Masked Autoencoders for Microscopy are Scalable Learners of Cellular Biology

Supplementary Material

A. Appendix

A.1. Datasets

RxRx1 [62] is a publicly-available proprietary Cell Painting dataset with 125,510 images of 4 human cell types under 1,108 different siRNA perturbations across 51 experimental batches. A unique feature of this dataset is that it is comprised entirely of siRNA perturbations, which are known to have severe off-target effects silencing hundreds of genes [36] causing very distinct phenotypes.

RxRx1-2M is a private version of RxRx1 containing over 1.6 million images across 16 different cell types and uses the same set of siRNA perturbations in RxRx1 from additional experimental batches.

RxRx3 [24] is a publicly-available proprietary Cell Painting dataset with over 2.2 million images of HUVEC cells each perturbed with one of 17,063 CRISPR knockouts (using one of six different guides) or 1,674 compounds across 180 experimental batches. This is the largest publicly available whole-genome HCS image set. CRISPR is a much more accurate technique for knocking out genes compare to siRNA and produces subtler phenotypes by targeting individual genes [4].

RPI-52M (Recursion Phenomics Imageset) is a private dataset with approximately 52 million proprietary images spanning 6,638 experimental batches and 40 cell types. This is a superset of the preceeding three datasets.

RPI-93M is a private dataset with approximately 93 million proprietary images spanning over 10,000 experimental batches and 41 cell types. To our knowledge, this is the largest HCS dataset collected for model training purposes. This is a superset of the preceding four datasets.

Train and Validation splits

All of the datasets are split such that model evaluation is performed on a non-overlapping set of *experiments*, i.e. groups of multi-well plates containing replicates of perturbations in randomized layouts per plate, to avoid data-leakage.

A.2. Model hyperparameters

Models trained on RxRx1 and RxRx1-2M were trained for 100 epochs, on RxRx3 for 50 epochs, and on RPI-52M and RPI-93M for up to 50 epochs, with early stopping depending on when validation performance plateaued. All models (except those using AdaBN) use random sampling without replacement over the full dataset to create training batches. Readers are encouraged to read [62] for more details on batch construction for AdaBN models.

A.2.1 Weakly supervised learning

All WSL models were initialized from Image-Net pretraining weights. For the DenseNet-161-based classifiers, we searched over different batch sizes, learning rates, and optimizers. We empirically found that a batch size of 4,096 with standard SGD+momentum optimization performs best on the classification task, one-cycle learning rate schedule with cosine decay and a 10% warm-up, a maximum learning rate of 0.32768, momentum of 0.9, and weight decay of 0.00001. For ViT-based classifiers, we used a batch size of 4,096, AdamW optimizer with a learning rate of at most 1e-3 using a one-cycle learning rate schedule with cosine decay and a 10% warm-up, $\beta_1 = 0.9$ and $\beta_2 = 0.95$, and a weight decay of 0.05.

All non-AdaBN classifiers used weighted random sampling based on the perturbation labels in the dataset, whereas AdaBN models used a custom batch sampler to ensure that batches were sampled from the same experimental plate. For DenseNet-161-based classifiers, we used a sub-batch-size of 16 for GhostBN.

A.2.2 Masked U-nets

MU-Nets trained on RxRx3 used a global batch size of 4,096, while those trained on RPI-52M and RPI-93M used a global batch size of 16,384. Each was trained using the AdamW optimizer [47] with $\beta_1 = 0.9$ and $\beta_2 = 0.95$, weight decay of 0.05, maximum learning rate 1e-3, cyclic cosine learning rate schedule, and no gradient clipping. We experimented with different mask ratios (25%, 50%, 75%) and kernel sizes (3, 5). We compared the performance on the recall of biological relationships, similar to Table 6, for these values. Changing the mask ratio or kernel size did not seem to effect the performance.

