This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CellRep: Multichannel Image Representation Learning Model

Lawrence Phillips
Artificial Intelligence & Digital Research,
Novo Nordisk
lwph@novonordisk.com

Rory Donovan-Maiye
Artificial Intelligence & Digital Research,
Novo Nordisk
rzdm@novonordisk.com

Abstract

Reliable feature extraction from multichannel microscopy images is crucial for biological discovery, but existing models typically require fixed channel architectures or artificial RGB compositing. We introduce CellRep, a channel-invariant foundation model that generates consistent feature representations across varying experimental conditions. By employing content-aware patch embedding and channel-mixing transformer encoding, CellRep learns to identify and represent biological structures independent of channel position or type. Our evaluations demonstrate CellRep's strong performance as a microscopy image featurizer for perturbation prediction, particularly when generalizing to novel cell types, imaging techniques, and channel configurations not seen during training.

1. Introduction

A fundamental challenge in biological research is quantifying cellular responses to genetic and chemical perturbations at scale. Understanding how cells respond to targeted interventions—whether through genetic modifications, chemical compounds, or environmental factors—provides critical insights into disease mechanisms and potential therapeutic targets. Microscopy remains a cornerstone technique for understanding cellular biology, with multichannel imaging providing vital information about cellular structures and responses to experimental interventions. High-content screening (HCS) combines automated fluorescence microscopy with computational image analysis to simultaneously measure multiple cellular features across many samples [10]. These systems typically capture multichannel images where each channel reveals specific cellular components through distinct fluorescent markers or stains, more clearly revealing cellular phenotypes.

Extracting the features with the richest possible biological signal from microscopy images is essential for advancing our understanding of complex cellular processes and enabling the discovery of novel biomarkers, drug targets, and disease mechanisms through unbiased feature extraction. Traditional tools like CellProfiler [5] rely on predefined feature extraction algorithms, which may miss subtle or complex patterns that neural networks can detect. Deep learning models have shown superior performance in many cell imaging tasks, including phenotype classification [12]. Promisingly, representation learning models excel at learning hierarchical representations that could correspond to biological structures at different scales. However, they face challenges when applied to multichannel microscopy data. This is because most state-of-the-art computer vision models were developed for natural RGB images, where the three channels exhibit high redundancy and information correlation. This presents a fundamental mismatch with cell staining assays in cellular microscopy imaging, where each channel captures distinct biological information through different fluorescent markers or imaging modalities. Using RGB-based models requires artificial channel compositing through tools like CellProfiler [5], introducing unnecessary complexity and potential artifacts into image processing pipelines. Furthermore, compositing down to RGB is inherently a form of lossy compression, which could limit these models' ability to generalize to unseen conditions across different experimental conditions, cell types, or imaging protocols, where channel structures may differ substantially.

This challenge of handling independent channels extends beyond just cell microscopy. Other examples in biological imaging include FISH (Fluorescence In Situ Hybridization), where different DNA/RNA sequences are labeled with distinct fluorescent probes, and immunofluorescence screens, whereby multiple antibodies tagged with different fluorophores can show the distribution of different proteins. Satellite imaging captures multiple spectral bands that each highlight different surface features, from vegetation health to thermal signatures. Medical imaging modalities like MRI generate multiple contrast weightings (T1, T2, FLAIR) that provide complementary anatomical information. Materials science techniques such as Energy Dispersive X-ray Spectroscopy produce channel-specific maps of different chemi-

cal elements. In each case, channels contain fundamentally different information rather than the correlated color components found in natural RGB images.

We introduce CellRep, a channel-invariant foundation model that generates strong feature representations. By employing content-aware patch embedding and channel-mixing encoding, CellRep learns to identify and represent biological structures independent of channel position or type. Our quantitative evaluations demonstrate CellRep's superior performance as a microscopy image featurizer, particularly when generalizing to novel cell types, imaging techniques, and channel configurations not seen during training.

2. Related Work

Like other areas of machine learning, computer vision has recently undergone a revolution due to the Transformer [14] and self-supervised learning methods. Notable advances include masked autoencoder (MAE) [9] and self-distillation models using Vision Transformers (ViTs) [7] as backbones, most prominently DINOv2 [13]. DINOv2 has emerged as the leading approach for general computer vision tasks, demonstrating exceptional performance across diverse applications. As mentioned before, DINOv2 was designed for natural RGB images, meaning its use in HCS requires compositing multichannel images into RGB composites, which leads to information loss and potential distortion of biological signals.

