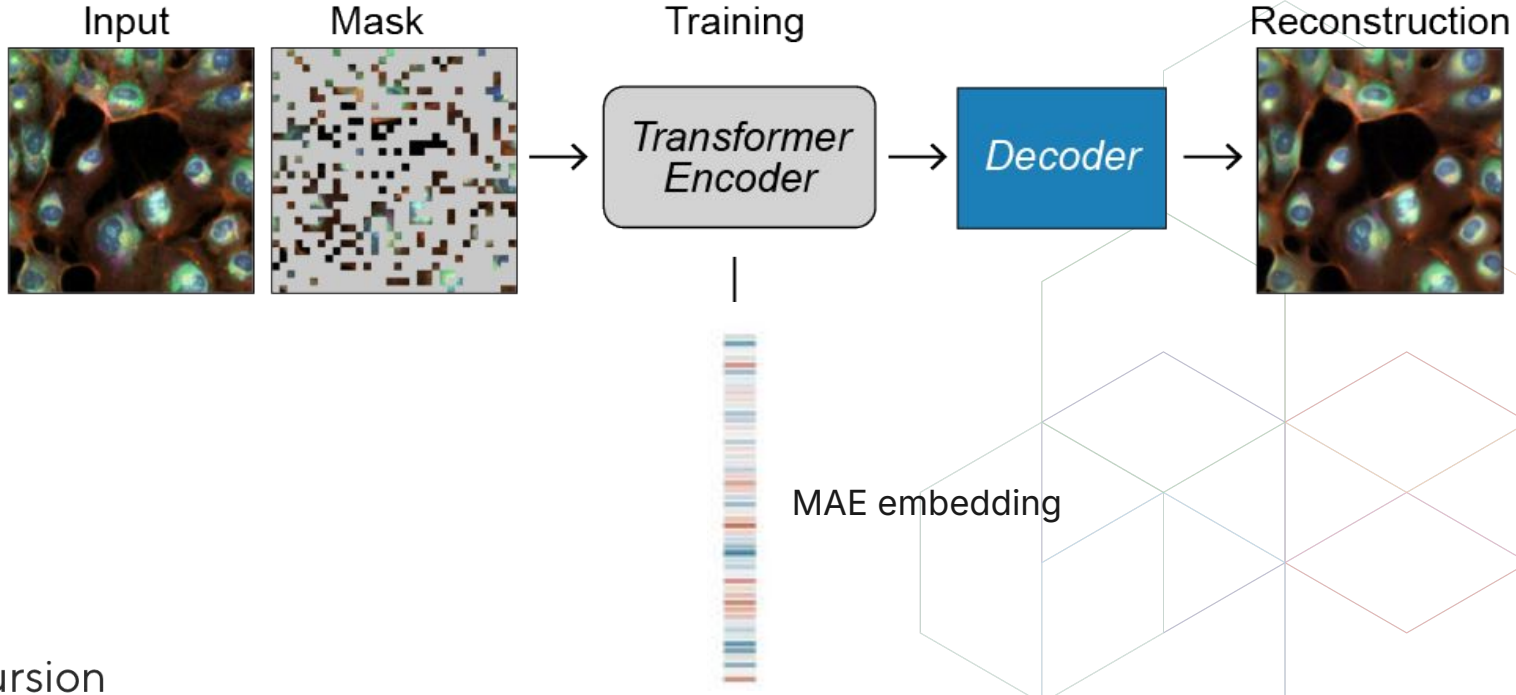
ViTally Consistent: Scaling Biological Representation Learning for Cell Microscopy

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

Kian Kenyon-Dean @ NeurIPS FM4Sci Workshop 2024
Staff Machine Learning Engineer, Recursion

