Towards scientific discovery with dictionary learning

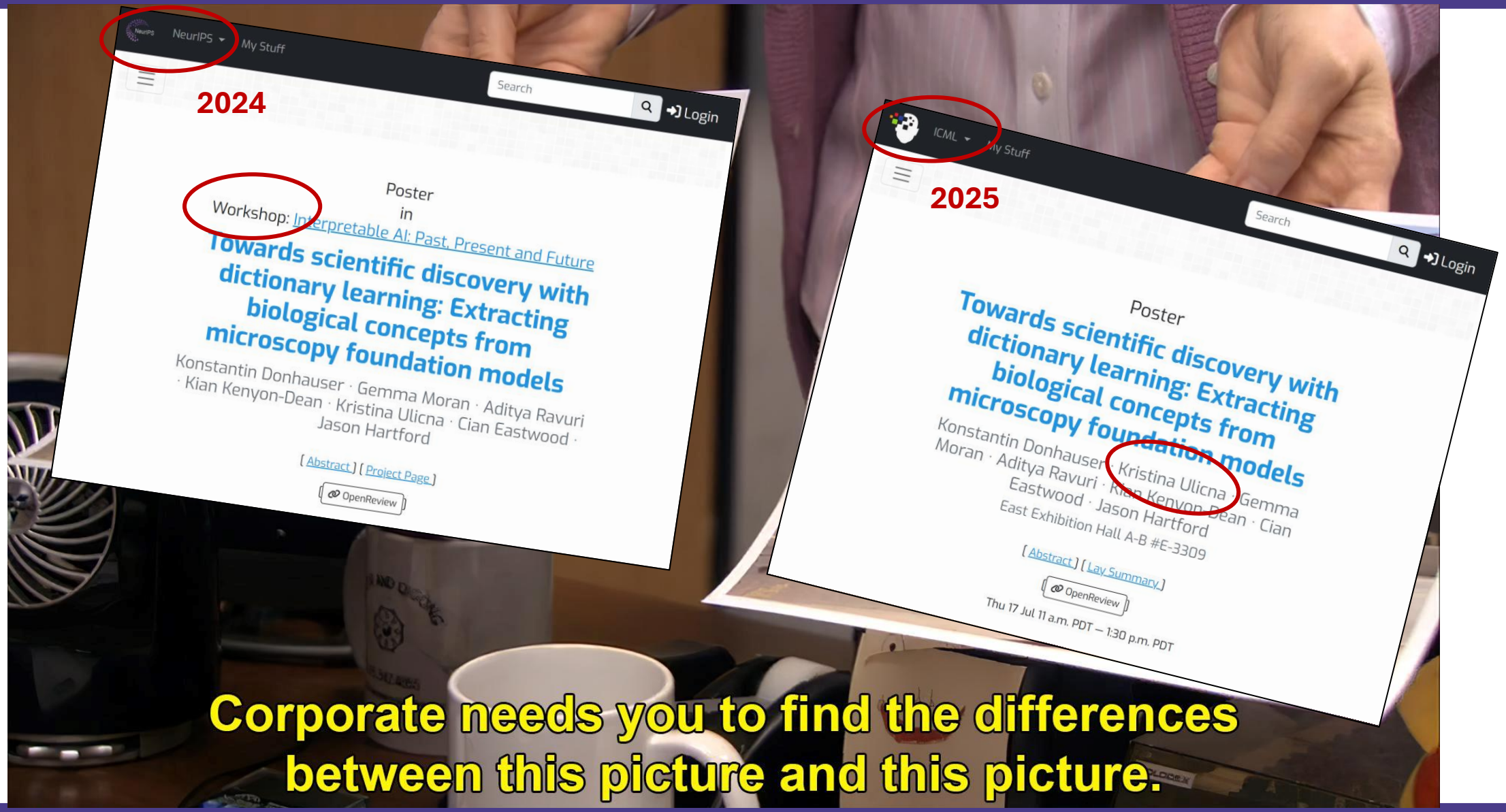
Extracting biological concepts from microscopy foundation models

Bartosz Kochański

13.10.2025

ICML
International Conference
On Machine Learning
2025

The two papers



Authors' affiliations

Konstantin Donhauser [*] [†] [1] Kristina Ulicna [*] [2] [3] Gemma Elyse Moran [4] Aditya Ravuri [†] [5] Kian Kenyon-Dean [3]
Cian Eastwood [2] [3] Jason Hartford [2] [3] [6]

* Equal contribution
† Work done while interning at Valence Labs.
1 ETH Zürich
2 Valence Labs
3 Recursion
4 Rutgers University
5 University of Cambridge
6 University of Manchester.

Valence Labs
Powered by Recursion

Recursion®

About the company

Recursion®

- **BioTech company**
- **Focused on feedback loop between:**
  - **Robotic wet labs**
  - **AI-based drug discovery**

Valence Labs
Powered by Recursion

- **Start-up acquired by Recursion**
- **Currently research center**
- **Launch announced at ICML 2023 (bronze sponsor exhibitor) with $1M scholarship program**
- **ICML 2024: launch of Polaris benchmarking platform**

Authors

Konstantin Donhauser [*†1] **Kristina Ulicna** [*23] **Gemma Elyse Moran** [4] **Aditya Ravuri** [†5] **Kian Kenyon-Dean** [3]
Cian Eastwood [23] **Jason Hartford** [236]



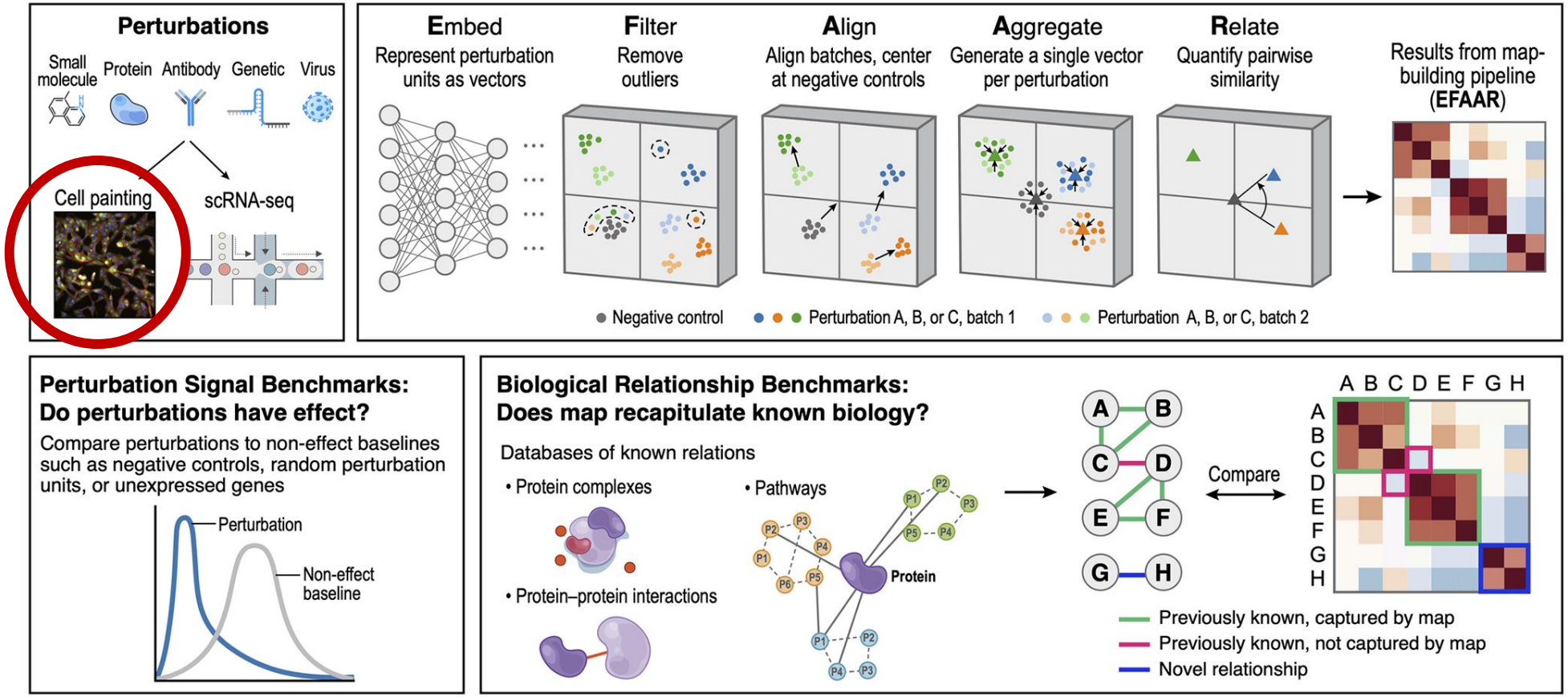




- Causal representation learning
- Active learning
- Postdoc @ Yoshua Bengio
- H: 11, regular pubs start @ 2020

The context

- Modeling how **genetic perturbation** affects **cells and tissues**
- Good model facilitates automatic closed loop of perturbation and evaluation
- This is also why in general causal modelling is core subject and computer vision remains auxiliary



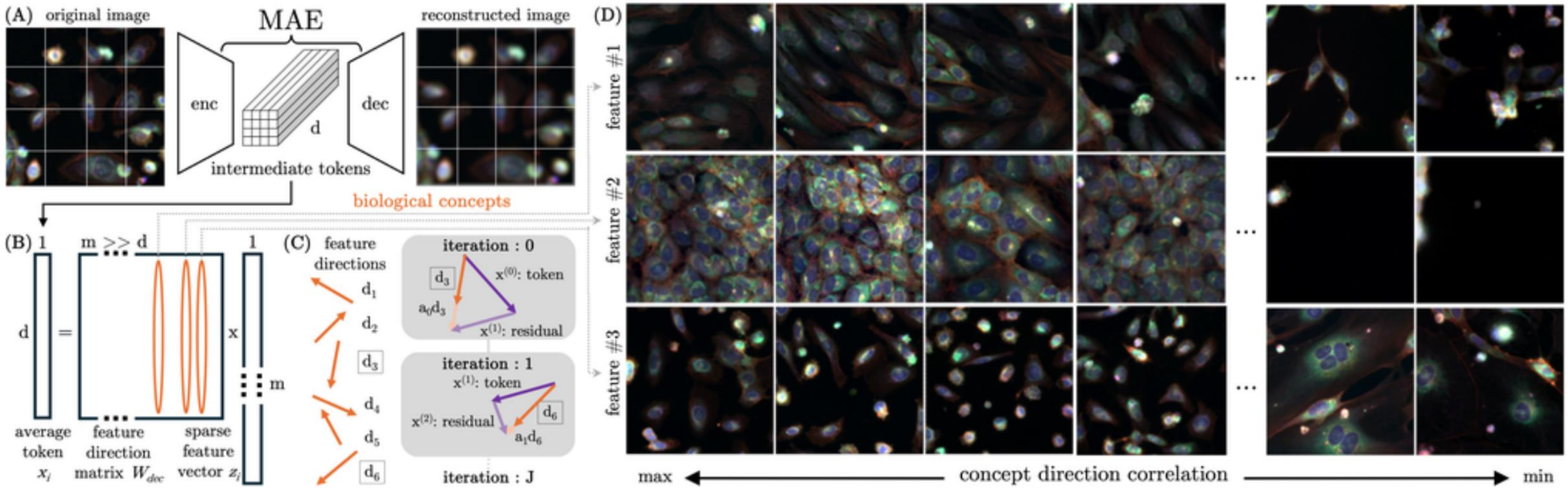
S. Celik *et al.*, "Building, benchmarking, and exploring perturbative maps of transcriptional and morphological data," 2024

