Towards scientific discovery with dictionary learning: Extracting biological concepts from microscopy foundation models

Konstantin Donhauser[†]
ETH Zurich

Gemma Elyse Moran
Rutgers University

Aditya Ravuri[†]
University of Cambridge

Kian Kenyon-Dean
Recursion

Kristina Ulicna, Cian Eastwood, Jason Hartford
Valence Labs

Abstract

Dictionary learning (DL) has emerged as a powerful interpretability tool for large language models. By extracting known concepts (e.g., Golden-Gate Bridge) from human-interpretable data (e.g., text), sparse DL can elucidate a model's inner workings. In this work, we ask if DL can also be used to discover *unknown* concepts from less human-interpretable scientific data (e.g., cell images), ultimately enabling modern approaches to scientific discovery. As a first step, we use DL algorithms to study microscopy foundation models trained on multi-cell image data, where little prior knowledge exists regarding which high-level concepts should arise. We show that sparse dictionaries indeed extract biologically-meaningful concepts such as cell type and genetic perturbation type. We also propose a new DL algorithm, Iterative Codebook Feature Learning (ICFL) and combine it with a pre-processing step which uses PCA whitening from a control dataset. In our experiments, we demonstrate that both ICFL and PCA improve the selectivity or "monosemanticity" of extracted features compared to TopK sparse autoencoders.

1 Introduction

Large scale machine learning systems are extremely effective at generating realistic text and images. However, these models remain black boxes: it is difficult to understand how they produce such detailed reconstructions, and to what extent they encode semantic information about the target domain in their internal representations. One approach to better understanding these models is to investigate how models encode and use high-level, human-interpretable concepts. A challenge to this endeavor is the "superposition hypothesis" (Bricken et al. 2023), which states that neural networks encode many more concepts than they have neurons, and as a result, one cannot understand the model by inspecting individual neuron. One hypothesis for how neurons encode multiple concepts at once is that they are low-dimensional projections of some high-dimensional, sparse feature space. Quite surprisingly, there is now a large body of empirical evidence that supports this hypothesis in language models [Mikolov et al., 2013, Elhage et al., 2022, Park et al., 2023], games [Nanda et al., 2023] and multimodal vision models [Rao et al., 2024], by showing that high-level features are typically predictable via *linear* probing. Further, recent work has shown that model representations can be decomposed into human-interpretable concepts using a dictionary learning model, estimated via sparse autoencoders [Templeton, 2024, Rajamanoharan et al., 2024b,a, Gao et al., 2024].

[†]Work done while interning at Valence Labs. Email: `konstantin.donhauser@ai.ethz.ch`

Interpretable AI: Past, Present and Future Workshop at NeurIPS 2024

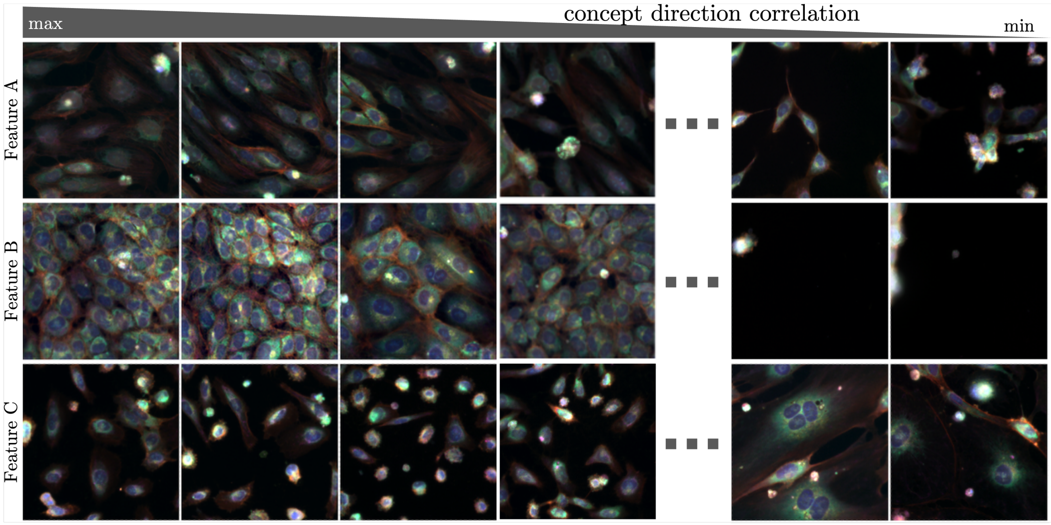


Figure 1: Cell images ranked according to the correlation strength with three selected features learned by our dictionary learning algorithm. Each feature captures distinct cellular morphologies: Feature A activates for cells with an elongated, spindle-like shape (left) and anti-correlates for sparser or aggregated cells (right); Feature B activates for cells that are densely packed with closely arranged nuclei (left) and deactivates when cell density drops (right); and Feature C activates for small-shaped, compact, brights cells without cell-cell contacts almost entirely made up from just nuclei (left), in contrast to multi-nucleated cells which occupy larger areas (right).

