Masked Autoencoders for Microscopy are Scalable Learners of Cellular Biology

Oren Kraus[1] Kian Kenyon-Dean[1] Saber Saberian[1] Maryam Fallah[1] Peter McLean[1]
Jess Leung[1] Vasudev Sharma[1] Ayla Khan[1] Jia Balakrishnan[1] Safiye Celik[1]
Dominique Beaini[2] Maciej Sypetkowski[2] Chi Vicky Cheng[1] Kristen Morse[1]
Maureen Makes[1] Ben Mabey[1] Berton Earnshaw[1,2]
[1]Recursion [2]Valence Labs

Abstract

Featurizing microscopy images for use in biological research remains a significant challenge, especially for large-scale experiments spanning millions of images. This work explores the scaling properties of weakly supervised classifiers and self-supervised masked autoencoders (MAEs) when training with increasingly larger model backbones and microscopy datasets. Our results show that ViT-based MAEs outperform weakly supervised classifiers on a variety of tasks, achieving as much as a 11.5% relative improvement when recalling known biological relationships curated from public databases. Additionally, we develop a new channel-agnostic MAE architecture (CA-MAE) that allows for inputting images of different numbers and orders of channels at inference time. We demonstrate that CA-MAEs effectively generalize by inferring and evaluating on a microscopy image dataset (JUMP-CP) generated under different experimental conditions with a different channel structure than our pretraining data (RPI-93M). Our findings motivate continued research into scaling self-supervised learning on microscopy data in order to create powerful foundation models of cellular biology that have the potential to catalyze advancements in drug discovery and beyond. Relevant code and select models released with this work can be found at: https://github.com/recursionpharma/maes_microscopy.

1. Introduction

A fundamental challenge in biological research is quantifying cellular responses to genetic and chemical perturbations and relating them to each other [53, 66]. Image-based experiments have proven to be a powerful approach for exploring cellular phenotypes induced by millions of perturbations [5]. High Content Screening (HCS) systems, which combine automated microscopy with robotic liquid handling technologies, have enabled assaying cellular responses to perturbations on a massive scale. Recent public releases of HCS image sets, like RxRx3 [24] and JUMP-CP [14], consist of millions of cellular images across 100,000s of unique chemical and genetic perturbations and demonstrate the scalability of this approach.

The size of recent HCS experiments presents a unique challenge and opportunity for extracting biologically meaningful representations from these datasets. HCS images are often analyzed with customized cell segmentation, feature extraction, and downstream analysis pipelines [7]. Despite the many discoveries made using this approach [5], developing robust segmentation and feature extraction pipelines

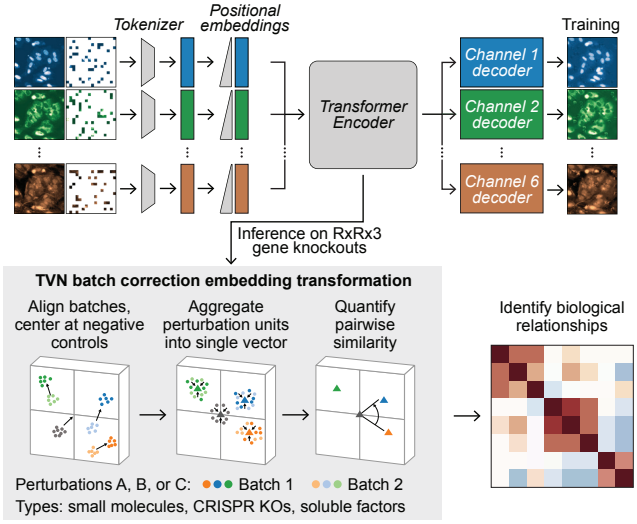

Figure 1. General depiction of the approach taken in this work. MAEs (channel-agnostic architecture depicted) learn to reconstruct HCS images, perform inference on RxRx3 [24] to obtain genomic representations, and apply TVN batch correction on the embeddings to predict biological relationships.

[†]An earlier version of this work appeared at the NeurIPS 2023 Generative AI and Biology Workshop [39].

[‡]Correspondence: oren.kraus@recursion.com, berton.earnshaw@recursion.com, info@rxrx.ai.

using proprietary or open-source software packages [10, 60] remains challenging [12].

Alternatively, representation learning approaches do not require prior knowledge of cellular morphology and have the potential to perform significantly better on practical biological research objectives, e.g., inferring relationships between perturbations [49]. Current SOTA approaches use *weakly supervised learning* (WSL) [71] to train models that predict the perturbations used to treat the cells in an image [8, 49]. However, the performance of WSL models has been found to be sensitive to the strength of perturbations used [49], potentially limiting the applicability of WSL to large scale datasets.

In order to overcome these limitations, we develop an alternative framework for learning representations of HCS datasets based on self-supervised learning (Fig. 1). Specifically, we train masked autoencoders (MAEs) [31] with U-Net and vision transformer (ViT) backbones on progressively larger HCS image sets. We show that these models, particularly MAE ViTs, are scalable learners of cellular biology, outperforming previous SOTA methods at inferring known biological relationships in whole-genome HCS screens. Specifically, we show that

- for MAEs, **recall of known biological relationships scales** with increasing model and training set sizes, while recall degrades when naively scaling WSL,
- a **Fourier domain reconstruction loss** stabilizes MAE training of large ViT backbones, and
- employing a **novel channel-agnostic MAE ViT** helps generalize to microscopy datasets with different channel configurations.

2. Related Work

Deep learning models have been successfully trained to perform cell segmentation [48, 61, 65] and phenotype classification [23, 40, 41, 51], however these supervised learning tasks require the costly creation of segmentation masks and other labels. Inspired by the successful use of embeddings obtained from ImageNet-trained models for other datasets and tasks [54], researchers have used models trained on natural images to featurize HCS data with varying results [1, 52]. Others [8, 49, 57, 62] have used WSL to train convolutional networks to classify labels obtained from experimental metadata (e.g., perturbation class). Despite obtaining SOTA results when trained on small, highly-curated image sets, we show that the performance of WSL models does not necessarily improve on larger datasets.

