-CellS-ProtS-Pool (orange) models. (**C**) Evaluation of our models trained with different channel combinations for the protein localization prediction task and cell-line classification task on the Kaggle test set.*

SubCell encodes biologically relevant features

Unbiased profiling of cellular morphology

To further evaluate SubCell's capability to encode meaningful morphological information, we performed an analysis based on established biological relationships akin to the one performed in (*5*). Initially, we aggregated single-cell features across the different biological groups (cell lines and organelles) to build morphological profiles from the image embeddings. We then computed cosine similarity across these groups and performed clustering to elucidate their phenotypic relations. These relationships were then compared to those derived from an independent data modality: transcriptional profiles obtained by RNA-sequencing (RNAseq) for cellular phenotypes and expert-annotated hierarchical groupings of organelles (Figure 3A-C).

We aggregated the features across 35 human cell lines from the HPA dataset to build morphological profiles from the image data. Mantel statistic (*26*) was used to compare them against the similarity matrix obtained from the transcriptional (RNAseq) profiles of the same cell lines (*27*). The SubCell models exhibited the strongest correlation with the RNAseq similarity matrix with the Mantel statistic of 0.486 for the MAE-CellS-ProtS-Pool model and 0.484 for ViT-ProtS-Pool (Figure 3A). In comparison, DINO4Cells-HPA features only achieved correlations of 0.444 and bestfitting 0.406. These results show that the SubCell models encode the morphological cell line variations identified by a different data modality (transcriptomic), as shown in Figure 3B.

Additionally, we aggregated the features of 26 protein localization categories in the HPA dataset and organized them according to a ground truth hierarchy annotated by experts (Supplementary Table 2). The localization categories were grouped into seven major and 13 minor clusters (Figure 3C). Our models achieve lower albeit comparable Mantel statistic (0.460, 0.421 for MAE-CellS-ProtS-Pool and ViT-ProtS-Pool, respectively) to the supervised models (0.529 for bestfitting) and greatly outperform the DINO4Cells-HPA model (0.297) (Figure 3A). In conclusion, we demonstrate that SubCell's best models, the ViT-ProtS-Pool and the MAE-CellS-ProtS-Pool models, learned features of subcellular architecture as well as cell line characteristics.

Biologically meaningful subcellular attention

ViTs convert an image into a sequence of patches and leverage an attention mechanism to learn complex features. The attention generated for each patch provides valuable insights into the model's behavior and decision-making process. Our analysis focused on the patch attention generated by the model's last layer and the patch importance derived from the attention pooling module.

Figure 3F shows the attention captured by the twelve attention heads and two pooled attention heads for various protein localization patterns for the MAE-CellS-ProtS-Pool model. Interestingly, most attention heads attend to specific subcellular regions, many resembling patterns of organelles, for example, AH-4 to nucleoli, AH-8 to microtubules, PAH-2 to the Endoplasmic Reticulum, AH-12 to the plasma membrane, and AH-10 to nucleoplasm. Notably, the attention across all attention heads is always focused only on regions in the cell and not on the background. This emphasizes the success of introducing the augmentation strategy using the cell masks, which guides the model to focus only on the complete cells in the crops. To quantify the relationship between the attention heads and the protein channel, we aggregated the correlation profiles over different localization categories and created a cluster map to see how the attention profiles are organized (Figure 3D). The attention profiles are organized into the major localization clusters defined in the previous

section. Furthermore, the plot highlights the variability in the attention patterns for different protein localization categories. For example, AH-4 and AH-10 seem to attend to the nuclear pattern only when the protein is present.

Visualizing the correlation profiles of the input channels with the attention maps also reveals interesting patterns in the attention learned by the model (Figure 3E). A similar plot for the ViT-ProtS-Pool model is shown in Supplementary Figure S12B. The major axes of variations in the profiles align orthogonally for the protein localizations (vertically) and cell phenotype (horizontally), respectively. Altogether, these results show that the model has learned biologically meaningful relations and effectively manages to disentangle morphological information (cell line) from subcellular patterns.

Modeling subcellular protein-protein interactions

After establishing that SubCell has learned features relevant to classifying protein localization and cell morphology, we wanted to test SubCell's performance on another task – modeling protein-protein interactions. In this evaluation, we asked if similarity in the SubCell embedding is predictive of interactions, since SubCell has learned the spatial embeddings of proteins and because spatial proximity of proteins is correlated to physical interactions. Specifically, we tested if embeddings of proteins in the HPA image dataset could be used to model protein-protein interactions as defined by the STRING database. We calculated the correlation between the average feature embeddings of the 11,925 proteins common in the HPA and STRING database and calculated the correlation between them. Overall, 3.35 million interactions were compared. The correlation of the profiles was plotted against the STRING Score, where the higher score signified stronger interaction. The analysis shows that the SubCell-derived embeddings of the ViT-ProtS-Pool model showed the best ability to predict protein-protein interaction data from STRING (Figure 3G). These results demonstrate the ability of SubCell to make biological associations beyond what is possible with the human eye.