However, all of these successes have relied on some form of text supervision, either directly through next-token prediction or indirectly via contrastive objectives like CLIP [Radford et al., 2021], which align text and image representations. Further, these successes appear in domains which are naturally human-interpretable (i.e. text, games and natural images), and as a result, one may worry that high-level features can be extracted only in settings that we already understand. This raises a natural question: can we extract similarly meaningful high-level concepts from completely unsupervised models in domains where we lack strong prior knowledge? For example, in computational biology, masked autoencoders (MAE) trained on cellular microscopy images have been shown to be very effective at learning representations that recover known biological relationships [Kraus et al., 2024]. However, it is not known whether analogous high-level concepts can be extracted from these large MAEs. These settings are precisely where extracting high-level concepts could be most valuable: given that models can detect subtle differences in images (even those that are very challenging for human experts to interpret), we might hope that we can use these techniques to better understand subtle differences.

We study the extraction of high-level concepts from large-scale MAEs trained on microscopy images of cells that have been perturbed in genetic and small molecule perturbations screens [Fay et al., 2023]. Understanding the morphological changes induced by genetic and small molecule perturbations is an inherently difficult and fundamental problem that plays a crucial role in drug discovery [Celik et al., 2022]. Recent progress in this field using machine learning has been made by building similarity maps of genetic perturbations via cosine-similarities of post-processed representations from MAEs [Kraus et al., 2024, Celik et al., 2022, Lazar et al., 2024]. However, a limitation of these deep learning-based methods is that we only gain limited insights about the morphological changes arising from the perturbations: we can tell whether two perturbations are similar (or dissimilar) via cosine similarity, but we cannot tell *why* (or the ways in which) they are different. That is, we collapse the multidimensional similarities and dissimilarities down to a single score.

In this paper, we train dictionary learners on top of intermediate representations of large-scale MAEs [Kraus et al., 2024] and find features correlated with single concepts such as individual cell types or genetic perturbations in an unsupervised manner. Moreover, via linear probing, we show that the reconstructed representations from the sparse features preserve significant amounts of biologically-meaningful information. Through this research, we make several key contributions:

- We show that dictionary learning can be used to extract biologically-meaningful concepts from microscopy foundation models (see Figure 1), opening the path to scientific discovery using tools from mechanistic interpretability.

Algorithm 1 Iterative Codebook Feature Learning

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

- We propose a new dictionary learning algorithm—Iterative Codebook Feature Learning (ICFL)—which naturally avoids "dead" features (Section 4).
- We further show how PCA whitening on a control dataset can act as a form of weak supervision for dictionary learning (Section 5), resulting in more meaningful features.
- We demonstrate empirically that both ICFL and PCA improve the selectivity or "monosemanticity" of extracted features compared to TopK sparse autoencoders (Section 6).

2 Related work

The disentanglement and causal representation literature (CRL) share the goal of learning high-level, interpretable concepts [Bengio et al., 2013, Kulkarni et al., 2015, Higgins et al., 2017, Chen et al., 2016, Eastwood and Williams, 2018, Schölkopf et al., 2021]. Two key differences with the dictionary learning approach are: (i) disentanglement/CRL methods consider low-dimensional representations to capture the factors of variation in data, whereas overcomplete dictionary learning seeks a higher-dimensional representation to capture a large set of sparsely-firing concepts; and (ii) disentanglement/CRL methods aim to be inherently interpretable, whereas this paper considers a post-hoc approach to interpret pre-trained models. Related work on post-hoc explainability also learns "concept vectors" in neural network internal states [Kim et al., 2018, Ghorbani et al., 2019]; a key difference is that these methods use class-labeled data, whereas this paper uses an unsupervised approach to discover concepts. Additionally, feature-visualization works aim to interpret internal states/neurons by finding the data points (or gradient-optimized inputs) that lead to maximal activation [Mordvintsev et al., 2015, Olah et al., 2017, Borowski et al., 2021].

3 Background

The superposition hypothesis. Let $x_i \in \mathbb{R}^d$ denote a representation for token i; as an example, x_i may be the embedding of token i after a transformer layer. Bricken et al. [2023] hypothesize that (i) such token representations $x_i \in \mathbb{R}^d$ are linear combinations of concepts; (ii) the number of available concepts M significantly exceed the dimension of the representation d; and (iii) each token representation is the sum of a sparse set of concepts. These desiderata are satisfied by the following model that is widely studied in compressed sensing and dictionary learning:

$$x_i \approx W z_i \qquad \text{where } \|z_i\|_0 \ll d \tag{1}$$

where $W \in \mathbb{R}^{d \times M}$ is a latent concept matrix and $z_i \in \mathbb{R}^M$ is a sparse latent concept-selector (resp. feature) vector.