Vision models pretrained with *self-supervised learning* (SSL) often outperform supervised models on downstream tasks [9, 15, 31]. Unlike supervised pretraining [38], SSL is readily applied to large datasets where labels are lacking or heavily biased. This is useful for HCS datasets, as they contain a wide range of cellular phenotypes that are difficult for human experts to interpret and annotate. For example, DiNO [9] is an SSL method that has been applied to HCS [17, 20, 29, 37, 58] data, however it relies on augmentations inspired by natural images, which may not be applicable to HCS image sets. Alternatively, *masked autoencoders* (MAEs) [31] are trained by reconstructing masked patches conditioned on unmasked patches of an image (Fig. 2). MAEs have been successfully applied to images [31], audio [35], video [25] and multimodal audio-video datasets [34]. However, previous attempts to train MAEs on HCS datasets have had limited success [37, 68], likely due to limitations in compute resources and dataset size.

3. HCS Datasets

We investigate the scaling properties [69] of MAE and WSL pretraining by evaluating increasingly larger models trained on five HCS microscopy datasets of different sizes, as summarized in Table 1 (see Appendix A.1 for additional details). In curating these datasets, we aimed to cover a broad range of biological and experimental factors that could impact a deep learning model's ability to learn transferable representations of the images. These datasets contain images captured using a six-channel proprietary implementation of the Cell Painting imaging protocol [6], which multiplexes fluorescent dyes to reveal eight broadly relevant cellular components. The RPI-52M and RPI-93M (Recursion Phenomics Imageset) datasets also include several million images obtained with Brightfield microscopy imaging. RPI-52M is a superset of RxRx1, RxRx1-2M, and RxRx3, and RPI-93M is a superset of RPI-52M.

4. Methods

This section discusses the strategies we used to train deep computer vision models on our HCS image datasets (Table 1). During pretraining, each model receives as input 256 x 256 crops randomly sampled from images in the training set, preprocessed with channel-wise self-standardization [62]. See Appendix A.2 for more details on training and hyperparameters.

4.1. Weakly supervised learning

We train WSL models to classify *perturbations*; i.e., to predict the genetic or chemical perturbation applied to the cells (e.g., siRNA knockdown, CRISPR knockout, or small molecule) in a random crop of a training image as input.

We reimplement the 28-million parameter DenseNet-161 backbone proposed in Sypetkowski et al. [62], trained to predict cellular perturbations and producing 128-dimensional embeddings from a two-layer MLP neck before the classification logits. We also trained model variants that produce 1,024-dimensional embeddings. We

Pretraining dataset	Imaging modality	Perturbation type(s)	# images	# perturbations
RxRx1 [62]	Cell Painting	gene KD (siRNA)	125,510	1,108
RxRx1-2M	Cell Painting	gene KD (siRNA)	1,650,319	1,108
RxRx3 [24]	Cell Painting	gene KO (CRISPR), SMC	2,222,096	113,517
RPI-52M	Cell Painting, Brightfield	gene KD/KO/OX, SMC, SF	51,516,177	2,345,638
RPI-93M	Cell Painting, Brightfield	gene KD/KO/OX, SMC, SF	92,764,542	3,957,400

Table 1. Summary of the HCS datasets explored for pre-training in this work. Each image in each dataset is 2,048 x 2,048 x 6 pixels. Genetic perturbations include knock-down (KD), knock-out (KO), and overexpression (OX). Non-genetic perturbations include small-molecule compounds (SMC) and soluble factors (SF; e.g. cytokines, biologics). RPI- datasets include genetic perturbations generated with siRNA, CRISPR, and other genetic manipulation technologies.

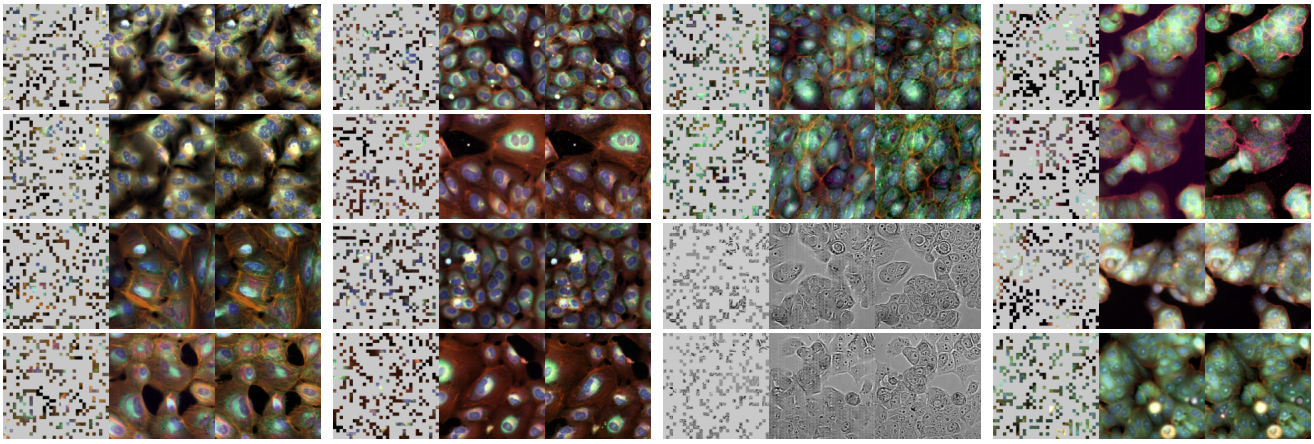

Figure 2. Visualizing MAE ViT-L/8+ (trained on RPI-93M) reconstructions on random *validation* images from four datasets – RxRx1, RxRx3, RPI-52M, and RPI-93M. For each dataset column, we show a triplet of the masked input (left), the reconstruction (middle), and the original (right); for this model, we randomly mask 75% of the 1,024 8x8 patches constructed from the 256 x 256 center crop of the full image. Images are taken from wells on the same experimental plate, rows alternate between randomly sampled negative control and perturbation conditions (see Fig. 1).

trained such models with and without adaptive batch normalization (AdaBN), an architectural technique to enable domain adaptation [44]. Our AdaBN-based DenseNet-161 classifiers are implemented with Ghost BatchNorm [33] in order to train with larger batch sizes.

We also trained WSL models with vision transformers (ViT-B/16 and ViT-L/16) [21], described further in the following sections. Our ViT classifiers use the embedding of the class token from the final layer as the representation of the image crop (we observed minimal difference in downstream performance between using the class token embedding versus averaging over patch embeddings).