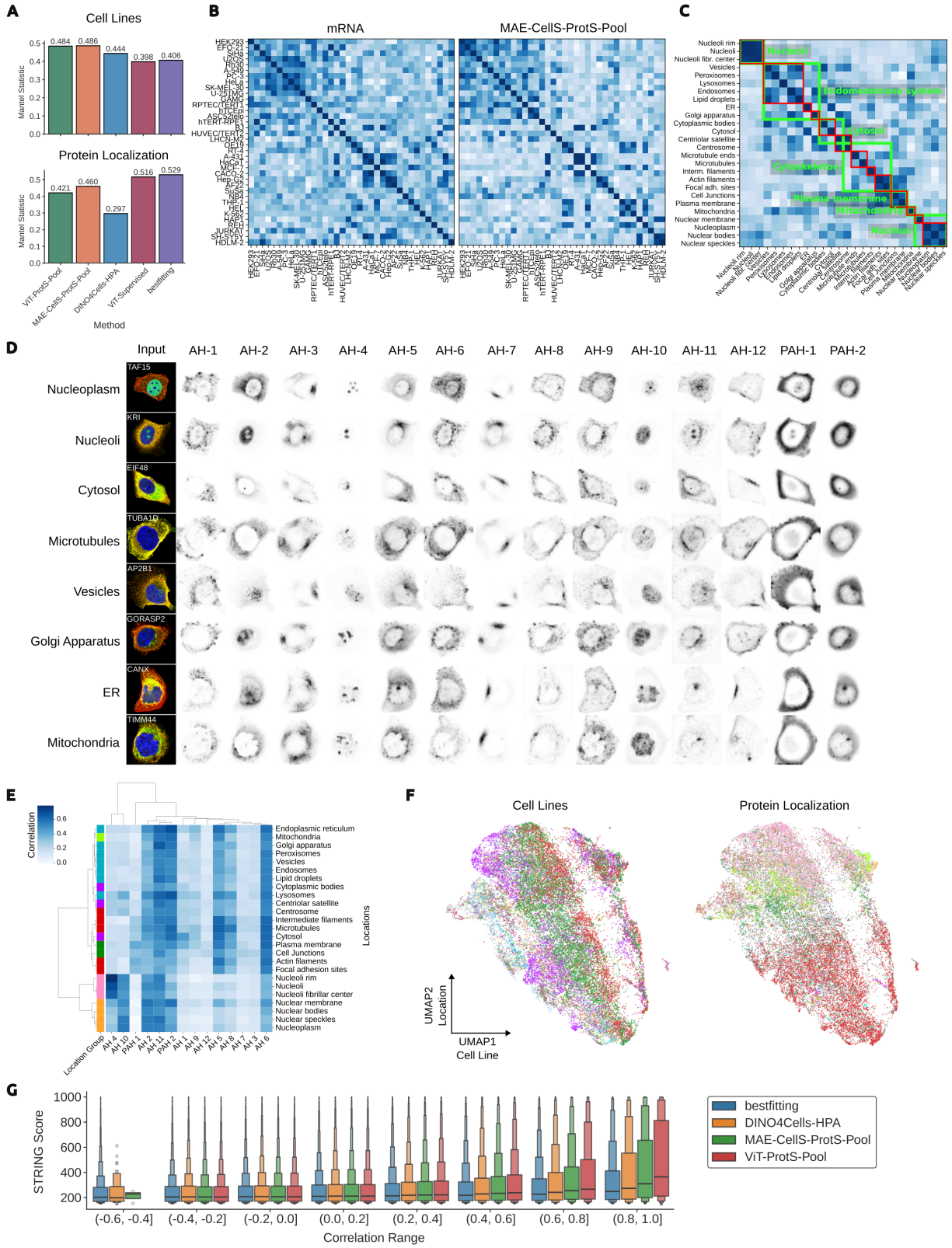


Figure 3. ***Feature validation of the SubCell models****. (**A**) Mantel statistic for the correlation between the aggregated feature profiles of subcellular localization patterns from the models and the hierarchical profiles annotated by experts (top) and Mantel statistic for the correlation between the aggregated feature profiles of cell lines from the models and mRNA profiles (bottom). (**B**) The cosine similarity among the MAE-CellS-ProtS-Pool features aggregated by the localization categories. Clusters highlighted in green are the localizations in seven major categories, and the clusters highlighted in red are the minor categories. (**C**) Profiling*

*of the cell lines using bulk mRNA levels (left) and aggregated MAE-CellS-ProtS-Pool features (right). The matrices show the cosine similarity among the cell lines based on mRNA levels or model features. The order of rows and columns follow the groups determined by the hierarchical clustering of the mRNA similarities. (**D**) Examples of fluorescent images and the attention captured by the attention heads and pooled attention heads in the MAE-CellS-ProtS-Pool model for different protein localizations. The twelve attention heads of the vision transformers are marked by the prefix AH- and the pooled attention heads by the prefix PAH-. (**E**) A cluster map showing the aggregated correlation profiles of the attention heads with the protein channel for different localization categories. The order of rows and columns is set by the hierarchical clustering of the correlation profiles. The positions of major location groups are shown in different colors on the left side of the plot. (**F**) The UMAP visualization of the correlation profiles of all the fluorescent channels with the attention maps. The profiles were aggregated over the FOVs, and the colors represent protein localization groups (top) and cell lines (bottom). (**G**) Boxen (Letter value) plot representing STRING scores divided into bins based on feature correlation between the gene pairs from the HPA dataset.*

SubCell generalizes to a variety of tasks on other image datasets

Predicting drug perturbations in breast cancer cells

To further explore the generalization capacities of SubCell in novel use cases and cell types, we investigated the performance in quantifying and predicting drug effects in a dataset we generated for the Cell Maps for Artificial Intelligence (CM4AI) project of the Bridge2AI program (*17*). This dataset aims to unveil changes in the subcellular architecture of triple-negative breast cancer cells treated with the anti-cancer drugs Vorinostat or Paclitaxel. Neither this cell line nor any drug-perturbed cells were included in model training. The dataset contains images acquired as in the HPA for 454 different proteins; HPA-style manual protein localization annotations are available on the image level for 91 of these proteins (*17*). For protein localization on the annotated images, our models outperformed DINO4Cell-HPA and bestfitting models by significant margins (Figure 4A).

To assess SubCell's ability to detect drug perturbations from the cell images, we split the dataset into three treatment categories: Paclitaxel, untreated, and Vorinostat. We then trained multilayer perceptron (MLP) classifiers in predicting treatment categories on unseen data using a ten-fold cross-validation scheme based on proteins. Our models achieved significantly better performance (0.979, 0.978 for MAE-CellS-ProtS-Pool and ViT-ProtS-Pool, respectively) than both the bestfitting (0.707) and DINO models (0.935) (Figure 4B).

Further analysis revealed how treatments were represented in the feature space (Figure 4C, D). We observed that for proteins that substantially change subcellular localization upon treatment, such as the apparent translocation of HDAC8 from the cytosol to nucleoplasm upon Paclitaxel treatment, the distance between the images obtained from the different treatments was large in the feature space (cosine distances: 0.158 (untreated vs. Paclitaxel) and 0.119 (untreated vs. Vorinostat)). In contrast, proteins that maintained a stable localization upon drug treatment, like TRIM24, exhibited shorter distances in feature space (cosine distances: 0.027 (untreated vs. Paclitaxel) and 0.041 (untreated vs. Vorinostat)). Additionally, for proteins that markedly changed their expression levels between untreated and treated conditions, such as AKR1B1 (cosine distances: 0.130 (untreated vs. Paclitaxel) and 0.116 (untreated vs. Vorinostat)), the feature space distance scaled with the difference in expression level. Notably, even though both Paclitaxel- and Vorinostat-treated cells showed localization of AKR1B1 to the cytoplasm, the morphological phenotype of cells undergoing Paclitaxel and Vorinostat treatment was different. We observed a rounder morphology and less single-cell variation in protein expression level with Paclitaxel treatment compared to Vorinostat treatment. This difference still manifests in small distances in the feature space, where the field of views for either treatment appeared in clearly different regions of the UMAP. The same applies to TRIM24 treatment, where small distances between the different treatments are visible in the UMAP even though protein localization and expression levels appear

stable across the different treatment conditions. We conclude that our model readily generalizes to new cell types and treated cells with vastly different morphology and can finely detect multifaceted aspects of drug treatment, such as changes in protein localization, morphology, and expression level. This performance positions SubCell as a powerful tool for automated and accurate quantification of cellular perturbations.