Rationale of work - broadly

- **Goal in broader picture:** Generation of interpretable features for assessment of genetic perturbation

- **Means to the goal:** Semantic analysis of vision foundation models developed by Recursion

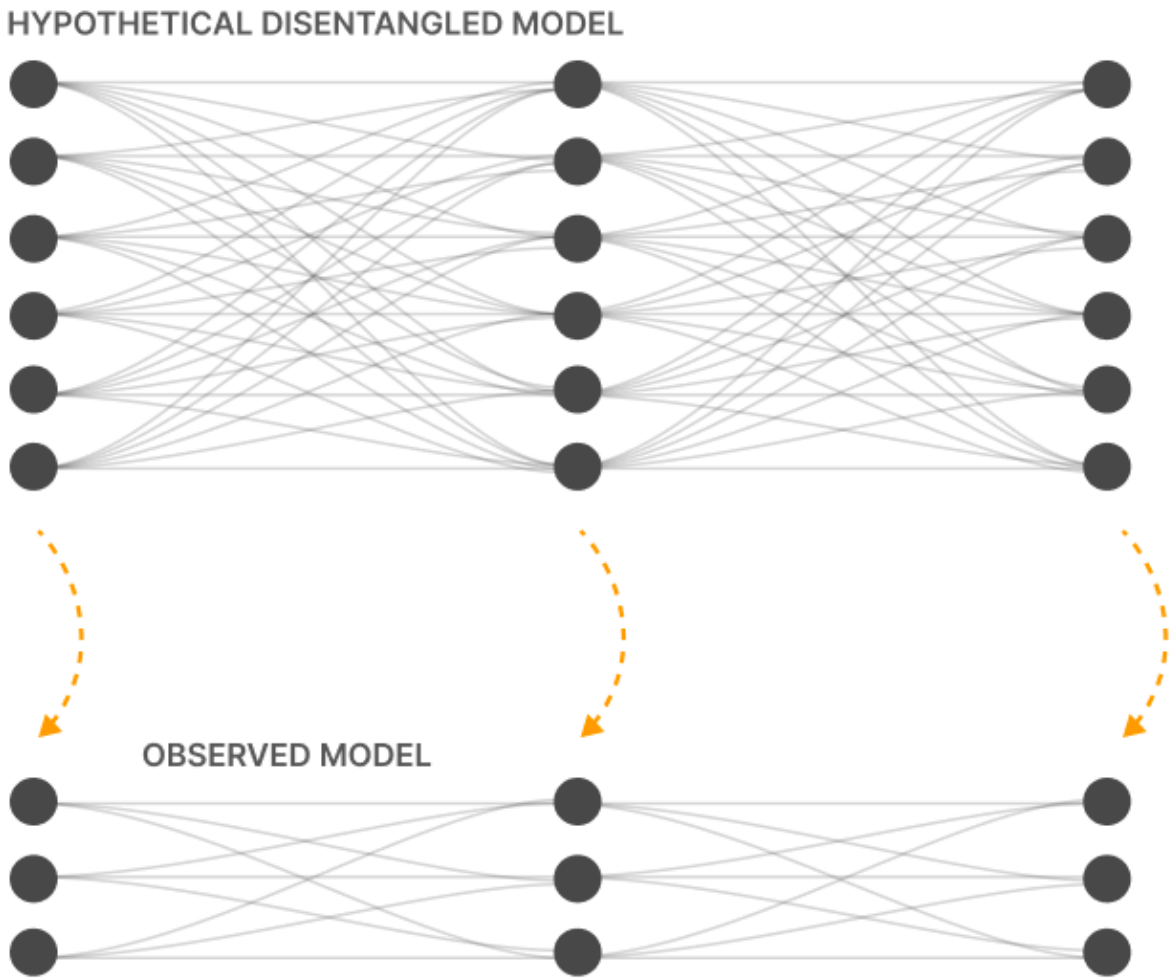
Rationale of work – graphical abstract



Intro – semantic inform. in model internal repres.

- To what extent large ML models encode semantic information about the target domain in their internal representations?
- **Superposition hypothesis** (Bricken et al., 2023)
  - Neural networks encode many more concepts than they have neurons.
  - Thus, one cannot understand the model by inspecting individual neurons.
- How neurons encode multiple concepts at once?
  - Hypothesis: they are low-dimensional projections of some high-dimensional, **sparse feature space**.
  - Literature supports this by:
    - showing that high-level features are typically predictable via **linear probing**.
    - Model representations can be decomposed into human-interpretable concepts using **dictionary learning** models, estimated via **sparse autoencoders** (SAEs)

Intro – superposition hypothesis

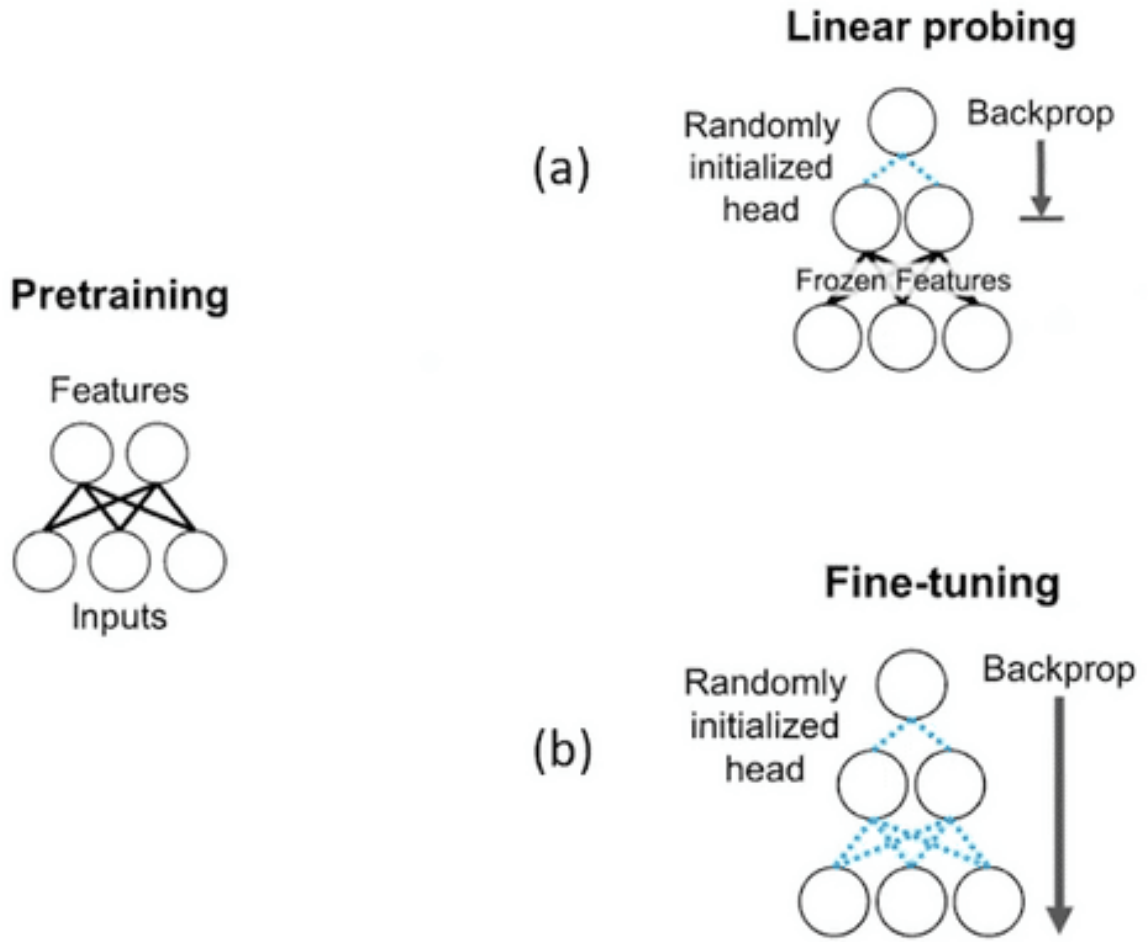


Under the superposition hypothesis, the neural networks we observe are **simulations of larger networks** where every neuron is a disentangled feature.

These idealized neurons are **projected** on to the actual network as "almost orthogonal" vectors over the neurons.

The network we observe is a **low-dimensional projection** of the larger network. From the perspective of individual neurons, this presents as polysemanticity.

Intro – linear probing



Intro – current successes and research gap

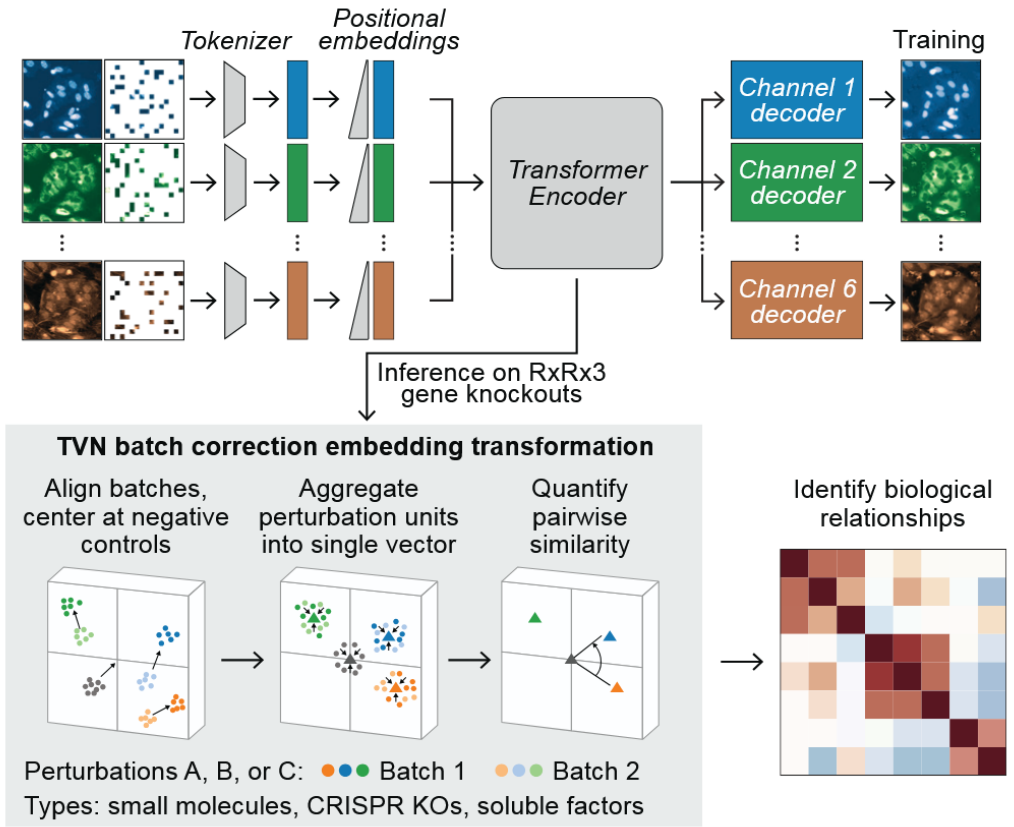
- Successes rely on some form of **text supervision**
- Successes also appear in **„naturally human-interpretable" domains** (e.g. natural images)
- Can we extract similarly meaningful high-level features from completely unsupervised models in domains where we lack strong prior knowledge? (like cell microscopy)