4.2. Masked autoencoders

We train and evaluate MAEs with convolutional and transformer backbones of different sizes, depending on the scale of the training set. We provide example reconstructions on our pretraining validation sets in Figure 2, and additional reconstructions in the Appendix A.4.

We adapt U-Nets [56] for use as masked autoencoders (MU-Nets) by training to reconstruct masked sections of input images. We train MU-Nets as described in Xun et al. [68] and report results for MU-Net-M and MU-Net-L, which have 52- and 135-million parameters, respectively. MU-Net-M's downsampling schedule is 32/64/128/256/512, while MU-Net-L incorporates an additional block of size 1,024. In each case, the decoder mirrors the encoder.

We train vision transformers [19, 21, 59, 69] as MAEs following the implementation in He et al. [31]. We report results for ViT-S, ViT-B, and ViT-L encoders [21], containing 22-, 86-, and 304-million parameters, respectively, and producing 384-, 768-, and 1,024-dimensional embeddings respectively. We explore the use of 8x8 and 16x16 patch sizes and 75% and 25% mask ratios (Fig. 2), respectively. A 25-million parameter decoder [31] is used for patch reconstructions. Note that 8x8 patches induce a sequence length 4 times greater than 16x16 patches and are thus more computationally expensive. Our MAE ViTs use the average of patch embeddings from the final layer of the encoder as the

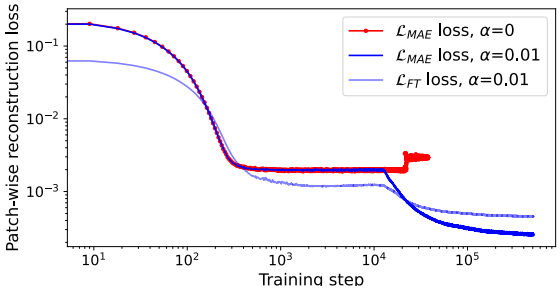

Figure 3. Example reconstruction loss curves (log-log scale) training a CA-MAE ViT-L/16, with and without Fourier domain reconstruction loss (same random seed), on RPI-93M; similar results hold for other large MAE ViTs across multiple runs. Training with $\mathcal{L}_F$ at $\alpha = 0.01$ (Eq. 3) enables surpassing the saddle-point region.

embedding of the image crop.

We observed (Fig. 3) an interesting behavior when training large MAE-ViTs on our largest datasets. Early in training, after a steep initial descent in loss, the model encountered an apparent saddle point region in the parameter landscape. When trained long enough, we could surpass that region and "double-dip" the loss curve after many million crops are seen (depending on model and dataset size). We found that training dynamics and downstream performance benefited from large batch sizes of up to 16,384 image crops and using the Lion optimizer [16], versus the typical choices of batch size and AdamW optimizer [3].

4.2.1 Fourier domain reconstruction loss

Even with the training strategies described above, our largest models with many tokens, such as ViT-L/8, diverged early during training. We also observed that reconstructions lacked the kind of texture prediction that characterize microscopy images, consistent with the original MAE results in which high-frequency textures were not reconstructed well [31]. We therefore added an additional reconstruction loss in the Fourier domain [67] to encourage the model to better reconstruct the textures of cellular morphology, which also facilitated more reliable navigation of the loss landscape for reconstruction in general.

MAEs are trained with mean squared error (L_2) reconstruction loss at the patch level only on the masked patches. Formally, given P masked patches for an individual sample, the patch's image pixels y_p and the model's reconstruction of the patch y'_p:

$$\mathcal{L}_{MAE} = \frac{1}{P}\sum_{p=1}^{P} L_2(y_p, y'_p). \tag{1}$$

We incorporated an additional loss term based on the fast Fourier transformation, $\mathcal{F}$, following the standard reconstruction loss in Eq. 1, calculated on masked patches only:

$$\mathcal{L}_{FT} = \frac{1}{P}\sum_{p=1}^{P} L_1(|\mathcal{F}(y_p)|, |\mathcal{F}(y'_p)|). \tag{2}$$

This loss term incentivizes the model to minimize the mean absolute error (L_1) between the original and reconstructed patches in the frequency domain.

Finally, we combine Eqs. 1 and 2 as follows:

$$\mathcal{L}_{MAE+} = (1-\alpha)\mathcal{L}_{MAE} + \alpha\mathcal{L}_{FT}, \tag{3}$$

where the hyperparameter $\alpha \in (0, 1)$. All models indicated with a + (e.g., ViT-L/8+) are trained using this loss function. We found that setting $\alpha = 0.01$ worked effectively. As illustrated in Figure 3, we found that training with this loss term consistently resulted in a stable double-descent in loss.

4.2.2 Channel-agnostic MAEs

Microscopy images captured by HCS can vary significantly across experiments and labs, often containing different numbers of channels and different cellular objects stained in each channel. Although many labs have aligned on the Cell Painting protocol [6], there are still variations between experimental implementations, with some protocols having 5 or 6 of the fluorescent morphology stains, and others adding brightfield or experiment-specific channels. Standard convolutional- [42] or vision transformer-based [21] architectures require input images to have a consistent set of channels between training and test settings.

In an effort to develop an architecture that can transfer to a different number and set of channels at test time, we developed the *channel-agnostic ViT* architecture (CA-MAE). This architecture was inspired by recent work on multimodal MAEs [2, 26], specifically Bachmann et al. [2], in which RGB images, scene depth and semantic segmentation are considered separate modalities that train a single ViT-based MAE. Our implementation treats each channel as a separate modality, creating $C \times N$ tokens where C is the number of channels and N is the number of patches defined by $N = HW/P^2$, where (H, W) is the resolution of the original image, and (P, P) is the resolution of each image patch. To make the model agnostic to the number and set of channels at test time, we apply a single shared linear projection and the same positional embeddings to all channels based on the standard sine-cosine functions [31]. We apply the masking ratio to the resulting $C \times N$ tokens, producing different masks for each channel. During training, we use separate decoders for each channel similar to the separate decoders used for each modality in Bachmann et al. [2]. We use a 75% (ViT-B/16) or 85% (ViT-L/16) masking ratio. Figure 4 describes this architecture in detail.