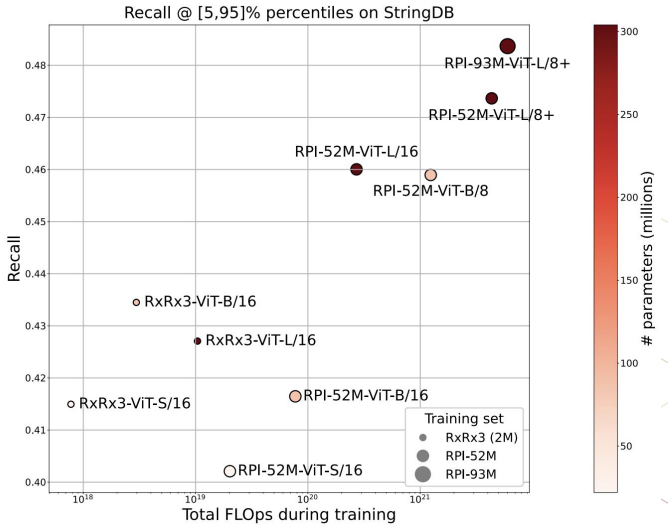
Context & related work

Masked Autoencoding (MAE): *effective self-supervised learning when scaled*



Phenom-1: Large Vision Transformer (ViT-L)

- ViT-Large/8, 330 million parameters (1024+1 tokens per sample)
- MAE trained on RPI-93M for ~40 epochs



Masked Autoencoders for Microscopy are Scalable Learners of Cellular Biology

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

CVPR
JUNE 17-21, 2024

Masked Autoencoders are Scalable Learners of Cellular Morphology

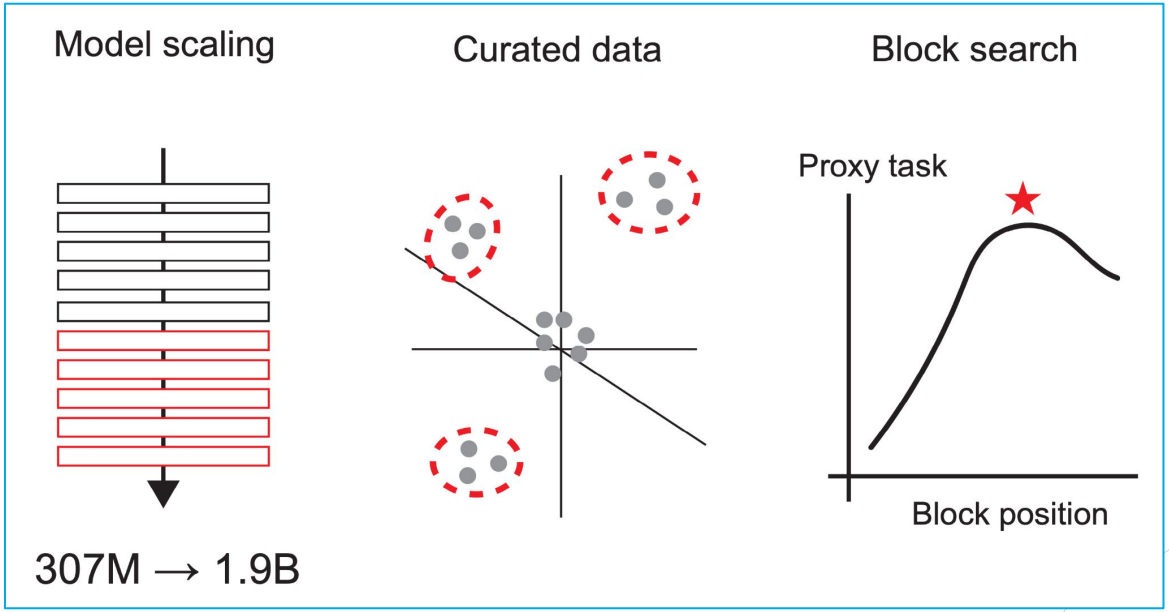
Oren Kraus* Kian Kenyon-Dean* Saber Saberian Maryam Fallah Peter McLean
Jess Leung Vasudev Sharma Ayla Khan Jia Balakrishnan Safiye Celik
Maciej Sypetkowski Chi Vicky Cheng Kristen Morse Maureen Makes
Ben Mabey Berton Earnshaw