Intro – emerging foundation models for study

Masked Autoencoders for Microscopy are Scalable Learners of Cellular Biology

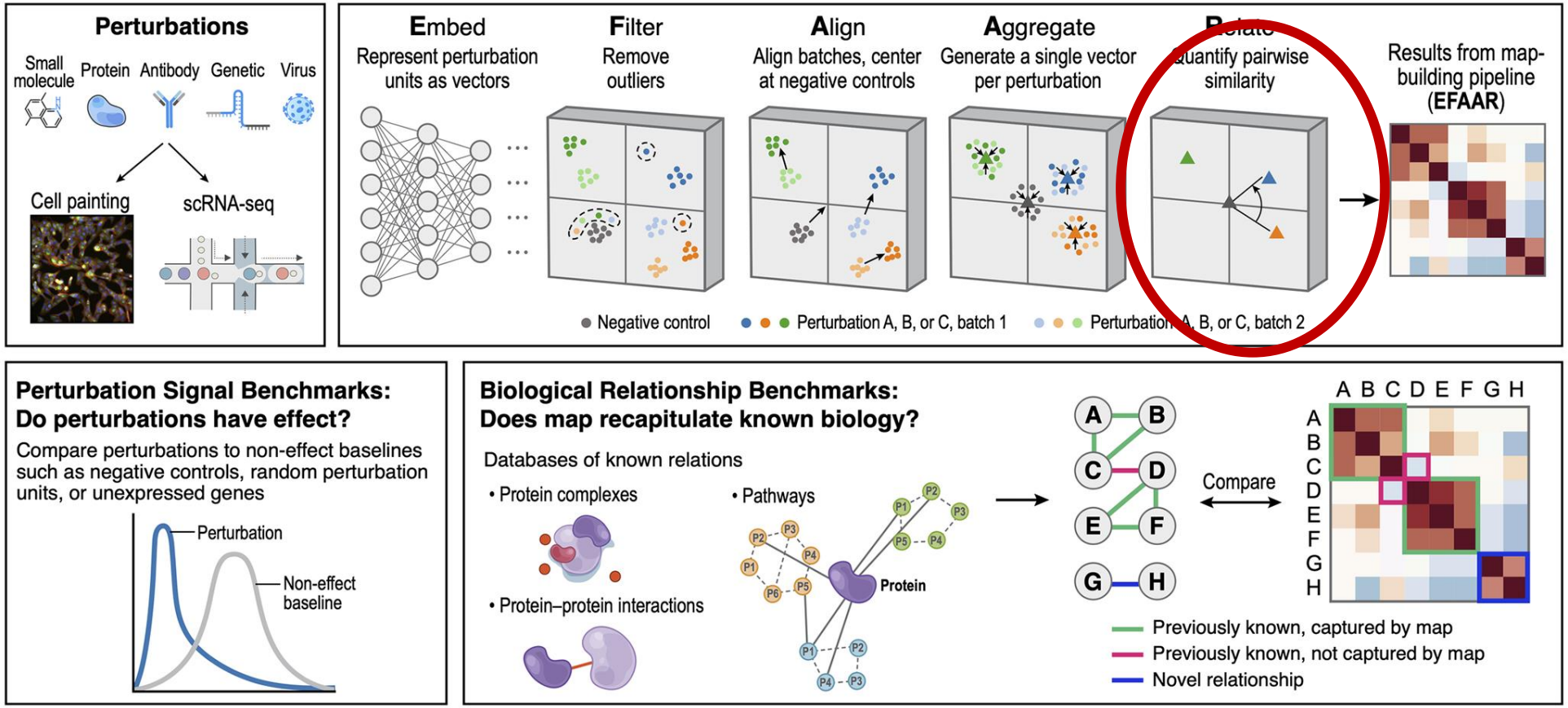
Oren Kraus[1] Kian Kenyon-Dean[1] Saber Saberian[1] Maryam Fallah[1] Peter McLean[1]
Jess Leung[1] Vasudev Sharma[1] Ayla Khan[1] Jia Balakrishnan[1] Safiye Celik[1]
Dominique Beaini[2] Maciej Sypetkowski[2] Chi Vicky Cheng[1] Kristen Morse[1]
Maureen Makes[1] Ben Mabey[1] Berton Earnshaw[1,2]
[1]Recursion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

CVPR 2024 Spotlight



State of the art

- Currently FM representations are used **very coarsely** (clustering, cosine similarity)
- Cosine similarity between representations does **not answer why** they are dissimilar



Research questions and contributions

- **How do we extract sparse features from MAEs?**
  - Iterative Codebook Feature Learning (ICFL)
    - Avoids „dead features”
    - Better reconstrution and more selective features than TopK SAE
- **Do we learn interesting concepts at all?**
  - Yes - some are relatively simple, like light leaks or dead cells, while others capture biologically-meaningful concepts
- **How do we evaluate the biological content of the sparse features?**
  - directly involving domain experts to assess feature quality and relevance
  - we introduce a carefully curated linear probing benchmark to show that ICFL extracts features of a comparable interpretability level to biology-informed hand-crafted features.

Method

- **Superposition hypothesis**

$$x \approx Wz = \sum_{m=1}^{M} z_m W_m \qquad \text{where } \|z\|_0 \ll d \qquad (1)$$

where $W \in \mathbb{R}^{d \times M}$ is a latent dictionary matrix and $z \in \mathbb{R}^M$ is a sparse latent vector. In this paper, we will refer to the columns W_m as "feature directions" and z as "features".

Method

- **TopK SAE**
  - Given a set of representations $\{x\}_{i=1}^{N}$, learning both W and $\{z\}_{i=1}^{N}$ is a dictionary learning or sparse coding problem
  - In the context of mechanistic interpretability, the dominant choice for learning these parameters are two-layer sparse autoencoders (SAEs)
    - One layer – encoder, second – decoder
    - the model for tokens i ∈ [N] is:

$$x_i = W_{\text{dec}} z_i + b_{\text{pre}}, \quad \text{with } z_i = \text{TopK}(W_{\text{enc}} x_i - b_{\text{pre}}),$$

where $\text{TopK}(\cdot)$ is an operator that sets all but the K largest elements to zero. The parameters $\{W_{\text{dec}}, W_{\text{enc}}, b_{\text{pre}}\}$ are learned by minimizing the reconstruction loss:

$$L(W_{\text{dec}}, b_{\text{pre}}) := \sum_{i=1}^{N} \|x_i - \widehat{x}_i\|_2^2, \quad \text{where} \qquad (2)$$

$$\widehat{x}_i = W_{\text{dec}} \text{TopK}(W_{\text{enc}} x_i - b_{\text{pre}}) + b_{\text{pre}}. \qquad (3)$$

Method

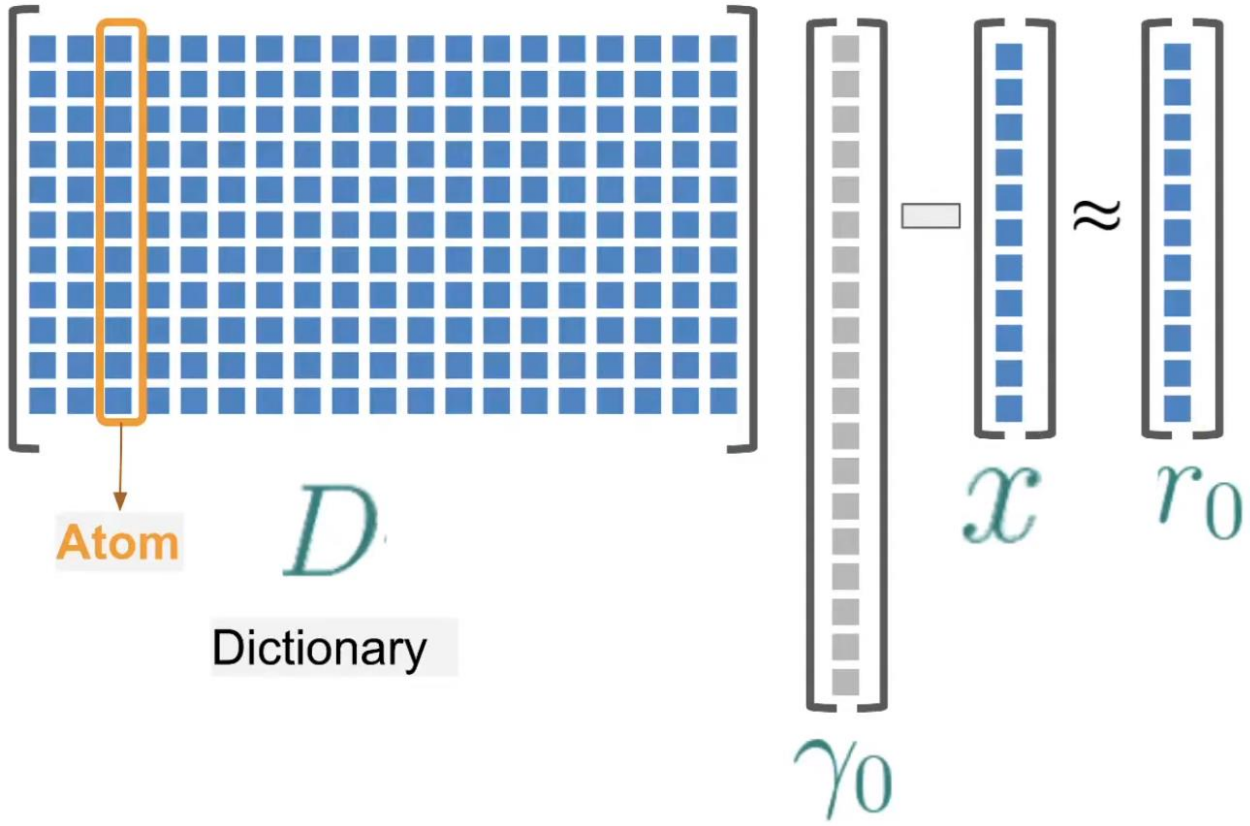
- **OMP – Orthogonal Matching Pursuit**
  - It assumes that the dictionary, W , is given, and after initializing $x^{(0)} = x$ it iteratively finds a k-sparse vector $\hat{z}$ by repeating the following steps:

1. Find the column in W that solves $w_i = \arg\max_i |W_i^\top x^{(t)}|$ and append it to the matrix W_s of selected columns.
2. Solve $z^{(t)} = \arg\min_z \|x^{(t)} - W_s z\|^2$
3. Update $x^{(t+1)} := x^{(t)} - W_s z^{(t)}$

and terminate when either $x^{(t+1)}$ is sufficiently small (i.e., the algorithm has converged) or $z^{(t)}$ is k-sparse for some pre-specified k.