Recent work has attempted to address the limitations of standard computer vision models for multichannel microscopy, notably ChannelViT [1] and Phenom-Beta [11] [10]. ChannelViT employs learnable channel embeddings that are added to patch embeddings, allowing the model to process variable numbers of channels. While innovative, this approach has a critical limitation: the channel embeddings are position-specific and learned during training for predetermined channel types. Though it can handle missing channels from its training set, it cannot meaningfully process new channel types or channel positions, limiting its generalizability. Additionally, ChannelViT incurs substantial computational overhead as the sequence length grows linearly with channel count, resulting in quadratic growth in attention computation costs[1]

Phenom-Beta takes a different approach, using an MAE specifically designed for multichannel cellular microscopy images by randomly masking patches across all channels simultaneously. The model is trained to reconstruct the masked regions while preserving channel-specific information through separate decoder heads for each channel type. While it avoids some of ChannelViT's limitations, it inherits the performance gap between MAE-based models and state-of-the-art self-distillation approaches like DINOv2. More crucially, it lacks explicit mechanisms for adapting to novel channel types not seen during training.

To address these limitations, we developed CellRep with a focus on channel-invariant feature representation that eliminates the need for channel compositing while maintaining computational efficiency.

[1]While the model employs Hierarchical Channel Sampling (dropping out channels) during training to help mitigate computational costs, this does not fully solve the high context length issue; it saves no cost during inference, and even though the effective sequence length during training is a fraction of the total sequence, it is still higher than that of a standard ViT.

3. Method

Our approach builds on the DINOv2 framework while introducing key modifications to handle arbitrary channel inputs. The architecture consists of three main modifications to the DINOv2 architecture: (1) content-aware patch embedding that preserves channel-specific information, (2) a channel-mixing transformer encoder that enables cross-channel feature sharing, and (3) an efficient pooling mechanism that enables near computational cost parity with DINOv2.

3.1. Content-Aware Patch Embedding

We introduce a content-aware patch embedding approach that processes multi-channel inputs while preserving channel-specific information. Unlike traditional patch embedding methods that treat all channels uniformly, our approach independently processes each channel and incorporates channel-specific context into the patch embeddings.

The method consists of two main components: a channel encoder that captures channel-specific features, and a content-aware patch embedder that combines channel information with spatial patch embeddings. The complete process is detailed in Algorithm 1.

The channel encoder processes each channel independently through a small convolutional network followed by global average pooling. This captures channel-specific features that are then projected to the embedding dimension. These channel embeddings encode the global context of each channel.

The patch embedding process begins by extracting patches from each channel independently using a convolutional layer with kernel size and stride equal to the patch size. This generates patch embeddings that preserve channel-specific spatial information. The channel embeddings from the channel encoder are then added to all patches from their respective channels, allowing the model to incorporate both local patch information and global channel context.

A key advantage of this approach is its flexibility with input dimensions, as it removes the need for fixed image sizes while maintaining the ability to process channel-specific features. This makes it particularly suitable for applications

