Masked Autoencoders are Scalable Learners of Cellular Morphology

Oren Kraus* **Kian Kenyon-Dean*** **Saber Saberian** **Maryam Fallah** **Peter McLean**

Jess Leung **Vasudev Sharma** **Ayla Khan** **Jia Balakrishnan** **Safiye Celik**

Maciej Sypetkowski **Chi Vicky Cheng** **Kristen Morse** **Maureen Makes**

Ben Mabey **Berton Earnshaw**

Abstract

Inferring biological relationships from cellular phenotypes in high-content microscopy screens provides significant opportunity and challenge in biological research. Prior results have shown that deep vision models can capture biological signal better than hand-crafted features. This work explores how self-supervised deep learning approaches scale when training larger models on larger microscopy datasets. Our results show that both CNN- and ViT-based masked autoencoders significantly outperform weakly supervised baselines. At the high-end of our scale, a ViT-L/8 trained on over 3.5-billion unique crops sampled from 93-million microscopy images achieves relative improvements as high as 28% over our best weakly supervised baseline at inferring known biological relationships curated from public databases. 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 complex cellular phenotypes and relating them across genetic and chemical perturbations [41, 53]. Image-based profiling has proven to be a powerful approach for exploring cellular phenotypes induced by genetic and chemical perturbations [3]. These experiments use *high content screening* (HCS) systems combining automated microscopy with high throughput technologies to assay perturbations on a massive scale. Recent public releases of HCS image sets, like RxRx3 [19] and JUMP-CP [9], consist of millions of cellular images across 100,000s of unique perturbations and demonstrate the scalability of this approach.

HCS image sets are often analyzed with customized cell segmentation, feature extraction, and downstream analysis pipelines [4]. Despite the many discoveries made using this approach [3], developing robust segmentation and feature extraction pipelines using open-source software packages [6, 47] remains challenging [8]. Alternatively, representation learning approaches do not require prior knowledge of cellular morphology and perform significantly better on practical biological research objectives, e.g. inferring relationships between perturbations [7]. In contrast to previous approaches employing weakly supervised pretraining [37], in this work we train masked autoencoders (MAEs) [24] on progressively larger HCS image sets and show that these models are scalable learners of

*Equal contribution.

All authors contributed to this article during their employment with Recursion. For correspondence, email: `oren.kraus@recursionpharma.com`, `berton.earnshaw@recursionpharma.com`, or `info@rxrx.ai`.

NeurIPS 2023 Generative AI and Biology Workshop.

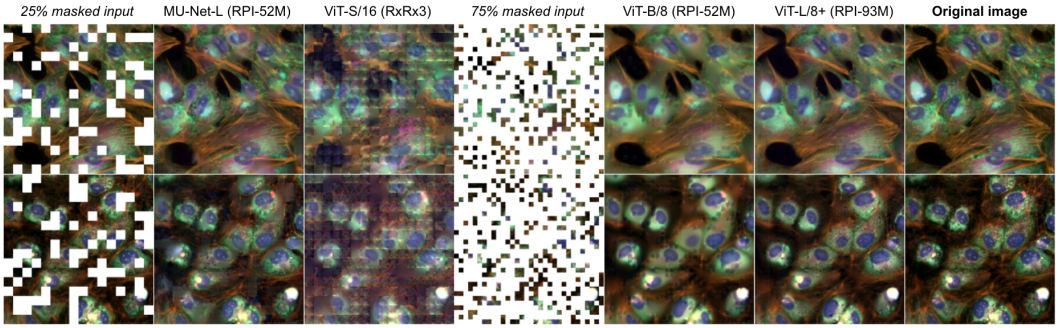

Figure 1: Visualizing reconstructions from masked random *validation* images for different MAEs.

cellular morphology, outperforming previous state-of-the-art methods at inferring known biological relationships in whole-genome HCS screens.