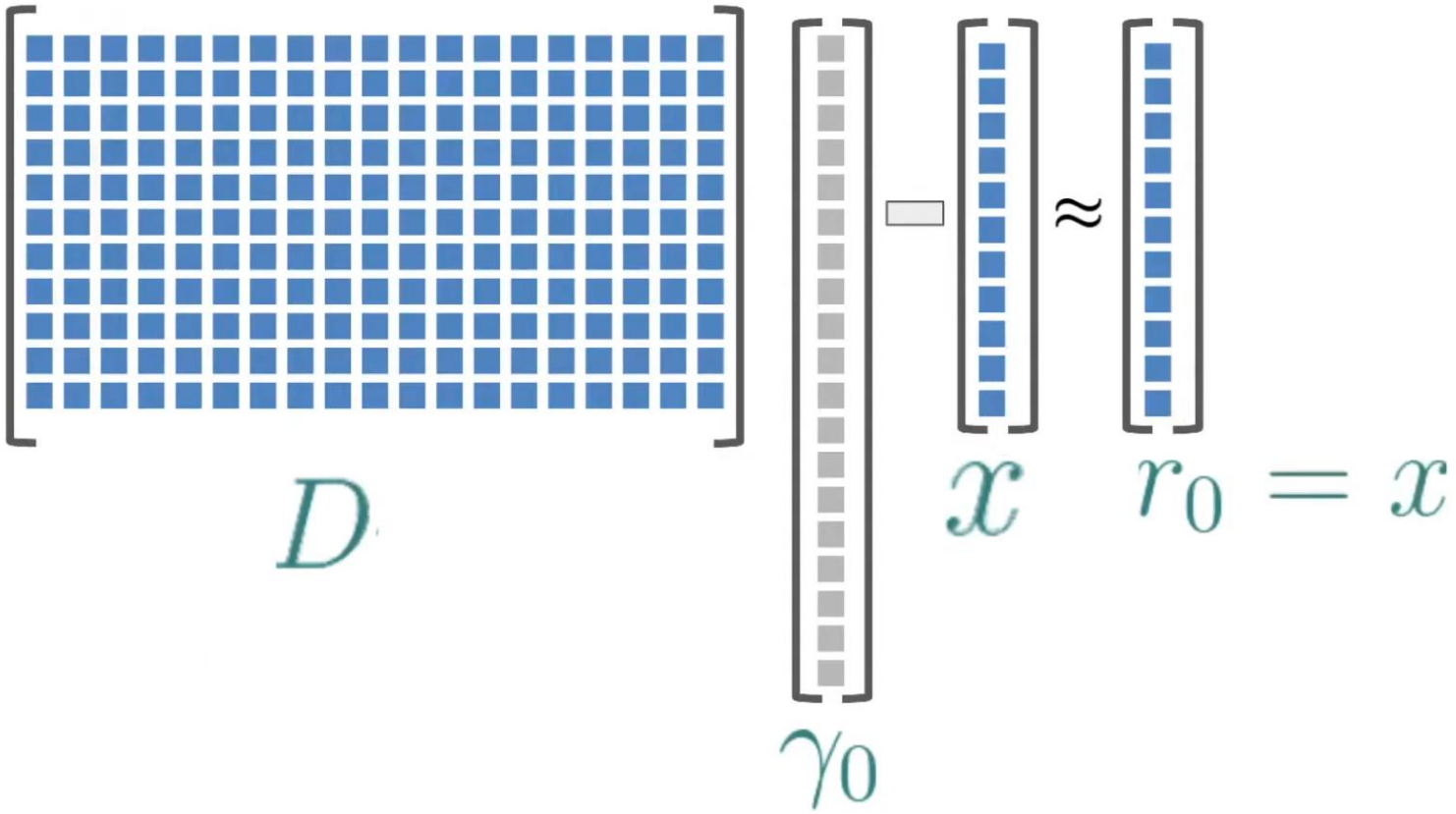
Method

- **Matching Pursuit**



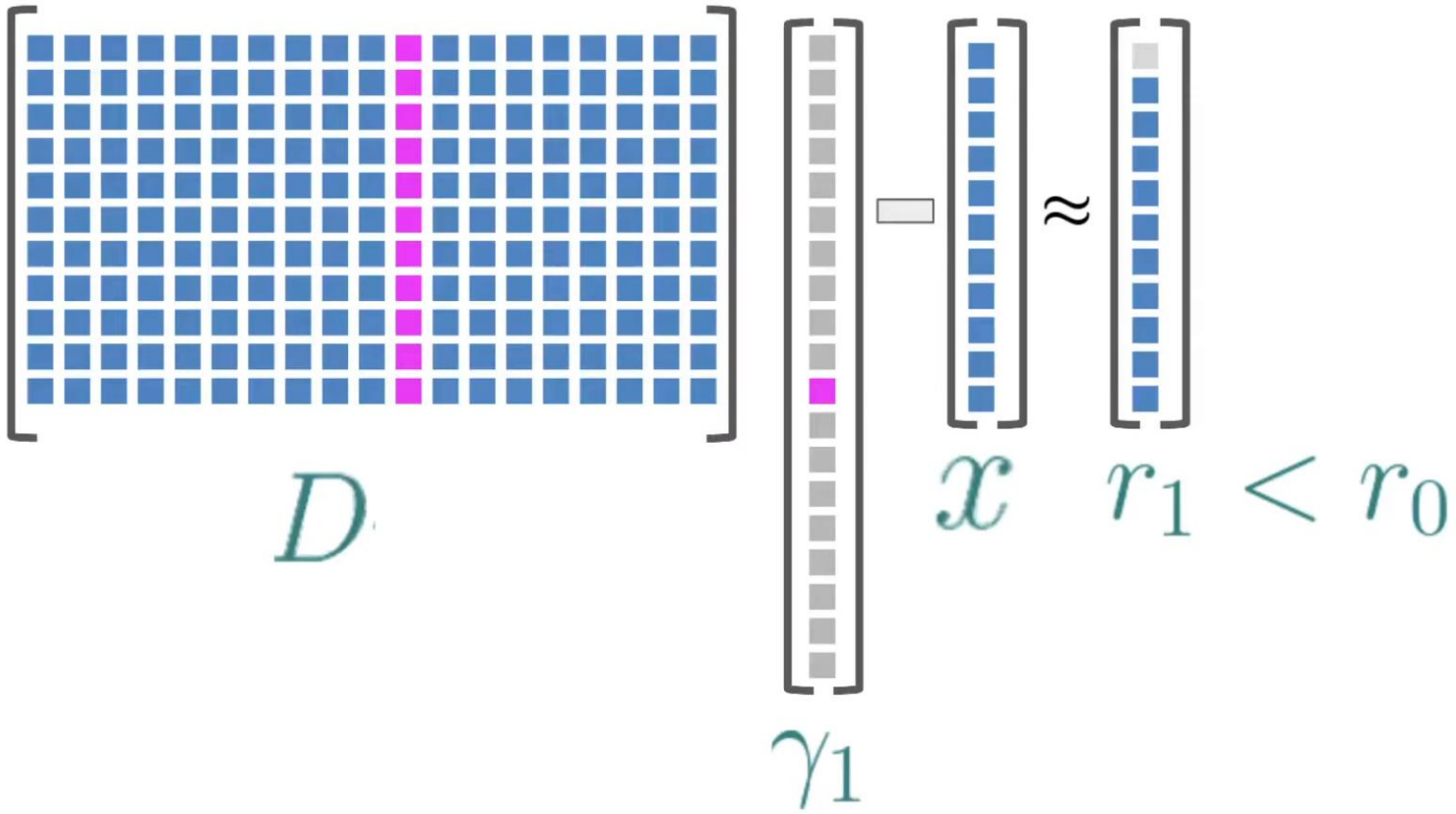
Method

- **Matching Pursuit**



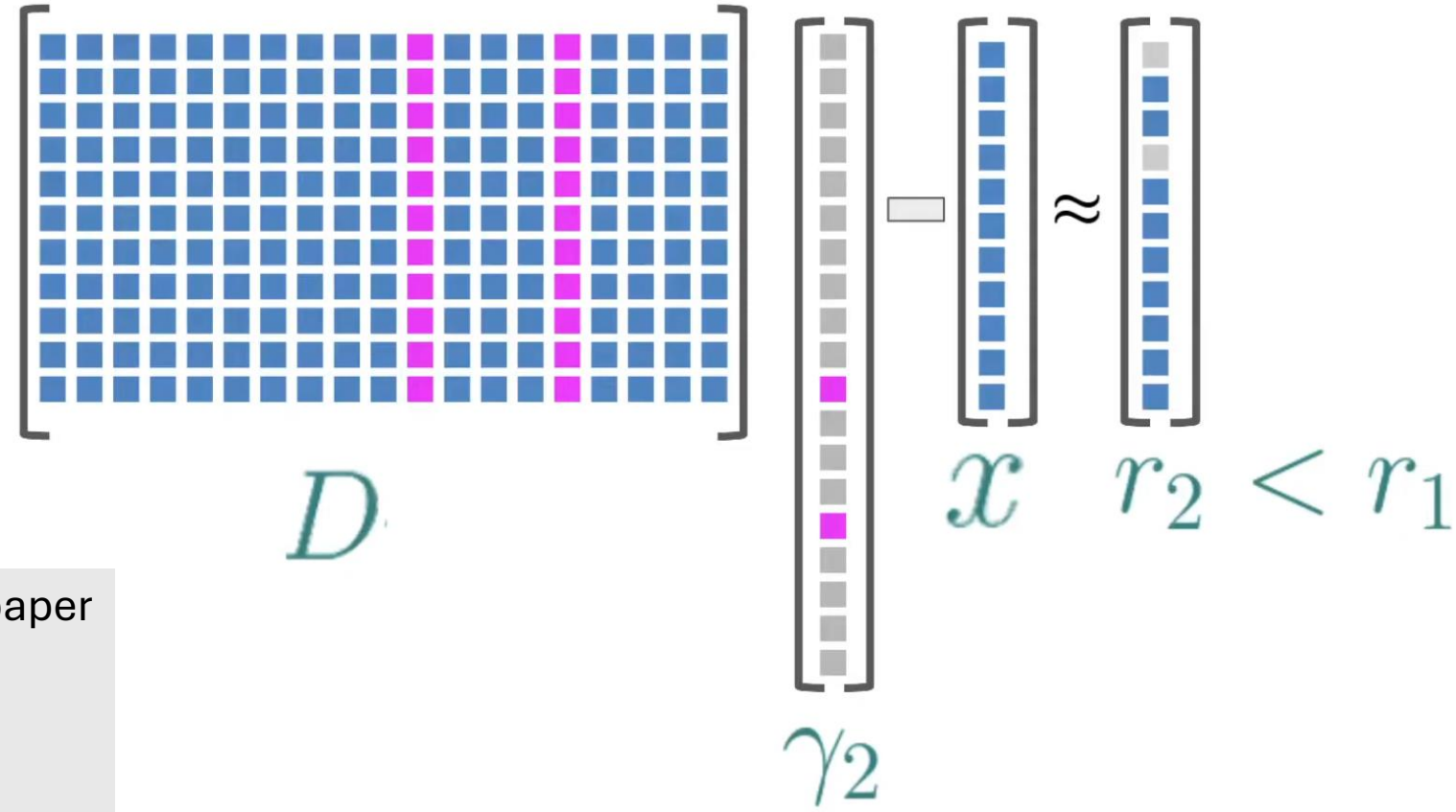
Method

- **Matching Pursuit**



Method

- **Matching Pursuit**



Side-note: further in the paper authors use this notation:

$D \rightarrow D_{dec}$

$y \rightarrow z$

$x \rightarrow x$

Method

- **ICFL – Iterative codebook feature learning**
  - Treat OMP as encoder in TopK SAE
    - Without pre-specified dictionary
      - Use W_{dec} instead
    - Batched instead of sequential
- Advantages over SAE:
  - early iterations subtract dominant feature directions from x
  - later iterations may select a broader set of feature directions that are not as correlated with the main features in x.