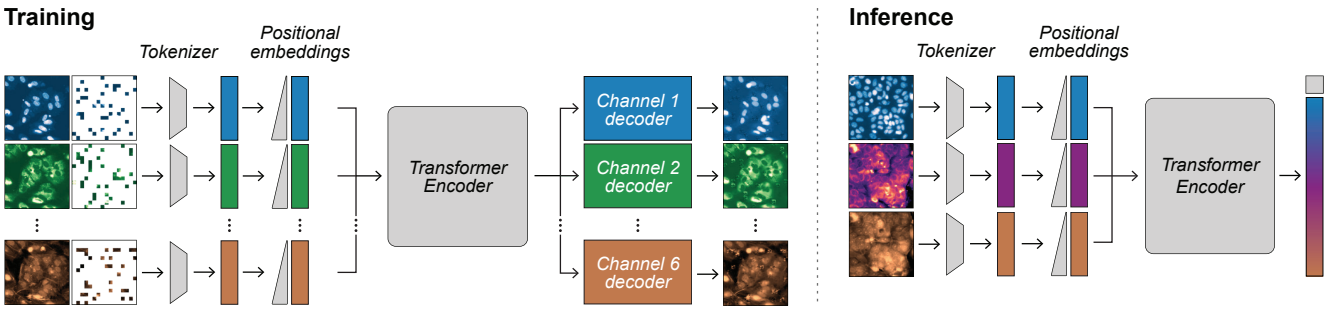

Figure 4. Channel-agnostic MAE (CA-MAE). This architecture enables transferring ViT encoders trained using MAEs from one set of channels to another. *Left*: CA-MAE training (ViT-L/16+, 85% mask) in which an input tensor is split into individual channels and a shared linear projection (Tokenizer) is applied to each channel, followed by the addition of positional embeddings per channel. *Right*: the trained ViT encoder can then be used to embed images with different sets, ordering, and/or numbers of channels (3 shown here) by using the class token, averaging all the patch embeddings, or averaging the patch embeddings from each channel separately and concatenating them.

Table 2. Impact of batch correction methods for RPI-93M MAE ViT-L/8+; findings are similar for other models. Recall of known relationships in top and bottom 5% of cosine similarities on CORUM/hu.MAP/Reactome/StringDB databases.

Transformation method	Recalls
No transformation	.124/.124/.096/.135
PCA	.126/.122/.102/.134
Center by plate	.449/.361/.184/.350
Center by experiment	.455/.365/.186/.353
Standardize by plate	.456/.367/.187/.359
Standardize by experiment	.460/.370/.188/.359
PCA+Standardize by plate	.614/.435/.261/.477
PCA+Standardize by experiment	.614/.435/.258/.477
TVN	**.622/.443/.267/.484**

5. Results

We evaluated our models based on their ability to identify biological relationships as well as predict aggregated single cell features [60].

5.1. Predicting biological relationships

A valuable use of large-scale HCS experiments is to perform large-scale inference of biological relationships between various types of perturbations. We evaluate each model's ability to recall known relationships by using the multivariate metrics described in Celik et al. [11]. We correct for batch effects using *Typical Variation Normalization* (TVN) [1, 11], and also correct for possible chromosome arm biases known to exist in CRISPR-Cas9 HCS data [43]. Table 2 shows the impact of other batch correction techniques on relationship prediction.

To predict biological relationships, we compute the aggregate embedding of each perturbation by taking the spherical mean over its replicate embeddings. We use the cosine similarity of a pair of perturbation representations as the relationship metric, setting the origin of the space to the mean of negative controls. We compare these similarities with the relationships found in the following public databases: CORUM [28], hu.MAP [22], Reactome [27], and StringDB [63] (with >95% combined score).

Table 3 reports the recall of known relationships amongst the top and bottom 5% of all cosine similarities between CRISPR knockout representations in RxRx3 [24]. This required embedding approximately 140 million image crops and aggregating them by gene. As expected, random baselines recall ∼10% of known relationships in each database (since recall is calculated from 10% of all cosine similarities). A baseline using 30 different pixel intensity statistics as image features already recalls relationships surprisingly well compared to random. Just as surprising, pretrained ImageNet models outperform most WSL models trained on HCS datasets. The one exception is ViT-L/16 trained on RxRx1-2M. RxRx1-2M is a dataset carefully curated to contain a large number of distinct perturbations with strong, consistent phenotypes across many cell types. The relative improvement this model achieves over training on RxRx3 suggests that implementing WSL on HCS data requires the training dataset to be curated for high-quality classes. However, this is resource intensive, experimentally and computationally, and would need to be repeated for every new HCS assay.

As previously described, we train MU-Nets and MAE ViTs of various sizes on increasingly larger datasets. Table 3 shows that MAEs outperform the pretrained ImageNet and WSL models, especially when we scale up to larger model and training set sizes. For example, our best MAE model, ViT-L/8+ trained on RPI-93M, achieves a 11.5% relative improvement over the best WSL model, ViT-L/16 trained on RxRx1-2M, when recalling known biological relationships in hu.MAP. For reasons mentioned in the previous paragraph, we did not train WSL models on datasets

Table 3. Recall of known relationships in top and bottom 5% of cosine similarities by model, pretraining set, and database. All results are computed on RxRx3 after applying TVN and chromosome arm bias correction. Results include simple baselines, intermediate model checkpoints, ablations, and performant WSL/SSL models. MAEs with + are trained with Fourier domain reconstruction loss, $\alpha = 0.01$ (Eq. 3).