Feature learning using TopK SAEs. Given a set of token representations $\{x_i\}_{i=1}^N$, learning both W and $\{z_i\}_{i=1}^N$ is a *dictionary learning* or *sparse coding* problem Olshausen and Field [1997], with a long history of works proposing efficient algorithms with provable guarantees [Aharon et al., 2006, Arora et al., 2014, 2015]. In the context of mechanistic interpretability, the dominant choice for learning these parameters are two-layer sparse autoencoders. In this paper, we compare to the state-of-the-art method called TopK SAE, originally proposed by Makhzani and Frey [2013] and recently studied by Gao et al. [2024]. Following their notation, the model is:

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

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

$$L(W, b) := \sum_i \|x_i - \widehat{x}_i\|_2^2, \quad \text{where } \widehat{x}_i = W_{\mathrm{dec}} \mathrm{TopK}(W_{\mathrm{enc}} x_i - b_{\mathrm{pre}}) + b_{\mathrm{pre}} \tag{2}$$

A problem with the above optimization is that some concept vectors $W_{\mathrm{dec},m}$ are barely used; that is, features $z_{im} = 0$ for almost all $i \in [N]$. This is called the "dead feature" phenomenon. To reduce the amount of dead features, Gao et al. [2024] introduce an additional reconstruction error term using only these concept vectors to encourage their usage in the model (see Table 1).

4 Iterative Codebook Feature Learning (ICFL)

Sparse autoencoders such as TopK SAEs face two major limitations: (i) they require regularization to avoid "dead features" after training [Gao et al., 2024, Bricken et al., 2023] and (ii) some concepts may be overrepresented in the samples $\{x_i\}_{i=1}^N$, biasing the estimation. To overcome these limitations, we propose Iterative Codebook Feature Learning (ICFL). ICFL retains the decoder of TopK SAEs, however, instead of using an encoder to learn the features z, ICFL updates z using a variant of the orthogonal matching pursuit algorithm of Mallat and Zhang [1993] as described in Algorithm 1. Specifically, given the current decoder/feature matrix W_{dec}, we first select the top-k columns most aligned with $x^{(1)} = x$. Then, we learn the features $z^{(1)}$ that best reconstruct $x \approx W_{\mathrm{dec}} z^{(1)}$, using only these columns (i.e. $z^{(1)}$ is K-sparse). Next, to obtain $z^{(2)}$, we repeat this step, but replace x with the residual $x^{(2)} = x - W_{\mathrm{dec}} z^{(1)}$. Repeating this process, the final output z is taken to be $z = \sum_{t=1}^J z^{(t)}$. Consequently, z is at most Jk-sparse.

The key idea of ICFL is that early iterations subtract dominant concepts from x, allowing the algorithm in later iterations to select a broader set of concepts that are not as correlated with the main concepts in x. After updating z as detailed in Algorithm 1, the decoder parameters $\{W_{\mathrm{dec}}, b_{\mathrm{pre}}\}$ are updated to minimize the reconstruction loss from equation 2 with $\widehat{x} = W_{\mathrm{dec}} z + b_{\mathrm{pre}}$. As z is fixed in this gradient step, the algorithm does not propagate gradients through z. Consequently, the algorithm results in very few "dead" features. As a result, we do not require any additional regularization to address this "dead feature" issue that often hinders SAEs, as shown in Table 1.

In practice, we leverage random resets to ensure that the columns of W_{dec} are not too correlated. To prevent the collapse of the feature directions (columns of W_{dec}), after every 100 stochastic gradient descent steps, we take every pair of columns of W_{dec} that have cosine-similarity above 0.9 and randomly initialize one of the pairs with a vector selected uniformly at random from the hypersphere. Before running Algorithm 1, we always center the representations by the average representation with unperturbed samples from the control distribution. By doing so, we center the representations such that the origin represents the unperturbed state. Finally, we normalize the representations before applying the dictionary learner.

	w/o	w/
ICFL	55	341
TopK	7640	8026

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

5 Experimental Setup

Data source and foundation model We evaluated our dictionary learning approach on two large-scale masked autoencoders trained on cellular microscopy Cell Painting image data using 256x256x6 pixel crops as input and a patch size of 8, following the same procedures as those described in Kraus et al. [2024], Kenyon-Dean et al. [2024]. These models were trained on data from multiple cell types that were perturbed with both CRISPR gene-knockouts and small molecule perturbations. Both models used the architecture hyperparameters from Kraus et al. [2024], Kenyon-Dean et al. [2024], with the smaller of the two using the ViT-L/8 configuration, while the larger model used the ViT-G/8 configuration. We refer to these models as *MAE-L* and *MAE-G*, respectively. We obtain a single token per input crop by aggregating all patch tokens (excluding the class token). For both the residual stream and the attention output (after the out-projection), the dimension d of the tokens (representations) are 1024 and 1664 for MAE-L and MAE-G, respectively. All the visualizations used Cell Painting microscopy images from the public RxRx1 [Sypetkowski et al., 2023] and RxRx3 [Fay et al., 2023] datasets.

Task	Cell Type	Experiment Batch	siRNA Perturbation	CRISPR Perturbation	Functional Gene Group
# Classes	23	272	1 138	5	39
# Samples	110,971	80,000	81,224	79,555	57,863
Bal. Test Acc.	97.2%	87.8%	51.6%	94.6%	32.1%

Table 2: The five classification tasks and the test bal. acc. for linear probes trained on well-level aggregated representations from the residual stream from an intermediate layer from *MAE-G*.