Algorithm 1 Batched-OMP

Input: Parameters $W_{\text{dec}}, b_{\text{pre}}$; model representation x; # sparse features L per iteration and # iterations J
Initialize $x^{(1)} := x - b_{\text{pre}}$
for $t = 1$ TO J **do**
  Select top L columns of W_{dec} which maximize $\langle W_{\text{dec},m}, x^{(t)} \rangle$
  Solve $z^{(t)} = \text{argmin}_z \|x^{(t)} - W_{\text{dec}} z\|_2^2$ with z non-zero only for L selected columns
  Update $x^{(t+1)} := x^{(t)} - W_{\text{dec}} z^{(t)}$
end for
Output: Sparse features $z := \sum_{t=1}^{J} z^{(t)}$

Method

- **ICFL – Iterative codebook feature learning**
  - After encoder step (optimizing z):
    - update the decoder parameters {W_{dec}, b_{pre}} via batched gradient descent minimizing the reconstruction loss:

$$L_{\mathrm{ICFL}}(W_{\mathrm{dec}}, b) := \sum_i \|x_i - W_{\mathrm{dec}} z_i - b_{\mathrm{pre}}\|_2^2. \qquad (4)$$

  - W_{dec} – standard initialization for a linear transform
  - Random resets to ensure that the columns of W_{dec} are not too correlated
    - after every 100 stochastic gradient descent steps, one of every pair of columns of W_{dec} that have cosine-similarity above 0.9
  - Before running OMP representations *x* are centered by subtracting the average representation of unperturbed samples
    - Thus - the origin represents the unperturbed state.
  - Before ICFL – normalize the representations to have unit norm.

Method

- **PCA whitening**
  - Target of learning is minimization of reconstruction loss of model representations
  - Thus learned features *z* are biased towards dominant directions
  - Dominant directions typically correspond to information common for all experiments (e.g. cell-cycle changes)
  - Solution:
    - Apply transformation **T** to experimental data before normalization, where:

$$\mathbf{T} : \mathbf{X} \rightarrow W(\mathbf{X} - \mu)$$

      - W – PCA matrix with downweighted dominant directions
      - X – matrix of representations of n control samples

Experimental setup - general

- **Foundation models:** 2 large MAEs trained on data from multiple cell types that were perturbed with CRISPR gene knockout perturbations
  - ViT-L/8 (MAE-L)
  - ViT-G/8 (MAE-G) - bigger
- **Data:** 256x256x6 microscopic images (1 data point = 1 „crop”)
- **Representations:**
  - Single token per input crop gained by aggregating all patch tokens.
  - Dimensions of tokens (representations) are 1024 for MAE-L and 1664 for MAE-G
  - Tokens extracted from layers 16 (MAE-L) and 33 (MAEG)
    - More likely to capture abstract high level concepts that are used by the model **internally** to solve the self-supervised learning task
    - These layers have been selected by finding the **maximized linear probing performance** on the functional group task from the original embeddings

Experimental setup - probing

Task	Classes	Samples
Cell type	23	110,971
Batch	272	80,000
siRNA	1,138	81,224
CRISPR	5	79,555
Group	39	57,863

- Probing based on 5 classificantion tasks
  1. 23 different cell types
  2. 272 different experiment batches
  3. 1138 siRNA perturbations
  4. 5 single-gene CRISPR perturbation knockouts which induce strong and consistent morphological profiles across cell types, known as "perturbation signal benchmarks„
  5. 39 functional gene groups composed of CRISPR single-gene knockouts
- Tasks 3 and 5 are non-trivial
- Probing done on raw and sparse features

Experimental setup – Dict. Learn. model training

- Sparsity of K = 100 for TopK SAEs and J = 20, L = 5 (resulting in a max sparsity of 100) for ICFL
- M = 8192 features
- apply the PCA whitening
- use representations from the residual stream
- train the models using 40M tokens (one token per image crop) with a batch size of 8192 for 300k iterations
- learning rate is 5 × 10−5

Results: Features are correlated with biological concepts

Two approaches:

1. Accuracy of linear probing
   - Due to heavy class imbalance (particularly for Task (1)), we train our linear probes using logistic regression on a class-balanced cross-entropy loss and report balanced test accuracy (BTA)
2. Selectivity
   - For each feature two metrics are computed:
     - **avg selectivity** score, which is the % of times that the feature is active given that label i occurs minus the % of times the feature is active given any other label
     - **max selectivity** score, that subtracts the maximum % for any other label

Results: Features are correlated with biological concepts

Task	Classes	Samples	BTA	Threshold 0.5	Threshold 0.2
Cell type	23	110,971	97.2%	73	455
Batch	272	80,000	87.8%	11	77
siRNA	1,138	81,224	51.6%	0	141
CRISPR	5	79,555	94.6%	0	2
Group	39	57,863	32.1%	0	37

- Threshold = number of features exhibiting an average selectivity greater than a given threshold for at least one label.
- Dominant concepts, such as cell types, batch effects, and siRNA perturbations that induce strong morphological changes, comprise a substantial portion of features displaying high selectivity

Results: Feature directions vs raw representations

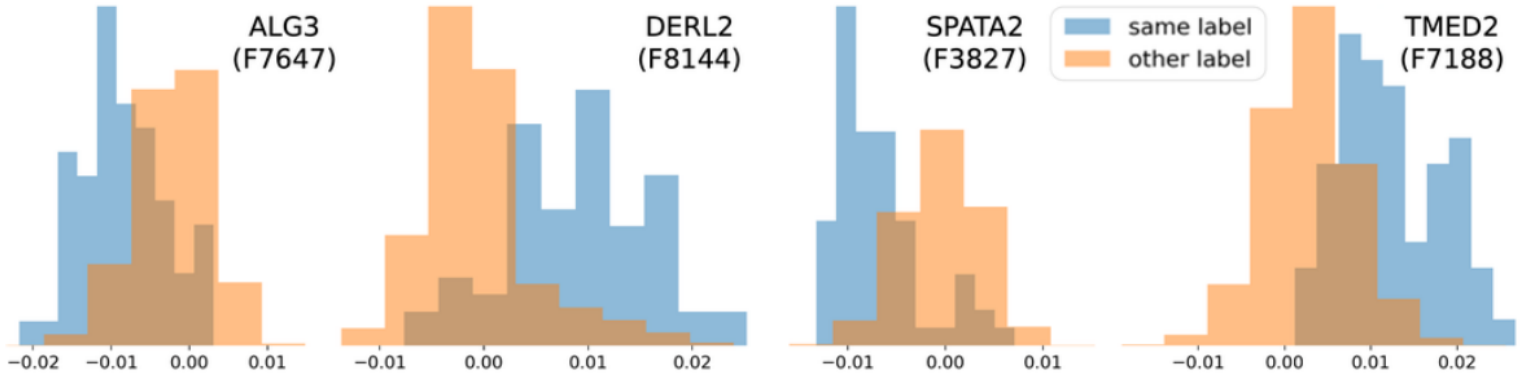


Figure 3. **Cosine similarity histograms for selected pairs of representations** from Task (3) and features directions, if its associated perturbation is applied (blue) *vs.* any other perturbation (orange).

Results: ICFL vs TopK SAE and raw representations

How much relevant biological information is lost when chooding sparse features?

How sparsity and PCA influence this?

What is general reconstruction quality?

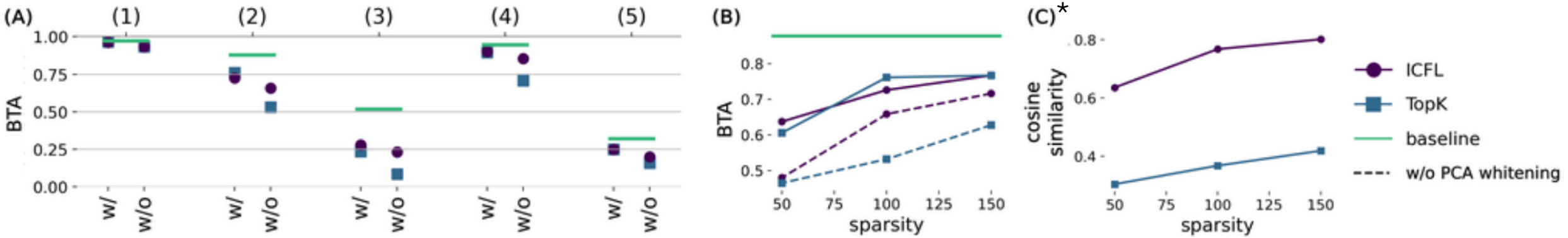


Figure 4. **Comparison of ICFL with TopK SAEs. (A)** Balanced test accuracy (BTA) of linear probes trained on the original representation (solid line) and reconstructions from ICFL and TopK SAEs with and without PCA whitening for five tasks from § 5. **(B)** Balanced test accuracy (BTA) as a function of the sparsity (dashed line is the original representation x) for classification Task (5). **(C)** Cosine similarity of reconstruction and original representations as a function of sparsity for tokens from a hold-out validation dataset.

* On Figure 4C the w/o PCA condition probably yielded better results for TopK

TopK vs ICFL – selectivity and dead features

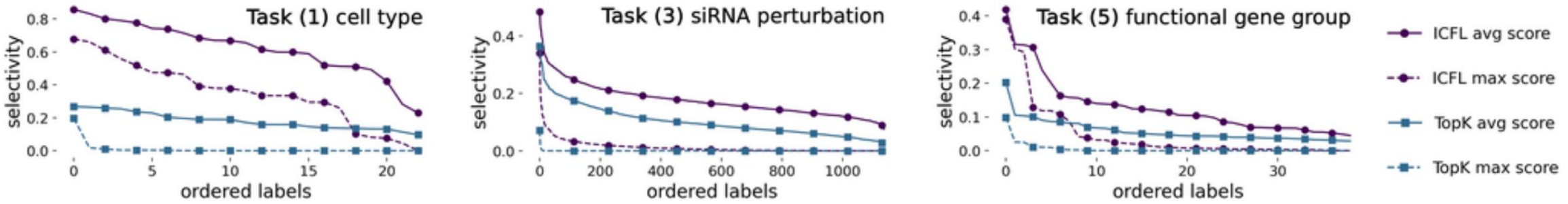


Figure 8. **The highest selectivity scores among all features for each label**, ordered separately for each line starting with maximum score.

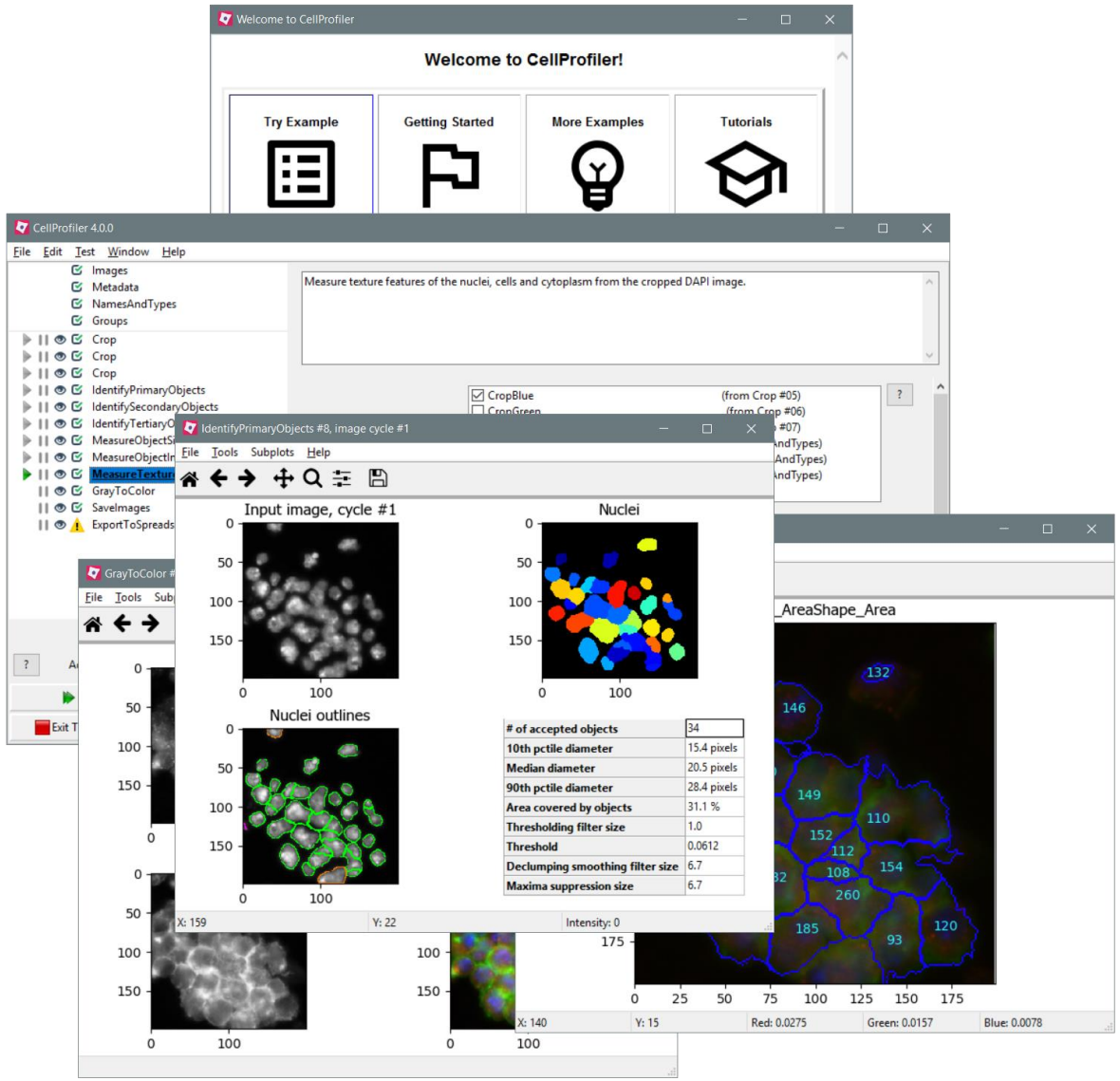
	w/o PCA whitening	w/ PCA whitening
ICFL	55	341
TopK	7640	8026