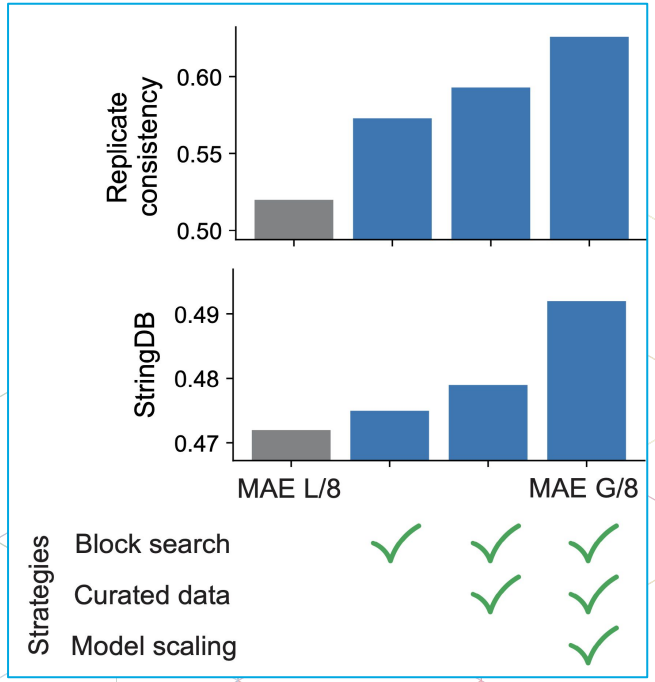
NEURAL INFORMATION PROCESSING SYSTEMS

Recursion

Our contributions in this work

General foundation model training approach



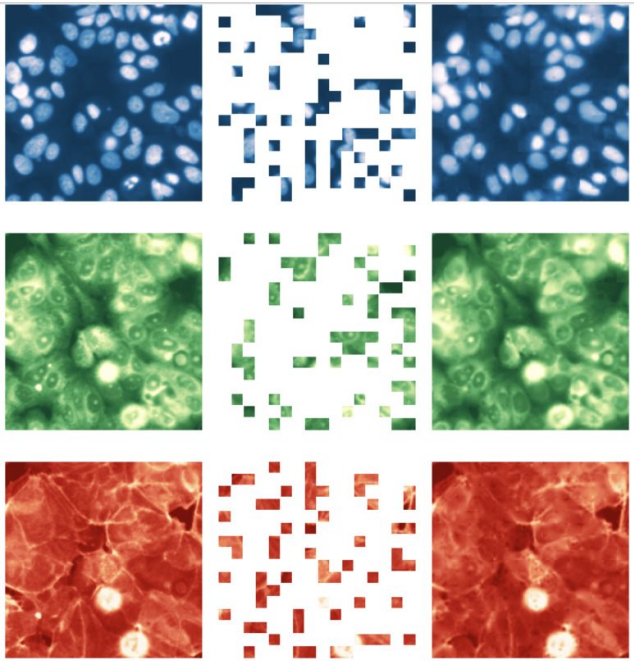


Recursion

	Phenom-1 (MAE-L/8)	Phenom-2 (MAE-G/8)
Training Regime	Self-supervised MAE	Self-supervised MAE
Model architecture	Vision Transformer (ViT-L/8) 24 blocks	Vision Transformer (ViT-G/8) **48 blocks**
Model parameters	330 million	**1.9 BILLION (+6x)**
Dataset base	93M unique well images	**Specially curated Phenoprints-16M (-6x)**
Dataset sampling	3.5 billion crops == 3.6 trillion tokens	**8 billion crops (+2.2x)** **== 8.2 trillion tokens**
Training time	20,000 A100 hours	**43,000 H100 hours (+4x)** (i.e., >5 GPU-years!)