A

Experiment	Macro AP	Micro AP
MAE-CellS-ProtS-Pool	**0.524**	0.835
ViT-ProtS-Pool	0.520	**0.836**
DINO4Cells-HPA	0.327	0.771
bestfitting	0.335	0.771

B

Experiment	Macro AP	Micro AP
MAE-CellS-ProtS-Pool	**0.978 ± 0.003**	**0.979 ± 0.003**
ViT-ProtS-Pool	**0.977 ± 0.003**	**0.978 ± 0.003**
DINO4Cells-HPA	0.931 ± 0.009	0.935 ± 0.009
bestfitting	0.683 ± 0.027	0.707 ± 0.025

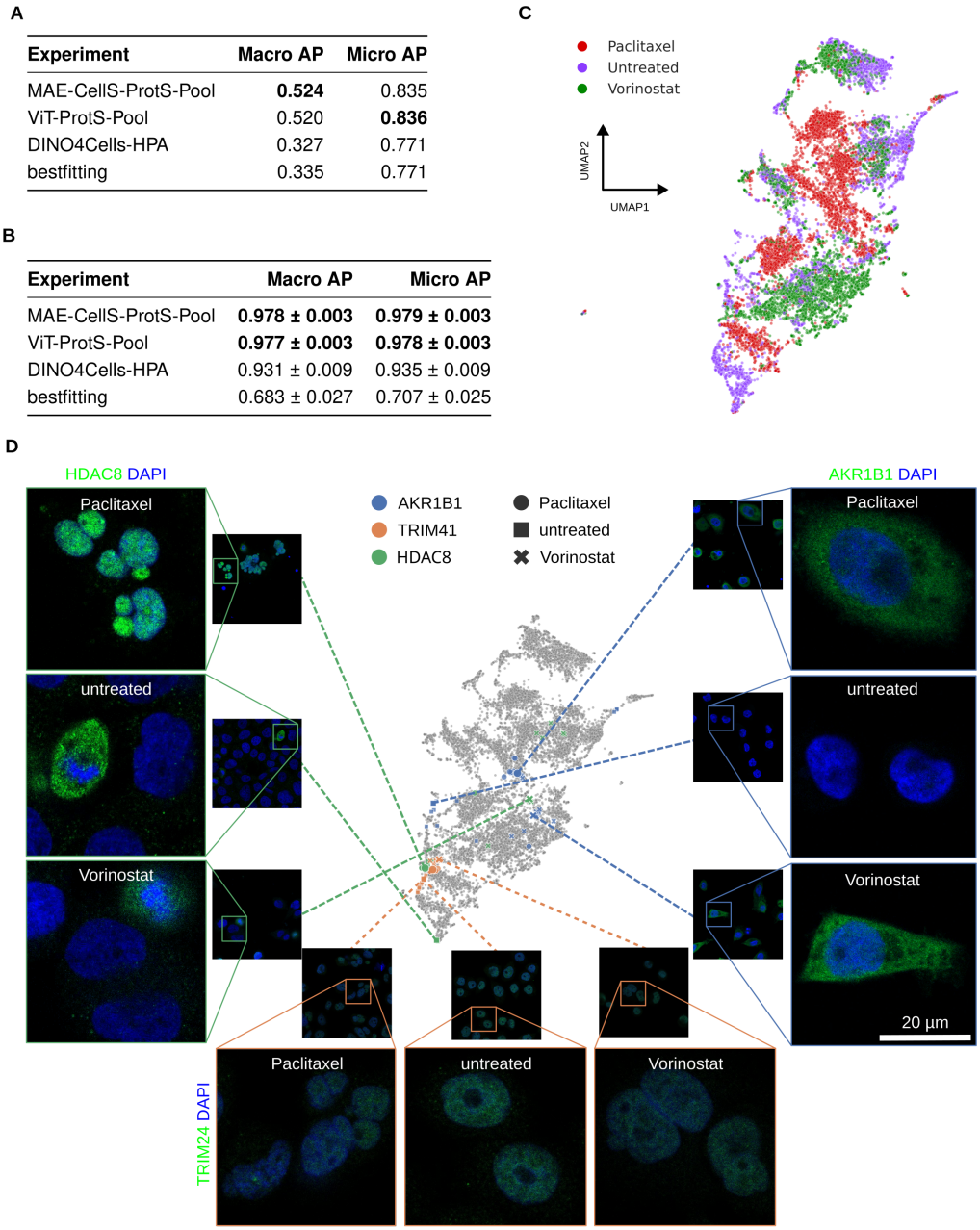


Figure 4. ***Evaluation of the protein localization and drug prediction in the image dataset of drug-treated breast cancer cells derived from the Bridge2AI-CM4AI program.*** *(A) The results of protein localization prediction on the expert-annotated subset of the Bridge2AI-CM4AI dataset at the FOV level.* ***(B)*** *The results of 10-fold cross-validation for drug prediction on the single-cell features extracted from the different models.* ***(C)*** *The UMAP plot of the FOV features highlighting the different treatments with Paclitaxel (in red), untreated (in violet), and Vorinostat (in green).* ***(D)*** *A vignette showing the UMAP plot of the FOV features and the FOV images highlighting different effects of the drugs captured by the MAE-CellS-ProtS-Pool model on three genes, namely, HDAC8, TRIM24, and AKR1B1.*

Classifying protein localization with fewer reference channels

We further evaluated the performance of our models on the OpenCell dataset (16). Given that this dataset comprises only two channels (DNA and Protein), we used the features extracted from the DNA-Protein version of SubCell and compared the feature representation with Cytoself (15), a self-supervised model trained on the OpenCell dataset. We performed k-means clustering on the features aggregated over the FOVs and evaluated the agreement of the resulting clusters with the OpenCell annotation labels to compare the feature representations between the models. SubCell performed better than Cytoself as measured by different clustering metrics (e.g., adjusted rand index of 0.331 and 0.339 for MAE-CellS-ProtS-Pool and ViT-ProtS-Pool, respectively vs 0.262 for CytoSelf) (Figure 5A). Furthermore, SubCell showed better separation between different classes (Fig. 5B), for example, between nucleolus and nucleoplasm. This reveals that SubCell can learn subtle localization patterns even with limited reference channels.