2 Related Work

Supervised learning on HCS image sets. Deep learning models have been successfully trained to perform cell segmentation [52, 36, 48] and phenotype classification [31, 32, 39, 18], 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 [42], researchers used models trained on natural images to to featurize HCS data with varying results [1, 40]. Others [37, 49, 44] have trained convolutional networks to classify labels obtained from experimental metadata (e.g., perturbation class), a technique called *weakly supervised learning* (WSL) [57]. 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.

Self-supervised learning. Vision models pretrained with self-supervised learning (SSL) often outperform supervised models on downstream tasks [24, 5, 10]. Unlike supervised pretraining [30], 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 [5] is an SSL method that has been applied to HCS [12, 23, 45, 29, 15] data, however it relies on augmentations inspired by natural images, which may not be applicable to HCS image sets. Alternatively, masked autoencoders [24] are trained by reconstructing masked patches from unmasked patches of an image (Fig. 1). MAEs have been successfully applied to images [24], audio [28], video [20] and multimodal audio-video datasets [27]. However, previous attempts to train MAEs on HCS datasets have had limited success [55, 29], due in part to limitations in compute resources and dataset size. The present work shows that MAE training scales with both model and training set size.

3 Methods

Datasets. We investigate the scaling properties [56] of cellular image sets by evaluating models trained on the following four microscopy datasets. **RxRx1** [49] 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. **RxRx3** [19] is a publicly-available proprietary Cell Painting dataset with over 2.2 million images of HUVEC cells under 17,063 CRISPR knockouts (over 6 guides) or 1,674 compounds across 180 experimental batches. **RPI-52M** and **RPI-93M** (Recursion Phenomics Imageset) are private datasets with 52 million and 93 million proprietary Cell Painting and Brightfield images, respectively. To our knowledge, these are the largest HCS datasets collected for model training purposes. All evaluations are performed on RxRx3, which is the largest publicly available whole-genome HCS image set.

Weakly supervised learning. As a baseline, we employ the 28-million parameter DenseNet-161 backbone implemented in [49], trained to predict cellular perturbations and producing 128-dimensional embeddings, with and without adaptive batch normalization (AdaBN) [34].

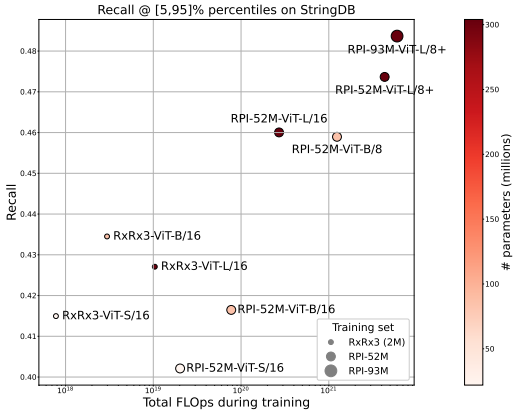

Figure 2: StringDB recall as a function of training FLOps. Similar results hold for other databases.

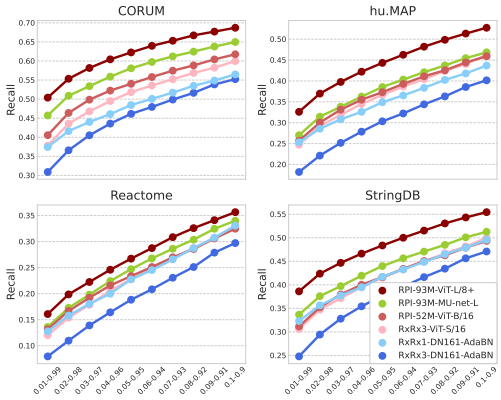

Figure 3: Recall across different cosine similarity percentiles for each database.

U-Nets. We adapt U-Nets [43] for masked autoencoding (MU-Nets) by training to reconstruct masked sections of input images. We train MU-Nets as described in Xun et al. [55] and report results for MU-Net-M and MU-Net-L, which have 52- and 135-million parameters, respectively. MU-Net-M's downsampling scale is set to 32/64/128/256/512. MU-Net-L incorporates an additional scale of 1024. In each case, the decoder mirrors the encoder's scale configuration. After an initial hyperparameter search (see A.1.2), we trained both models with a mask ratio of 25% and kernel size of 5.