Table 1. The number of "dead features" (out of 8192) activated less than a fraction of $10^{-5}\times$ during the last 1000 training steps, for both TopK and ICFL with and without PCA whitening (§ 5).

Results – Cell Profiler features

CellProfiler™
cell image analysis software

- CP extracts features for each cell;
  - for a multi-cell image, we calculate an image-level CP feature by averaging the CP features from each cell in the image.
- The CP features are not sparse;
  - to make them comparable with the sparse ICFL features, we thresholded the CP features at α and 1 − α quantiles, with α chosen such that the average number of non-zeros was ≈ 100.
  - A CP feature was considered to be “activated" when its value, under perturbation conditions, exceeded these quantiles.



Results – Comparisons with Cell Profiler

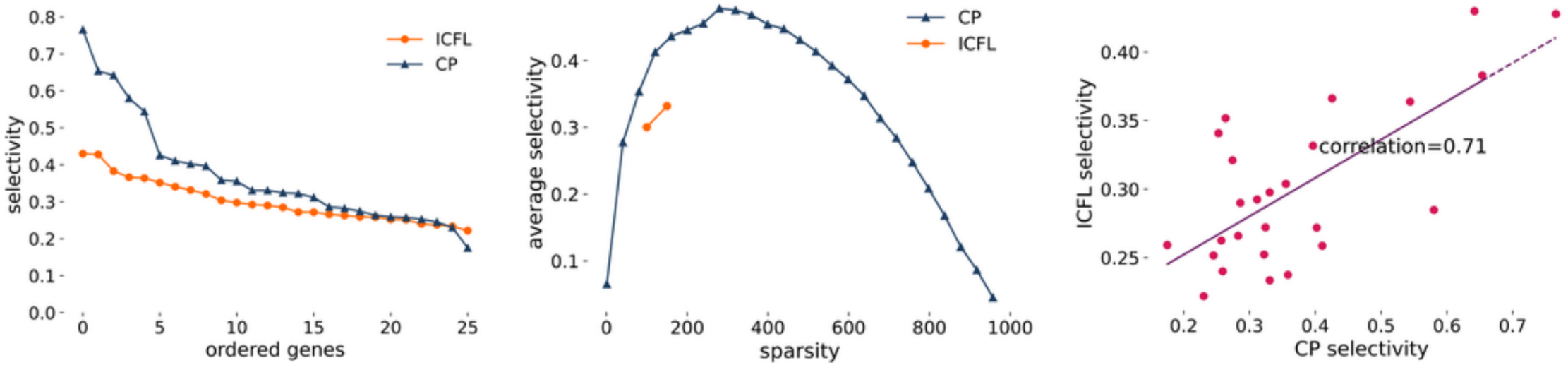
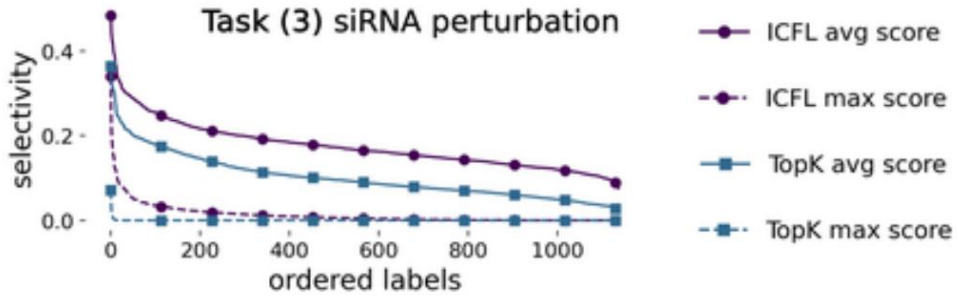


Figure 2. **Comparison of feature average selectivity scores** from *CellProfiler* (CP) and ICFL for a subset of Task (3). **(A)** Max avg selectivity scores for each label in descending order, **(B)** averaged across labels at different thresholds for CP and sparsity levels for ICFL, as a function of the average number of non-zero values. **(C)** Correlation of max avg selectivity scores for each label between CP and ICFL.



Interpretability analysis of cell imaging data

- We illustrate **non-trivial patterns captured by selected features**, exemplifying how domain experts can interpret and validate biological concepts learned by DL.
- To understand what parts of each image are most “aligned” with a particular feature, we generate **heatmaps for each image** by calculating the **cosine similarity of the feature direction** with each of the **8 × 8 image patch token representations**.

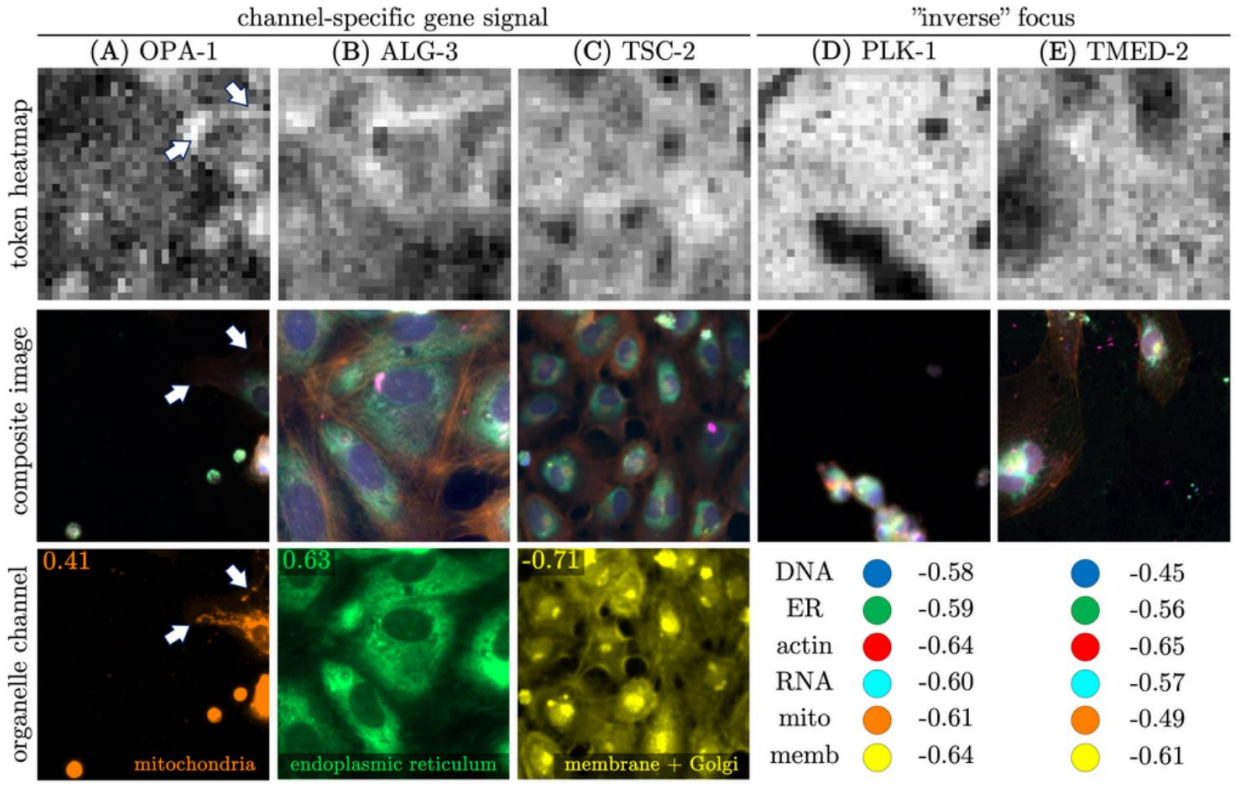
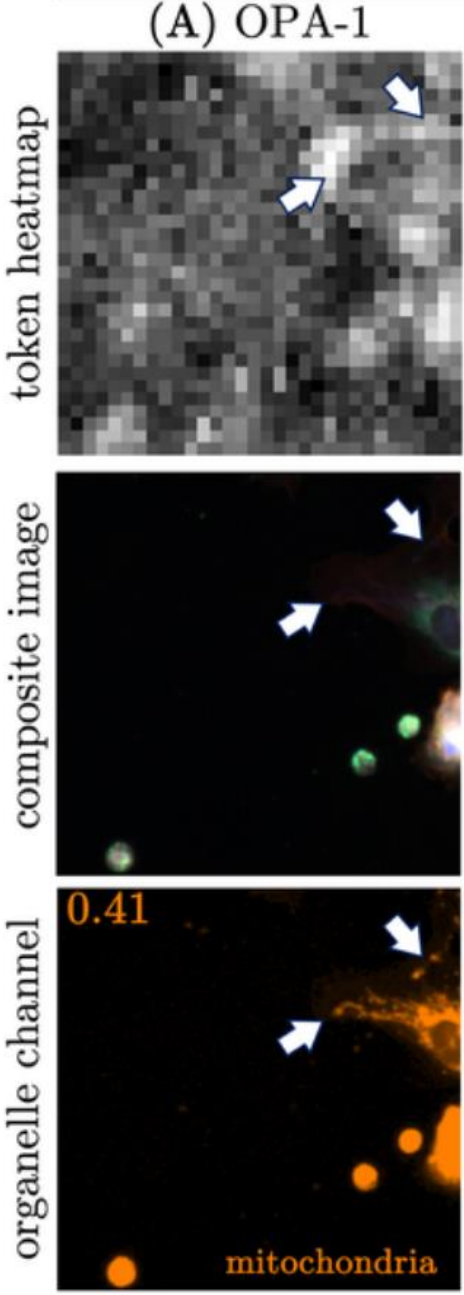


Figure 5. **Channel-specific features visualization** from selected single-gene perturbations. Token heatmaps (top) are plotted above the composite 6-channel images (middle) and channel-specific staining images (bottom) of selected subcellular compartments, along with the channel-heatmap Pearson correlation coefficients.