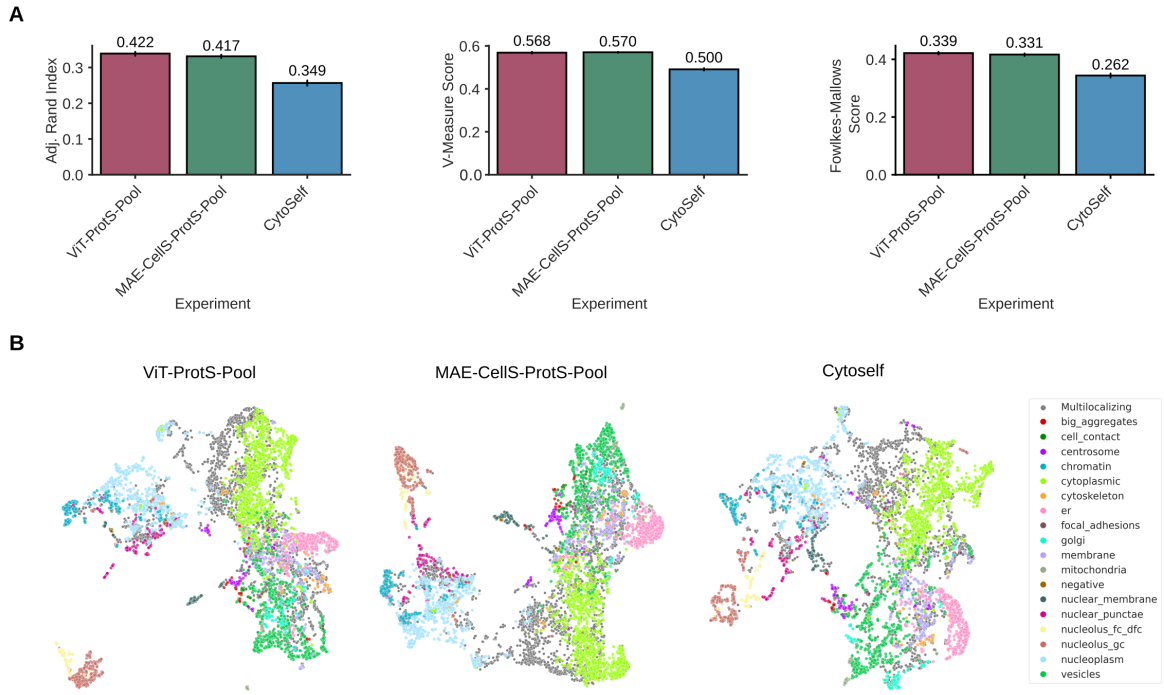


*Figure 5. **(A)** The clustering performance evaluation of the models on various metrics: top (left to right) V-measure score, adjusted mutual information score and bottom (left to right) adjusted rand index, and Fowlkes-Mallows score. **(B)** The UMAP of the FOV features of the models: from left to right, ViT-ProtS-Pool, MAE-CellS-ProtS-Pool, and Cytoself.*

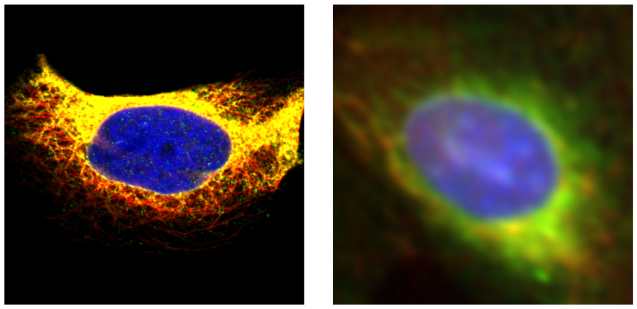
Morphological profiling from drug perturbation (Cell Paint) screens

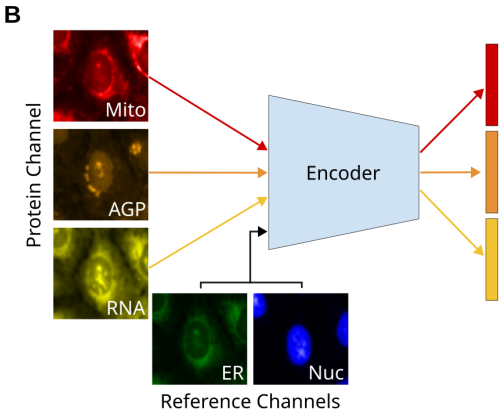
We evaluated our models on data that is even further out-of-distribution from the training data in that it 1) has a lower resolution than HPA (Figure 6A), 2) includes channels not included in HPA, and 3) includes atypical cell morphologies due to drug perturbations. Specifically, we tested SubCell's ability to quantify phenotypic drug effects in images from "Cell Painting" assays provided by the JUMP1 dataset (*3*). Cell Painting is a high-throughput and low-cost image-based assay widely used at large-scale in phenotypic drug discovery to characterize the morphological effects of chemical and genetic perturbations. It consists of six dyes recorded in five different

image channels that reveal eight cellular compartments (DNA, ER, Nucleoli, Golgi Apparatus, F-Actin, Plasma Membrane and Mitochondria). Cell Painting images are typically quantified by creating single-cell crops and calculating predefined image statistics from them, such as intensity, texture, and channel correlation, with an automated image analysis pipeline with tools such as CellProfiler. The single-cell profiles are then aggregated for each perturbation and used in various downstream tasks such as mechanisms of action (MoA) identification, off-target effect detection, and toxicity prediction.

We used our models to profile single cells in the dataset and evaluate the utility of these representations with two standard tasks: Replicate retrieval - identifying wells treated with the same compound across experimental batches; and MoA identification - revealing compounds that share the same MoA. We compared our models to the three current state-of-the-art methods used for Cell Painting data: (1) CellProfiler , a bespoke pipeline consisting of handcrafted features, (2) DeepProfiler , a weakly-supervised model trained to predict compound treatment labels, and (3) DINO4Cells-CP (*5*), a self-supervised model. We used the ER-DNA-Protein version of SubCell models since they have two common channels (ER, DNA) with the Cell Painting data. For each image, we consecutively profiled three Cell Painting channels at a time: the two reference channels and the third unseen channel, which was analyzed by SubCell's Protein channel. The three resulting embeddings were appended together to create a final single-cell profile (Figure 6B). The post-processing pipelines typically include three steps: single-cell aggregation, feature selection, and normalization. We evaluated several combinations and orders of techniques for these steps. For each model, we report metrics from the post-processing pipeline that resulted in optimal performance on replicate retrieval (Figure 5C) and MoA prediction (Figure 6D).

SubCell models excelled in replicate retrieval, with MAE-CellS-ProtS-Pool achieving the highest compound mean average precision (mAP) of 0.387, significantly better than the other models trained on the Cell Painting data. In contrast, the performance gap for the mechanism of action (MoA) prediction task was narrower. Here, CellProfiler led with a mAP of 0.178, while ViT-ProtS-Pool followed closely with a mAP of 0.177. We would like to emphasize that, unlike DINO4Cells-CP and DeepProfiler, the SubCell models were never trained on Cell Painting data and were therefore naive to the unseen channels. Additionally, the SubCell models were trained with HPA images that feature a much higher resolution in x and y and, notably, confocal z resolution, whereas the JUMP1 data are recorded at much lower resolution and with normal epi-fluorescence microscopy, resulting in off-focus blur in the images (Figure 6A). Considering these major differences between the datasets, we reach exceptional performance without fine-tuning. We conclude that SubCell models are generalizable for capturing morphological features from different channels and at different scales. We expect further fine-tuning and training of our models with lower-resolution images to boost performance even more.





