ViTally Consistent: Scaling Biological Representation Learning for Cell Microscopy

Kian Kenyon-Dean [1] **Zitong Jerry Wang** [1] **John Urbanik** [1] **Konstantin Donhauser** [2] **Jason Hartford** [2 3] **Saber Saberian** [1] **Nil Sahin** [1] **Ihab Bendidi** [2] **Safiye Celik** [1] **Juan Sebastián Rodríguez Vera** [1] **Marta Fay** [1] **Imran S Haque** [1] **Oren Kraus** [1]

Abstract

Deriving insights from experimentally generated datasets requires methods that can account for random and systematic measurement errors and remove them in order to accurately represent the underlying effects of the conditions being tested. Here we present a framework for pretraining on large-scale microscopy datasets that includes three steps: (1) curating a set of diverse and self-consistent training samples, (2) scaling training of an appropriate foundation model architecture on this dataset, (3) evaluating intermediate layers of the trained model to identify the best representation for downstream tasks. Using this strategy, we present the largest foundation model for cell microscopy data to our knowledge, a new 1.9 billion-parameter ViT-G/8 MAE trained on over 8 billion microscopy image crops. Compared to a previous published ViT-L/8 MAE, our new model achieves a 60% improvement in linear separability of genetic perturbations and obtains the best overall performance on whole-genome relationship recall, batch correction replicate consistency, and compound-gene activity prediction benchmarks.[1]

1. Introduction

Microscopy images capture detailed information about a cell's responses to perturbations: indeed many early biological discoveries were made by biologists looking through microscopes studying how cells behave (Hooke, 1665). But with the advent of high throughput screening platforms that capture cells' responses to tens or hundreds of thousands of perturbations, we can no longer rely on people to study these images. Instead, we need representations of the cell images that capture the biologically meaningful information and capture the similarities and differences in cells' response to perturbations.

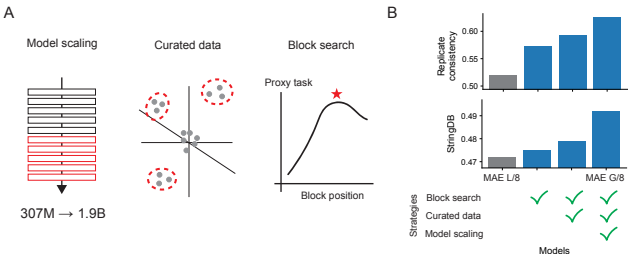


Figure 1: (A) Overview of performance gain from different MAE pretraining and inference strategies. (B) Example whole-genome results for **replicate consistency** and **biological relationship recall** on StringDB for models trained with different combinations of strategies, by model name and dataset (left to right):

State-of-the-art (SOTA) deep learning methods for learning representations of microscopy leverage Vision Transformers (ViT) (Dosovitskiy et al., 2020) trained with self-supervised learning (SSL) techniques (Balestriero et al., 2023) from large-scale screens (Doron et al., 2023; Kim et al., 2023; Bourriez et al., 2024). They have been shown to be very effective at representing the subtle changes in cell morphology in response to perturbations, and masked autoencoders (He et al., 2022) in particular exhibit scaling laws for recovering known biological relationships (Kraus et al., 2024). However, given that this data has to be collected experimentally, there is a limit to how far we can scale these approaches before one runs out of data. To continue to improve these representations, we need to use experimental data as efficiently as possible. In this paper, we show how curated training datasets combined with care-

[1]Recursion [2]Valence Labs [3]University of Manchester. Correspondence to: Kian Kenyon-Dean <kian.kd@recursion.com>, Oren Kraus <oren.kraus@recursion.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

[1]An earlier version of this work appeared at the NeurIPS 2024 Foundation Models for Science Workshop (Kenyon-Dean et al., 2024).

ful probing of model representations allows us to scale to larger models that significantly improve recall and separation of biological phenotypes.

Careful curation of data has enabled better data efficiency in language models (Llama3, 2024; Eldan & Li, 2023; Javaheripi et al., 2023). However, unlike language, we cannot simply rely on prior knowledge about the data source to weigh data (e.g. by up-weighting trusted sources like arXiv or wikipedia). Instead, we curate a training dataset that is relatively uniform with respect to the observed phenotypes. Because many perturbations will not induce a morphological change, a random sample of images across all perturbations will be dominated by cells that look like unperturbed cells. We show that one can use weaker models to filter images to perturbations that induce a distinct morphological phenotype, thereby reducing the absolute training set size but increasing the relative diversity. The resulting dataset, *Phenoprints-16M*, has $5\times$ fewer images than the dataset used to train Kraus et al.'s masked autoencoder, but we find that it leads to better performance when combined with scaled compute through longer self-supervised training.

Data curation gives one way of improving a model, but we can also optimize over the layer that we use to construct a representation. Traditionally, the finally hidden layer has been used (Doron et al., 2023; Kim et al., 2023; Kraus et al., 2024), but there is some evidence from mechanistic interpretability (Templeton et al., 2024) that earlier layers may learn more abstract features, which may provide more useful features for biological tasks. We show that using a linear probing proxy task, we are able to cheaply find the best performing intermediate block for many models. We find that using intermediate layers leads to better performance on downstream whole-genome relationship recall and perturbational replicate consistency at a lower computational inference cost, and that these results are consistent across a variety SSL ViTs trained on either microscopy or natural images. Our proxy task involves fitting a set of biological linear probing tasks to evaluate representations learned by intermediate ViTs blocks for microscopy data (§ 4). We find this task correlates extremely well with biological recall benchmarks which allows us to reliably probe the model's representations.

In summary, we present a series of data curation and model probing techniques that allow us to more effectively use experimental microscopy data. By combining these insights, we are able to train a **new foundation model, MAE-G/8**, a 1.9 billion parameter ViT-G/8 MAE trained on Phenoprints-16M for 48,000 H100 GPU hours on more than 8 billion samples drawn from the curated dataset (Figure 1A, § 3.2) resulting in significant improvements across a range of challenging biological benchmarks. These results indicate that the scaling properties first identified by Kraus et al. (2023) extend to the multi-billion parameter model regime across a wide variety of newly examined biologically-motivated benchmark tasks.

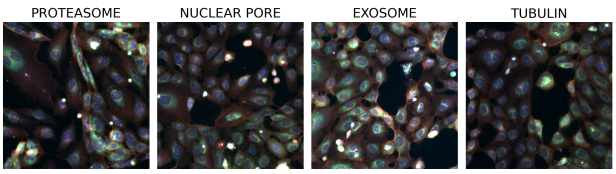


Figure 2: Samples for subset of groups in Anax 40-class functional gene group classification task.

2. Related work

Dataset Curation for Foundation Models. Dataset curation is crucial for enhancing the efficiency of foundation models, especially in large-scale contexts. Usual approaches to dataset construction are inspired by the image retrieval community (Weinzaepfel et al., 2022; Radenović et al., 2018; Berman et al., 2019). Existing methods often utilize pre-trained models for filtering and pruning, such as vision-language models to discard irrelevant pairs (Schuhmann et al., 2021), semantic deduplication to remove redundancy (Abbas et al., 2023), and prototypicality-based approaches to retain representative data (Sorscher et al., 2022). However, these techniques are less effective for HCS, where redundancy, variability, and subtle morphological differences make conventional filtering challenging. Our work addresses these limitations by building on Celik et al. (2024)'s *perturbation consistency* framework to curate a balanced dataset of images across semantic classes, which is vital for effective learning under the masked objectives (Zhang et al., 2022).

Layer-wise Analysis of Deep Neural networks. Recent work suggests that intermediate layers (or, blocks) in large ViTs may achieve superior performance on certain linear probing tasks compared to the final encoder layer (Evci et al., 2022; Dehghani et al., 2023). Alkin et al. (2024) reported that intermediate layers in large MAE-ViTs (ViT-L, ViT-H) have superior ImageNet-1K k-NN accuracy, likely because later encoder layers become more optimized for the reconstruction task.

Evaluating Representations for Drug Discovery. Evaluating the quality of biological representation learning methods for drug discovery remains challenging, as ground truth data is sparse, noisy, biased to well-studied diseases and pathways, and poorly annotated. Metrics have been proposed that use mean average precision (Kalinin et al., 2024) or AUC ROC (Sivanandan et al., 2023) to assesses how similar related samples are represented, in-

cluding replicates of the same perturbation or different perturbations with similar annotated biological activities. Recently, Celik et al. (2024) introduced terminology for describing perturbative "maps of biology", in which representations of perturbations in HCS data can be placed in unified, relatable embedding spaces allowing for the generation of genome-scale sets of pairwise comparisons. Here we leverage the *biological relationship recall* benchmark proposed by Celik et al. (2024), which assess how well known relationships between pairs of perturbations are recalled among the most similar or dissimilar embeddings. Computing reliable versions of these relationship benchmarks with HCS data is particularly expensive as they require genome-wide embeddings to be inferred for hundreds of millions of image crops from the genome-wide RxRx3 microscopy screen (Fay et al., 2023).

3. Vision Transformers for Microscopy Images

We train and evaluate various vision transformers (ViTs, Table 4) as encoders to extract feature embeddings from $256\times256\times6$ (HxWxC) microscopy image crops (Figure 2).