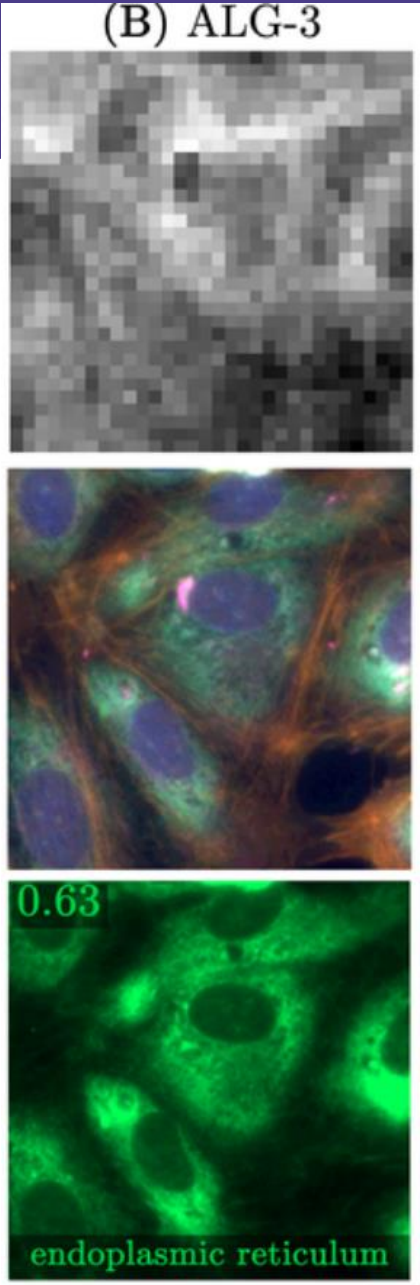
Interpretability analysis of cell imaging data

OPA-1

- The **mitochondrial channel** shows that most correlated tokens are overlaid with distant regions where enlarged mitochondria are present (white arrows).
- Qualitative channel examination highlights this delicate detail which is not obvious from the composite images, confirming that our approach identifies channel-specific level of detail.



Interpretability analysis of cell imaging data



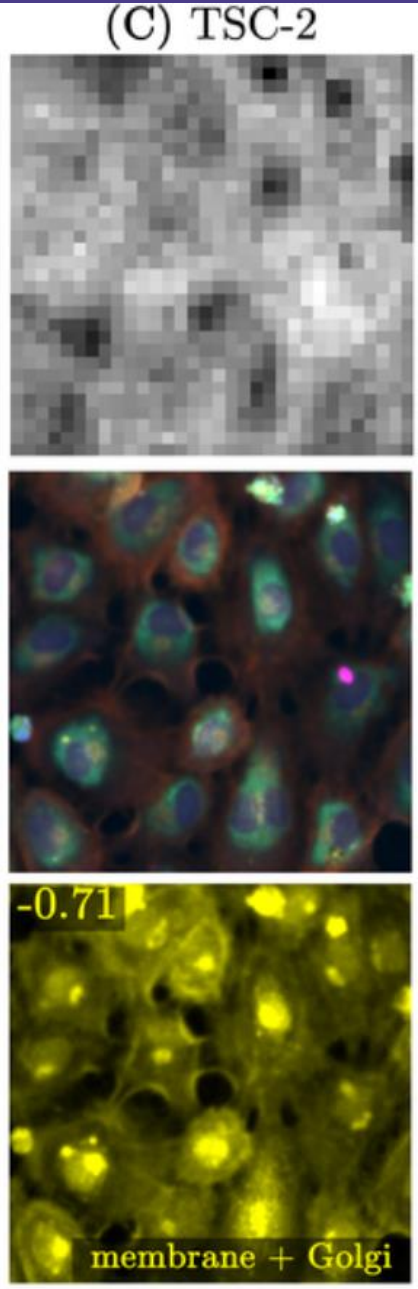
ALG-3

- The most aligned tokens appear specific to regions of **endoplasmic reticulum (ER)** and RNA with which ALG3 co-localizes.
- This dense image suggests that our token heatmap is prevalently focused on ER-specific information, which is consistent with ALG-3 function in aiding the attachment of sugar-like groups to proteins.

Interpretability analysis of cell imaging data

TSC-2

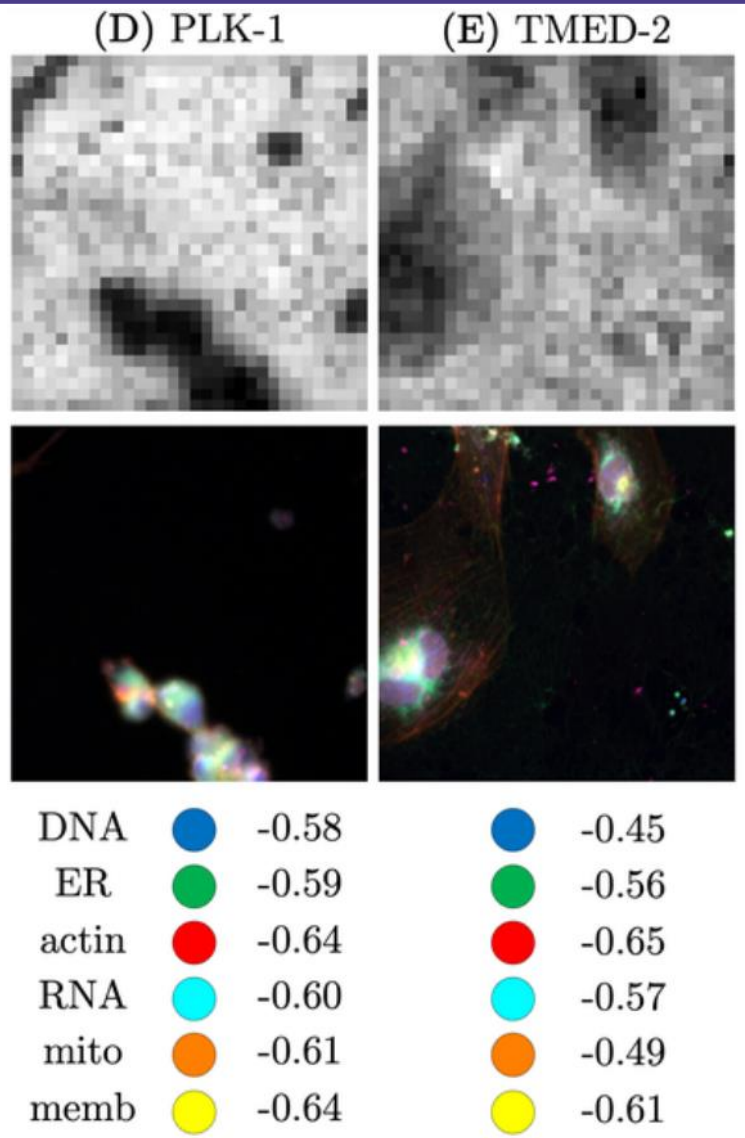
- We examine the **plasma membrane- and Golgi apparatus-specific channel** to relate perturbed cell size control to the token alignment.
- We confirm that this channel correlates most strongly with the queried concept direction, but this time in a **negative direction**. As the plasma membrane — and, hence, cytoplasmic area — are the most extensive from the cell center, the mostly aligned tokens appear to focus specifically on regions which are not covered by the cell membrane.
- Although this behavior is harder to interpret, it is suggestive of that the salient feature for these perturbations is the lack of cell density in a well.



Interpretability analysis of cell imaging data

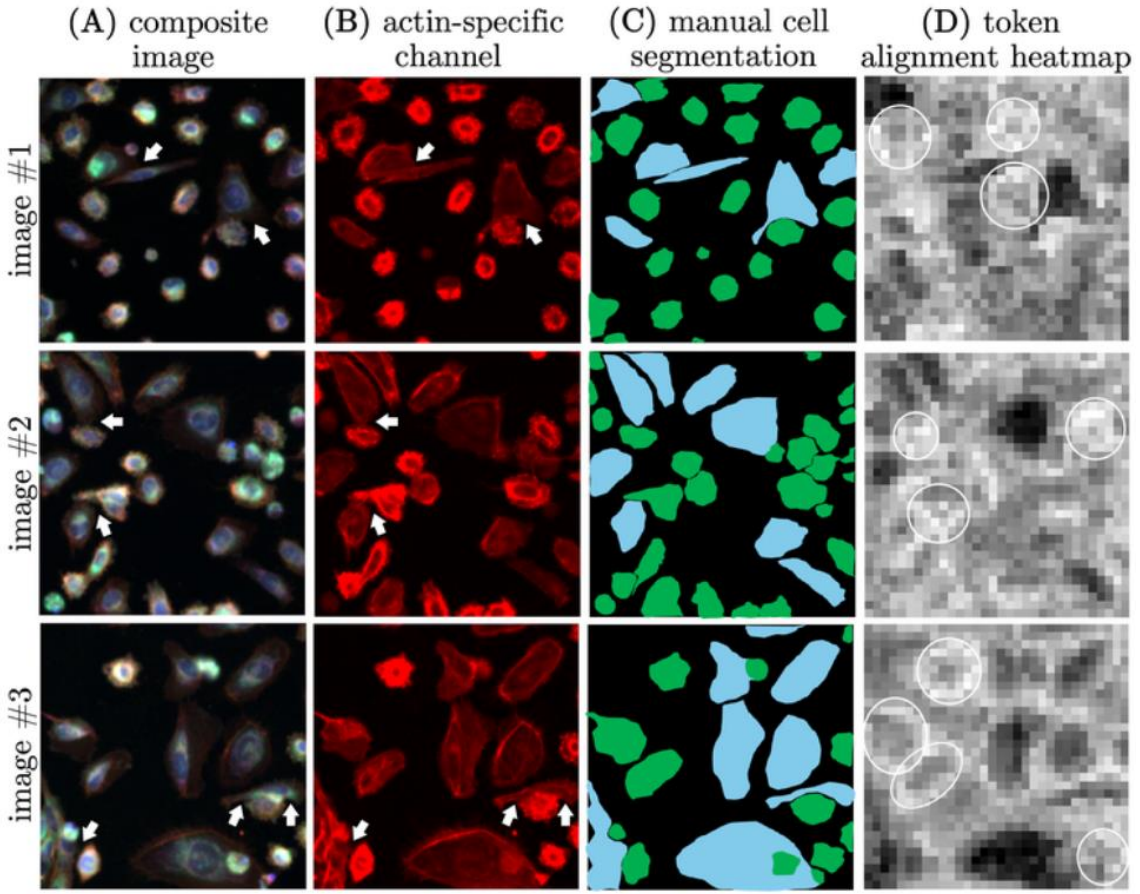
Inverse focus

- Monitoring of the lack of channel-specific signal is a plausible explanation in other images where the tokens are not always co-localized with regions occupied by cells.
- Several genes appear to follow an “inverse” trend, including **PLK-1**, which enables cell cycle progression through mitosis, and **TMED-2**, which helps to regulate intracellular protein transport.
- While both of these gene perturbations render the cells in a characteristic affected state (small, clumped cells struggling to divide vs. large, spread out, actively dividing cells), their most aligned tokens correspond to areas not covered by cells, confirmed by highly negative correlations across all channels.