A.2.3 Masked Autoencoder Vision Transformers

MAE-ViTs on RxRx3 trained with a global batch size of 4,096, while those trained on RPI-52M and RPI-93M used a global batch size of 16,384. Each used the Lion optimizer [16] with $\beta_1 = 0.9$ and $\beta_2 = 0.95$, weight decay of 0.05, and no gradient clipping (based on the AdamW optimizer settings from He et al. [31]). We found that training dynamics and downstream performance was significantly better with large batch sizes and the Lion optimizer versus using the recommended batch size and AdamW settings presented by Balestriero et al. [3]. All ViT-S and ViT-B encoders were trained with a maximum learning rate of 1e-4 and all ViT-L encoders were trained with a maximum learning rate of 3e-5 (cosine decay schedule), based on initial experiments and recommended Lion learning rate settings presented in [16]. All MAE-ViTs were trained with stochastic depth [3], LayerScale [3], flash attention [18], parallel scaling blocks [19], QK-normalization [19], and no QK-bias [19]. Stochastic depth was set to 0.1 for ViT-S and ViT-B, and 0.3 for ViT-L. All models were initialized with random weights, as initial experiments found no benefit starting from pre-trained ImageNet weights.

A.3. Training and Inference

We scaled training based on the results of smaller models trained on smaller datasets [19, 32, 50, 69], as visualized in Figure 5 (total FLOps is based on Touvron et al. [64]). Our most computationally intensive model, ViT-L/8+ (using the loss function described in Eq. 3), was trained for over 20,000 GPU hours, learning on over 3.5 billion image crops sampled from RPI-93M.

Models were trained with data-distributed parallel (DDP) training and PyTorch 2.0 for up to 100 epochs on up to 256 NVIDIA 80GB A100 GPUs, depending on the size of the model and dataset. 256 x 256 x 6 image crops were randomly sampled from 2048 x 2048 x 6 images, augmenting with random horizontal and vertical flips. For each dataset, we use a validation set of center-cropped images from full experiments unseen during training. All image crops are preprocessed with channel-wise self-standardization [62] before being passed into the deep learning models.

Inference was performed on a large-scale distributed kubernetes T4 GPU cluster. The results in Section 5 are calculated on the gene knockout experiments of RxRx3 [24]. Each *well* in a biology experiment is loaded as a 2048 x 2048 x 6 int8 tensor. We tile over this image, obtaining 64 unique 256 x 256 x 6 crops. Each *crop* is fed-forward through the encoder, and the resultant 64 embeddings are averaged to produce a final *well-aggregated embedding*. Each genetics-only *experiment* in RxRx3 has 9 plates, and each *plate* has 1380 wells; therefore, nearly 800,000 samples need to be fed-forward through the encoder for each experiment. Given the 175 genetics-only experiments in RxRx3, this yields roughly 140 million individual samples fed-forward through each encoder in order to obtain genomic representations from the model. Note that the AdaBN-based weakly supervised models require careful mini-batch construction during both training and inference, whereas the rest of our models are deterministic in producing embeddings of individual samples.

A.4. Additional reconstructions

Additional visualizations of the reconstructed masked input images using MAE ViT-L/8+ on the JUMP-CP dataset, for both Cell Painting and Brightfield channels, are shown in Figure 7. Recall that JUMP-CP was not included in any training set, thus this data is OOD. Nevertheless, the MAE reconstruction generalizes well to this dataset, especially for the Cell Painting samples.

A.5. Additional results

Calculation of FLOps. In Figure 8 we include the scaling plots as in Figure 5, for the other three benchmark databases (CORUM, hu.MAP, and Reactome). Floating point operations (FLOps) are approximated based on the FLOp counts presented in Table 1 from Touvron et al. [64], which presents FLOps for ViT-S/B/L/16 on a 224x224x3 image. We adjust flop counts by a factor of $(\frac{16*16}{14*14})^2 = 1.69$ to account for the changed crop size, and then for 8x8 patching models we multiply by a factor of 16 to account for the 4x more tokens and the quadratic impact this has on the attention head computations. We lastly multiply the FLOps by the number of image crops seen during training for each model.