Channel-Agnostic MAE baseline

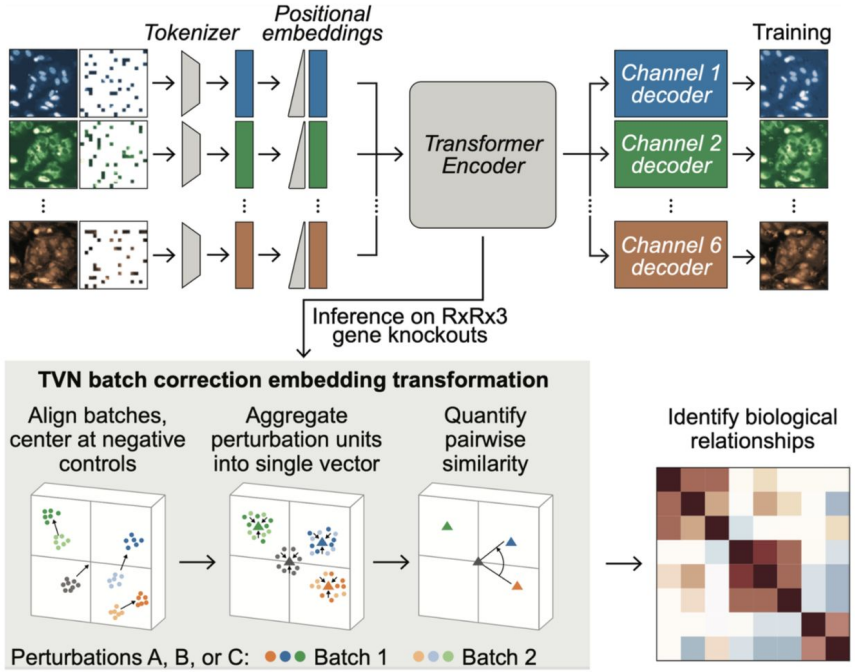
- CA-MAE-S/16 (1536+1 tokens per sample)
  - Trained on **RxRx3** for 100 epochs
- **OpenPhenom** - new open-source publicly available model on HuggingFace 🤗
  - CA-MAE-S/16 but also trained on RxRx3 + public JUMP-CP data.



Recursion

Evaluation: how "good" (in terms of batch effect correction / *replicate consistency* and *relationship prediction*) are your model's **genetic representations** derived from images?

- **Inference regime**: 36 center crops (256 × 256 × 6) per well (2048 × 2048 × 6 pixels) → **80 million RxRx3 images fed forward through a single trained model** to have a comprehensive **whole-genome** evaluation $$



Whole-Genome analysis is expensive and labour-intensive, so let's **linearly probe** the model on a small dataset to find *the best layer*.