Vision transformers. We train vision transformers (ViTs) [16, 46, 14, 56] as MAEs following the implementation in He et al. [24]. We report results for ViT-S, ViT-B, and ViT-L encoders [16], containing 22-, 86-, and 304-million parameters, respectively, and producing 384-, 768-, and 1024-dimensional embeddings, respectively. We explore the use of 8x8 and 16x16 patch sizes and 75% and 25% mask ratios (Fig. 1). A 25-million parameter decoder [24] is used for patch reconstructions. Note that 8x8 patches induce a sequence length 4 times greater than 16x16 patches and is thus more computationally expensive.

Training. Models were trained on Recursion's HPC cluster, BioHive-1, for up to 100 epochs on as many as 128 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.

Scaling to ViT-L/8+. We scale training based on the results of smaller models trained on smaller datasets [14, 25, 38, 56], as visualized in Figure 2 (total FLOps is based on Touvron et al. [51]). Our largest model, ViT-L/8+, was trained for over 20,000 GPU hours, learning from over 3.5 billion image crops sampled from RPI-93M. Inspired by [54], we added a term to the loss function to prevent divergence and improve texture reconstruction.

Inference. The metrics of Section 4 are calculated on the gene knockout experiments of RxRx3 [19], requiring the embedding of ~140 million image crops for each encoder. See A.2 for details.

4 Results

An important use of HCS data is the accurate inference of biological relationships amongst genetic and chemical perturbations. We evaluate each model's ability to capture known relationships using the multivariate metrics described in Celik et al. [7]. Briefly, each model's embeddings are first aligned across experimental batches using TVN (typical variation normalization) [1], fitted to the negative experimental controls across all batches. Following TVN, we correct for possible chromosome arm biases known to exist in CRISPR-Cas9 HCS data [33]. We compute the 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 a relationship metric, setting the origin of the space to the mean of negative experimental controls. We compare these similarities with the annotated

Table 1: Recall of known relationships in top and bottom 5% of cosine similarities by model backbone and training set, with results for each database (CORUM/hu.MAP/Reactome/StringDB). DenseNet-161 backbones are trained via WSL, all others via SSL. See Fig. 3 for recall at other percentiles.

Model backbone	RxRx1 [49]	RxRx3 [19]	RPI-52M	RPI-93M
DenseNet-161	.38/.31/.19/.33	.36/.27/.17/.32	–	–
DenseNet-161 w/ AdaBN	.48/.35/.23/.42	.46/.30/.19/.38	–	–
MU-Net-M	–	.56/.38/.23/.42	–	–
MU-Net-L	–	.57/.37/.23/.43	.58/.39/.24/.44	.58/.39/.25/.44
MAE ViT-S/16	–	.52/.37/.23/.41	.51/.36/.22/.40	–
MAE ViT-B/16	–	.57/.39/.23/.43	.54/.37/.23/.42	–
MAE ViT-B/8	–	–	.60/.40/.25/.46	–
MAE ViT-L/16	–	.56/.37/.23/.43	.61/.41/.26/.46	–
MAE ViT-L/8+	–	–	.61/.42/**.27**/.47	**.62/.44/.27/.48**

Table 2: Recall (at 5% false positive rate) of StringDB relationships for select models on three different gene sets defined in Sivanandan et al. [45].

Training dataset	**Model backbone**	PoC-124	MoA-300	DG-1640
RxRx1 [49]	WSL DenseNet-161 w/ AdaBN	.79	**.24**	.15
RxRx3 [19]	MAE ViT-S/16	.74	.19	.14
RPI-52M	MU-Net-L	.79	.20	.15
RPI-93M	MAE ViT-L/8+	**.80**	.23	**.17**
CP-1640 [45]	DiNO ViT-S/8	.53	.12	.14

relationships found in the following public databases: CORUM [22], hu.MAP [17], Reactome [21], and StringDB [50] (with >95% combined score).