C

Model	FOV Aggregation	Feature Select	Normalization	Compound mAP	MoA mAP	Compound NN-acc	MoA NN-acc
CellProfiler	mean	Yes	PCA-cor → MAD robustize	0.301	**0.178**	0.438	0.182
DeepProfiler	mean	Yes	ZCA-cor → MAD robustize	0.263	0.155	0.463	0.182
DINO4Cells-CP	mean	Yes	MAD robustize → PCA-cor	0.229	0.175	0.435	**0.193**
ViT-ProtS-Pool	mean	No	PCA-cor → MAD robustize	0.336	0.177	**0.477**	0.177
MAE-CellS-ProtS-Pool	median	No	PCA-cor → MAD robustize	**0.351**	0.175	0.461	0.167
Random Baseline	-	-	-	0.007	0.045	0.003	0.009

D

Model	FOV Aggregation	Feature Select	Normalization	Compound mAP	MoA mAP	Compound NN-acc	MoA NN-acc
CellProfiler	mean	Yes	MAD robustize	0.321	0.156	0.478	0.125
DeepProfiler	mean	Yes	standardize → PCA	0.271	0.155	0.463	0.182
DINO4Cells-CP	mean	No	PCA → MAD robustize	0.330	**0.167**	0.518	**0.188**
ViT-ProtS-Pool	median	Yes	ZCA → MAD robustize	0.350	0.163	0.493	0.167
MAE-CellS-ProtS-Pool	mean	No	ZCA-cor → MAD robustize	**0.387**	0.167	**0.522**	0.161
Random Baseline	-	-	-	0.007	0.045	0.003	0.009

Figure 6. ***Evaluation of compound matching and mechanism of action prediction (MoA) on the JUMP Cell Painting chemical perturbation dataset. (A)*** *Single-cell crop from HPA. The nucleus is shown in blue, microtubules in red, ER in yellow, and protein of interest in green. Single-cell crop from JUMP1. Nucleus is shown in blue, ER in green and Mitochondria in red. Both images have the same physical pixel size.* ***(B)*** *Diagram of how CellPaint images were fed into SubCell models.* ***(C)*** *Results of the post-processing pipeline for each model resulted in the best compound mean average precision (mAP)* ***(D)*** *Results of the post-processing pipeline for each model that resulted in the best MoA mAP. For both tables the best performance for each metric is highlighted in bold, and the second best is highlighted in italics.*

SubCell enables vision-based multiscale cell maps of proteome-wide cellular substructures

Spatial proteomics based on high-resolution confocal imaging, as performed in the subcellular section of HPA, reveals the subcellular location of a protein at the resolution limit of light diffraction. In our analysis of the HPA images, we focused on distinguishing 35 high-level subcellular organelles or compartments. However, it is important to note that even more known and possibly unknown subcellular compartments and protein condensates exist. Recently, we demonstrated a computational framework for the integration of image and protein-protein interaction to create a multiscale integrated cell (MuSIC) map defining the cellular protein assemblies from nanometer to whole cell scale (doi:10.1038/s41586-021-04115-9). Our MuSIC map identified new subcellular complexes. Due to the limitation of protein-protein interaction data, this MuSIC model was limited to less than 700 proteins. It was also constructed using the supervised bestfitting model trained to distinguish the 35 high-level compartments. Given that the SubCell models are self-supervised and have learned biologically relevant features to the extent that they can be used to predict aspects beyond what is capable by the human eye, such as protein-protein interactions, we explored if SubCell can be used to construct a vision-only-based multiscale map of the cellular proteome architecture (*9*).

Defining cellular substructures with protein-complex resolution

We applied SubCell to represent all HPA images for one cell line (U-2 OS) in an embedding space and segregated the embedding space by a hierarchical clustering approach to reveal cellular compartments and protein assemblies at different scales by their pattern similarity (Figure 7A). We annotated the different clusters using the most significant Gene Ontology (GO) terms (Figure 7B) and observed a clear resolution into the different subcellular compartments down to very specific protein complexes, such as the cytosolic ribosome (Figure 7C). In the images underpinning the cytosolic ribosome cluster (cluster 6.31) we observed a characteristic spotty cytosol pattern (Figure 7C). Protein-protein interaction data available in STRING validated this cluster to, in fact, contain proteins that form the cytoplasmic ribosome (cluster 7.59, Figure 7C, left). This exemplifies that SubCell allows to define subcellular compartments from images beyond what humans can realistically annotate. We conclude that our purely vision-based multiscale map can achieve a resolution of detecting large individual protein complexes.

Revealing proteins with similar function within the vision-based multiscale cell map

SubCell refines our view on widely known subcellular patterns, such as the staining of cytoskeletal microtubules. The SubCell-based multiscale map distinguished microtubule cytoskeleton staining (cluster 3.4) in the cell into two major classes (4.8 and clusters 4.10), which we found to represent a spotty (cluster 4.8) or continuous microtubule staining (cluster 4.10) (Figure 7B) and – by human annotators – would be jointly annotated as microtubule-staining. The continuous staining pattern was clearly linked to common microtubule cytoskeletal proteins and was not further dispersed at increasing resolution (Figure 7B), while the spotty staining pattern cluster was dispersed at high resolution yielding a cluster relating to cilia assembly (Cluster 7.27, Figure 7B). Notably, U-2 OS cells have never been observed to form a cilium – the cilium is a microtubule-based structure extending like an antenna from the mother centriole into the extracellular space (doi:10.1038/s41576-023-00587-9). However, cilium-related proteins are expressed in U-2 OS, and SubCell revealed a characteristic spotty staining pattern for these proteins, suggesting they're transported on microtubules even though they cannot reach the cilium. SubCell distinguished three types of spotty cilium-related microtubule clusters, two of which could not be annotated with GO even though they contained established cilium-related proteins, such as ARL13B. Validating SubCell's representation, also visually observed these three clusters to represent three distinct types of spotty microtubule staining: one exhibiting similar intensity spots across the whole cell

(Cluster 7.85, Figure 7D), another with increased intensity spots around the Golgi apparatus and additional Golgi-staining (Cluster 7.27, Figure 7D), and a third with increased intensity spots around the centrosome, which could represent centriolar satellites (Cluster 7.57, Figure 7D). By exploring the proteins grouped within these clusters, we may be able to associate them to specific functions. For instance, cluster 7.27, GO-annotated as cilium assembly and including Golgi staining, contains proteins related to transport of proteins to the cilium and the ciliary membrane (e.g., the small GTPase ARL13B known to be enriched in all cilia, the intraflagellar transport protein IFT22 which was related to transporting small GTPase dependent cargo to cilia, two subunits of the Phosphoinositite 3 kinase PIK3, and the Inositol polyphosphate 5-phosphatase OCRL). Conversely, cluster 7.57, related to centriolar-satellite-related staining, contains proteins known to play roles in ciliogenesis through control of the ciliary microtubule cytoskeleton (such as the Tubulin ligase TTLL9 or the tektin TEKT2). This cluster also contained NEK8, known to engage in ciliary transport and – when mutated – to trigger cilium-related diseases (Ciliopathies). Cluster 7.85 contains 16 lesser-studied proteins, such as WDR11, a protein ascribed to modulate ciliogenesis and signaling. Investigating the functional similarities of these proteins will be a subject of interest for future studies.

In conclusion, the multiscale map built from SubCell embeddings not only catalogs subcellular compartment architecture but also elucidates the functional similarity of proteins based on their precise subcellular distribution.

Distinguishing clusters of dynamic and stable proteins

A similar case was observed for the nuclear envelope, where SubCell discerned two major clusters (6.34 and 6.22) with two subclusters each on level 7. The nuclear membrane cluster 6.34 contains two clusters (7.65 and 7.77) underpinning the nuclear membrane structurally. Cluster 7.77 comprises proteins known to form the nuclear envelope (Fig 7E, Suppl. Table S3). An unannotated cluster (7.65) is distinguished from 7.77 containing multiple nuclear lamina proteins. The other nuclear membrane cluster (6.22) relates more to dynamic processes across the nuclear membrane, containing cluster 7.82 with proteins driving mRNA processing, splicing, and export, whereas the unannotated cluster 7.36 contained nuclear exporting and proteins that exert nuclear enzymatic functions. Visually, we found the images in cluster 7.36 to show envelope staining and confined spots in the nucleus and in the cytoplasm. We suggest that these might be proteins in their different transport stages. We conclude that the multiscale map built from SubCell embeddings can discern dynamic (e.g., shuttling) proteins from stable (e.g., structural) proteins within subcellular compartments.