Model backbone	**Pretraining dataset**	CORUM	hu.MAP	Reactome	StringDB
Simple baselines					
Random 1024-dim embeddings	N/A	.100	.100	.100	.100
Pixel intensity statistics	N/A	.280	.260	.160	.270
ImageNet-pretrained classifiers					
ViT-S/16	Imagenet-21k [55]	.494	.348	.213	.388
ViT-B/16	Imagenet-21k [55]	.511	.344	.216	.395
ViT-B/8	Imagenet-21k [55]	.472	.324	.203	.369
ViT-L/16	Imagenet-21k [55]	.531	.360	.228	.409
Weakly supervised models					
DenseNet-161	RxRx1 [62]	.383	.307	.190	.330
DenseNet-161 w/ AdaBN	RxRx1 [62]	.485	.349	.228	.417
DenseNet-161 w/ AdaBN	RxRx3 [24]	.461	.303	.188	.377
DenseNet-161 w/ AdaBN (1024-dim)	RxRx1 [62]	.502	.363	.220	.422
DenseNet-161 w/ AdaBN (1024-dim)	RxRx3 [24]	.520	.350	.207	.413
ViT-B/16	RxRx1 [62]	.505	.348	.218	.408
ViT-L/16	RxRx3 [24]	.532	.353	.196	.402
ViT-L/16	RxRx1-2M	.568	.397	.255	.472
MU-Nets					
MU-Net-L	RxRx3 [24]	.566	.374	.232	.427
MU-Net-L	RPI-52M	.576	.385	.238	.443
MU-Net-L	RPI-93M	.581	.386	.247	.440
Intermediate MAE ViT checkpoints					
MAE ViT-L/8+ (epoch 1)	RPI-52M	.524	.357	.216	.405
MAE ViT-L/8+ (epoch 25)	RPI-52M	.595	.411	.254	.461
MAE ViT-L/8+ (epoch 46)	RPI-52M	.605	.424	**.267**	.474
MAE ViTs					
MAE ViT-B/16	RxRx3 [24]	.565	.387	.232	.435
MAE ViT-B/16	RPI-52M	.540	.373	.234	.416
MAE ViT-B/8	RPI-52M	.601	.404	.251	.459
MAE ViT-L/16	RxRx3 [24]	.560	.374	.231	.427
MAE ViT-L/16	RPI-52M	.607	.414	.258	.460
MAE ViT-L/16+	RPI-52M	**.626**	.425	.260	.468
MAE ViT-L/8+	RPI-93M	.622	**.443**	**.267**	**.484**
Channel-agnostic MAE ViTs					
CA-MAE ViT-B/16	RPI-52M	.587	.404	.257	.459
CA-MAE ViT-B/16+	RPI-52M	.586	.398	.249	.455
CA-MAE ViT-L/16+	RPI-93M	.614	.424	.264	.478

larger than RxRx3. We also show the performance of intermediate MAE ViT checkpoints and observe that, as training progresses, both the reconstruction of validation images (training loss for epochs 1, 25, and 46 was 2.4e-3, 4.4e-4, and 4.1e-4, respectively) and recall of known biological relationships improve. This indicates that image reconstruction is an appropriate proxy task for capturing biological information for use in downstream tasks of interest.

CA-MAE. Table 3 shows results for three channel-agnostic MAEs (Sec. 4.2.2). Note that CA-MAE ViT-B/16 significantly outperforms the MAE ViT-B/16 when trained on RPI-52M, suggesting that these architectures can offer

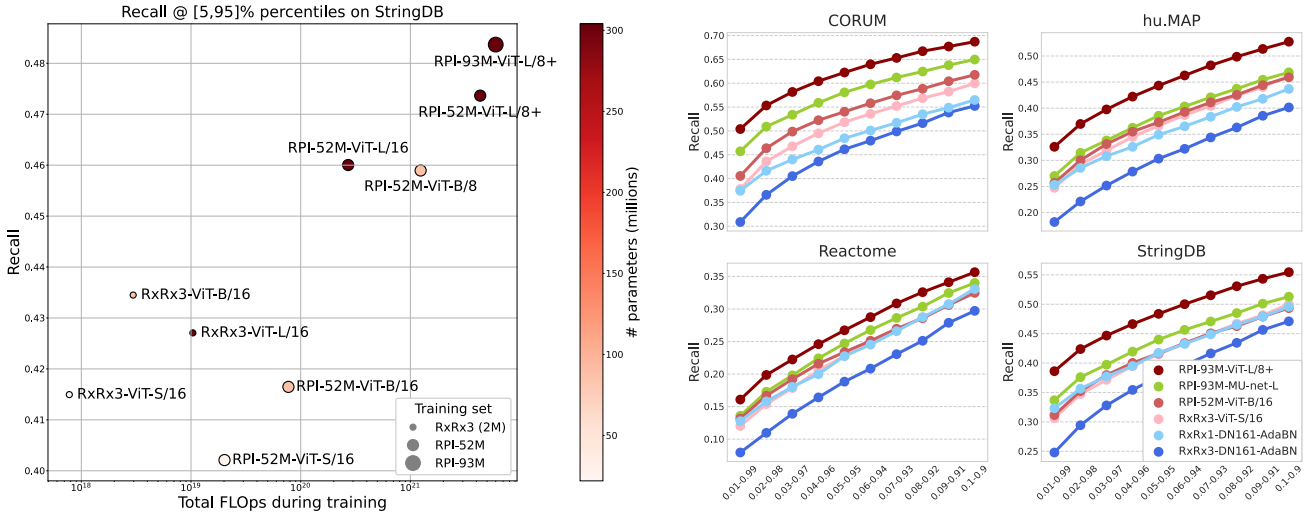

Figure 5. Results for select MAE ViTs taken from Table 3. *Left*: StringDB recall as a function of number of training FLOps. *Right*: Recall across different cosine similarity percentiles on each database. Similar results hold for other models on other datasets.

improved performance over standard MAE ViTs. Moreover, CA-MAEs enable generalizing to datasets with different numbers of channels (see Sec. 5.3). We did not scale CA-MAE to the best performing MAE ViT-L/8+ architecture due to the large number of tokens generated by this architecture (6,144 for 6-channel images). We leave exploring techniques to address large token sequences in training MAEs (e.g., SWIN [45, 46, 70] or dilated attention [30]) to future work.

5.2. MAEs are scalable learners of cellular biology

In Figure 5 we see that recall strongly correlates with the number of training FLOps, a function of both model and training set size (see Appendix A.5 for similar trends on other databases). We also see that the relative performance of different pretrained models on this metric is preserved for different choices of similarity percentiles. Our overall best model, RPI-93M MAE ViT-L/8+, is an MAE ViT-L using 8 x 8 patching, 75% mask ratio, and trained with the Fourier domain reconstruction loss (Eq. 3) on 128 A100 GPUs for over 20,000 GPU hours on the largest dataset, RPI-93M.