A.6. CellProfiler feature prediction

We tested the ability of two models and model architectures, RxRx1 DenseNet-161 w/ AdaBN (WSL), and RPI-93M ViT-L/8+ (MAE) to predict CellProfiler (CP) features using linear regression. Training was performed on one internal experiment representing 12 plates of 1380 wells each, for a total of 16,560 wells. Testing was performed with a different internal experiment of the same size representing 1,160 different CRISPR knock-out perturbations (with 121 control perturbations in common, equaling $<$ 10% reagent overlap between train and test experiments). 955 CP features were extracted over the categories of area-shape, intensity, neighbors, radial-distribution, and texture, and averaged to the well-level. Highly-skewed CP feature distributions were transformed by log scaling (skew $>$ 0.5) or by squaring (skew $<$ -0.5) to make them more normal then all features were centered to 0 and scaled to unit variance. 1,024-dimensional embeddings for both models were similarly averaged to the well-level, centered to 0, and scaled to unit variance. All feature predictors were trained as single-task linear regressors using scikit-learn's ElasticNetCV estimator class. A grid-search over a small range of L1/L2 ratios (0.1, 0.6, 0.9, 0.95, 0.99) and alphas (auto-determined) with a 5-fold cross-validation schedule was used. The best-fit parameters were then used to predict and score the independent experiment test set using the coefficient of determination (Fig. 6, Supp. Fig. 9, Supp. Table 7).

A.7. JUMP-CP benchmarks

The un-aggregated data for Table 4 are presented in Table 8 and Table 9.

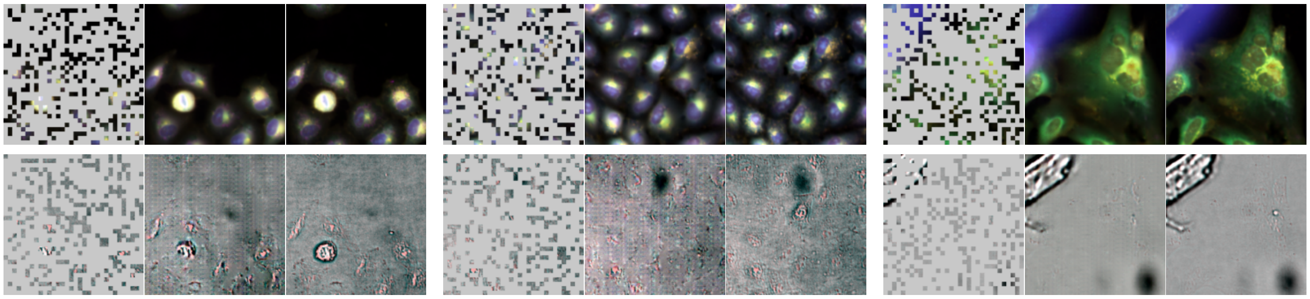

Figure 7. Visualizing MAE ViT-L/8+ (trained on RPI-93M) 75% masked reconstructions on randomly selected out-of-domain JUMP-CP [14] image crops. Rows alternate between Cell Painting and Brightfield images obtained from the same well. Note that the wells in JUMP-CP were imaged using different assays, channel composition, microscopes, and labs compared to the well images we used for pre-training.