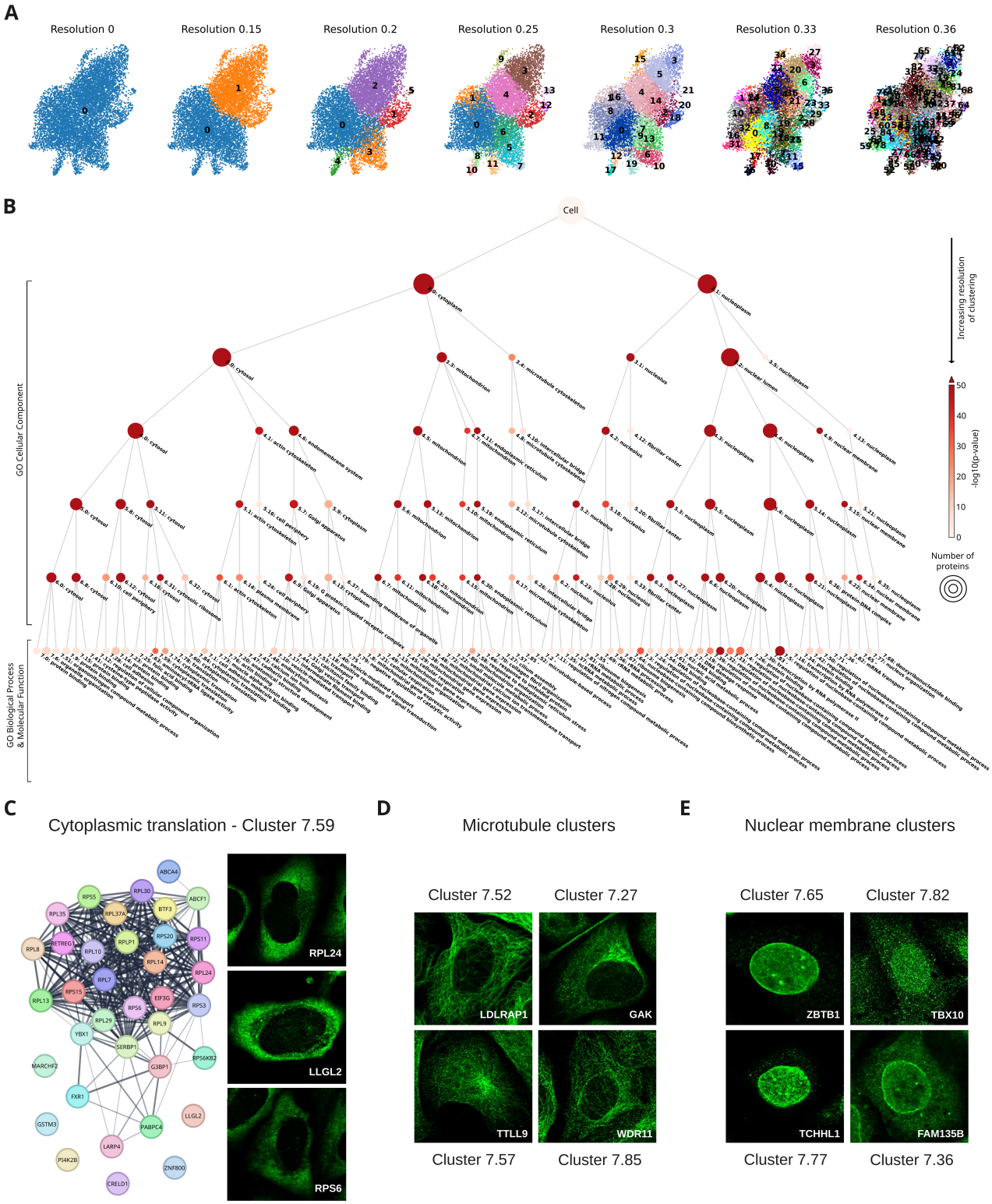


Figure 7. ***A vision-based multiscale map of cellular protein architecture in U-2 OS cells.*** *(**A**) HPA protein embedding UMAPs showing the resulting clusters of Leiden subclustering at increasing resolutions. (**B**) Multiscale hierarchical map obtained by subclustering of the protein representations of HPA images. Clusters and subclusters were annotated by GO Cellular Compartment (levels 1-6) or GO Biological Process / Molecular Function (level 7) enrichment analysis – the most significant term is shown. Each node is a cluster from A at the given subclustering / resolution level. The node size indicates the number of proteins in the cluster; the node color shows the -log10(P-value) of the functional enrichment. (C) STRING network of the proteins contained in*

*cluster 7.59 annotated as cytoplasmic translation (left). Exemplary HPA immunofluorescence microscopy images of proteins images of proteins contained in the cluster showing the distinct staining pattern. (**D and E**) Representative images for the four different microtubule related clusters (D) and the four different nuclear membrane related clusters (E) displaying the different staining patterns in the respective clusters.*

Discussion

This paper introduces SubCell, a suite of biological vision foundation models for encoding single-cell representations from microscopy imaging data, designed as a multi-task, self-supervised learning framework. SubCell outperforms state-of-the-art methods across a diversity of tasks relevant to single-cell biology, including localization prediction, cell phenotyping, morphological phenotyping, drug treatment prediction, and mechanism of action prediction and is readily generalizable to other datasets of different imaging modalities and resolutions.

During the design and optimization of SubCell, we observed that typical learning frameworks like masked autoencoders were unable to encode complex protein localization patterns compared to the other supervised and self-supervised approaches. Introducing a protein-specific objective to maximize the similarity in representation between the cells treated with the same antibody boosted the performance in protein localization prediction. To increase the versatility of our software suite, we trained models based on multiple combinations of marker channels, broadly increasing the applicability of SubCell to existing and future datasets. Additionally, our novel masking strategy during training eliminates the need for cell segmentation masks during inference, which further improves the usability of our suite.

We showed that SubCell generalizes to unseen datasets acquired using different microscopy methods, novel cell models, and drug treatments. It also outperforms state-of-the-art models, such as Cytoself (*15*), a self-supervised model trained on the OpenCell dataset in protein localization clustering. We showed that our models predict cancer cell treatment effects better than the state-of-the-art models using the Bridge2AI-CM4AI dataset (*17*), paving the way for improved quantification of protein relocalization upon cell perturbation. Given recent studies demonstrating that pervasive mislocalization of pathogenic coding variants commonly underly human disorders (doi: 10.1016/j.cell.2024.09.003), as well as demonstrating novel druggable approaches through protein relocalization (doi: 10.1038/s41586-024-07950-8), we speculate that the ability to detect and disentangle subtle spatial differences in protein subcellular localization across cells of different morphology will become increasingly important to understand disease.