New task: Anax

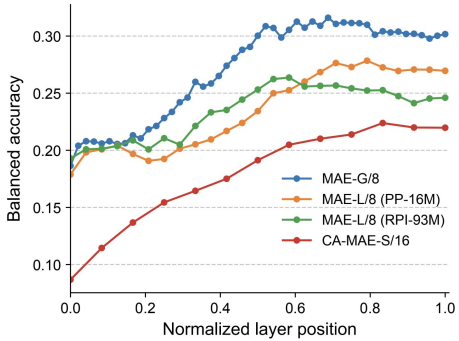
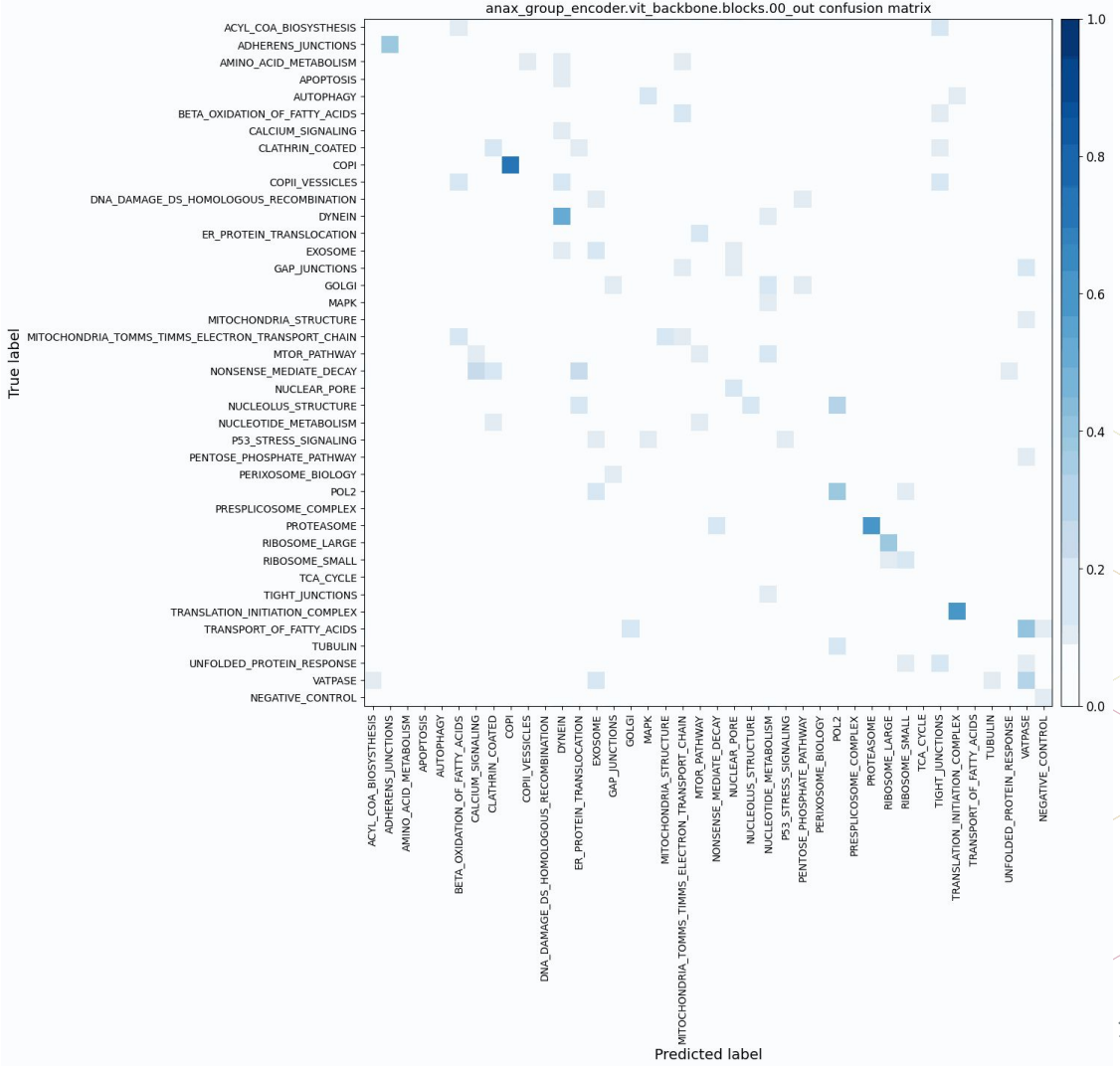
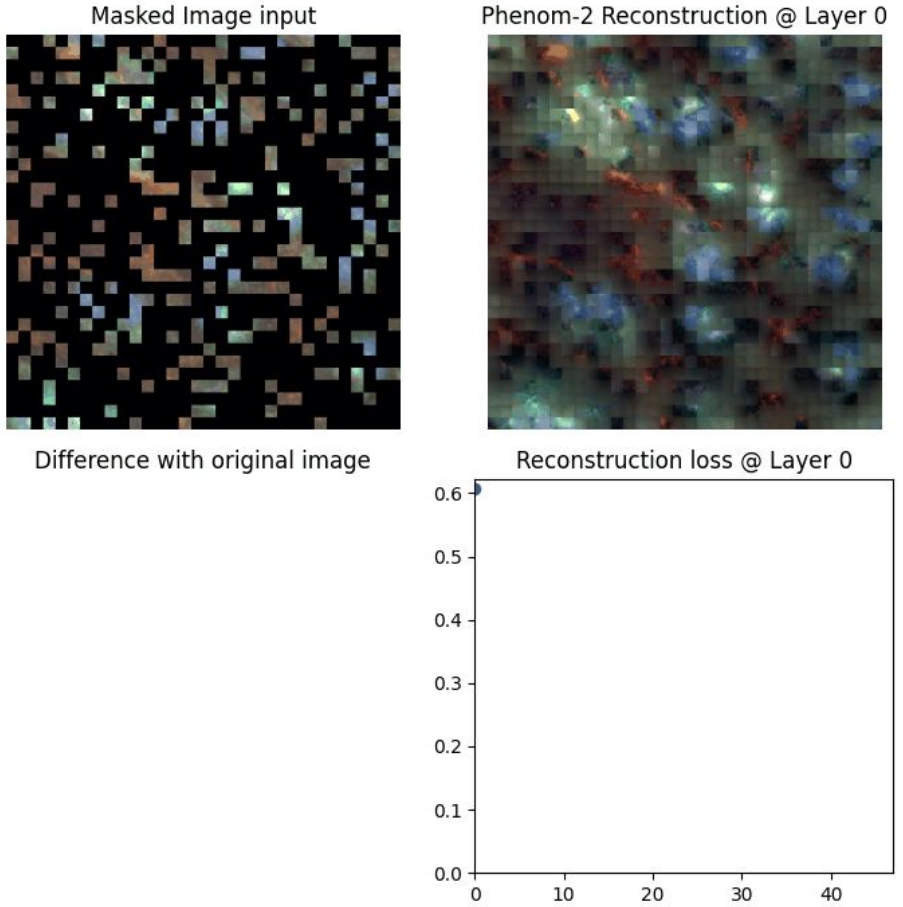
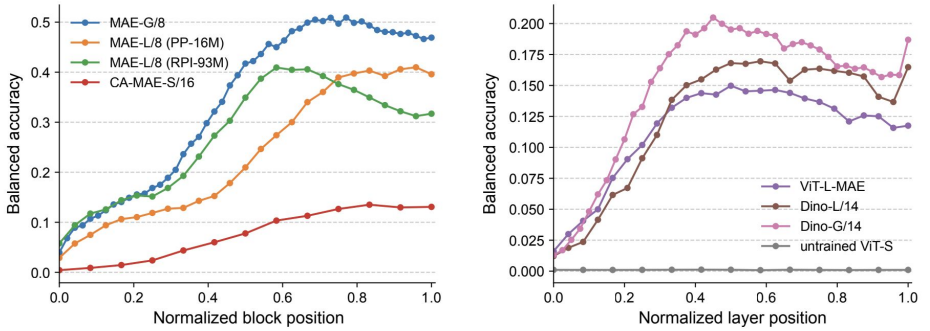