5.3. Transfer to JUMP-CP

To further evaluate the transferability of our models, we inferenced CPJUMP1, a subset of the JUMP-CP [14] dataset, and ran the corresponding benchmarking tasks introduced in Chandrasekaran et al. [13]. This dataset includes Cell Painting and Brightfield images of two different cell types with ∼130K unique perturbations and consists of two primary tasks, perturbation retrieval and sibling retrieval, where siblings represent similar but distinct perturbations. For both tasks, cosine similarity between samples is measured for individual perturbations or siblings, and Average

Table 4. Perturbation detection and siblings retrieval on the JUMP-CP dataset, measured in fraction retrieved. Values are averaged (± standard deviation) over cell types, modalities, and time-points.

Model backbone, dataset	Pert.	Siblings
CellProfiler [60]	.53 ±.30	**.13** ±.07
ViT-L/16, ImageNet-21k [55]	.88 ±.09	.06 ±.03
WSL ViT-L/16, RxRx1-2M	.84 ±.08	.02 ±.02
MAE ViT-L/8+, RPI-93M	.78 ±.13	.03 ±.03
CA-MAE ViT-L/16+, RPI-93M	**.95** ±.05	.02 ±.02

Precision (AP) is measured against a null of negative control samples. Permutation testing is used to establish the significance of the AP values, which are then false discovery rate-adjusted to yield q values with a cut-off of 5% for being considered as retrieved.

Some adaptations for image embedding and data normalization were necessary compared to Chandrasekaran et al. [13], including our use of TVN on the negative controls to normalize the embeddings rather than *robustize MAD*. Additionally, use of the WSL ViT-L/16 and MAE ViT-L/8+ models required mapping the JUMP-CP stains to those of the training set and duplicating one channel to match the model's expected six. Meanwhile, the CA-MAE model jointly embedded the five Cell Painting channels and three Brightfield channels, despite being only trained on unpaired six-channel inputs.

We observe significantly improved performance of deep learning models on the perturbation retrieval task compared to CellProfiler [60], while having smaller variability across cell types, modalities, and time-points, indicating that normalized embeddings from these models consistently rep-

Table 5. Recall (at 5% false positive rate) of StringDB relationships for select models on three different gene sets PoC-124/MoA-300/DG-1640 as defined in Sivanandan et al. [58].

Model backbone	Training data	Recalls
WSL DN161 w/ AdaBN	RxRx1 [62]	.79/.**24**/.15
MAE ViT-S/16	RxRx3 [24]	.74/.19/.14
MU-net-L	RPI-52M	.79/.20/.15
MAE ViT-L/8+	RPI-93M	**.80**/.23/.**17**
DiNO ViT-S/8 [58]	CP 1640	.53/.12/.14

resent perturbations despite plate and well variations (Table 4).

In contrast, we note the lower performance of the normalized MAE model embeddings on the sibling retrieval task, where experimentally related pairs of perturbations are less similar compared to CellProfiler features. These observations are consistent with the hypothesis that MAE-trained models produce highly-resolved representations of cellular images that, in this case, are also capable of differentiating even biologically or chemically related perturbations. This illustrates the need to further develop fine-tuning strategies, or alignment methods techniques to increase performance on application-specific tasks, such as relatability among similar reagents in spite of phenotypic variation (as seen here), or other biologically-relevant research objectives like identifying genetic interactors or compound mechanisms of action.

5.4. Comparison with external platforms

We compare these models with recent results from an alternative HCS platform combining pooled CRISPR screening with Cell Painting [58]. Table 5 reports recall at 5% FPR in StringDB on three gene sets defined in Sivanandan et al. [58]. The ViT-L/8+ MAE trained on RPI-93M yields a minimum 20% relative improvement in gene set performance over CP-DiNO 1640 (ViT-S/8), which was trained on ∼1.5M single-cell images. We note the significant differences in assay technology, cell lines, and modeling methodology between the two platforms, making their direct comparison impossible using this metric. Nonetheless, we hope this comparison brings the field closer to an accepted set of benchmarks for evaluating models trained on HCS datasets.

5.5. Predicting morphological features

To determine whether models of different architectures were able to learn a diverse array of morphological characteristics, we used linear regression to predict 955 CellProfiler (CP) features spanning area-shape, texture, radial distribution, intensity, and neighbor categories [10]. Although many of these features are highly correlated and display highly skewed distributions in practice, they nonetheless quantify a diverse set of specific morphological characteristics that can be used to assess the richness of model embeddings. Specifically, we observe that MAE model embeddings (RPI-93M ViT-L/8+) are better predictors of CP extracted morphological features than WSL model embeddings (RxRx1 DenseNet-161 w/ AdaBN), as measured by the coefficient of determination of predicted features from an independent experimental dataset (Fig. 6; see also Appendix A.6). For example, improvements offered by this MAE over the WSL model range from a 14% relative improvement in predicting the AreaShape features (.456 vs .401) to a 148% improvement in predicting the Intensity feature (.737 vs .297), based on the median R^2. These observations suggest that MAEs can produce representations that more effectively capture a wide range of morphological features compared to the most performant WSL model proposed by Sypetkowski et al. [62].

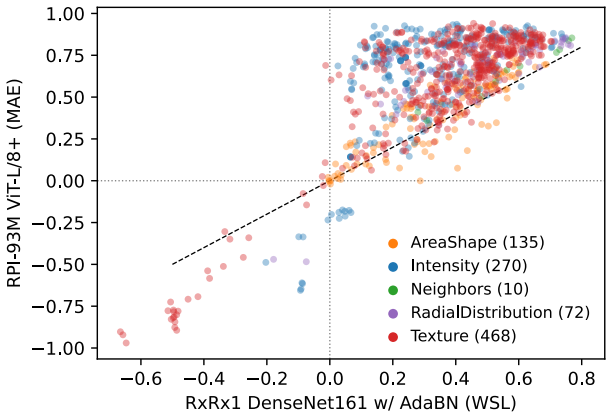

Figure 6. Single-task linear regression illustrates how an MAE-trained embedding model outperforms a WSL-trained model in predicting CellProfiler features across all categories.

6. Conclusion

This work demonstrates that scaling properties [69] apply to learning microscopy-based representations of cellular biology that can accurately infer known biological relationships. Unlike previous approaches that use weakly supervised learning [49, 62] on small, curated datasets, we show that the performance of self-supervised MAEs on biologically meaningful benchmarks scales to massive HCS image sets. Additionally, we introduce a novel reconstruction loss based on the Fourier transform which stabilizes large MAE training, and a channel-agnostic MAE architecture that generalizes to different channel configurations and offers promising directions for future work.