3.1. Training Dataset Curation

Many academic and industry labs have adopted the Cell Painting imaging protocol (Bray et al., 2016), which multiplexes fluorescent dyes to reveal eight broadly relevant cellular components. The datasets used here contain a six-channel implementation of Cell Painting (Figure 2), as well as brightfield images, spanning 100,000s of chemical and genetic perturbations applied to dozens of cell types (Kraus et al., 2024). In these datasets, cells that look like unperturbed cells tend to be very over-represented because many perturbations do no induce a morphological change. Some morphological changes are also far more common (e.g. many perturbations will kill cells, resulting in a relatively high proportion of dead cell morphological phenotype). This results in significant imbalance in the morphological phenotypes that the models learn to reconstruct.

To address this, we constructed an aggressively curated training dataset (§ A.1). To learn an initial representation, we began by reproducing the MAE-L/8 model of Kraus et al. (2024) on a dataset of similar size consisting of 93 million HCS images. Using this representation, we first filtered perturbations that did not induce consistent morphological changes to cells. To perform this filtering, we utilized Celik et al. (2024)'s non-parametric perturbation consistency test (§ A.3) after correcting for batch effects using Typical Variation Normalization (Ando et al., 2017; Kraus et al., 2024). This test was applied within each experiment for computational efficiency, and we restricted the analysis to wells containing single perturbations. This consistency was computed for CRISPR guides, siRNAs, and particular concentrations of small molecules across replicates of the same perturbation. P-values were computed for each gene and each (perturbation, concentration) pair. When multiple experiments existed for the same condition, we combined p-values using the Cauchy Combination test (Liu & Xie, 2018).

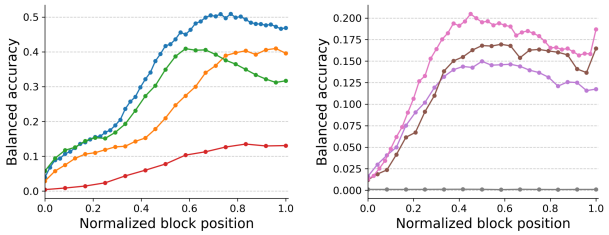
We repeated this procedure with a weakly supervised learning (WSL) model trained on RxRx1 (Sypetkowski et al., 2023) and filtered to perturbations where any condition had a p-value < 0.01 in either the MAE-L/8 or WSL model. This process reduced our original dataset of 93M samples to 16M, which we refer to as Phenoprints-16M. While some redundancy remains when distinct perturbations have the same effect, the proportion of samples with that differ from negative controls increased substantially with little decrease in overall diversity. We believe that iteratively repeating this process with the best models from previous iterations to guide data selection for subsequent models may be a viable strategy.

3.2. Models

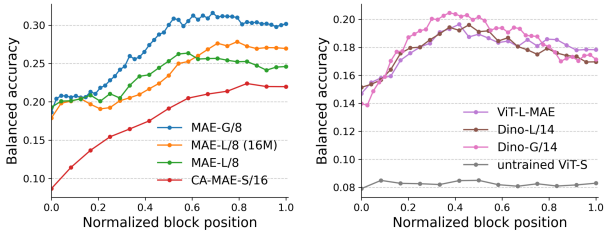
Baselines. We compare to several non-finetuned baseline ViT image encoders: three different Dino-v2 backbones (Oquab et al., 2024) (with 4 register tokens (Darcet et al., 2024)) trained on a curated non-biological natural image dataset; a MAE ViT-L/16 trained on Imagenet-21k (He et al., 2022); and an untrained ViT-S/16. We found that channel-wise self-standardization worked best as the image normalization preprocessing for these baselines, and that the class token was slightly better than the global pool of the patch tokens (except for MAE). Convolutional weights in the patch embedding layer were repeated to embed 6 channel images when using models trained on RGB datasets (Wightman, 2019).

Prior work. Our primary point of comparison is with respect to the best pretrained foundation model presented by Kraus et al. (2024), the MAE-ViT-L/8+ trained on RPI-93M. This MAE-L/8 was trained for approximately 40 epochs, learning from over 3.5 billion image crops, using the L2 mean squared error loss function plus an additional Fourier domain reconstruction loss term.

CA-MAE-S/16 trained on RxRx3. We trained a new channel-agnostic MAE (Kraus et al., 2024) ViT-S/16 on the RxRx3 dataset (Fay et al., 2023) for 100 epochs. Channel-agnostic ViTs tokenize each image channel separately with shared patch embedding weights and leverage the dynamic sequence length of transformers with repeated positional encodings to train ViTs that can process images with varying numbers of channels (Bao et al., 2024; Bourriez et al., 2024; Kraus et al., 2024). Kraus et al. (2024) demonstrate



(a) RxRx1 siRNA knockdown classification.



(b) Anax functional gene group classification.

Figure 3: Block-wise validation set **linear probe results** comparing ViT models pretrained on cell microscopy images (left) versus natural images (right); note the difference in y-axes. (a) 1139-class RxRx1 SiRNA knockdown classification (Sypetkowski et al., 2023); (b) 40-class Anax functional gene group classification on HUVEC cell images from RxRx3 CRISPR knockouts (Fay et al., 2023).

that the large MAEs with 8x8 patch size perform either better or the same as the 16x16 channel-agnostic variants for consistently 6-channel data, so we opted to train standard MAEs for the following two new models since they require fewer tokens at inference time.

MAE-L/8 trained on Phenoprints-16M. Holding the model backbone constant compared to the MAE-ViT-L/8 by (Kraus et al., 2024), we assess the impact of our curated dataset in contrast to the 93M dataset by training a new ViT-L/8 MAE for 500 epochs on Phenoprints-16M.

MAE-G/8 trained on Phenoprints-16M. Holding the dataset constant compared to MAE-L/8 above, we assess the impact of increased model scale in terms of parameters by training a new ViT-Gigantic MAE with nearly 1.9 billion parameters for 500 epochs on Phenoprints-16M. Training this model required 256 H100 GPUs running in parallel for over 1 week. See § A.2 for other hyperparameter settings we used for model training.

4. Linear probing representation learning across ViT blocks

We improve the quality of our learned image representations by leveraging previous findings that suggest intermediate blocks within an encoder can provide better representation compared to the final block (Alkin et al., 2024). Unfortunately, it is infeasible to search for the best block by simply performing whole-genome evaluation on each block of a large model because the evaluation is extremely time-consuming and resource intensive. For example, evaluating the final block of MAE-G/8 required 4,000 L4 GPU hours just for inference (§ 5). We demonstrate that using block-wise linear probes provides insights into the quality of biological features extracted by these models in their intermediate blocks, allowing us to trim the model to an earlier block to both reduce inference costs and improve representation quality.

Our block-wise search consists of training a logistic regression model (linear probe) on the output features of each transformer block to predict either the gene that was perturbed or the functional group that the gene belongs to, and test performance on held-out experiments (§ A.4). We define the optimal block b^* for a probing task as the block whose output features achieve the highest test balanced accuracy when trained on the probing task, across all N blocks of the encoder,

$$b^* = \underset{b \in \{1,2,\ldots,N\}}{\arg\max} \ \text{BalancedAccuracy}(\mathbf{z}^{(b)}), \qquad (1)$$

where $\mathbf{z}^{(b)}$ are output features from block b of a ViT. Performance on our linear probing tasks can be viewed as a measure of linear separability of a feature space across experimental batches.

RxRx1 1139-class siRNA genetic perturbation classification. We expect high quality representations of cell images to generate similar embeddings for cells with the same perturbation, hence a simple linear probe should be able to predict gene perturbation from these representation reasonably well. We train linear probes on the publicly-available RxRx1 dataset (Sypetkowski et al., 2023) which consists of 125,510 high-resolution fluorescence microscopy images of human cells under 1,138 siRNA-induced gene knockdowns (plus unperturbed controls) across four cell types (HepG2, HUVEC, U2OS, RPE). These gene knockdowns produce strong phenotypes which makes the prediction task more feasible.