Table 4: Anax groups and their associated genes. This table presents a comprehensive list of gene groups and their corresponding genes.

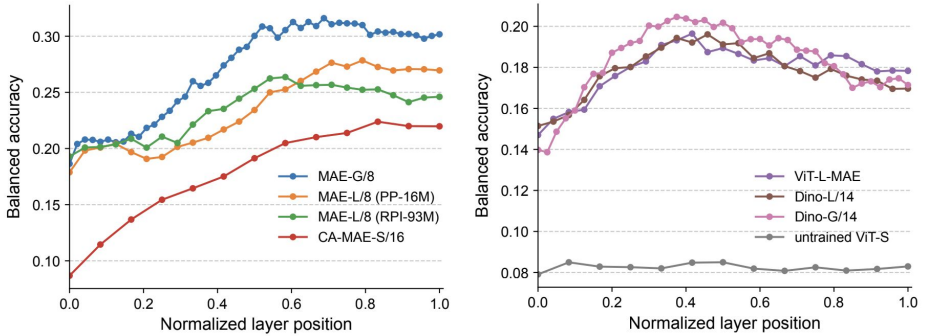
Anax Group	Genes
Acyl Coa Biosynthesis	ELOVL2, ELOVL5, ELOVL6, HACD1, HACD2, HSD17B12, SCD, SCD5, TECR
Adherens Junctions	ACTB, ACTG1, AFDN, CDH1, CTNNA1, CTNNB1, CTNND1, NECTIN1, NECTIN3, NECTIN4
Amino Acid Metabolism	ALDH4A1, ARG2, CKB, CKMT2, CPS1, DAO, OTC, PYCR2, PYCR3, SAT1
Apoptosis	CFLAR, DFFB, 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
Presplicesome 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







(a) RxRx1 siRNA knockdown classification.