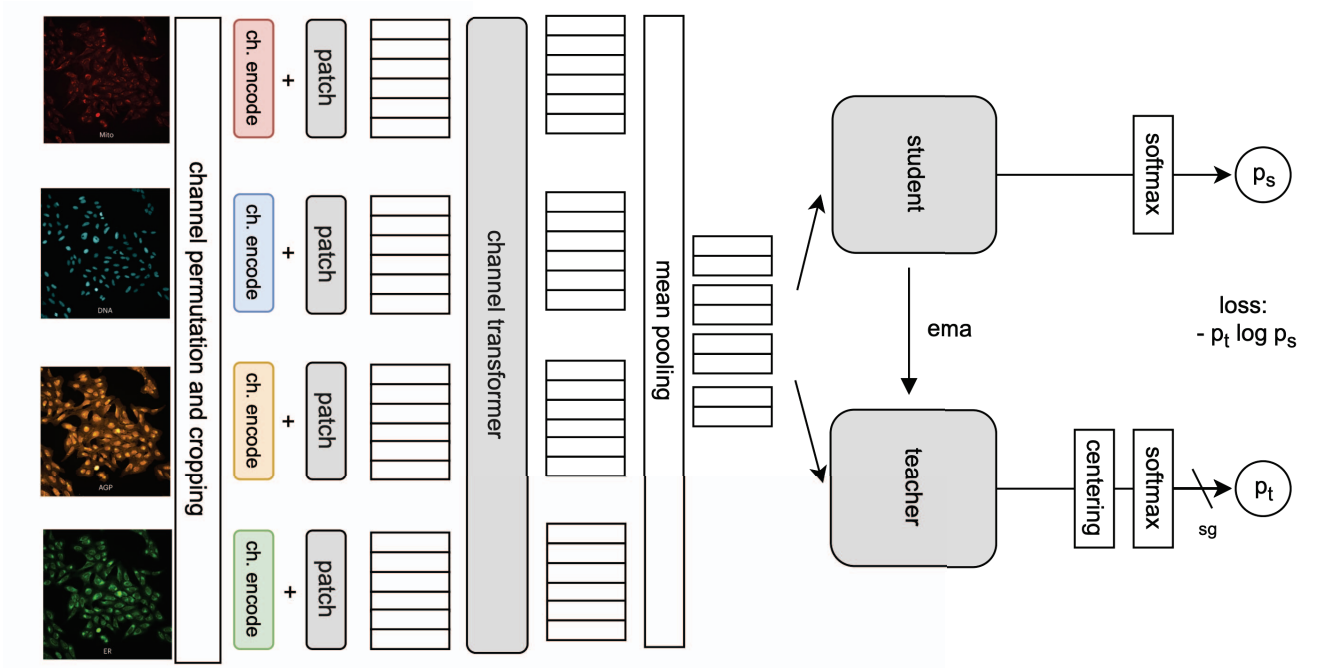

Figure 1. The CellRep architecture first randomly permutes channel order and normalizes channel pixel values. For each image, global and local views are cropped at consistent spatial locations across all channels. Each channel crop is processed in two parallel paths: it passes through a patch embedding layer to generate spatial patch embeddings, while simultaneously, a channel encoder assigns a unique learned vector to each channel type. The channel-specific encoding vector is then added to all patch embedding vectors derived from that same channel, allowing the model to maintain channel identity information while processing spatial features. The resulting patch embeddings are then processed by a transformer encoder that enables feature sharing across all channels and spatial locations. The output embeddings undergo average pooling to reduce dimensionality by a factor of the number of channels. These pooled representations are then used in the DINOv2 student-teacher self-distillation framework, . The teacher network is updated through a momentum-based exponential moving average (ema) of the student's parameters, with gradients flowing only through the student network during backpropagation, as a stop gradient (sg) is applied to the teacher network. All components are jointly optimized during training. Colorized channel samples are taken from [6], and student-teacher depiction is adapted from [4].

involving multi-spectral imaging or datasets with varying image dimensions.

3.2. Channel Transformer and Mean Pooling

To get the model to learn a channel-agnostic representation, we feed the patch embeddings to a small transformer encoder. The same positional embeddings are added to each channel's patch embeddings, ensuring spatial relationships are preserved while maintaining channel independence. Through self-attention mechanisms, the encoder learns to identify and combine relevant features across both spatial locations and channels, enabling the model to process arbitrary channel combinations without requiring channel-specific parameters.

The resulting fused embeddings are then down-sampled using average pooling with a window size equal to the number of channels. The resulting fused embeddings form a sequence of length $L * C$, where L is the number of patch embeddings per channel and C is the number of channels. We apply mean pooling to reduce this sequence to length L. The transformer encoder has already learned to combine relevant information across channels, making this averaging operation more information-preserving. Furthermore, this reduction ensures the sequence length fed to the student and teacher networks matches that of standard DINOv2, allowing us to maintain the same computational efficiency while handling multi-channel inputs.

3.3. Self-distillation

We employ a very similar self-distillation setup to DINOv2. The framework uses a student network that processes both global and local crops and a teacher network that processes only global crops. The student learns to predict the teacher's output distribution for the corresponding views of the same image. Following DINOv2, we use a momentum-updated teacher and center the output distributions using an exponential moving average. Although DINOv2 incorporates iBOT's patch-wise masking as an auxiliary task, we forgo this component as our experiments showed a slightly negative performance impact for fluorescently stained cellular

Algorithm 1 Channel Encoder and Patch Embedding in PyTorch-like pseudocode

procedure ENCODECHANNELS(x, d)
  $B, C \leftarrow x$.shape[: 2]
  $f \leftarrow$ conv2d(reshape(x, ($B \cdot C$, $1, H, W$)), ch)
  $f \leftarrow$ avgpool(relu(f))
  return linear(reshape(f, (B, C, ch)), d)
end procedure
procedure PATCHEMBED(x, p, d)
  $B, C, H, W \leftarrow x$.shape
  patches $\leftarrow$ stack([conv2d(x[:, $c : c + 1$], $d, k = p, s = p$) for c in range(C)])
  patches $\leftarrow$ reshape(patches, (B, C, $-1, d$)).transpose($-2, -1$)
  patches $\leftarrow$ patches + unsqueeze(EncodeChannels(x, d), 2)
  return layer_norm(reshape(patches, ($B, -1, d$)))
end procedure

images.