We extract the tokens from layer 16 (*MAE-L*) and layer 33 (*MAE-G*), respectively. The motivation for using intermediate instead of final layers is that these tokens are more-likely to capture abstract high level concepts that are *internally used* by the model to solve the SSL task [Alkin et al., 2024]. We selected this layer by finding the layer which maximized linear probing performance on the functional group tasked (described below) from the original embeddings.

Preserving linear probing signals To investigate whether the features found by sparse dictionary learning retain important information from the original representation, we define five different classification tasks, summarized in Table 2. For each classification task, we use a separate (potentially overlapping) dataset and split it into train and test data to distinguish labels across:

(1) 23 different cell types which are almost perfectly distinguishable via linear classification.

(2) 272 different experiment batches. Even in controlled conditions, subtle changes in experimental conditions can induce strong *batch effects*, *i.e.* changes in experimental outcomes due to experiment-specific variations unrelated to the perturbation that is being tested.

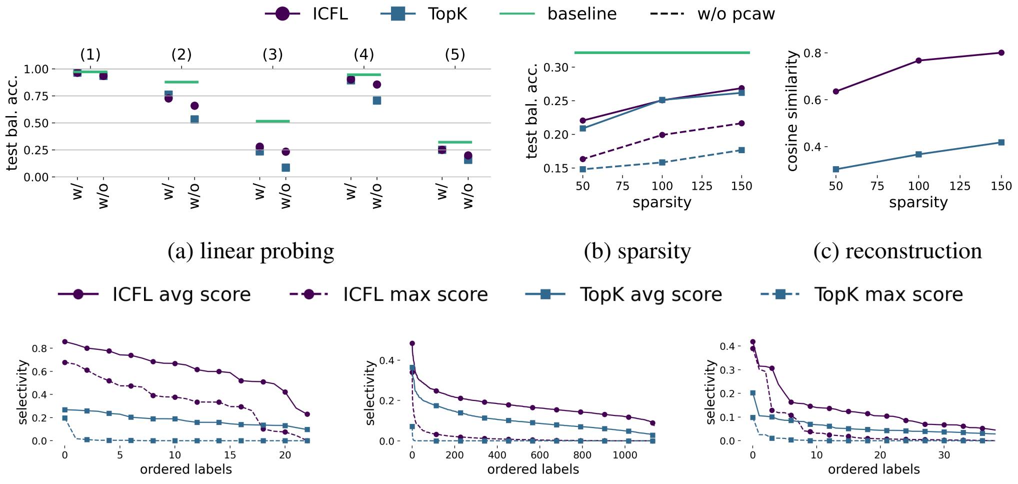
(3) 1138 siRNA perturbations from the RxRx1 dataset [Sypetkowski et al., 2023], where the single-gene expression (i.e. gene mRNA level) is partially (or completely) silenced using short interfering (si-)RNA. siRNA targets the gene mRNA for destruction via the RNA interference pathway [Tuschl, 2001]. As the extent of siRNA knock-downs is hard to quantify and prone to significant but consistent off-target effects, we also evaluated:

(4) 5 single-gene CRISPR perturbation knockouts which induce strong and consistent morphological profiles across cell types, known as "perturbation signal benchmarks" [Celik et al., 2024]. Unlike the siRNA approach, CRISPR cuts the gene DNA directly, which induces mutation in the sequence and represses the gene function. To evaluate whether our method retrieves signal which corresponds to similar phenotypes, we also assessed:

(5) 39 functional gene groups composed of CRISPR single-gene knockouts categorized by phenotypic relationships between the genes, including major protein complexes, metabolic and signaling pathways. Each gene group targets similar or related cellular process, which results in inducing morphologically similar changes in the cells [Celik et al., 2022].

To remove the impact of spurious correlations between perturbations and batch effects on the test accuracy, we always use mutually exclusive experiments for test and train data, except for (ii) where the task is to predict the experiment. Except for (i), all classification tasks use HUVEC cells and always use well-level aggregated representations: that is, we take the mean over tokens from all 36 non-edge crops from an image of a given well of cells. Because some of the classes are heavily imbalanced (particularly for Task (1)), we always report the *balanced test accuracy* and train our linear probes using logistic regression on a class-balanced cross-entropy loss.

PCA whitening using a control dataset As dictionary learners seek to minimize the Euclidean distance between the model representations x and their reconstructions $\hat{x} = Wz$, the learned features z are naturally biased towards capturing the dominant directions in the data (i.e., those that explain the most variance). Unfortunately, these directions often do not align with meaningful concepts. To address this, we use a dataset of control samples as a form of weak supervision, downweighting dominant directions in this control dataset as we know they do not correspond to the biological perturbations of interest. In particular, we learn a PCA-and-centerscale transform on this control dataset and apply it to the entire dataset *before normalization*. For our multi-cell data, unperturbed HUVEC-cell images act as our control dataset. Note that similar PCA whitening on a control dataset has been used to improve the quality of the learned multi-cell image representations [Kraus et al., 2024].