We found that, for MAE-G/8 , the best features came from intermediate block $b^* = 38$ (out of 48) of the encoder, achieving a balanced accuracy (0.51) that is 8.5% greater compared to its final block's output features (Figure 3a, left). Additionally, these features achieved 60% greater accuracy than the typically used final block of MAE-L/8+ (Kraus et al., 2024). We observed similar trends for ViT models pretrained on natural images. For example, DINO-G/14 and ViT-L/16 MAE trained on non-biological

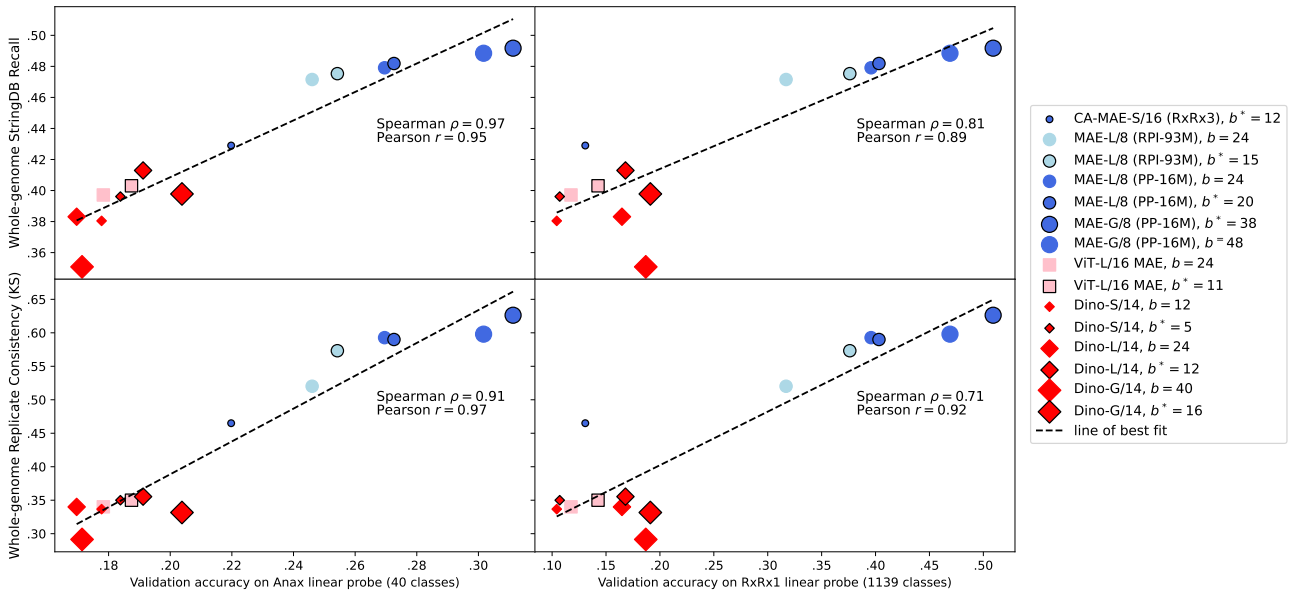


Figure 4: **Correlations** between validation set linear probing (Figure 3) on Anax and RxRx1 for best and last blocks (Eq. 1) compared to downstream whole-genome benchmarks (Table 1) for biological relationship recall on StringDB at 0.05-0.95 gene-gene cosine similarity threshold and replicate consistency KS statistic. Models with **bold** borders are *trimmed*, red are natural image baseline models and blue are trained on microscopy.

natural image data have their best features at blocks that are positioned within the first half of the encoder. For ViT-L/16 MAE, the performance of the best block is 27% higher compared to its final block output features that are typically used for downstream tasks. The higher performance observed for intermediate blocks does not appear to be an intrinsic feature of the ViT architecture as an untrained ViT did not exhibit such a parabolic trend (Figure 3a, right).

Anax 40-class functional gene group classification. Biologically meaningful representation of microscopy images of genetically perturbed cells should capture functional relationships between genes, hence a simple linear probe should be able to predict functional gene groups when trained on these representations. We curated a small subset of 80,000 wells from RxRx3 (Fay et al., 2023) to evaluate linear probes on functional group prediction. We also evaluated similar whole genome knockout screens with ARPE-19 and an additional population of HUVEC cells with soluble TNFα added to all wells. We manually curated Anax, a set of 40 functionally-diverse gene groups containing 348 genes, with details provided in (§ A.8). Examples of groups include major protein complexes (e.g. proteasome, ribosome-small/large), metabolic pathways (e.g. Krebs cycle) and signaling pathways (e.g. calcium signaling) (Figure 2). These groups span broad biological processes that are conserved across cell types – linear separability of these groups would likely indicate that representations are biologically meaningful regardless of cell type.

As shown in Figure 3b, MAE-G/8 significantly outperforms other models in Anax group linear probe classification. The best representations once again are obtained from an intermediate block, achieving a balanced accuracy (0.32) that is 5% greater compared to its final block. We observed similar trends for ViT models pretrained on natural images and representations computed from microscopy images of other cell types/conditions (§ A.5, Figure 6).

In Figure 4, we observe that performance on this novel linear probing task correlates strongly with downstream whole-genome benchmarks across all models (Table 1), whether they are trained on microscopy data or natural images, achieving an overall rank correlation $\rho = 0.97$ with whole-genome StringDB recall and $\rho = 0.91$ with whole-genome replicate consistency. This strong correlation is crucial as it allows us to trim our model to the block with the best linear probe performance as a way to improve the quality of our representations for the whole-genome (Table 1).

5. Whole-genome benchmarking

Table 1 presents our benchmarks computed across the whole-genome. These evaluate the genomic representations obtained for each model by aggregating millions of embeddings of cell images spanning >100,000 of genetic knockout perturbations (17,063 genes $\times$ 6 single guide

Table 1: Multivariate **known biological relationship recall** and univariate **replicate consistency** benchmarks by model, encoding block b, aggregated recall over a total of 145,447 possible gene-gene relationships annotations accumulated across five benchmark databases, and replicate consistency test statistics. The *trimmed* models used linear probes to select an earlier block as the feature encoder (Fig. 3). Each benchmark is computed over TVN-aligned gene-aggregated model embeddings, for each one higher is better, and the best overall result is in **bold**. We report the **Recall %** of biological relationships obtained below the 5th and above the 95th percentiles (0.05-0.95) of all gene-gene cosine similarities against the null distribution, as per (Celik et al., 2024; Kraus et al., 2024), with mean $\pm$ the standard deviation computed over 3 different random seeds of sampling the null distribution in each benchmark run. For replicate consistency, we report the Komologorov-Smirnov (**KS**) and Cramer-von Mises (**CM**) test statistics.

Model backbone	b	**Recall** %	**KS**	**CM**
Baseline ViTs				
ViT-S/16 Untrained	12	34.2 $\pm$ .08	.30	4.3
ViT-L/16 ImgNet MAE	24	37.4 $\pm$ .06	.34	5.1
trimmed	11	37.8 $\pm$ .05	.35	5.8
Baseline Dino ViTs				
ViT-S/14 Dino-V2	12	35.6 $\pm$ .09	.34	5.6
trimmed	5	37.2 $\pm$ .16	.35	6.0
ViT-L/14 Dino-V2	24	35.7 $\pm$ .05	.34	5.3
trimmed	12	38.8 $\pm$ .04	.36	5.9
ViT-G/14 Dino-V2	40	32.7 $\pm$ .10	.29	3.8
trimmed	16	37.5 $\pm$ .13	.33	5.2
Microscopy MAEs				
CA-MAE-S/16 RxRx3	12	39.6 $\pm$ .10	.47	10.4
MAE-L/8 RPI-93M	24	44.4 $\pm$ .12	.52	12.3
trimmed	15	44.3 $\pm$ .12	.57	15.2
MAE-L/8 PP-16M	24	44.4 $\pm$ .12	.59	16.2
trimmed	20	44.7 $\pm$ .06	.59	16.2
MAE-G/8 PP-16M	48	45.4 $\pm$ .07	.60	16.4
trimmed	38	**45.4** $\pm$ .15	**.63**	**18.2**

RNAs each) on HUVEC cells from RxRx3 (Fay et al., 2023). Computing these benchmarks for HCS screens typically requires inferring 140 million crops from the genome-wide RxRx3 microscopy screen (Kraus et al., 2023) (64 tiled crops per each of the 2.2 million wells), but, to reduce compute costs, we discard the outer ring of crops, leaving the 36 center non-edge crops for each well. This requires 80 million forward passes to comprehensively evaluate a new encoder. After inference, we use typical variation normalization (Ando et al., 2017) and chromosome arm bias correction (Lazar et al., 2024) to post-process the embeddings and aggregate them to the gene-level.

We present the multivariate **biological relationship recall** benchmarks proposed by Celik et al. (2024) and originally evaluated for MAEs by Kraus et al. (2023; 2024). These metrics evaluate how many annotated pair-wise relationships are recalled from public databases (CORUM, hu.MAP, Reactome-PPI, Signor, StringDB) in the extremities of a ranked list of cosine similarities of all pair-wise post-processed embeddings (details in § A.6).

To ensure embeddings represent technical replicates of perturbations consistently, we also evaluate model performance on **replicate consistency** based on the experimental design used in the RxRx3 dataset. Specifically, we compare the similarity of the embedding for corresponding wells across different experiments via a non-parametric statistical test. The test statistic measures the difference between the perturbation replicates' similarity distribution and an empirical null distribution, with larger values indicating greater consistency (§ A.7). To compare models, we summarize the resulting statistics over all technical replicates in RxRx3 by taking their median, reported in columns KS and CM (Kolmogorov–Smirnov and Cramer-von Mises test statistics) in Table 1.

Not suprisingly, even the smallest CA-MAE-S/16 trained on microscopy data outperforms the largest baselines ViTs trained on natural images. Training self-supervised on our novel carefully curated Phenoprints-16M dataset improves the performance of the MAEs, as does trimming to an early layer detected by linear probes, while the most scaled model MAE-G/8 achieves the best overall performance in all respects when *trimmed*. Compared to the best previously SOTA model at its normally used final block, MAE-L/8 RPI-93M (Kraus et al., 2023), MAE-G/8 *trimmed* obtains: (a) the highest statistically significant gain in relationship recall over the previous SOTA, achieving a z-score of improvement of 5.21; (b) a 21% improvement in the KS statistic (.52→.63); and, (c) a 48% improvement in the replicate consistency CM statistic (12.3→18.2).