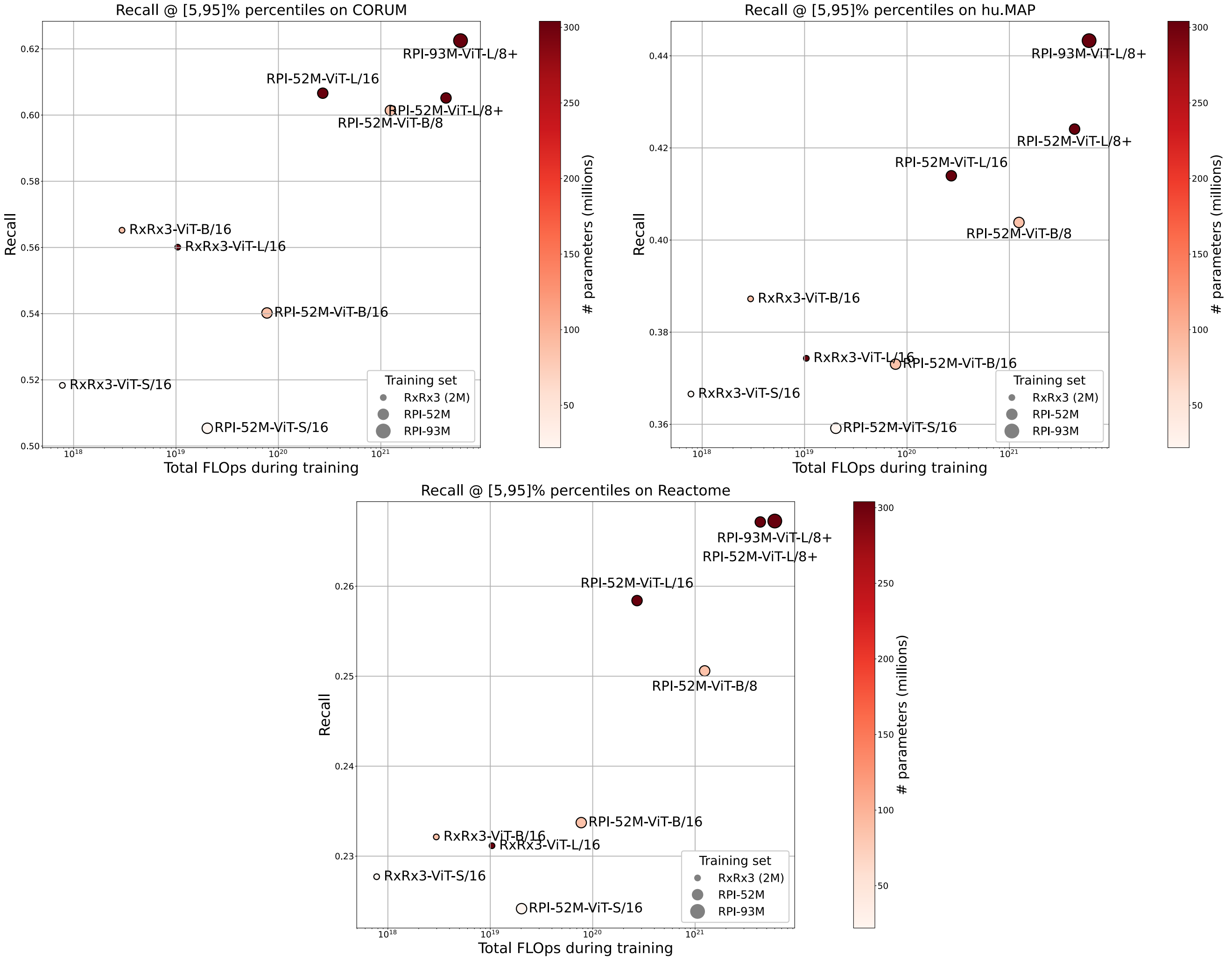

Figure 8. CORUM, hu.MAP, and Reactome recalls for ViTs as a function of training FLOps.

Table 6. Summary of results discussed in Section 5, including additional results for smaller models. Recall of known relationships in top and bottom 5% of cosine similarities by model, training set, and database (CORUM/hu.MAP/Reactome/StringDB).

Model backbone / Pretraining dataset	RxRx1 [62]	RxRx3 [24]	RPI-52M	RPI-93M
WSL				
DenseNet-161	.383/.307/.190/.330	.359/.271/.174/.319	–	–
DenseNet-161 w/ AdaBN	.485/.349/.228/.417	.461/.303/.188/.377	–	–
DenseNet-161 w/ AdaBN (1024-dim)	.502/.363/.220/.422	.520/.350/.207/.413	–	–
SSL models				
MU-net-M	–	.557/.382/.236/.432	–	–
MU-net-L	–	.566/.374/.232/.427	.576/.385/.238/.443	.581/.386/.247/.440
MAE ViT-S/16	–	.518/.367/.228/.415	.505/.359/.224/.402	–
MAE ViT-B/16	–	.565/.387/.232/.435	540/.373/.234/.416	–
MAE ViT-B/8	–	–	.601/.404/.251/.459	–
MAE ViT-L/16	–	.560/.374/.231/.427	.607/.414/.258/.460	–
MAE ViT-L/8+	–	–	.605/.424/**.267**/.474	**.622/.443/.267/.484**

Table 7. Median R^2 (± median absolute deviation) for CellProfiler predictions across feature categories.