4. Results

4.1. Training Data

Our training data is composed of two large-scale cell painting datasets from the Broad Institute: CDRP-BBBC047-Bray [3] and LINCS-Pilot [8]. After filtering, our training set comprised approximately 1.2 million five-channel microscopy images of cancer cells, namely U2OS and A549 cells, respectively. The Cell Painting assay [2] used in these datasets captures distinct cellular components through the following channels:

- RNA/nucleoli and cytoplasmic RNA (SYTO 14)
- ER/endoplasmic reticulum (concanavalin A)
- AGP/actin, Golgi and plasma membrane (phalloidin and WGA)
- Mito/mitochondria (MitoTracker Deep Red)
- DNA/nucleus (Hoechst 33342)

These datasets contain images of cells treated with diverse chemical compounds, providing a rich set of morphological phenotypes for model training. For both training and testing datasets for all models, we applied our full normalization pipeline to ensure consistent processing across all experiments.

4.2. Model Training and Comparison

We evaluate Cellrep versus two other architectures: DINOv2 and Phenom-Beta[2]. We pretrained all three models from scratch on the training data described above. For CellRep, we directly fed the individual normalized channels as input. For DINOv2, which requires RGB input, we created channel composites using CellProfiler's standard compositing process, assigning colors to channels evenly spaced around the color wheel; we also applied selected image augmentations[3] that enabled noticeably better performance than DINOv2's standard augmentations.

All models were implemented using the ViT-Large backbone architecture. DINOv2 and CellRep were trained for 64 epochs, while Phenom-Beta was trained for 50 epochs following the authors' protocol [10].

We pretrain all models from scratch using the same data mix described above for the following reasons.

1. To conduct a fair methodological comparison of architectural choices for cell microscopy analysis
2. To avoid potential local optima from natural image pretraining - our preliminary experiments showed that initializing DINOv2 with their pretrained weights actually degraded performance on the held-out test set, likely due to the significant domain shift between natural and microscopy images

In addition to these baseline models, we also evaluated a variant called DINOv2 Finetuned, where the pretrained DINOv2 model underwent additional self-supervised learning on the set of internal lipocyte images (both composites and brightfield), which includes the lipocyte plates used in evaluation. This variant allows us to assess the strength of CellRep's adaptability by comparing it directly to a base that has had exposure to evaluation data.

4.3. Evaluation Framework

We evaluated our models using three distinct benchmarks designed to test both the generalization capability and biological relevance of the learned representations. For all benchmarks, we extracted embeddings from each model and trained a logistic classifier on either mechanism of action (MoA) or perturbation labels using consistent train/test splits.

Held-out set: Our primary benchmark uses CDRP-bio-BBBC036-Bray, a held-out subset of 124,416 images from CDRP-BBBC047-Bray containing known bioactive compounds. Each compound in this dataset has an annotated. As multiple compounds can share the same MoA, this helps test if they learn biologically meaningful features within the same assay rather than memorizing compound-specific artifacts or batch effects. To ensure statistical reliability, we restrict our evaluation to the 40 most frequent MoA classes in CDRP-bio-BBBC036-Bray.

[2]We trained the channel-agnostic MAE version of Phenom-Beta, as it enables inference on unseen numbers of channels as is required for some of our benchmarks.

[3]Specifically, we randomly vary the brightness (up to 25%), contrast (up to 10%), saturation (up to 10%), and hue (such that we randomly spin the color wheel).

Table 1. **Classification Performance**. Held-out set evaluates MoA classification on CDRP-bio-BBBC036-Bray. Lipocyte benchmarks evaluate perturbation classification on novel cell types and staining protocols. We show the top-1 accuracy and weighted average precision scores.