Linear probing to select optimal ViT blocks leads to significant improvements even when applied to Dino-V2 and MAE models pretrained on natural images. Dino-V2 ViT-G is dramatically better in terms of recall and consistency with the early block embeddings at $b^* = 16$ rather than the final embedding from $b = 40$ (which performs worse than a random untrained ViT-S). Dino-V2 ViT-S also observes improvements by using $b^* = 5$ rather than $b = 12$ and outperforms Dino-V2 ViT-G in replicate consistency, while the *trimmed* Dino-V2 ViT-L obtains the best recall among the baselines. This finding is important for the scientific community to consider when applying ML foundation models to experimental data, as it is common to take embeddings from the final layer as a default strategy even when processing potentially out-of-distribution im-

Table 2: Comparing MAEs to manually extracted CellProfiler (McQuin et al., 2018) features on biological relationship recall. Reported at 0.05-0.95 cosine threshold on the public JUMP-CP image dataset (Chandrasekaran et al., 2023), which is generated by completely different labs and assay protocols compared to the images used for pretraining. Each result has a standard deviation $\leq \pm.0023$, spanning gene-gene relationships across nearly 8,000 gene-knockouts.

Model backbone	b	Pretraining data	**CORUM**	**hu.MAP**	**Reactome**	**StringDB**
CellProfiler	-	N/A	.219	.184	.131	.191
CA-MAE-S/16	12	RxRx3	.233	.199	.154	.214
MAE-L/8	24	RPI-93M	.248	.208	.160	.226
MAE-G/8 *trimmed*	38	Phenoprints-16M	**.264**	**.215**	**.165**	**.235**

ages (Lastufka et al., 2024), which, as we have shown, can significantly hinder results.

5.1. Performance on external data (JUMP-CP)

In order to validate that the MAEs generalize to entirely novel data, we evaluated a subset of models on completely external public data generated by different assays and from a variety of different labs as produced by the JUMP-CP consortium (Chandrasekaran et al., 2023). Table 2 presents these results using the relationship recall benchmarks of (Celik et al., 2024), noting that only a subset of 7,976 gene-knockouts are covered by this dataset. For post-processing embeddings, we use PCA with center-scaling for standard batch correction alignment. We observe that the MAEs perform better than the CellProfiler manual feature extraction baseline (§ A.11) (Carpenter et al., 2006; Kamentsky et al., 2011), and that the general trend is maintained with the trimmed MAE-G/8 obtaining the best recall overall. Notably, recall on JUMP-CP is considerably lower than on RxRx3 (Table 1) likely due to different assay protocols and more variance in the data.

6. RxRx3-core benchmarking

RxRx3-core[2] (Kraus et al., 2025) is a publicly available benchmarking dataset for assessing biological capabilities of computer vision models. RxRx3-core includes labeled images (compressed to JPEG-2000) of 735 genetic knockouts and 1,674 small-molecule perturbations across eight concentrations drawn from 222,601 wells ($512 \times 512 \times 6$ pixel center-crops) drawn from the larger RxRx3 dataset.

We evaluate a random choice baseline, CellProfiler, the CA-MAE-S/16 model, the MAE-L/8 model from previous work (Kraus et al., 2023), and the MAE-G/8. We evaluate both the trimmed and full-length version of the latter two models to determine the impact of our model-trimming strategy for the most performant models in this context.

[2] https://huggingface.co/datasets/recursionpharma/rxrx3-core

We also evaluate a preliminary version of a SSL ViT-L/16 we trained with the Dino-V2 algorithm on the RxRx3 microscopy dataset; however, we found that it underperformed compared to the CellProfiler 4 baseline (§ A.11) (Kamentsky et al., 2011) and, unlike the MAEs we trained, significantly overfit during pretraining (§ A.9). As such, we did not pursue additional analysis on this model. We leave further development of Dino-V2 for microscopy data to future work, as we were unable to determine an effective recipe for applying Dino SSL pretraining on large-scale microscopy data.

To evaluate each model, we first inference all 222,601 $\times$ 4 crops and then average the 4 embeddings to the well-level. Then, to perform standard batch correction alignment, we use the "EMPTY", unperturbed wells as our control population. We fit PCA on those control embeddings, use it transform the rest, and then fit a separate standard-scaler on each batch's controls to transform the rest. This simplified alignment strategy empirically performed better than TVN on this dataset.

We present results for the benchmark measuring zero-shot prediction of compound-gene activity using cosine similarities between embeddings (Figure 5). This measures, for each compound, whether the cosine similarities from a model's embeddings correctly rank the compound's known target genes higher than a randomly sampled set of other genes from a ground truth dataset. Table 3 provides exact values along the *max* axis, which captures the strongest potential interaction regardless of concentration. The relative ranking of model performance, holds as expected from the results in § 4 and § 5, and trimming benefits both MAE-L/8 and MAE-G/8, with MAE-G/8 offering a 42% (3.77 $\rightarrow$ 5.38) relative improvement over the method from previous work in predicting compound-gene activity.

7. Discussion and Conclusions

This results in this work demonstrate that: (1) within the context of biological imaging, trimming many ViTs to an earlier block leads to stronger biological linearity and im-

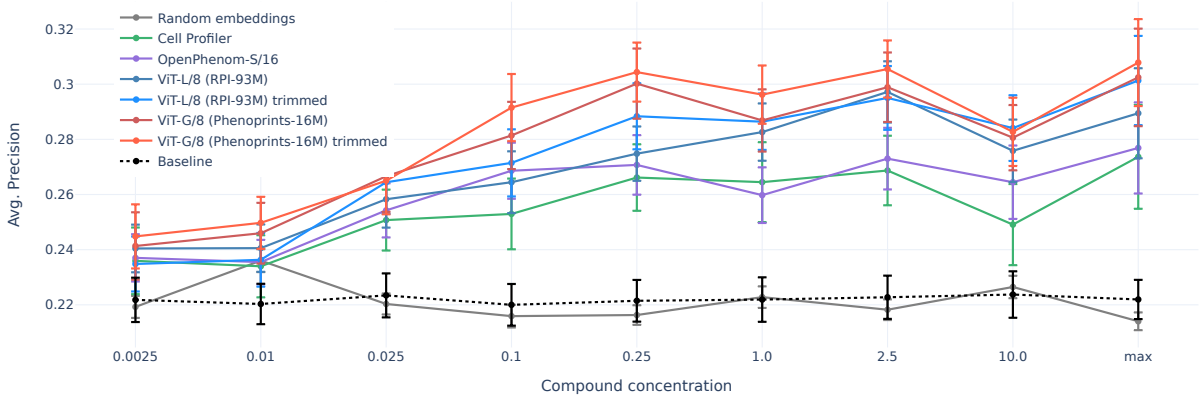


Figure 5: Mean average precision performance on RxRx3-core public benchmark in predicting compound activity against annotated gene targets, across all compound concentrations with error bars for 100 runs of the benchmark with different random seeds (Table 3).

Table 3: Performance on public **RxRx3-core compound-gene benchmark**, measuring mean (± STD over 100 random seeds) average precision in predicting compound activity against target genes with its z-score of improvement over the random baseline.

Model	**Avg. prec. ↑**	**Z-score ↑**
Random baseline	0.222 ± .007	0.00
CellProfiler	0.274 ± .019	2.55
Dino-V2 ViT-L/16, RxRx3	0.258 ± .015	2.13
CA-MAE-S/16, RxRx3	0.273 ± .016	2.90
MAE-L/8, RPI-93M	0.290 ± .017	3.77
trimmed	0.299 ± .016	4.49
MAE-G/8, PP-16M	0.302 ± .015	4.79
trimmed	**0.309** ± .015	**5.38**

proved performance on downstream tasks in addition to cheaper inference costs (Figure 3); (2) linear probing performance on a subset of genetic perturbations correlates strongly with downstream performance on whole-genome benchmarks and can be used to optimize which block is selected for representing the whole-genome (Figure 4); (3) the most scaled model, MAE-G/8 trained on the specially curated dataset Phenoprints-16M for 500 epochs, obtains the overall best performance across all benchmarks and linear probes, providing further evidence for the scaling hypothesis in biological image data (Table 1, Table 2, Table 3, § A.10). We have found that intentionally scaling training compute and parameters of MAEs for microscopy on curated data can benefit a wide variety of biologically relevant tasks, even in comparison to a model trained on a $5\times$ larger dataset. Indeed, even after training the MAE-G/8 on more than eight billion crops from a 16-million-image dataset, the validation reconstruction loss continued to improve. This suggests that increasing the model's parameters further on this data may continue to yield improvements.

More broadly, this work proposes a reusable recipe for training and extracting optimal representations from fully self-supervised models trained on experimental data. The pattern we use can be applied to other domains that contain data from repeated experiments but without accurate ground truth labels. Specifically, we recommend: (a) curating the training set by identifying diverse sets of samples that are represented consistently, e.g., by using a pre-existing model to select such samples; (b) training a scaled transformer-based model using a self-supervised learning technique, such as masked autoencoding; and, (c) inspecting the performance of the trained transformer, and all baselines, at every block to identify the optimal layer for representing the data.