Model Backbone	AreaShape	Intensity	Neighbors	RadialDistribution	Texture
RxRx1 DN161 w/ AdaBN (WSL)	0.401 ±0.127	0.297 ±0.121	0.583 ±0.142	0.484 ±0.127	0.413 ±0.112
RPI-93M ViT-L/8+ (MAE)	0.456 ±0.162	0.737 ±0.120	0.674 ±0.137	0.711 ±0.093	0.705 ±0.133

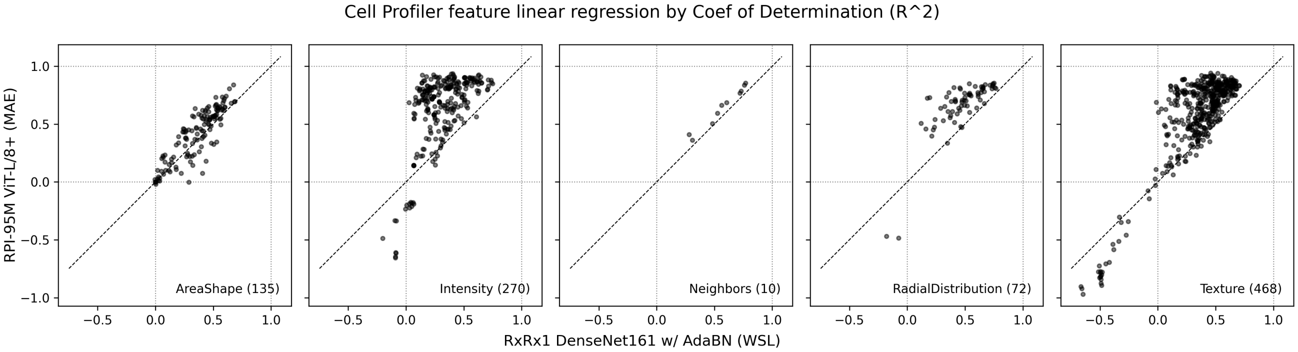

Figure 9. Single-task linear regression illustrates how an MAE-trained embedding model outperforms a WSL-trained model in predicting CellProfiler features across all categories.

Table 8. Perturbation retrieval on the JUMP-CP dataset, measured in fraction retrieved.

			Model backbone, dataset			
Cell type	Modality	Time-point	CA-93M-ViT-L	CA-93M-ViT-L-8chans	ViTL-Image-net	cellprofiler
A549	compound	long	**1.00**	0.99	0.99	0.95
		short	0.98	**0.99**	0.93	0.76
	crispr	long	0.89	**0.95**	0.90	0.68
		short	0.88	**0.97**	0.90	0.68
	orf	long	**0.84**	0.83	0.71	0.05
		short	0.63	**0.93**	0.78	0.06
U2OS	compound	long	0.98	**0.99**	0.94	0.66
		short	0.88	**0.97**	0.88	0.78
	crispr	long	0.91	**0.96**	0.94	0.46
		short	0.91	**0.98**	0.94	0.67
	orf	long	0.65	**0.89**	0.75	0.20
		short	0.79	0.89	**0.90**	0.37

Table 9. Siblings retrieval on the JUMP-CP dataset, measured in fraction retrieved. Note that ORF's do not have siblings.

			Model backbone, dataset			
Cell type	Modality	Time-point	CA-93M-ViT-L	CA-93M-ViT-L-8chans	ViTL-Image-net	cellprofiler
A549	compound	long	0.05	0.04	0.13	**0.17**
		short	0.13	0.04	0.08	**0.14**
	crispr	long	0.06	0.01	0.07	**0.12**
		short	0.04	0.01	0.04	**0.11**
U2OS	compound	long	0.12	0.00	0.03	**0.25**
		short	0.06	0.02	**0.05**	0.04
	crispr	long	0.03	0.02	0.03	**0.18**
		short	0.03	0.02	0.02	**0.07**