(a) linear probing (b) sparsity (c) reconstruction

(d) Cell type (e) siRNA perturbation (f) Functional gene group

Figure 2: ***Top row:*** a) Test bal. acc. of linear probes trained on the original representation (solid line) and reconstructions from ICFL and TopK SAEs in combination with PCA whitening and with out. b) Test bal. acc. as a function of the sparsity (dashed line is the original representation) 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. ***Bottom row:*** The highest selectivity scores among all features for each label. We separately order the labels for each line starting with the maximum score. We plot the avg (solid) and max (dashed) selectivity scores.

Training the DL models By default, we always choose a sparsity of $K = 100$ for TopK SAEs and $J = 20, k = 5$ (resulting in a max sparsity of 100) for ICFL as described in Section 4, and use a total of 8192 features. Unless otherwise specified, we always apply the PCA whitening described in Section 5 and use representations from the residual stream. We train the sparse autoencoders using 40M tokens (one token per crop) with a batch size of 8192 for 300k iterations. Our learning rate is 5×10^{-5} for all experiments. Similar to Gao et al. [2024], we observed that changing the learning rate has a limited impact on the outcome.

6 Experimental Results

In this section we present our experimental results. If not further specified, we always use features extracted from ICFL in combination with PCA whitening.

6.1 Dictionary features are correlated with biological concepts

Preserving linear probing signals By comparing linear probes on the representations and reconstructions from ICFL sparse features, we can measure how much "biologically-relevant" information is lost when extracting sparse features. Figure 2a shows that almost the entire signal is preserved for simple concepts such as cell types (1), batch effects (2) and perturbations with strong morphological changes (4). For the difficult tasks of distinguishing between many genetic perturbations (3,5), a substantial amount of the linear signal is preserved. Both TopK SAEs and ICFL features yield a similar linear probing accuracy, while we can see a clear drop if no PCA whitening is used during pre-processing. We further present in Figure 2b an ablation for the sparsity of the extracted feature vector. While increasing the number of non-zeros improves the accuracy, the effect is limited compared to PCA whitening.

Reconstruction loss To evaluate the quality of unsupervised DL, the cosine similarity (or ℓ_2-error) has been often used as a benchmark [Rajamanoharan et al., 2024a, Gao et al., 2024]. Figure 2c shows that the reconstruction quality of ICFL is much higher than TopK SAE for the same sparsity constraints when using PCA whitening.

Selectivity of features for biological concepts As a third experiment, we investigate how strongly correlated the features are with labels from the classification tasks in Table 2. For each dataset associated with a classification task, we extract from every image a feature vector using the center

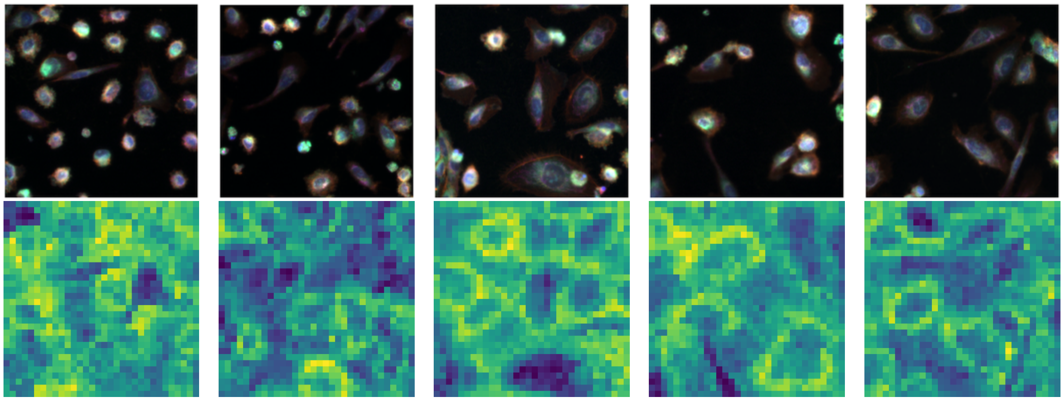


Figure 3: Visualization of images strongly correlated with a selected codebook feature. We plot the original image and the histogram of the inner products of the individual tokens with the selected linear feature direction.

crop as input to the MAE. For each feature, we then compute two selectivity scores: the **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. As a stronger notion of correlation, we also use the **max selectivity** score, that subtracts the maximum % for any other label. The selectivity score has been originally proposed in the context of neuroscience [Hubel and Wiesel, 1968] and has also been used by Madan et al. [2022] to measure the "monosemanticity" of neurons.

We plot in Figure 2d-2f the selectivity scores for both ICFL features and TopK SAEs. We see that ICFL features consistently achieve higher selectivity scores than TopK SAE features. Moreover, especially for cell types, we observe a high *max* selectivity across almost all cell types, while for more complex features we still observe a moderate selectivity score of more than 0.1 across all labels.