Table 1 reports the recall of known relationships amongst the top and bottom 5% of all cosine similarities between CRISPR knockout representations in RxRx3. Note how both recall and image reconstruction (see Fig. 1) improve with larger models, larger training sets, smaller patches, and larger mask ratio. In Figure 2 we see that recall strongly correlates with training FLOps, a function of both model and training set size (see A.3 for similar results trends on other databases). Figure 3 shows similar trends in recall for other similarity percentiles. In contrast, the performance of re-implemented WSL baselines [49] decreases when the dataset is scaled from RxRx1 to RxRx3, which could be due to the chromosome arm bias present in CRISPR-Cas9 systems [33] or other factors such as the increased size of the label set.

We compare these models with recent results from an alternative HCS platform combining pooled CRISPR screening with Cell Painting [45]. Table 2 reports recall at 5% FPR in StringDB on three gene sets defined in Sivanandan et al. [45]. The ViT-L/8+ MAE trained on RPI-93M yields a minimum 20% relative improvement in gene set performance over CP-DiNO 1640 (a ViT-S/8), which was trained on ~1.5 million 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 Conclusion

This work demonstrates that scaling properties [56] apply to learning representations of cellular morphology that can accurately infer known biological relationships. Unlike previous approaches that use weakly supervised learning [37, 49] on small, curated datasets, we showed that the performance of MAEs on biologically meaningful benchmarks scales to massive HCS image sets. In future work, we will continue to scale model and training set sizes even further. We will also explore new applications of this technology beyond predicting biological relationships, with the ultimate goal of creating general-purpose foundation models of cellular biology.

Acknowledgements

This work reflects the combined efforts of many current and former Recursion employees. Special thanks to the Recursion lab team for design and execution of the HCS experiments which fueled our datasets. Additional thanks to the Recursion HPC team for their dedicated support in keeping our cluster, BioHive-1, running effectively. We would especially like to thank the following individuals for their contributions toward this work: Dominique Beaini, Jordan Christensen, Joshua Fryer, Brent Gawryluik, Imran Haque, Jason Hartford, Alex Timofeyev, and John Urbanik.

References

[1] D. Michael Ando, Cory Y. McLean, and Marc Berndl. Improving Phenotypic Measurements in High-Content Imaging Screens. *bioRxiv*, page 161422, 2017. doi: 10.1101/161422.

[2] 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. doi: 10.48550/arxiv.2304.12210.

[3] Michael Boutros, Florian Heigwer, and Christina Laufer. Microscopy-Based High-Content Screening. *Cell*, 163(6):1314–1325, 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.11.007.

[4] 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. ISSN 1548-7091. doi: 10.1038/nmeth.4397.

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

[6] 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. ISSN 1465-6906. doi: 10.1186/gb-2006-7-10-r100.

[7] 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.

[8] 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. ISSN 1474-1776. doi: 10.1038/s41573-020-00117-w.

[9] 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.

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

[11] 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.

[12] 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. doi: 10.48550/arxiv.2209.07819.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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. ISSN 1744-4292. doi: 10.15252/msb.20167490.

[18] 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. doi: 10.1038/s41467-017-00623-3.

[19] 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.

[20] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked Autoencoders As Spatiotemporal Learners. *arXiv*, 2022. doi: 10.48550/arxiv.2205.09113.

[21] 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. ISSN 0305-1048. doi: 10.1093/nar/gkab1028.

[22] 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. ISSN 0305-1048. doi: 10.1093/nar/gky973.

[23] Johan Fredin Haslum, Christos Matsoukas, Karl-Johan Leuchowius, Erik Müllers, and Kevin Smith. Metadata-guided Consistency Learning for High Content Images. *arXiv*, 2022. doi: 10.48550/arxiv.2212.11595.