Limitations and reproducibility

In this work, we evaluated publicly available SSL ViTs as baselines and trained new MAE models trained on a specially curated microscopy dataset. Our preliminary attempts to train ViTs with Dino on this microscopy data encountered suboptimal performance (§ 6). Consequently, we allocated our time and compute budget to investigate scaling MAE to ViT-G/8 on curated data. We recognize the potential for other SSL pretraining regimes and fine-tuning

strategies (Singh et al., 2023; Lehner et al., 2024; Hondru et al., 2024; Khan & Fang, 2024; Alkin et al., 2024) oriented for microscopy data to lead to future improvements on these tasks. We publicly release the inference, reconstruction visualization and benchmarking code[3], along with the full weights for CA-MAE ViT-S/16[4].

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning in its applications to the sciences. There are many potential societal consequences of our work, especially relating to the discovery of new biological relationships between genes, potential drug treatments for diseases, and overall accelerating the process of drug development. At the same time, the predictions of these models are not guaranteed to be correct, which is why utmost care must be taken to validate safety and efficacy in orthogonal assays and early-stage clinical trials.

References

Abbas, A. K. M., Tirumala, K., Simig, D., Ganguli, S., and Morcos, A. S. Semdedup: Data-efficient learning at web-scale through semantic deduplication. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023.

Alkin, B., Miklautz, L., Hochreiter, S., and Brandstetter, J. Mim-refiner: A contrastive learning boost from intermediate pre-trained representations. In *NeurIPS 2024 Workshop: Self-Supervised Learning-Theory and Practice*, 2024.

Ando, D. M., McLean, C. Y., and Berndl, M. Improving Phenotypic Measurements in High-Content Imaging Screens. *bioRxiv*, pp. 161422, 2017. doi: 10.1101/161422.

Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., Schwarzschild, A., Wilson, A. G., Geiping, J., Garrido, Q., Fernandez, P., Bar, A., Pirsiavash, H., LeCun, Y., and Goldblum, M. A Cookbook of Self-Supervised Learning. *arXiv*, 2023. doi: 10.48550/arxiv.2304.12210.

Bao, Y., Sivanandan, S., and Karaletsos, T. Channel vision transformers: An image is worth 1 x 16 x 16 words. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11*, 2024.

[3] https://github.com/recursionpharma/maes_microscopy

[4] https://huggingface.co/recursionpharma/OpenPhenom

Berman, M., Jégou, H., Vedaldi, A., Kokkinos, I., and Douze, M. Multigrain: a unified image embedding for classes and instances, 2019. URL https://arxiv.org/abs/1902.05509.

Bourriez, N., Bendidi, I., Cohen, E., Watkinson, G., Sanchez, M., Bollot, G., and Genovesio, A. Chada-vit : Channel adaptive attention for joint representation learning of heterogeneous microscopy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Bray, M.-A., Singh, S., Han, H., Davis, C. T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S. M., Gibson, C. C., and Carpenter, A. E. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11 (9):1757–1774, 2016. ISSN 1754-2189. doi: 10.1038/nprot.2016.105.

Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I. H., Friman, O., Guertin, D. A., Chang, J. H., Lindquist, R. A., Moffat, J., et al. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7:1–11, 2006.

Celik, S., Hütter, J.-C., Carlos, S. M., Lazar, N. H., Mohan, R., Tillinghast, C., Biancalani, T., Fay, M. M., Earnshaw, B. A., and Haque, I. S. Building, benchmarking, and exploring perturbative maps of transcriptional and morphological data. *PLOS Computational Biology*, 20(10):1–24, 10 2024. doi: 10.1371/journal.pcbi.1012463. URL https://doi.org/10.1371/journal.pcbi.1012463.

Chandrasekaran, S. N., Ackerman, J., Alix, E., Ando, D. M., Arevalo, J., Bennion, M., Boisseau, N., Borowa, A., Boyd, J. D., Brino, L., et al. Jump cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, pp. 2023–03, 2023.

Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Pham, H., Dong, X., Luong, T., Hsieh, C.-J., Lu, Y., et al. Symbolic discovery of optimization algorithms. In *Advances in Neural Information Processing Systems*, volume 36, pp. 49205–49233, 2023.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11*, 2024.

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A. P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pp. 7480–7512. PMLR, 2023.

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2020.

Drew, K., Lee, C., Huizar, R. L., Tu, F., Borgeson, B., McWhite, C. D., Ma, Y., Wallingford, J. B., and Marcotte, E. M. 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.

Eldan, R. and Li, Y. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

Evci, U., Dumoulin, V., Larochelle, H., and Mozer, M. C. Head2toe: Utilizing intermediate representations for better transfer learning. In *International Conference on Machine Learning*, pp. 6009–6033. PMLR, 2022.

Fay, M. M., Kraus, O., Victors, M., Arumugam, L., Vuggumudi, K., Urbanik, J., Hansen, K., Celik, S., Cernek, N., Jagannathan, G., et al. RxRx3: Phenomics map of biology. *bioRxiv*, pp. 2023–02, 2023.

Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., Deng, C., Varusai, T., Ragueneau, E., Haider, Y., May, B., Shamovsky, V., Weiser, J., Brunson, T., Sanati, N., Beckman, L., Shao, X., Fabregat, A., Sidiropoulos, K., Murillo, J., Viteri, G., Cook, J., Shorser, S., Bader, G., Demir, E., Sander, C., Haw, R., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D687–D692, 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1028.

Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Ruepp, A. 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.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Hondru, V., Croitoru, F. A., Minaee, S., Ionescu, R. T., and Sebe, N. Masked image modeling: A survey. *arXiv preprint arXiv:2408.06687*, 2024.

Hooke, R. *Micrographia: Or Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses with Observations and Inquiries Thereupon*. Jo. Martyn and Ja. Allestry, Printers to the Royal Society, London, 1665.

Javaheripi, M., Bubeck, S., Abdin, M., Aneja, J., Bubeck, S., Mendes, C. C. T., Chen, W., Del Giorno, A., Eldan, R., Gopi, S., et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1 (3):3, 2023.

Kalinin, A. A., Arevalo, J., Vulliard, L., Serrano, E., Tsang, H., Bornholdt, M., Rajwa, B., Carpenter, A. E., Way, G. P., and Singh, S. A versatile information retrieval framework for evaluating profile strength and similarity. *bioRxiv*, pp. 2024.04.01.587631, 4 2024. doi: 10.1101/2024.04.01.587631.

Kamentsky, L., Jones, T. R., Fraser, A., Bray, M.-A., Logan, D. J., Madden, K. L., Ljosa, V., Rueden, C., Eliceiri, K. W., and Carpenter, A. E. Improved structure, function and compatibility for cellprofiler: modular high-throughput image analysis software. *Bioinformatics*, 27 (8):1179–1180, 2011.

Kenyon-Dean, K., Wang, Z. J., Urbanik, J., Donhauser, K., Hartford, J., Saberian, S., Sahin, N., Bendidi, I., Celik, S., Fay, M., et al. Vitally consistent: Scaling biological representation learning for cell microscopy. In *Neural Information Processing Systems Workshop on Foundation Models for Science (NeurIPS FM4Science)*, 2024.

Khan, M. O. and Fang, Y. What is the best way to fine-tune self-supervised medical imaging models? In *Annual Conference on Medical Image Understanding and Analysis*, pp. 267–281. Springer, 2024.

Kim, V., Adaloglou, N., Osterland, M., Morelli, F., and Zapata, P. A. M. Self-supervision advances morphological profiling by unlocking powerful image representations. *bioRxiv*, pp. 2023–04, 2023.

Kraus, O., Kenyon-Dean, K., Saberian, S., Fallah, M., McLean, P., Leung, J., Sharma, V., Khan, A., Balakrishnan, J., Celik, S., et al. Masked autoencoders are scalable learners of cellular morphology. In *Neural Information Processing Systems Workshop on Generative AI and Biology (NeurIPS GenBio)*, 2023.

Kraus, O., Kenyon-Dean, K., Saberian, S., Fallah, M., McLean, P., Leung, J., Sharma, V., Khan, A., Balakrishnan, J., Celik, S., et al. Masked autoencoders for microscopy are scalable learners of cellular biology. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11757–11768, 2024.

Kraus, O., Comitani, F., Kenyon-Dean, K., Urbanik, J., Arumugam, L., Saberian, S., Wognum, C., Celik, S., and Haque, I. S. RxRx3-core: Benchmarking drug-target interactions in high-content microscopy. In *Learning Meaningful Representations of Life (LMRL) Workshop at ICLR 2025*, 2025.

Lastufka, E., Drozdova, M., Kinakh, V., and Voloshynovskiy, S. Vision foundation models: can they be applied to astrophysics data? In *Neural Information Processing Systems Workshop on Foundation Models for Science (NeurIPS FM4Science)*, 2024.

Lazar, N. H., Celik, S., Chen, L., Fay, M. M., Irish, J. C., Jensen, J., Tillinghast, C. A., Urbanik, J., Bone, W. P., Gibson, C. C., et al. High-resolution genome-wide mapping of chromosome-arm-scale truncations induced by crispr–cas9 editing. *Nature Genetics*, pp. 1–12, 2024.