References

[1] D. Michael Ando, Cory Y. McLean, and Marc Berndl. Improving Phenotypic Measurements in High-Content Imaging Screens. *bioRxiv*, page 161422, 2017. 2, 5

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer, 2022. 4

[3] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A Cookbook of Self-Supervised Learning. *arXiv*, 2023. 4, 1, 2

[4] Rodolphe Barrangou and Jennifer A Doudna. Applications of crispr technologies in research and beyond. *Nature biotechnology*, 34(9):933–941, 2016. 1

[5] Michael Boutros, Florian Heigwer, and Christina Laufer. Microscopy-Based High-Content Screening. *Cell*, 163(6): 1314–1325, 2015. 1

[6] 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. 2, 4

[7] Juan C Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S Vasilevich, Joseph D Barry, Harmanjit Singh Bansal, Oren Kraus, Mathias Wawer, Lassi Paavolainen, Markus D Herrmann, Mohammad Rohban, Jane Hung, Holger Hennig, John Concannon, Ian Smith, Paul A Clemons, Shantanu Singh, Paul Rees, Peter Horvath, Roger G Linington, and Anne E Carpenter. Data-analysis strategies for image-based cell profiling. *Nature Methods*, 14(9):849–863, 2017. 1

[8] Juan C. Caicedo, Claire McQuin, Allen Goodman, Shantanu Singh, and Anne E. Carpenter. Weakly Supervised Learning of Single-Cell Feature Embeddings. *bioRxiv*, page 293431, 2018. 2

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. *arXiv*, 2021. 2

[10] 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. 2, 8

[11] Safiye Celik, Jan-Christian Huetter, Sandra Melo, Nathan Lazar, Rahul Mohan, Conor Tillinghast, Tommaso Biancalani, Marta Fay, Berton Earnshaw, and Imran S Haque. Biological cartography: Building and benchmarking representations of life. In *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*, 2022. 5

[12] Srinivas Niranj Chandrasekaran, Hugo Ceulemans, Justin D. Boyd, and Anne E. Carpenter. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nature Reviews Drug Discovery*, 20(2):145–159, 2021. 2

[13] Srinivas Niranj Chandrasekaran, Beth A Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, John G Doench, Briana Fritchman, Adam Skepner, Michelle Melanson, et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Biorxiv*, pages 2022–01, 2022. 7

[14] Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric Alix, D Michael Ando, John Arevalo, Melissa Bennion, Nicolas Boisseau, Adriana Borowa, Justin D Boyd, Laurent Brino, et al. Jump cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, pages 2023–03, 2023. 1, 7, 3

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv*, 2020. 2

[16] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, et al. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*, 2023. 4, 1, 2

[17] Jan Oscar Cross-Zamirski, Guy Williams, Elizabeth Mouchet, Carola-Bibiane Schönlieb, Riku Turkki, and Yinhai Wang. Self-Supervised Learning of Phenotypic Representations from Cell Images with Weak Labels. *arXiv*, 2022. 2

[18] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022. 2

[19] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023. 3, 2

[20] Michael Doron, Théo Moutakanni, Zitong S Chen, Nikita Moshkov, Mathilde Caron, Hugo Touvron, Piotr Bojanowski, Wolfgang M Pernice, and Juan C Caicedo. Unbiased single-cell morphology with self-supervised vision transformers. *bioRxiv*, pages 2023–06, 2023. 2

[21] 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 (ICLR)*, 2020. 3, 4

[22] Kevin Drew, Chanjae Lee, Ryan L Huizar, Fan Tu, Blake Borgeson, Claire D McWhite, Yun Ma, John B Wallingford, and Edward M Marcotte. Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular Systems Biology*, 13(6): 932, 2017. 5

[23] Philipp Eulenberg, Niklas Köhler, Thomas Blasi, Andrew Filby, Anne E. Carpenter, Paul Rees, Fabian J. Theis, and

F. Alexander Wolf. Reconstructing cell cycle and disease progression using deep learning. *Nature Communications*, 8 (1):463, 2017. 2

[24] Marta M Fay, Oren Kraus, Mason Victors, Lakshmanan Arumugam, Kamal Vuggumudi, John Urbanik, Kyle Hansen, Safiye Celik, Nico Cernek, Ganesh Jagannathan, et al. Rxrx3: Phenomics map of biology. *bioRxiv*, pages 2023–02, 2023. 1, 3, 5, 6, 8, 2, 4

[25] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked Autoencoders As Spatiotemporal Learners. *arXiv*, 2022. 2

[26] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. 4

[27] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, Chuan Deng, Thawfeek Varusai, Eliot Ragueneau, Yusra Haider, Bruce May, Veronica Shamovsky, Joel Weiser, Timothy Brunson, Nasim Sanati, Liam Beckman, Xiang Shao, Antonio Fabregat, Konstantinos Sidiropoulos, Julieth Murillo, Guilherme Viteri, Justin Cook, Solomon Shorser, Gary Bader, Emek Demir, Chris Sander, Robin Haw, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D687–D692, 2021. 5

[28] Madalina Giurgiu, Julian Reinhard, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Research*, 47(Database issue): D559–D563, 2019. 5

[29] Johan Fredin Haslum, Christos Matsoukas, Karl-Johan Leuchowius, Erik Müllers, and Kevin Smith. Metadata-guided Consistency Learning for High Content Images. *arXiv*, 2022. 2

[30] Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer. *arXiv preprint arXiv:2209.15001*, 2022. 7

[31] 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, 3, 4, 1

[32] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017. 2

[33] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30, 2017. 3

[34] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Haoqi Fan, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, and Christoph Feichtenhofer. MAViL: Masked Audio-Video Learners. *arXiv*, 2022. 2

[35] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked Autoencoders that Listen. *arXiv*, 2022. 2

[36] Aimee L. Jackson and Peter S. Linsley. Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. *Nature Reviews Drug Discovery*, 9 (1):57–67, 2010. 1

[37] Vladislav Kim, Nikolaos Adaloglou, Marc Osterland, Flavio Morelli, and Paula Andrea Marin Zapata. Self-supervision advances morphological profiling by unlocking powerful image representations. *bioRxiv*, pages 2023–04, 2023. 2

[38] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big Transfer (BiT): General Visual Representation Learning. *arXiv*, 2019. 2

[39] Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, et al. Masked autoencoders are scalable learners of cellular morphology. *arXiv preprint arXiv:2309.16064*, 2023.