We further showed that our models can also be used as general out-of-the-box feature extractors in fluorescence microscopy and highlighted this on the JUMP1 dataset (*1*). Remarkably, our models not only surpassed the performance of existing deep learning-based models trained on the cell painting images in replicate retrieval tasks, but also matched the performance for MoA prediction – all without finetuning. This underscores the versatility and robustness of our models, positioning them as essential tools for advancing high-throughput cellular analysis.

We believe that SubCell performs well on a variety of tasks and datasets because it has truly learned the intricacies of subcellar architecture through a diversity of protein localization patterns and an understanding of cell morphology through a wide range of cell lines. This is thanks to the unique strength of the training dataset, the HPA image dataset, which is a standardized image collection with a large diversity of subcellular staining patterns for nearly all non-secreted proteins. We believe this was a key to why the SubCell models learned biologically meaningful attention patterns, which provides exceptional generalizability to downstream tasks without any fine-tuning.

Biological images alone serve as the basis for the meaningful information learned by SubCell. This is demonstrated by the construction of the first vision-based multiscale map of subcellular protein architecture, which succeeds in categorizing subcellular protein localization by protein interaction, function, and dynamics. This map provides a breadth of new applications for decoding cellular phenotypes (doi: 10.1146/annurev-biodatasci-102423-113534), such as employing it as a vision-based data-driven alternative to GO for gene, protein, and pathway annotations, using it as an unbiased annotation of subcellular protein localization independent of human-defined localization classes, or projecting the effect of mutations and drug screenings on functional protein groups within it as a cellular map for biological reasoning. We also envision that this map can be integrated into transcriptome-based models for improved biological context and causal reasoning.

SubCell is a generalizable foundation model that pioneers quantitative image representation for microscopy image analysis independent of optimizing image analysis pipelines. To unlock broad and fair usability for SubCell for all researchers worldwide, this model is made freely available and easily accessible. We provide tutorials and a ready-to-use application online, as well as an interactive version of the proteome-wide hierarchical cellular map.

Beyond cell biology research, we envision SubCell will both aid progress in drug and perturbation screening by providing a new level of phenotypic screening far beyond human-crafted feature extraction and bring us one step closer to virtual experimentation across data modalities to accelerate the life science by ultimately facilitating the construction of a universal virtual cell model (doi: 10.48550/arXiv.2409.11654).

Code and data availability

The model and data are available at: https://virtualcellmodels.cziscience.com/

The HPA dataset is available in the Human Protein Atlas database (www.proteinatlas.org). The cropped image dataset used for training the models is available in the S3 bucket: s3://czi-subcell-public/

The code necessary for training all the models is available in these repositories
https://github.com/CellProfiling/subcell-embed
https://github.com/czi-ai/sub-cell-embed

The code necessary for inference is available in this repository:
https://github.com/czi-ai/SubCellPortable

The code necessary for reproducing the results in the analysis and figure generation is available in this repository: https://github.com/CellProfiling/subcell-analysis.

The proteome-wide map of cellular substructures is available as an interactive version in the Human Protein Atlas (to be released upon full publication)

Acknowledgements

Training and evaluation of SubCell models was enabled by computing resources provided by the Chan Zuckerberg Initiative and the National Supercomputer Centre at Linköping University. EL was supported by the Chan Zuckerberg Initiative, the Wallenberg Foundation (2021.0346), Schmidt Futures, Göran Gustafsson Foundation, the Bridge2AI Program (NIH Common Fund; OT2 OD032742), the Cancer Cell Map Initiative (NCI Center for Cancer Systems Biology; U54

CA274502), Stanford Institute for Human Centered AI, and Param Hansa Philanthropies. JNH was supported by a Postdoctoral Fellowship from the Wenner-Gren Foundations and by an EMBO Postdoctoral Fellowship (ALTF 556-2022).

Author Contributions

AG and EL conceptualized the project and designed the study. AG designed and trained the models. AG, ZW, KK, and JNH evaluated model performance across tasks. UA, JNH and WL generated and curated data for evaluations. FB prepared data for model training. AG, ZW, KK, JNH, AC, WL, and EL analyzed and interpreted the results. FB, DL, and TK contributed to building model-sharing frameworks. AG, ZW, KK, and JNH prepared figures. AG, ZW, KK, JNH, and EL wrote the paper. TK contributed with methodology discussions. EL supervised the project. All authors edited and commented on the manuscript.

Competing Interests

E.L. is an advisor for the Chan-Zuckerberg Initiative Foundation, Element Biosciences, Cartography Biosciences, Pfizer, Santa Ana Bio, and Pixelgen Technologies.

References

1. S. N. Chandrasekaran, B. A. Cimini, A. Goodale, L. Miller, M. Kost-Alimova, N. Jamali, J. G. Doench, B. Fritchman, A. Skepner, M. Melanson, A. A. Kalinin, J. Arevalo, M. Haghighi, J. C. Caicedo, D. Kuhn, D. Hernandez, J. Berstler, H. Shafqat-Abbasi, D. E. Root, S. E. Swalley, S. Garg, S. Singh, A. E. Carpenter, Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nat. Methods* **21**, 1114–1121 (2024).
2. J. C. Caicedo, J. Arevalo, F. Piccioni, M.-A. Bray, C. L. Hartland, X. Wu, A. N. Brooks, A. H. Berger, J. S. Boehm, A. E. Carpenter, S. Singh, Cell Painting predicts impact of lung cancer variants. *Mol. Biol. Cell* **33**, ar49 (2022).
3. M.-A. Bray, S. Singh, H. Han, C. T. Davis, B. Borgeson, C. Hartland, M. Kost-Alimova, S. M. Gustafsdottir, C. C. Gibson, A. E. Carpenter, Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **11**, 1757–1774 (2016).
4. P. J. Thul, L. Åkesson, M. Wiking, D. Mahdessian, A. Geladaki, H. Ait Blal, T. Alm, A. Asplund, L. Björk, L. M. Breckels, A. Bäckström, F. Danielsson, L. Fagerberg, J. Fall, L. Gatto, C. Gnann, S. Hober, M. Hjelmare, F. Johansson, S. Lee, C. Lindskog, J. Mulder, C. M. Mulvey, P. Nilsson, P. Oksvold, J. Rockberg, R. Schutten, J. M. Schwenk, Å. Sivertsson, E. Sjöstedt, M. Skogs, C. Stadler, D. P. Sullivan, H. Tegel, C. Winsnes, C.

Zhang, M. Zwahlen, A. Mardinoglu, F. Pontén, K. von Feilitzen, K. S. Lilley, M. Uhlén, E. Lundberg, A subcellular map of the human proteome. *Science* **356** (2017).

5. M. Doron, T. Moutakanni, Z. S. Chen, N. Moshkov, M. Caron, H. Touvron, P. Bojanowski, W. M. Pernice, J. C. Caicedo, Unbiased single-cell morphology with self-supervised vision transformers. *bioRxiv*, doi: 10.1101/2023.06.16.545359 (2023).