Visualizing token-level features ViTs produce embeddings on a per-token level, so for any given concept direction, we can ask how aligned the per-token embeddings are to the concept direction? Or in other words, can we find interpretable patterns in pixel space? In Figure 3 (top row), we plot for a selected feature five of the 30 most correlated crops from the subset of the images from task (5) contained in the public *RXRX3* [Fay et al., 2023] dataset. All crops contain small cells, suggesting that the feature is correlated with the cell size. We can support this hypothesis by plotting heatmaps of the inner product of the individual tokens with the linear feature direction Figure 3 (bottom row). The heatmaps show that tokens surrounding the cell centers are most correlated with the concept direction, which suggests that the concept corresponds to missing actin (rendered in red) surrounding the nucleus. Note that tokens are *not* aligned with the feature direction in cells where actin is present: e.g. the large cell in bottom-centre of the third image (see the large blue patch in the heatmap). We provide further evidence for the described interpretable pattern of this feature in Figure 4 (Feature 3), where we observe that the crops least correlated with the feature direction contain abnormally large cells. Additional examples are provided in Appendix A.

7 Conclusion

In this paper we have explored the extent to which dictionary learning can be used to extract biologically-meaningful concepts from microscopy foundation models. The results are encouraging: with the right approach, we were able to extract sparse features that are associated with distinct and biologically-interpretable morphological traits. That said, these sparse features are clearly incomplete: we see significant drops in their linear-probing performance on tasks that involve more subtle changes in morphology. It is not clear to what extent this is a limitation of our current dictionary learning techniques, the scale of our models, or whether these more subtle changes are simply not represented linearly in embedding space. Nonetheless, it is clear that the choice of dictionary learning algorithm matters to extract meaningful features.

References

Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11): 4311–4322, 2006.

Benedikt Alkin, Lukas Miklautz, Sepp Hochreiter, and Johannes Brandstetter. Mim-refiner: A contrastive learning boost from intermediate pre-trained representations. *arXiv preprint arXiv:2402.10093*, 2024.

Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Conference on Learning Theory*, pages 779–806. PMLR, 2014.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. In *Conference on learning theory*, pages 113–149. PMLR, 2015.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Judy Borowski, Roland Simon Zimmermann, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis, Matthias Bethge, and Wieland Brendel. Exemplary natural images explain {cnn} activations better than state-of-the-art feature visualization. In *International Conference on Learning Representations*, 2021.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2, 2023.

Safiye Celik, Jan-Christian Huetter, Sandra Melo, Nathan Lazar, Rahul Mohan, Conor Tillinghast, Tommaso Biancalani, Marta Fay, Berton Earnshaw, and Imran S Haque. Biological cartography: Building and benchmarking representations of life. In *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*, 2022.

Safiye Celik, Jan-Christian Hütter, Sandra Melo Carlos, Nathan H Lazar, Rahul Mohan, Conor Tillinghast, Tommaso Biancalani, Marta M Fay, Berton A Earnshaw, and Imran S Haque. Building, benchmarking, and exploring perturbative maps of transcriptional and morphological data. *bioRxiv*, 2024. doi: 10.1101/2022.12.09.519400.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 29, pages 2180–2188, 2016.

Cian Eastwood and Christopher K I Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Marta M Fay, Oren Kraus, Mason Victors, Lakshmanan Arumugam, Kamal Vuggumudi, John Urbanik, Kyle Hansen, Safiye Celik, Nico Cernek, Ganesh Jagannathan, et al. Rxrx3: Phenomics map of biology. *bioRxiv*, pages 2023–02, 2023.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *Neural Information Processing Systems*, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

Kian Kenyon-Dean, Zitong Jerry Wang, John Urbanik, Konstantin Donhauser, Jason Hartford, Saber Saberian, Nil Sahin, Ihab Bendidi, Safiye Celik, Marta Fay, et al. Vitally consistent: Scaling biological representation learning for cell microscopy. *arXiv preprint arXiv:2411.02572*, 2024.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, and Fernanda Viegas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, 2018.

Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, 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*, pages 11757–11768, 2024.

Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, volume 28, pages 2539–2547, 2015.

Nathan H Lazar, Safiye Celik, Lu Chen, Marta M Fay, Jonathan C Irish, James Jensen, Conor A Tillinghast, John Urbanik, William P Bone, Christopher C Gibson, et al. High-resolution genome-wide mapping of chromosome-arm-scale truncations induced by crispr–cas9 editing. *Nature Genetics*, pages 1–12, 2024.

Spandan Madan, Timothy Henry, Jamell Dozier, Helen Ho, Nishchal Bhandari, Tomotake Sasaki, Frédo Durand, Hanspeter Pfister, and Xavier Boix. When and how convolutional neural networks generalize to out-of-distribution category–viewpoint combinations. *Nature Machine Intelligence*, 4(2):146–153, 2022.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. *Google research blog*, 20(14):5, 2015.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*, 2024a.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024b.

Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. *arXiv preprint arXiv:2407.14499*, 2024.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Maciej Sypetkowski, Morteza Rezanejad, Saber Saberian, Oren Kraus, John Urbanik, James Taylor, Ben Mabey, Mason Victors, Jason Yosinski, Alborz Rezazadeh Sereshkeh, Imran Haque, and Berton Earnshaw. Rxrx1 a dataset for evaluating experimental batch correction methods. In *Proceedings of the IEEE CVF Conference on Computer Vision and Pattern Recognition*, pages 4285–4294, 2023.

Alex Tamkin, Mohammad Taufeeque, and Noah D Goodman. Codebook features: Sparse and discrete interpretability for neural networks. *arXiv preprint arXiv:2310.17230*, 2023.

Adly Templeton. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic, 2024.

Thomas Tuschl. Rna interference and small interfering rnas. *Chembiochem*, 2(4):239–245, 2001.

A Interpretable features

In this section, we present additional visualizations of crops strongly correlated with selected feature directions. In the spirit of recent works for LLMs [Bricken et al., 2023], we only present a qualitative analysis that aims to highlight non-trivial, complex, and interpretable patterns captured by these features.

For completeness, Figure 4 shows the same crops as Figure 1 but this time all 6 most correlated and anti-correlated crops. We further present in Figures 5 to 9 additional examples similar to Figure 3 for images strongly correlated with different features. In addition to the heat-map and the entire crop, we also plot the patches that are most strongly correlated with the feature. We make two important observations: a) we can see clear interpretable patterns for which patches are most strongly correlated with the cells, posing a promising area for future research on interpreting and validating concept directions found in large foundation models for microscopy image data; b) we see that the most correlated patches are robust to light artifacts, which can be seen best in the last column in Figure 5.

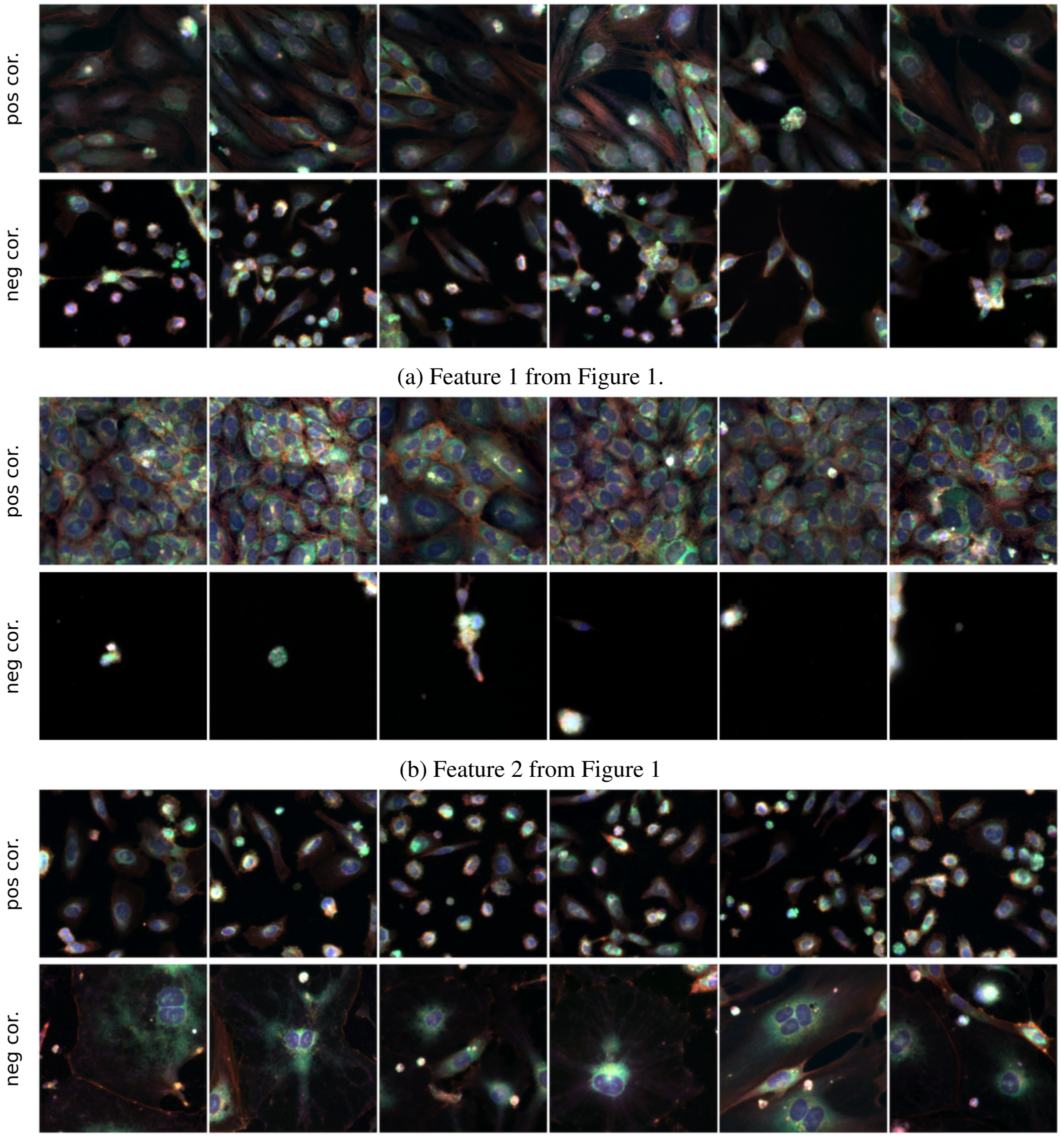
B Ablations

In this section we present ablations on the choice of the representations, as well as the model size.