Single-cell level interpretability analysis

- Showcase of feature strongly correlated with gene knockouts from the *adherens junctions* pathway (label from Task 5)
- The images most correlated with this feature direction reflect this disrupted cellular morphology
  - small, bright and isolated cells
  - Some cells not affected due to CRISPR imperfections
- As visible on alignment heatmaps:
  - “dark” areas consistently belong to cells whose actin meshwork extensively protrudes away from the cell center
  - „bright” regions (most correlated with the feature direction) belong to areas surrounding the perturbed, compact cells and appear to form a ring-like pattern around them



Single-cell level interpretability analysis

- Selected 5 cell images from the adherens junctions pathway
- Domain expert manually counted the cells, classified each cell instance as (un-)perturbed, and manually segmented them.
- The corresponding token-level heatmaps were then examined to annotate cell instances by whether they align with the overall feature direction (‘bright’) or not (‘dark’).

Image	**Manual Labeling**			**ICFL Recall**	
Cells:	*Total*	*Perturbed*	*Control*	*‘Control’*	*[%]*
A	28	22	6	6	100.0
B	34	26	8	6	75.0
C	20	17	3	3	100.0
D	23	16	7	6	85.7
E	19	12	7	6	85.7
Total:	**124**	**93**	**31**	**27**	**87.1**

Table 3. **Tokens align based on single-cell identities in multi-cell images.** ‘Control’ recall is calculated as percentage of control-like (‘dark’) cells in heatmaps over all human-labeled cells. Note that ‘perturbation’ recall is 100% in all images from Figure 10.

Single-cell level interpretability analysis

- Additional analysis of token alignment values in three segments: background, perturbed, unperturbed

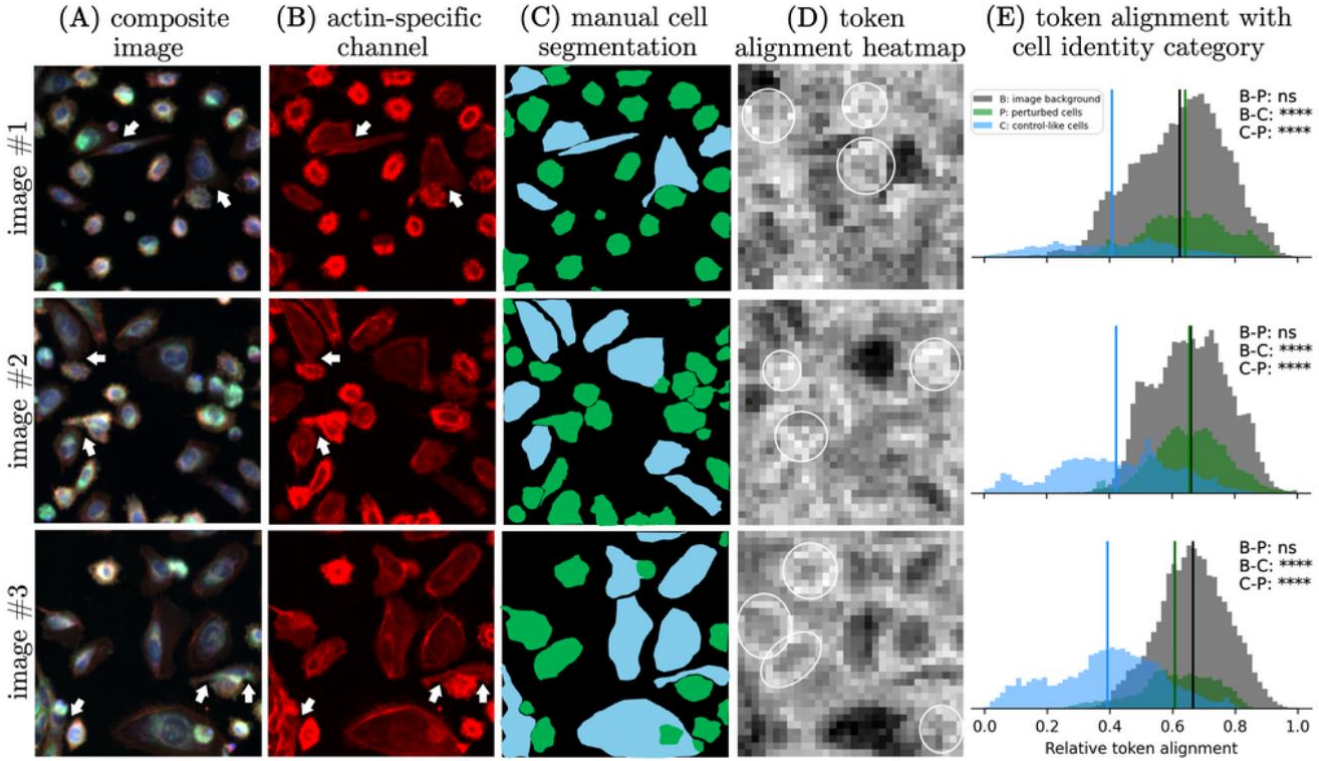


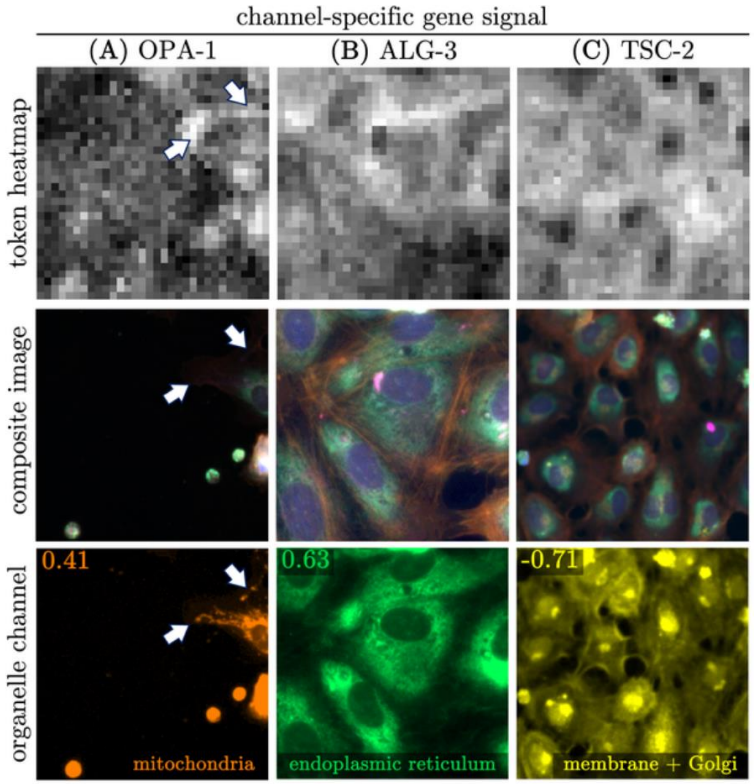
Figure 6. **Interpretation of single-cell identities from cell images** through **(A)** composite images and **(B)** their actin-specific channel, which strongly correlates with a feature from the *adherens junctions* gene group. **(C)** Segmented single cells classified based on morphology and interactions with neighbors (white arrows). Token-level alignment as **(D)** heatmaps highlighting cells with ring-like boundaries (white circles) and **(E)** category histograms with compared distributions (means as solid lines, one-sided Mann Whitney U test; *ns*, no significant difference; ****, $p < 0.0001$).

Authors' summary

- In biological systems, complex mixtures of cell morphologies are frequently observed; however, identifying which exact cell instance belongs to which population would require **deep expertise and time-consuming manual labeling at scale**.
- Our **token-level analysis presents a potential avenue for enabling modern approaches to scientific discovery**, leveraging methods from mechanistic interpretability to uncover previously unknown biological concepts at the resolution of individual cells.
- **Sparse features are clearly incomplete**: their linear probing performance significantly drops on tasks that involve more subtle changes in cell morphology.
  - It is not clear to what extent this limitation stems from **a)** our current dictionary learning techniques, **b)** the scale of our models, or **c)** whether these more subtle changes are simply not represented linearly in embedding space.
  - Nonetheless, it is evident that the **choice of the dictionary learning algorithm matters** to extract meaningful features
- **ICFL and PCA significantly improve the selectivity of extracted features, compared to TopK sparse autoencoders**

Summary

- **Still preliminary research** – only few examples analysed deeply
- **No open code** – no easy replication
- **Models only partially open**
- Hand-crafted **CellProfiler** features still seem to be **more selective**
- Some analyses are **qualitative** while they could be well parametrized
  - e.g. why there are no quantitative analyses of heatmaps vs perturbation specific channels →

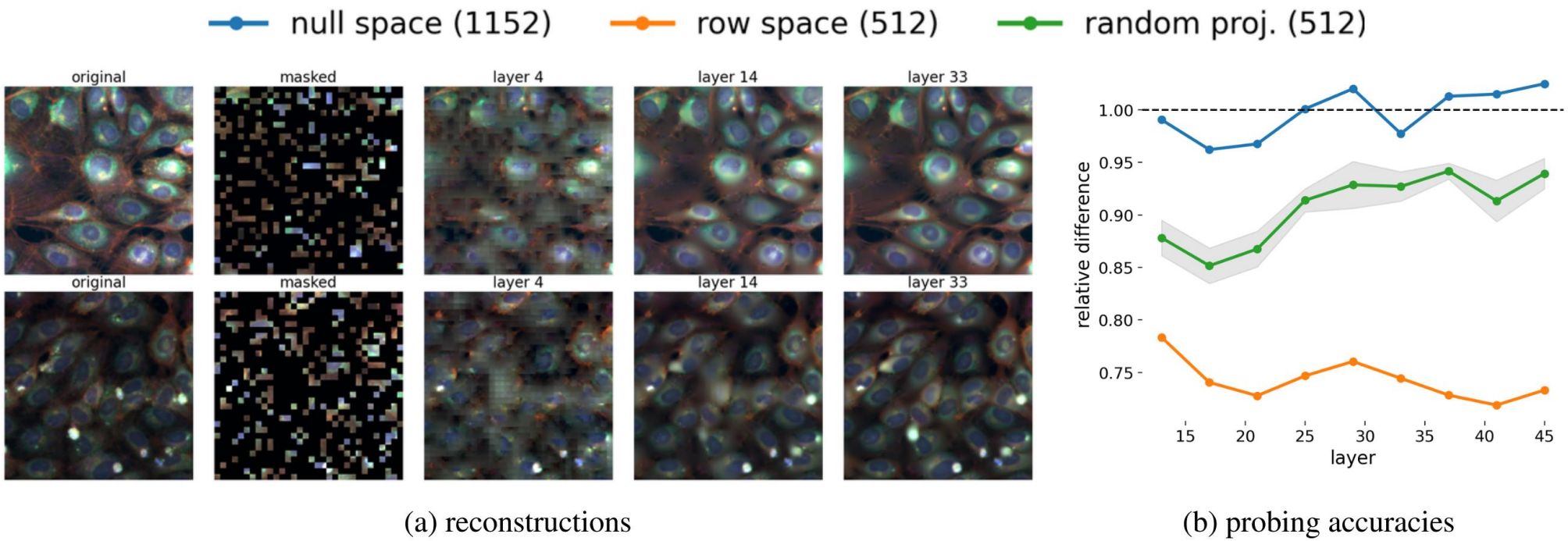


Reviews

- NeurIPS 2024 Workshop
  - non-public
  - non-archival
- ICLR 2025 – reject
- ICML 2025 – accept

Appendices – Role of linear concept directions

- Are linear directions actually inherently meaningful to the model?



(a) reconstructions

(b) probing accuracies

Figure 7. **Decoding of tokens from intermediate transformer layers.** (A) Sample reconstructed images when decoding from 3 different intermediate layers. (B) The relative linear probing accuracy when using the component from the null space, row space and a random 512-dimensional subspace as component compared to the full component. Both Figures use the MAE-G model.