Lehner, J., Alkin, B., Fürst, A., Rumetshofer, E., Miklautz, L., and Hochreiter, S. Contrastive tuning: A little help to make masked autoencoders forget. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2965–2973, 2024.

Liu, Y. and Xie, J. Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115:393 – 402, 2018.

Llama3, T. The Llama 3 Herd of Models. *arXiv*, 2024. doi: 10.48550/arxiv.2407.21783.

McQuin, C., Goodman, A., Chernyshev, V., Kamentsky, L., Cimini, B. A., Karhohs, K. W., Doan, M., Ding, L., Rafelski, S. M., Thirstrup, D., et al. Cellprofiler 3.0: Next-generation image processing for biology. *PLoS biology*, 16(7):e2005970, 2018.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2024.

Radenović, F., Tolias, G., and Chum, O. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41 (7):1655–1668, 2018.

Schuhmann, C., Kaczmarczyk, R., Komatsuzaki, A., Katta, A., Vencu, R., Beaumont, R., Jitsev, J., Coombes, T., and Mullis, C. LAION-400M: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*. Jülich Supercomputing Center, 2021.

Singh, M., Duval, Q., Alwala, K. V., Fan, H., Aggarwal, V., Adcock, A., Joulin, A., Dollár, P., Feichtenhofer, C., Girshick, R., et al. The effectiveness of mae prepretraining for billion-scale pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5484–5494, 2023.

Sivanandan, S., Leitmann, B., Lubeck, E., Sultan, M. M., Stanitsas, P., Ranu, N., Ewer, A., Mancuso, J. E., Phillips, Z. F., Kim, A., Bisognano, J. W., Cesarek, J., Ruggiu, F., Feldman, D., Koller, D., Sharon, E., Kaykas, A., Salick, M. R., and Chu, C. A Pooled Cell Painting CRISPR Screening Platform Enables de novo Inference of Gene Function by Self-supervised Deep Learning. *bioRxiv*, pp. 2023.08.13.553051, 2023. doi: 10.1101/2023.08.13.553051.

Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., and Morcos, A. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Sypetkowski, M., Rezanejad, M., Saberian, S., Kraus, O., Urbanik, J., Taylor, J., Mabey, B., Victors, M., Yosinski, J., Sereshkeh, A. R., et al. Rxrx1: A dataset for evaluating experimental batch correction methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4284–4293, 2023.

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., and Mering, C. v. 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.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

Weinzaepfel, P., Lucas, T., Larlus, D., and Kalantidis, Y. Learning super-features for image retrieval. In *International Conference on Learning Representations*, 2022.

Wightman, R. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Zhang, Q., Wang, Y., and Wang, Y. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems*, 35:27127–27139, 2022.

A. Appendix

A.1. Training dataset curation details

In order to produce Phenoprint-16M, we curated 93M using the following steps:

1. Filtering out data that did not pass data quality filters related to the focus of the image, quantity of dead cells, assay conditions, and presence of strong anomalous imaging artifacts.

2. Filtering out data with missing information about the perturbations applied, data with more than 3 perturbations applied, and data of unusual size (in the image dimension or number of channels).

3. Filtering out perturbation conditions that had been in less than 3 distinct experiments or 20 distinct wells so as to capture a variety of batch effects and have a broad sample of positives per class.

4. Under-sampling perturbation conditions that were clearly over-represented in the dataset. Our experiment designs contain positive controls, negative controls, and wells without perturbation within each experiment. At this step, we keep 10% of positive controls and wells without any perturbation, 30% of negative controls, and all other perturbation conditions.

5. Filtering out wells where none of the perturbation conditions had a phenoprint (§A.3) (across different map types) in any experiment it had been run in.

A.2. Training hyperparameters

Table 5 provides the hyperparameters used for training the new vision transformers presented in this work. Each model was trained using a 75% mask ratio and the standard decoder architecture for MAEs (He et al., 2022). Each model was trained with the standard L2 MAE loss and the Fourier-space loss function implemented by (Kraus et al., 2024) with a weight of $\alpha = 0.01$. We note, however, that the details presented by (Kraus et al., 2024) do not precisely correspond with the implementation provided in their Github repository; when reshaping the tokens to a shape compatible with the 2D Fourier transform, the permute operation resulted in adjacent pixels being from different channels of the input, resulting in the high frequency components of the loss being a function of the relationships between input channels. An initial investigation with a ViT-L/8 showed that changing the implementation to the one described in the paper did not dramatically change probing results. As such, we used the implementation as-is and leave additional analysis of loss function design for MAEs to future work.

A.3. Perturbation Consistency

In order to assess the consistency of the induced morphology on the cells by the perturbations, we used a non-parametric perturbation consistency test similar to the one introduced in (Celik et al., 2024). Let $x_{g,1}, x_{g,2}, \cdots, x_{g,n}$ be the embeddings for replicates of perturbation x_g on experiment (batch) e. As the test statistic for perturbation consistency, $\bar{s}_g^e$ is defined as the mean of the cosine similarities across all pairs of replicates of x_g.

$$\bar{s}_g^e = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\langle x_{g,i}, x_{g,j} \rangle}{||x_{g,i}|| ||x_{g,j}||}. \tag{2}$$

where $\langle . \rangle$ and $||.||$ denote dot product and L_2 norm.

Statistical significance of $\bar{s}_g^e$ is assessed using a permutation test comparing it against an empirical null distribution generated using the same statistic for a set of randomly selected perturbations in experiment e, $\{\bar{s}'_1, \cdots, \bar{s}'_K\}$. The p-value for $\bar{s}_g^e$ is computed as follows

$$p_g = \frac{\max\left\{ \#\{\bar{s}'_k \geq \bar{s}_g^e\}, 1 \right\}}{K}. \tag{3}$$

When multiple experiments existed for the same perturbation, we combined p-values using the Cauchy Combination test (Liu & Xie, 2018).

A.4. Training linear probes

In this section, we provide details about the training process and preprocessing steps used in our logistic regression models. These models were trained on output features derived from various Vision Transformer (ViT) blocks.

The data was split by experiments, ensuring that the test data originated from experiments distinct from those used for training. This approach helps to validate the generalization performance of our models across different experimental conditions.

For both RxRx1 gene prediction and Anax group prediction, we apply `StandardScaler` from the scikit-learn library as the only preprocessing step to standardize the features prior to training linear probes. `StandardScaler` transformation was fitted on data from the train split. We trained the logistic regression models using scikit-learn's `LogisticRegression` class. The following parameters and settings were used during model optimization:

- Solver: lbfgs
- Maximum Iterations: 2000
- Class Weight: balanced

Table 4: Overview of vision transformer (ViT) encoders used and evaluated in this work.

Model Name	Parameters	Blocks	Model Dim	Pretraining Data
Baselines				
Untrained ViT-S/16	25M	12	384	N/A
Dino-V2 ViT-S/14	25M	12	384	Natural images
Dino-V2 ViT-L/14	307M	24	1024	Natural images
Dino-V2 ViT-G/14	1,100M	40	1536	Natural images
ViT-L/16 MAE	307M	24	1024	Imagenet-21k
MAEs for microscopy				
CA-MAE-S/16	25M	12	384	RxRx3
MAE-L/8	307M	24	1024	RPI-93M
MAE-L/8	307M	24	1024	Phenoprints-16M
MAE-G/8	1,860M	48	1664	Phenoprints-16M

Table 5: Training hyperparameters for the new models presented in this work. Each used a one-cycle cosine learning rate decay schedule with 10% warm-up using the Lion optimizer from (Chen et al., 2023) with betas (0.9, 0.95) and weight decay of 0.05, with additional ViT settings such as LayerScale as proposed by (Dehghani et al., 2023). *Note that MAE-G/8 had multiple restarts during training due to challenges associated with massive model training on large-scale shared distributed compute clusters.

Hyperparameter	CA-MAE-S/16	MAE-L/8	MAE-G/8
Vision transformer backbone	ViT-S	ViT-L	ViT-G (Zhai et al., 2022)
Pretraining Data	RxRx3	Phenoprints-16M	Phenoprints-16M
Training epochs	100	500	500*
Learning rate	1e-4	3e-5	3e-5
Global batch size	2048	16384	8192
Stochastic depth	0.1	0.3	0.6
# GPUs	16 A100s	128 H100s	256 H100s
# GPU-hours	400	15,360	48,000

For RxRx1 gene prediction, we trained logistic regression models to predict one of 1139 possible perturbation labels (1138 genetic perturbation and non-perturbed control). For Anax group prediction, we trained logistic regression models to predict one of 40 possible function group labels (§ A.8). We report the balanced test accuracy as the main evaluation metric for all linear probing experiments.

A.5. Anax classification for other cell lines/treatment conditions: ARPE19 and HUVEC with TNFα background

We performed linear probing on imaging data obtained for a retinal pigment epithelia (RPE) cell line, ARPE19, and HUVEC cells treated with an inflammatory cytokine, TNFα. We similarly observed that intermediate blocks often have the most linearly separate features compared to the final block.