6. V. Kim, N. Adaloglou, M. Osterland, F. M. Morelli, M. Halawa, T. König, D. Gnutt, P. A. M. Zapata, Self-supervision advances morphological profiling by unlocking powerful image representations, *Bioinformatics* (2023). https://www.biorxiv.org/content/10.1101/2023.04.28.538691v2.full.pdf.

7. D. Mahdessian, A. J. Cesnik, C. Gnann, F. Danielsson, L. Stenström, M. Arif, C. Zhang, T. Le, F. Johansson, R. Schutten, A. Bäckström, U. Axelsson, P. Thul, N. H. Cho, O. Carja, M. Uhlén, A. Mardinoglu, C. Stadler, C. Lindskog, B. Ayoglu, M. D. Leonetti, F. Pontén, D. P. Sullivan, E. Lundberg, Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. *Nature* **590**, 649–654 (2021).

8. C. J. Jeffery, Protein moonlighting: what is it, and why is it important? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373** (2018).

9. T. Le, C. F. Winsnes, U. Axelsson, H. Xu, J. Mohanakrishnan Kaimal, D. Mahdessian, S. Dai, I. S. Makarov, V. Ostankovich, Y. Xu, E. Benhamou, C. Henkel, R. A. Solovyev, N. Banić, V. Bošnjak, A. Bošnjak, A. Miličević, W. Ouyang, E. Lundberg, Analysis of the Human Protein Atlas weakly supervised single-Cell Classification competition. *Nat. Methods* **19**, 1221–1229 (2022).

10. S. S. Husain, E.-J. Ong, D. Minskiy, M. Bober-Irizar, A. Irizar, M. Bober, Single-cell subcellular protein localisation using novel ensembles of diverse deep architectures. *Commun. Biol.* **6**, 489 (2023).

11. Q. Tang, R. Ratnayake, G. Seabra, Z. Jiang, R. Fang, L. Cui, Y. Ding, T. Kahveci, J. Bian, C. Li, H. Luesch, Y. Li, Morphological profiling for drug discovery in the era of deep learning. *Brief. Bioinform.* **25**, bbae284 (2024).

12. C. McQuin, A. Goodman, V. Chernyshev, L. Kamentsky, B. A. Cimini, K. W. Karhohs, M. Doan, L. Ding, S. M. Rafelski, D. Thirstrup, W. Wiegraebe, S. Singh, T. Becker, J. C. Caicedo, A. E. Carpenter, CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biol.* **16**, e2005970 (2018).

13. D. R. Wong, D. J. Logan, S. Hariharan, R. Stanton, D.-A. Clevert, A. Kiruluta, Deep representation learning determines drug mechanism of action from cell painting images. *Digit. Discov.* **2**, 1354–1367 (2023).

14. O. Kraus, K. Kenyon-Dean, S. Saberian, M. Fallah, P. McLean, J. Leung, V. Sharma, A. Khan, J. Balakrishnan, S. Celik, D. Beaini, M. Sypetkowski, C. V. Cheng, K. Morse, M.

Makes, B. Mabey, B. A. Earnshaw, Masked autoencoders for microscopy are scalable learners of cellular biology. *ArXiv* **abs/2404.10242** (2024).

15. H. Kobayashi, K. C. Cheveralls, M. D. Leonetti, L. A. Royer, Self-supervised deep learning encodes high-resolution features of protein subcellular localization. *Nat. Methods* **19**, 995–1003 (2022).

16. N. H. Cho, K. C. Cheveralls, A.-D. Brunner, K. Kim, A. C. Michaelis, P. Raghavan, H. Kobayashi, L. Savy, J. Y. Li, H. Canaj, J. Y. S. Kim, E. M. Stewart, C. Gnann, F. McCarthy, J. P. Cabrera, R. M. Brunetti, B. B. Chhun, G. Dingle, M. Y. Hein, B. Huang, S. B. Mehta, J. S. Weissman, R. Gómez-Sjöberg, D. N. Itzhak, L. A. Royer, M. Mann, M. D. Leonetti, OpenCell: Endogenous tagging for the cartography of human cellular organization. *Science* **375**, eabi6983 (2022).

17. T. Clark, J. Mohan, L. Schaffer, K. Obernier, S. Al Manir, C. P. Churas, A. Dailamy, Y. Doctor, A. Forget, J. N. Hansen, M. Hu, J. Lenkiewicz, M. A. Levinson, C. Marquez, S. Nourreddine, J. Niestroy, D. Pratt, G. Qian, S. Thaker, J.-C. Bélisle-Pipon, C. Brandt, J. Chen, Y. Ding, S. Fodeh, N. Krogan, E. Lundberg, P. Mali, P. Payne-Foster, S. Ratcliffe, V. Ravitsky, A. Sali, W. Schulz, T. Ideker, Cell Maps for Artificial Intelligence: AI-Ready Maps of Human Cell Architecture from Disease-Relevant Cell Lines, *Systems Biology* (2024). https://www.biorxiv.org/content/10.1101/2024.05.21.589311v1.full.pdf.

18. K. He, X. Chen, S. Xie, Y. Li, P. Doll'ar, R. B. Girshick, Masked Autoencoders Are Scalable Vision Learners. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 15979–15988 (2021).

19. T. Chen, S. Kornblith, M. Norouzi, G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations" in *Proceedings of the 37th International Conference on Machine Learning*, H. D. Iii, A. Singh, Eds. (PMLR, 13--18 Jul 2020)vol. 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607.

20. J.-B. Grill, F. Strub, F. Altch'e, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. '. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap Your Own Latent: A new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **abs/2006.07733** (2020).

21. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, *arXiv [cs.CV]* (2020). http://arxiv.org/abs/2010.11929.

22. M. Ilse, J. M. Tomczak, M. Welling, Attention-based deep multiple instance learning, *arXiv [cs.LG]* (2018). http://arxiv.org/abs/1802.04712.

23. Z. Huang, X. Jin, C. Lu, Q. Hou, M.-M. Cheng, D. Fu, X. Shen, J. Feng, Contrastive Masked Autoencoders are Stronger Vision Learners. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 2506–2517 (2024).

24. N. Moshkov, M. Bornholdt, S. Benoit, M. Smith, C. McQuin, A. Goodman, R. A. Senft, Y. Han, M. Babadi, P. Horvath, B. A. Cimini, A. E. Carpenter, S. Singh, J. C. Caicedo,

Learning representations for image-based profiling of perturbations. *Nat. Commun.* **15**, 1594 (2024).

25. S. N. Chandrasekaran, H. Ceulemans, J. D. Boyd, A. E. Carpenter, Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nat. Rev. Drug Discov.* **20**, 145–159 (2021).
26. N. Mantel, The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**, 209–220 (1967).
27. M. Marti-Solano, S. E. Crilly, D. Malinverni, C. Munk, M. Harris, A. Pearce, T. Quon, A. E. Mackenzie, X. Wang, J. Peng, A. B. Tobin, G. Ladds, G. Milligan, D. E. Gloriam, M. A. Puthenveedu, M. M. Babu, Combinatorial expression of GPCR isoforms affects signalling and drug responses. *Nature* **587**, 650–656 (2020).