[24] 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.

[25] 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.

[26] 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.

[27] 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. doi: 10.48550/arxiv.2212.08071.

[28] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked Autoencoders that Listen. *arXiv*, 2022. doi: 10.48550/arxiv.2207.06405.

[29] 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.

[30] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big Transfer (BiT): General Visual Representation Learning. *arXiv*, 2019. doi: 10.48550/arxiv.1912.11370.

[31] 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. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw252.

[32] 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. ISSN 1744-4292. doi: 10.15252/msb.20177551.

[33] 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.

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

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

[36] 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. ISSN 1548-7091. doi: 10.1038/s41592-019-0403-1.

[37] 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. doi: 10.1101/2022.08.12.503783.

[38] OpenAI. Gpt-4 technical report, 2023.

[39] 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. ISSN 1548-7091. doi: 10.1038/s41592-019-0658-6.

[40] 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. doi: 10.1101/085118.

[41] Laralynne Przybyla and Luke A. Gilbert. A new era in functional genomics screens. *Nature Reviews Genetics*, 23(2):89–103, 2022. ISSN 1471-0056. doi: 10.1038/s41576-021-00409-w.

[42] 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. doi: 10.1109/cvprw.2014.131.

[43] 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.

[44] 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.

[45] 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. doi: 10.1101/2023.08.13.553051.

[46] 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.

[47] 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. doi: 10.1186/s12859-021-04344-9.

[48] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106, 2021. ISSN 1548-7091. doi: 10.1038/s41592-020-01018-x.

[49] 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.

[50] 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. ISSN 0305-1048. doi: 10.1093/nar/gkaa1074.

[51] 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.

[52] 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. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1005177.

[53] 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. ISSN 1474-1776. doi: 10.1038/s41573-022-00472-w.

[54] 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.

[55] 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.

[56] 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.

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

A Appendix

A.1 Model hyperparameters

RxRx1 models were trained for 100 epochs, RxRx3 models for 50 epochs, and RPI-52M / 93M models were trained 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 [49] for more details on batch construction for AdaBN models.

Models were trained on Nvidia A100-80GB GPUs with data-distributed parallel (DDP) training and PyTorch 2.0. Each model was trained on 16 to 128 GPUs, depending on the size of the model and dataset. We accelerated training speed by training with large batch sizes, as described below.

A.1.1 Weakly supervised learning

Our weakly supervised CNN models were trained as standard classifiers to predict the perturbation applied in the cellular image, given a random crop as input. We use the DenseNet-161 backbone and a neck that outputs 128-dimensional embeddings for each sample, implemented exactly following the model architecture presented in [49]. We trained with a crop size of 256 x 256 x 6 and batch size of 4,096. We found that training with a specially-tuned SGD optimizer yielded the best performance for these models versus other optimizers. Furthermore, we observed that issues such as chromosome arm bias [33] become significantly more pronounced for WSL classifiers trained on large datasets like RxRx3.

Adaptive batchnorm (AdaBN) is an architectural technique to enable domain adaptation [34]. Our AdaBN-based DenseNet-161 classifiers are implemented with Ghost BatchNorm [26] in order to train with larger batch sizes. Such models can perform effective domain adaptation, but they require a specialized sampler during both training and inference time to ensure that each batch is constructed from the same plate in the experiment.

A.1.2 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 [35] 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 1, for these values. Changing the mask ratio or kernel size did not seem to effect the performance.