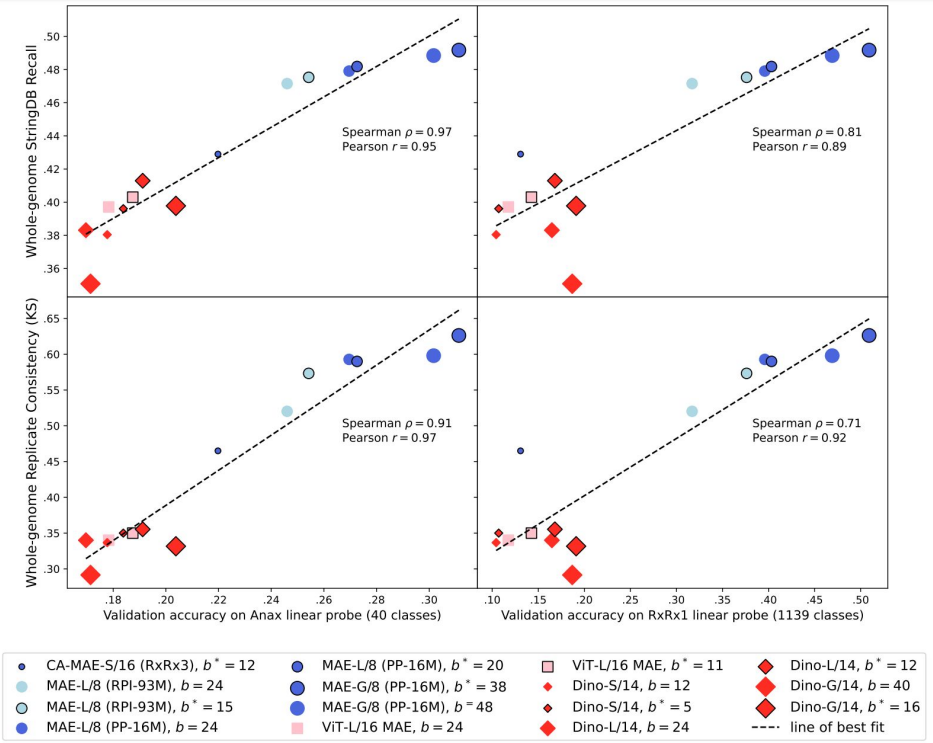


(b) Anax functional gene group classification.

Figure 3: Block-wise validation set **linear probe results** comparing ViT models pretrained on cell microscopy images (left) versus natural images (right). (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).

Final takeaways

- **Curated microscopy data** = awesome
- Nearly all **SSL ViTs** we evaluate (MAEs and Dino-v2 imagenet baselines) are **better at intermediate layers**
- **Linear separability on small datasets strongly correlates** to performance at the **whole-genome scale** and **transfer to new datasets** for these SSL models
- Scaled **MAE to 1.9 billion parameters** is **SOTA** across variety of **newly evaluated benchmarks**



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	38	Phenoprints-16M	**.264**	**.215**	**.165**	**.235**

Table 2: Biological relationship recall benchmarks at 0.05-0.95 cosine threshold on public JUMP-CP image data (Chandrasekaran et al., 2023) generated by completely different labs and assay protocols compared to the data used for pretraining. Each result has a standard deviation $\leq \pm.0023$, and spans nearly 8,000 gene-knockouts and are computed after applying PCA with center-scaling for embedding post-processing alignment.

Recursion

Questions?

kian.kd@recursion.com
oren.kraus@recursion.com
info@rxrx.ai