	Held-out Set		5-Channel Lipocyte		4-Channel Lipocyte		1-Channel Brightfield	
Model	Top-1	Precision	Top-1	Precision	Top-1	Precision	Top-1	Precision
CellRep	**0.16**	**0.18**	**0.35**	**0.37**	0.63	**0.68**	0.63	0.65
DINOv2	**0.16**	0.17	0.34	0.34	0.47	0.61	0.60	0.64
DINOv2 Finetuned	0.16	0.17	0.34	0.36	**0.64**	0.66	**0.69**	**0.72**
Phenom-Beta	0.05	0.06	0.05	0.06	0.08	0.29	0.11	0.13

Generalization to Novel Assays: To evaluate generalization to unseen cell and stain (as thus channel) types, we tested on three high quality internal datasets:

- 5-channel lipocyte fluorescent stains: Uses the same channel count but different stain types than the Cell Painting assay used in training
- 4-channel lipocyte fluorescent stains: Tests adaptation to both new stain types and a different channel count
- 1-channel myoblast brightfield: Tests adaptation to new cell type and imaging technique which is single-channel

These datasets include compound perturbation labels but not MoA annotations. Importantly, these assays contain cell types and channel configurations that are not present in the training data, providing a strong test of model generalization.

We do not include DINOv2 implemented with the ChannelViT backbone because it cannot accommodate unseen channel types as is required for the novel assay benchmarks. For Phenom-Beta, we followed the authors' procedure by tiling each image into multiple crops. Each crop was processed independently through the encoder, and the resulting embeddings were averaged to produce a final aggregated embedding representing the entire well.

4.4. Classification Performance

Table 1 presents the classification performance across our evaluation benchmarks. For clarity, mechanism of action (MoA) labels categorize compounds by their biological effect (e.g., "HDAC inhibitor" or "proteasome inhibitor"), while perturbation labels identify specific compounds applied or genetic manipulations performed on cells (e.g., "Melatonin", "Paclitaxel", or "siRNA knockdown of gene X"). We observe several key findings:

Both DINO and CellRep significantly outperform Phenom-Beta across all benchmarks. We hypothesize this gap stems from self-distillation approaches learning invariant features across different image views, while masked autoencoders focus on pixel-level reconstruction that may not capture subtle phenotypic differences.

On the held-out set evaluation, CellRep achieves comparable accuracy with DINOv2 but slightly lower precision.

On the 5-channel lipocyte benchmark with novel stain types, CellRep (0.35 accuracy, 0.37 precision) performs slightly better than DINOv2 (0.34 accuracy, 0.34 precision).

The performance difference becomes more noticeable in the 4-channel lipocyte benchmark, where CellRep (0.63 accuracy, 0.68 precision) achieves higher scores than DINOv2 (0.47 accuracy, 0.61 precision). This suggests potential adaptability to different channel configurations.

With 1-channel brightfield myoblast images, CellRep (0.63 accuracy, 0.65 precision) is marginally better than DINOv2 (0.60 accuracy, 0.64 precision), suggesting that CellRep may be slightly more robust to featurizing across different imaging modalities, though further tests are necessary.

The results show CellRep performs similarly to DINOv2 Finetuned on lipocyte plates (CellRep: 0.63 accuracy, 0.68 precision vs. DINOv2 Finetuned: 0.64 accuracy, 0.66 precision on 4-channel lipocyte), notable given that DINOv2 Finetuned was exposed to the evaluation lipocyte images during SSL training.

For myoblast brightfield evaluation, DINOv2 Finetuned achieved higher metrics than CellRep (DINOv2 Finetuned: 0.69 accuracy, 0.72 precision vs. CellRep: 0.63 accuracy, 0.65 precision). This difference may be attributed to the domain shift in brightfield imaging, which DINOv2 Finetuned encountered during training while CellRep did not.

These results validate CellRep's effectiveness in learning generalizable representations from multichannel microscopy data, demonstrating adaptability to novel cell types, staining protocols, and imaging methods.

5. Discussion

Our work demonstrates that CellRep successfully addresses a fundamental limitation in applying state-of-the-art computer vision models to multichannel microscopy. By combining content-aware patch embedding with channel-mixing encoding, CellRep achieves true channel invariance—allowing flexible processing of any number of chan-

nels in any order without performance degradation. When paired with DINOv2's self-distillation framework, CellRep achieves superior generalization across novel cell types and imaging conditions while maintaining computational efficiency. By eliminating channel compositing information loss, CellRep emerges as a best-in-class foundation model for biological image analysis, with applications across multichannel imaging domains.

In attempting to build a channel-invariant representation, we experimented with a variety of design choices. One such design choice was dropping out channels before showing them to the teacher network while showing the student the full multi-channel views. Despite speculation that this could lead to better biological understanding by forcing the student to learn representations robust to missing channels, this approach performed worse in practice. Similarly, when we tried dropping out channels shown to the student in the CellRep architecture while maintaining full information for the teacher, performance degraded. We hypothesize this is because the distillation objective is partially unachievable, which could entail a weaker learning signal.