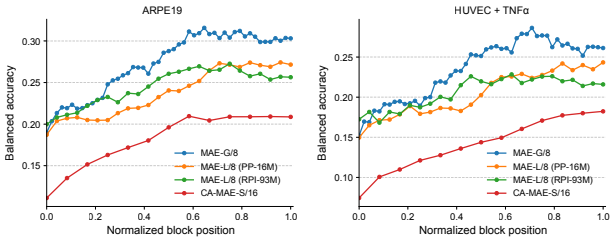


Figure 6: Layerwise validation set linear probe performance on Anax functional gene group classification beyond RxRx3: CRISPR knockouts in the ARPE-19 immortalized epithelial cell-line (left), and in HUVEC cells with a TNFα background (right).

A.6. Biological Relationship Recall

A valuable use of large-scale HCS experiments is to perform large-scale inference of biological relationships between genetic perturbations. We evaluate each model's ability to recall known relationships by using the *biolog-*

ical relationship recall benchmark described in Celik et al. (2024). First, we correct for batch effects using *Typical Variation Normalization* (TVN) (Ando et al., 2017), and also correct for possible chromosome arm biases known to exist in CRISPR-Cas9 HCS data (Lazar et al., 2024). To infer biological relationships, we compute the aggregate embedding of each perturbation by taking the spherical mean over its replicate embeddings across experiments. 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 (Giurgiu et al., 2019), hu.MAP (Drew et al., 2017), Reactome (Gillespie et al., 2021), and StringDB (Szklarczyk et al., 2020) (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 (Fay et al., 2023).

A.7. Replicate Consistency

In order to assess the reproducibility of the perturbations across their technical replicates, we compare the distributions of the similarities for same perturbations across replicates against an empirical null distribution. Specifically, for technical replicate experiments e_a^i and e_b^i, we calculate the cosine similarity between the embeddings of perturbation x_j in them, denoted as s^{x_j}.The query distribution q^{e_i} is constructed by computing the cosine similarities for all perturbations that have a matching well on experiments e_a^i and e_b^i. An empirical null distribution of identical cardinality is created by computing cosine similarity, r^{x_k,x_l}, between random pairs from e_a^i and e_b^i such that no pair corresponds to the same perturbation, $p_0^{e_i}$. Using non-parametric statistical tests, namely Kolmogorov-Smirnov (KS) and Cramer Von-Mises (CVM), we can evaluate the hypothesis that q^{e_i} and $p_0^{e_i}$ are drawn from the same distribution. Formally, let $Q^{e_i}(x)$ and $P_0^{e_i}(x)$ be the cumulative distribution functions for q^{e_i} and $p_0^{e_i}$ respectively, then the KS statistic for the two-sample case of technical replicate experiments e_a^i and e_b^i is defined as:

$$\text{KS}^{e_i} = \sup_x |Q^{e_i}(x) - P_0^{e_i}(x)|. \tag{4}$$

The Cramér–von Mises test statistic (CVM) for experiments e_a^i and e_b^i is computed as:

$$\text{CVM}^{e_i} = \frac{1}{2N^2} \sum_{m=1}^{N} \left[(r_m - m)^2 + (s_m - m)^2 \right] - \frac{4N^2 - 1}{12N}. \tag{5}$$

where N is the cardinality of q^{e_i} and $p_0^{e_i}$ and s_m and r_m are ranks of similarities s^{x_j} and r^{x_k,x_l} in the combined distribution of q^{e_i} and $p_0^{e_i}$ when ordered. In order compare models, we use the median of CVM^{e_i} and KS^{e_i} over all technical replicate experiment pairs e_i.

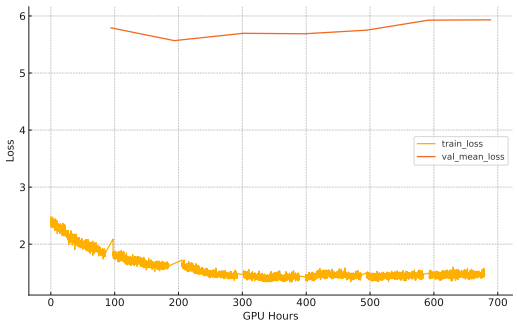


Figure 7: Loss curve when training Dino-v2 ViT-L/16 on RxRx3.

Since the pairs are randomly selected for $p_0^{e_i}$, the embeddings would be mostly orthogonal thus the distribution would be centered around 0.Given that not all CRISPR knockouts would induce a morphological change in the cells, it's plausible for distribution q^{e_i} to exhibit a peak around 0. As the model approaches the precision of an oracle, we would anticipate the mass situated around this peak to shift towards higher cosine similarity values.

A.8. Anax Group Prediction Details

The Anax probing task introduced in this paper is intended to balance capturing a diverse range of biology that is broadly conserved between cell types with a reduced cost of execution. The name "Anax" is a reference to Anaximander, the 6th century B.C. philosopher credited with making the first world map.

In curating these genes, we analyzed the sources listed in § A.6 as well as internal gene expression data to produce "functional" groups corresponding to biological processes, cellular components, and molecular functions. Not all genes within each group are expected to have the same knockout phenotype, but are classified by humans as having related function – linear separability of these genes would indicate that a model has learned similar concepts to those deemed significant by biologists.

The gene groups we use for the 40-class Anax group classification task (§ A.4) are listed in Table 7.

A.9. Dino-V2 pretraining on microscopy data

We attempted to train two Dino-v2 models on microscopy data. One ViT-L/16 from scratch on RxRx3, and another attempt of fine-tuning the MAE-L/8 on RPI-93M with the Dino-v2 losses. They had the following hyperparameter

Table 6: RxRx3-core benchmarks for our initial attempts to train Dino-V2 models on microscopy data. The latter was finetuned from the MAE-L/8 trained on RPI-93M. Results compare to Table 3.

DinoV2 model	Avg. Prec.	Z-score
ViT-L/16 RxRx3	$0.258 \pm .015$	2.13
ViT-L/8 (ft.) RPI-93M	$0.255 \pm .018$	1.76

settings which were tuned on another dataset: output-dim 65536, 2 global crops, 4 local crops, dino loss weight 1.0, koleo loss weight 0.1, ibot loss weight 1.0, Lion optimizer with max learning rate 1e-5, weight decay 0.05, betas 0.9 0.95, and cosine annealing. In both cases, we observed significant over-fitting of the loss from the start (Figure 7).

In Table 6 we show that both models fail to improve on the RxRx3-core benchmark metrics (i.e., z-scores over the random baseline) versus the CA-MAE-S/16 RxRx3 model which had Z-score of 2.90 on average precision for predicting compound activity. We have not found an effective recipe for training Dino on microscopy data. However, we note that theoretical evidence exists arguing that MAE learning is in some ways equivalent to contrastive learning (Hondru et al., 2024), so even if an appropriate Dino recipe is found it would remain to be seen if it differs substantially from MAEs for microscopy given the same training compute. As described in the **Limitations** section, we expect that future work would have to dedicate significant training ablations and creativity to determine the best possible training recipe for training Dino on microscopy data.

A.10. Correlation between model scale and benchmark results

In Figure 8 we show the correlations between training FLOps (floating point operations) and downstream results. Over all benchmarks we observe a very strong consistent linear trend where scaling training FLOps improves overall pwerformance. This work provides the next log step in scale as we enter into the billion-parameter model regime with MAE-G/8. These results therefore provide additional evidence that the trend initially discovered by (Kraus et al., 2023) between FLOps and relationship recall actually extends both to billion-parameter models and even moreso for other biologically meaningful benchmarks pertaining to linear probes on small experiments and to replicate consistency on the whole-genome.

A.11. CellProfiler features

CellProfiler bioimage analysis software (Carpenter et al., 2006; McQuin et al., 2018) was used to compute features using classical segmentation and feature extraction algorithms. Benchmarking results using CellProfiler features are reported for JUMP-CP (§ 5.1) and RxRx3-core (§ 6). JUMP-CP CellProfiler features were downloaded from https://cellpainting-gallery.s3.amazonaws.com/index.html#cpg0016-jump. CellProfiler features for RxRx3-core were computed using version 2.2.0 (Kamentsky et al., 2011). Single cells were segmented after applying illumination correction and shape, intensity, and texture features were extracted from each cell, resulting in 952 dimensional profiles. For RxRx3-core, we kept only cells in the center 512x512 crop of original 2048x2048 images and then mean aggregated them. 4,749 wells of the 222,601 RxRx3-core wells were missing CellProfiler features, so these were held out of the benchmark computations for all models on RxRx3-core.

Table 7: Anax groups and their associated genes. This table presents a comprehensive list of gene groups and their corresponding genes.