A.1.3 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 [11] 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. [24]). 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. [2]. 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 [11]. All MAE-ViTs were trained with stochastic depth [2], LayerScale [2], flash attention [13], parallel scaling blocks [14], QK-normalization [14], and no QK-bias [14]. 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.2 Inference

Inference runs on a large-scale distributed kubernetes T4 GPU cluster. Each *well* in a biology experiment is imaged as a 2048 x 2048 x 6 int8 tensor. We tile over this image, obtaining 64 unique

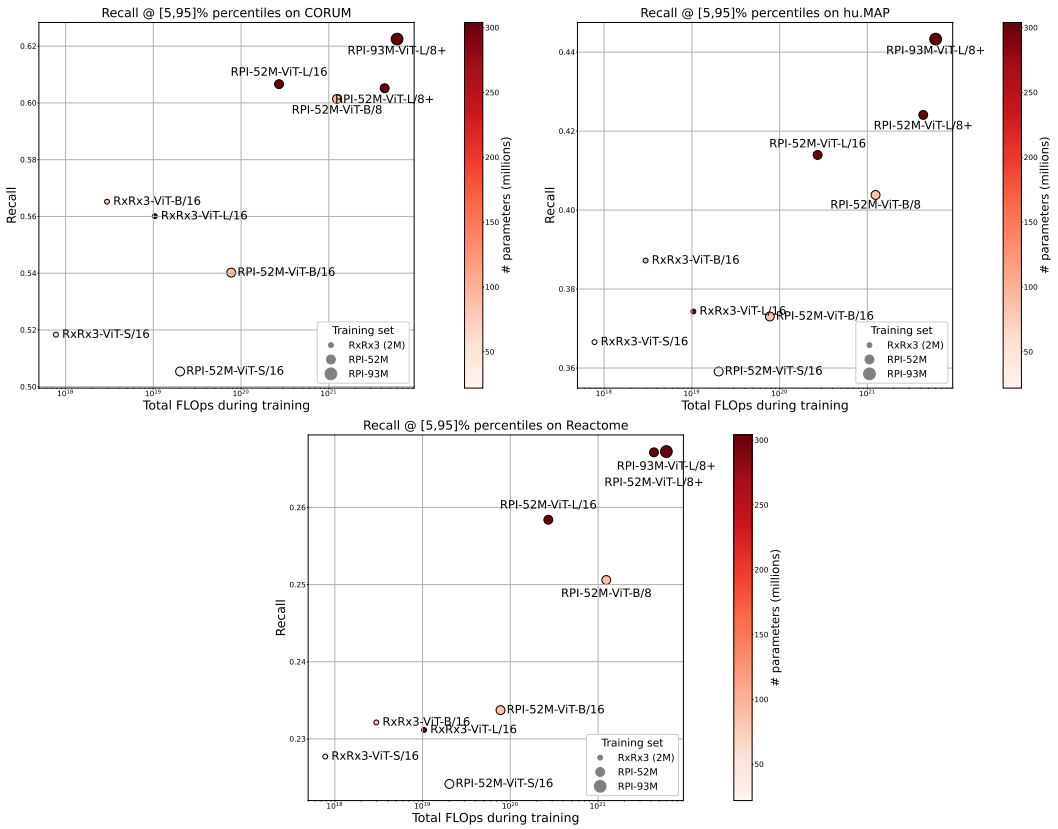

Figure 4: CORUM, hu.MAP, and Reactome recalls for ViTs as a function of training FLOps.

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 needing to be fed-forward through each encoder (64 crops per well x 1380 wells per plate x 9 plates per experiment x 175 experiments) in order to obtain whole-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.3 Additional results

Calculation of FLOps. In Figure 4 we include the scaling plots as in Figure 2, 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. [51], 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.

Recall of other baseline models. We computed the recall metrics reported in Table 1 for a number of rudimentary baselines. For example, we we generated random embeddings with different dimensions (128, 384, 512, 768, 1024), as well as randomly shuffled embeddings generated by the models illustrated in Figure 3. All of these random baselines recalled ~10% of the known relationships in each database, consistent with the fact that the metric considers only 10% of the total cosine similarities (top and bottom 5%). We also constructed a baseline with a simple 30-dimensional feature set built from pixel intensity statistics of each image and applied a TVN transformation to them. The recall

obtained for these features, .28/.26/.16/.27 for CORUM/hu.MAP/Reactome/StringDB databases, was better than random but significantly worse than all models considered in Table 1.