[40] Oren Z. Kraus, Jimmy Lei Ba, and Brendan J. Frey. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59, 2016. 2

[41] Oren Z Kraus, Ben T Grys, Jimmy Ba, Yolanda Chong, Brendan J Frey, Charles Boone, and Brenda J Andrews. Automated analysis of high-content microscopy data with deep learning. *Molecular Systems Biology*, 13(4):924, 2017. 2

[42] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 4

[43] Nathan H Lazar, Safiye Celik, Lu Chen, Marta Fay, Jonathan C Irish, James Jensen, Conor A Tillinghast, John Urbanik, William P Bone, Genevieve HL Roberts, et al. High-resolution genome-wide mapping of chromosome-arm-scale truncations induced by crispr-cas9 editing. *bioRxiv*, pages 2023–04, 2023. 5

[44] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018. 3

[45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7

[46] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 7

[47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[48] Erick Moen, Dylan Bannon, Takamasa Kudo, William Graf, Markus Covert, and David Van Valen. Deep learning for cellular image analysis. *Nature Methods*, 16(12):1233–1246, 2019. 2

[49] 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. 2, 8

[50] OpenAI. Gpt-4 technical report, 2023. 2

[51] Wei Ouyang, Casper F. Winsnes, Martin Hjelmare, Anthony J. Cesnik, Lovisa Åkesson, Hao Xu, Devin P. Sullivan, Shubin Dai, Jun Lan, Park Jinmo, Shaikat M. Galib, Christof Henkel, Kevin Hwang, Dmytro Poplavskiy, Bojan Tunguz, Russel D. Wolfinger, Yinzheng Gu, Chuanpeng Li, Jinbin Xie, Dmitry Buslov, Sergei Fironov, Alexander Kiselev, Dmytro Panchenko, Xuan Cao, Runmin Wei, Yuanhao Wu, Xun Zhu, Kuan-Lun Tseng, Zhifeng Gao, Cheng Ju, Xiaohan Yi, Hongdong Zheng, Constantin Kappel, and Emma Lundberg. Analysis of the Human Protein Atlas Image Classification competition. *Nature Methods*, 16(12):1254–1261, 2019. 2

[52] Nick Pawlowski, Juan C Caicedo, Shantanu Singh, Anne E Carpenter, and Amos Storkey. Automating Morphological Profiling with Generic Deep Convolutional Networks. *bioRxiv*, page 085118, 2016. 2

[53] Laralynne Przybyla and Luke A. Gilbert. A new era in functional genomics screens. *Nature Reviews Genetics*, 23(2): 89–103, 2022. 1

[54] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 512–519, 2014. 2

[55] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 6, 7

[56] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[57] M Sadegh Saberian, Kathleen P Moriarty, Andrea D Olmstead, Christian Hallgrimson, François Jean, Ivan R Nabi, Maxwell W Libbrecht, and Ghassan Hamarneh. Deemd: Drug efficacy estimation against sars-cov-2 based on cell morphology with deep multiple instance learning. *IEEE Transactions on Medical Imaging*, 41(11):3128–3145, 2022. 2

[58] Srinivasan Sivanandan, Bobby Leitmann, Eric Lubeck, Mohammad Muneeb Sultan, Panagiotis Stanitsas, Navpreet Ranu, Alexis Ewer, Jordan E Mancuso, Zachary F Phillips, Albert Kim, John W Bisognano, John Cesarek, Fiorella Ruggiu, David Feldman, Daphne Koller, Eilon Sharon, Ajamete Kaykas, Max R Salick, and Ci Chu. A Pooled Cell Painting CRISPR Screening Platform Enables de novo Inference of Gene Function by Self-supervised Deep Learning. *bioRxiv*, pages 2023–08, 2023. 2, 8

[59] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 3

[60] David R. Stirling, Madison J. Swain-Bowden, Alice M. Lucas, Anne E. Carpenter, Beth A. Cimini, and Allen Goodman. CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics*, 22(1):433, 2021. 2, 5, 7

[61] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106, 2021. 2

[62] Maciej Sypetkowski, Morteza Rezanejad, Saber Saberian, Oren Kraus, John Urbanik, James Taylor, Ben Mabey, Mason Victors, Jason Yosinski, Alborz Rezazadeh Sereshkeh, et al. Rxrx1: A dataset for evaluating experimental batch correction methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4284–4293, 2023. 2, 3, 6, 8, 1, 4

[63] Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, Lars J Jensen, and Christian von Mering. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1):D605–D612, 2020. 5

[64] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Hervé Jégou. Three things everyone should know about vision transformers. In *European Conference on Computer Vision*, pages 497–515. Springer, 2022. 2

[65] David A. Van Valen, Takamasa Kudo, Keara M. Lane, Derek N. Macklin, Nicolas T. Quach, Mialy M. DeFelice, Inbal Maayan, Yu Tanouchi, Euan A. Ashley, and Markus W. Covert. Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments. *PLoS Computational Biology*, 12(11):e1005177, 2016. 2

[66] Fabien Vincent, Arsenio Nueda, Jonathan Lee, Monica Schenone, Marco Prunotto, and Mark Mercola. Phenotypic drug discovery: recent successes, lessons learned and new directions. *Nature Reviews Drug Discovery*, 21(12):899–914, 2022. 1

[67] Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Masked frequency modeling for self-supervised visual pre-training. In *The Eleventh International Conference on Learning Representations*, 2022. 4

[68] Dejin Xun, Rui Wang, Xingcai Zhang, and Yi Wang. Microsnoop: a generalist tool for the unbiased representation of heterogeneous microscopy images. *bioRxiv*, pages 2023–02, 2023. 2, 3

[69] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022. 2, 3, 8

[70] Chaoning Zhang, Chenshuang Zhang, Junha Song, John Seon Keun Yi, Kang Zhang, and In So Kweon. A survey

on masked autoencoder for self-supervised learning in vision and beyond. *arXiv preprint arXiv:2208.00173*, 2022. 7

[71] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018. 2