Several promising avenues for future research emerge from this work:

First, while our current implementation uses mean pooling to compress channel-wise information, this presents an interesting trade-off space that warrants deeper investigation. Future work could systematically evaluate how different pooling ratios affect the performance-computation trade-off. More sophisticated learned compression approaches using attention weights could potentially preserve more information, though mean pooling may actually serve as beneficial regularization in low-data regimes.

The architecture could be extended to incorporate multi-scale patch embeddings, potentially allowing the model to better capture both fine-grained subcellular features and whole cell-level patterns simultaneously. This might be particularly valuable for applications involving varying microscopy magnifications or multi-scale biological phenomena.

References

[1] Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel vision transformers: An image is worth $1 \times 16 \times 16$ words. In *International Conference on Learning Representations*, 2024. 2

[2] Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9):1757–1774, 2016. 4

[3] Mark-Anthony Bray, Sigrun M Gustafsdottir, Mohammad H Rohban, Shantanu Singh, Vebjorn Ljosa, Katherine L Sokolnicki, Joshua A Bittker, Nicole E Bodycombe, Vlado Dančík, Thomas P Hasaka, et al. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the cell painting assay. *GigaScience*, 6(12):giw014, 2017. 4

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *CoRR*, abs/2104.14294, 2021. 3

[5] Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A Guertin, Joo Han Chang, Robert A Lindquist, Jason Moffat, Polina Golland, and David M Sabatini. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, 7(10):R100, 2006. 1

[6] Srinivas Niranj Chandrasekaran, Beth A. Cimini, and Anne E. Carpenter. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 21:1114–1121, 2024. 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[8] Sigrun M Gustafsdottir, Vebjorn Ljosa, Katherine L Sokolnicki, J Anthony Wilson, Deepika Walpita, Melissa M Kemp, Kathleen Petri Seiler, Hyman A Carrel, Todd R Golub, Stuart L Schreiber, Paul A Clemons, Anne E Carpenter, and Alykhan F Shamji. Multiplex cytological profiling assay to measure diverse cellular states. *PLoS ONE*, 8(12):e80999, 2013. 4

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[10] Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, Dominique Beaini, Maciej Sypetkowski, Chi Vicky Cheng, Kristen Morse, Maureen Makes, Ben Mabey, and Berton Earnshaw. Masked autoencoders for microscopy are scalable learners of cellular biology. *arXiv preprint arXiv:2306.04669*, 2023. 1, 2, 4

[11] Ben Mabey. Nothing short of phenomenal: New deep learning model available on nvidia's bionemo platform. Recursion Press Release, 2024. Accessed on February 25, 2025. 2

[12] Nikita Moshkov, Michael Bornholdt, Santiago Benoit, Matthew Smith, Claire McQuin, Allen Goodman, Rebecca A. Senft, Yu Han, Mehrtash Babadi, Peter Horvath, Beth A. Cimini, Anne E. Carpenter, Shantanu Singh, and Juan C. Caicedo. Learning representations for image-based profiling of perturbations. *bioRxiv*, page 2022.08.12.503783, 2022. 1

[13] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, 2017. 2

A. Appendix

A.1. Loss

The total training loss consists of three components:

$$\mathcal{L} = \mathcal{L}_{global} + \lambda_{local}\mathcal{L}_{local} + \lambda_{reg}\mathcal{L}_{kl}$$

The global loss $\mathcal{L}_{global}$ is computed between the teacher's prediction on a global view and the student's predictions on all other global views of the same image. Similarly, $\mathcal{L}_{local}$ is computed between the teacher's global view prediction and the student's predictions on all local views. Both losses use cross-entropy:

$$\mathcal{L}_{global} = -\sum_{i}\sum_{j\neq i} P_t^i \log P_s^j$$

$$\mathcal{L}_{local} = -\sum_{i}\sum_{k} P_t^i \log P_s^k$$

where P_t^i is the teacher's prediction on the i-th global view and P_s^j, P_s^k are the student's predictions on the j-th global and k-th local views respectively.

Following DINOv2, we also include the KL regularization term $\mathcal{L}_{kl}$ that encourages uniform output distributions, preventing collapse to trivial solutions:

$$\mathcal{L}_{kl} = D_{KL}(\frac{1}{K}\|\bar{P})$$

where $\bar{P}$ is the average output probability across the batch and K is the output dimension.