Anax Group	**Genes**
Acyl Coa Biosynthesis	ELOVL2, ELOVL5, ELOVL6, HACD1, HACD2, HSD17B12, SCD, SCD5, TECR
Adherens Junctions	ACTB, ACTG1, AFDN, CDH1, CTNNA1, CTNNB1, CTNND1, NECTIN1, NECTIN3, NECTIN4
Amino Acid Metabolism	ALDH4A1, ARG2, CKB, CKMT2, CPS1, DAO, OTC, PYCR2, PYCR3, SAT1
Apoptosis	CFLAR, DFFB, CASP6, CASP3, FASLG, BCL2, DFFA, XIAP, TNFSF10, AKT3
Autophagy	ATG12, ATG3, ATG4B, ATG4C, ATG7, GABARAP, PIK3C3, PIK3R4, PRKAA1, ULK1
Beta Oxidation Of Fatty Acids	ACAA2, ACADL, ACADM, ACADS, ACADVL, ECHS1, ECI1, HADH, HADHA, HADHB
Calcium Signaling	ADCY1, ADCY2, ADCY3, CALM1, CAMK2B, CAMK2D, PDE1B, PDE1C, PRKACG, PRKX
Clathrin Coated Vesicles	AP2A1, AP2A2, AP2B1, AP2M1, AP2S1
COPI	ARCN1, COPA, COPB1, COPB2, COPE, COPG1, COPZ1
COPII Vesicles	SEC13, SEC23A, SEC24B, SEC24D, SEC31A
DNA Damage Repair	BLM, BRCA2, EME1, NBN, POLD2, RAD51B, RAD51C, RAD51D, RPA1, XRCC2
Dynein	DYNC1H1, DYNC1I2, DYNC1LI1, DYNC1LI2, DYNLT1
ER Protein Translocation	SPCS3, SEC61A1, SRP14, SRP72, SPCS1, SRPRA, SEC11A, SRP68, SRPRB, SRP54
Exosome	DIS3, EXOSC10, EXOSC3, EXOSC4, EXOSC5, EXOSC6, EXOSC7, EXOSC8, EXOSC9, MPHOSPH6
Gap Junctions	ADCY8, DRD2, HTR2C, ITPR2, LPAR1, PDGFD, PDGFRB, PLCB3, TUBA1C, TUBB1
Golgi	ACTR10, ACTR1A, CAPZA3, COG4, CTSZ, PPP6C, RAB1B, SEC22C, SEC24C, TMED9
MAPK	DUSP4, EGF, FGF18, FGF20, HSPB1, MAP2K2, MAPKAPK5, RAC1, RAP1A, RASGRP3
Mitochondria Structure	APOOL, APOO, TMEM11, CHCHD6, ATP5ME, MICOS13, ATP5F1C, DNAJC11, DMAC2L, ATP5MF
Mitochondrial Transport	ATP5F1A, COA4, COA6, COX17, HSPA9, IDH3G, PITRM1, PMPCA, PMPCB, SLC25A4
mTOR Pathway	CAB39, CAB39L, EIF4EBP1, MLST8, PRKAA2, RPS6KB1, RPTOR, STK11, STRADA, TSC1
Nonsense Mediated Decay	CASC3, EIF4A3, MAGOH, MAGOHB, RBM8A
Nuclear Pore	NUP107, NUP133, NUP153, NUP188, NUP205, NUP37, NUP85, NUP93
Nucleolus Structure	FBL, NAT10, NOLC1, NOP58, UTP20
Nucleotide Metabolism	ADSL, ADSS1, ADSS2, ATIC, GMPS, IMPDH1, IMPDH2, PAICS, PFAS, PPAT
P53 Stress Signaling	ATM, ATR, CCNG1, CDK1, CHEK1, CHEK2, MDM2, MDM4, TP53, TP73
Pentose Phosphate Pathway	G6PD, TALDO1, DERA, RPE, PGM2, RBKS, PGD, PGLS, RPEL1, PRPS2
Peroxisome Biology	ACOT8, AGPS, BAAT, HMGCL, HSD17B4, MLYCD, PAOX, PEX12, PEX6, PIPOX
Prespliceosome Complex	ALYREF, AQR, CRNKL1, DDX5, HNRNPK, LSM2, PLRG1, PRPF4, SMNDC1, SRSF4
Proteasome	PSMA1, PSMA4, PSMB1, PSMB2, PSMB7, PSMA6, PSMA3, PSMB4, PSMA5, PSMB3
Ribosome Large	RPL13A, RPL11, RPL10, RPL23A, RPL30, RPL7A, RPLP2, RPL28, RPL5, RPL27A
Ribosome Small	RPS2, RPS6, RPS8, RPS16, RPS11, RPS3A, RPS19, RPS15, RPS4X, RPS9
RNA Polymerase II	POLR2A, POLR2B, POLR2C, POLR2G, POLR2I, POLR2L
TCA Cycle	ACO2, DLST, FH, IDH2, IDH3B, MDH2, OGDH, SDHB, SUCLA2, SUCLG2
Tight Junctions	CLDN14, CLDN17, CLDN18, CLDN19, CLDN4, CLDN8, CLDN9, MPP5, PARD6B, PRKCI
Translation Initiation Complex	EIF3G, EIF3A, EIF3D, EIF3I, EIF3K, EIF3M, EIF3B, EIF3H, EIF3E, EIF3L
Transport Of Fatty Acids	APOD, LCN12, LCN15, LCN9, SLC27A1, SLC27A4, SLC27A6
Tubulin	TUBA3C, TBCC, TBCD, TUBA4B, TUBA8, TUBAL3, TUBA1A, TUBB4B, ARL2, TUBA1B
Unfolded Protein Response	CXXC1, DNAJB11, EIF2S3, KHSRP, MBTPS1, SHC1, TATDN2, TLN1, TSPYL2, YIF1A
V-ATPase	ATP6V1A, ATP6V, ATP6V1D, ATP6V1E1, ATP6V1F, ATP6V1H

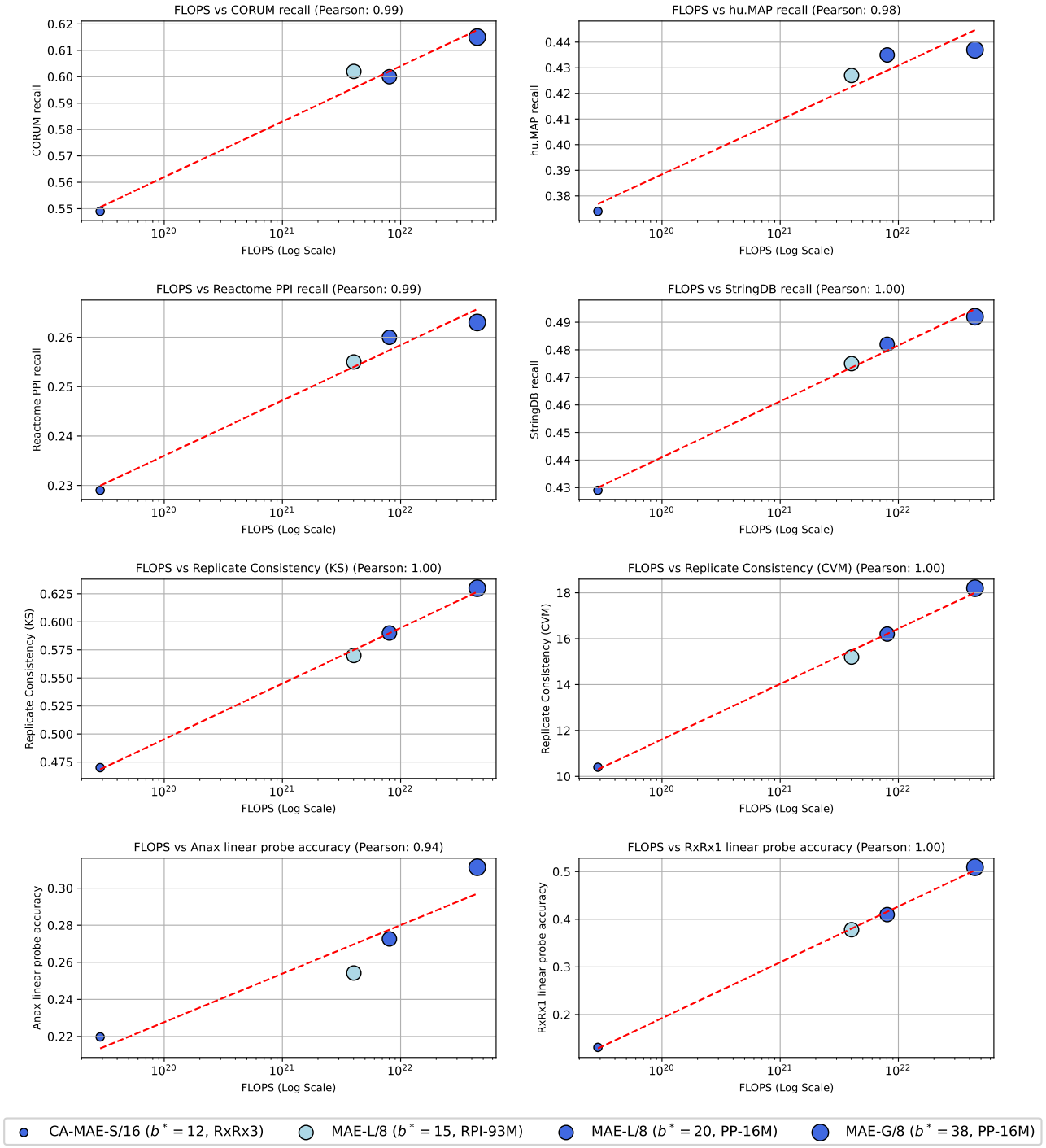


Figure 8: Relationship between FLOPs and benchmark evaluation results for the six whole-genome tasks (Table 1) and the two linear probing tasks (Figure 3).