Attention block It is common in the literature to use representations from the MLP output or the attention output [Bricken et al., 2023, Tamkin et al., 2023, Rajamanoharan et al., 2024a]. We compare in Table 3 the test balanced accuracy when taking representations from the residual stream and attention output. We observe that both result in similaraccuracies. We make the same observation in Figure 10a and 10b showing an ablation for the linear probes trained on the reconstruction using the same setting as described in Section 6. Moreover, we compare in Figure 11 the selectivity scores as in Figure 2, confirming further that the residual stream and the attention output show a similar behavior. The only exception is TopK for cell types, where the attention outputs result in significantly better selectivity scores, however, still substantially below the ones obtained by ICFL.

Residual stream	97.2%	87.8%	51.6%	94.6%	32.1%
Attention output	96.8%	85.8%	52.5%	94.6%	32.1%

Table 3: The test bal. acc. like in Table 3 for representations taken from the residual stream (Test. Bal. Acc. row from Table 3) and the attention output.



(a) Feature 1 from Figure 1.

(b) Feature 2 from Figure 1

(c) Feature 3 from Figure 1

Figure 4: For each row in Figure 1 we also include the crops that are the most correlated with the feature direction in the opposite direction. More precisely, for each feature we show the 6 most positively (first row) and negatively (second row) correlated crops. For each of the three features we observe a clear concept shift along the feature direction (going from negatively correlated to positively correlated).

Model size We further investigate the model size, as shown in Figures 10a and 10b, where we compare the linear probes for Ph2 (1.9B parameters) with the much smaller model Ph1 (330M parameters). We observe that for simple tasks like classifying cell types, both models yield similar performances. However, we observe consistent improvements on complex classification tasks (3,5), both for the probes trained on the original representations, as well as the reconstructions from ICFL and TopK. This demonstrates that dictionary learning benefits from scaling the model size.

We further plot in Figure 12 the selectivity scores. For ICFL, we consistently observe improvements when increasing the model size, while for TopK SAE, we see a significant drop. Interestingly, this drop does not occur for the probing accuracy on the reconstructions in Figures 10a and 10b. This suggests that, although capturing meaningful signals in the reconstructions, TopK SAE faces more difficulties in finding "interpretable" features with high selectivity scores from richer representations post-processed using PCA whitening.

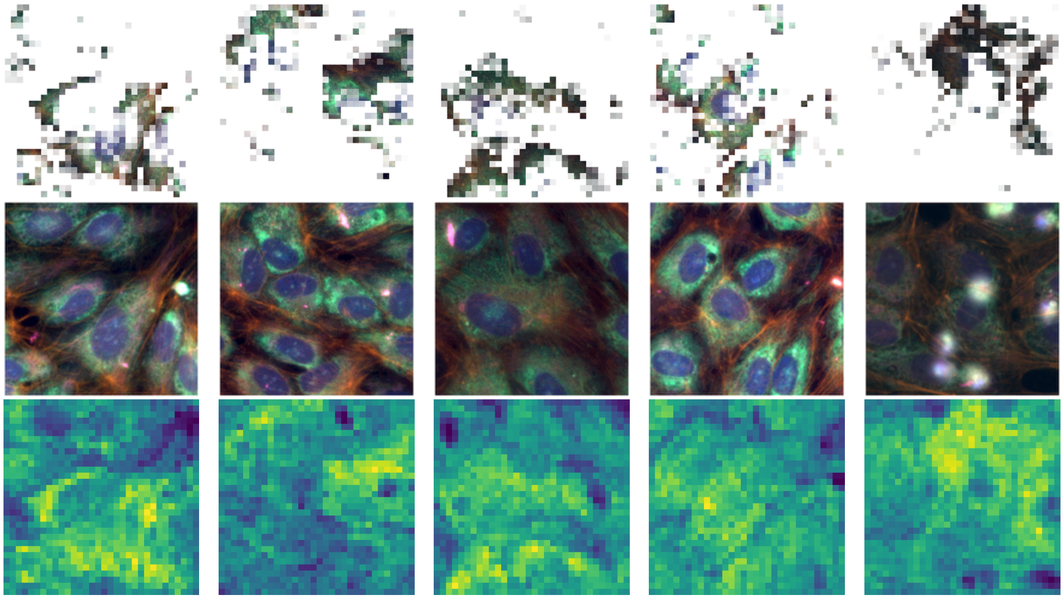


Figure 5: This feature appears to be focusing on the endoplasmic reticuli and nucleoli channel (cyan area) surrounding the nucleus. These are expanded relative to the usual morphology of HUVEC cells.

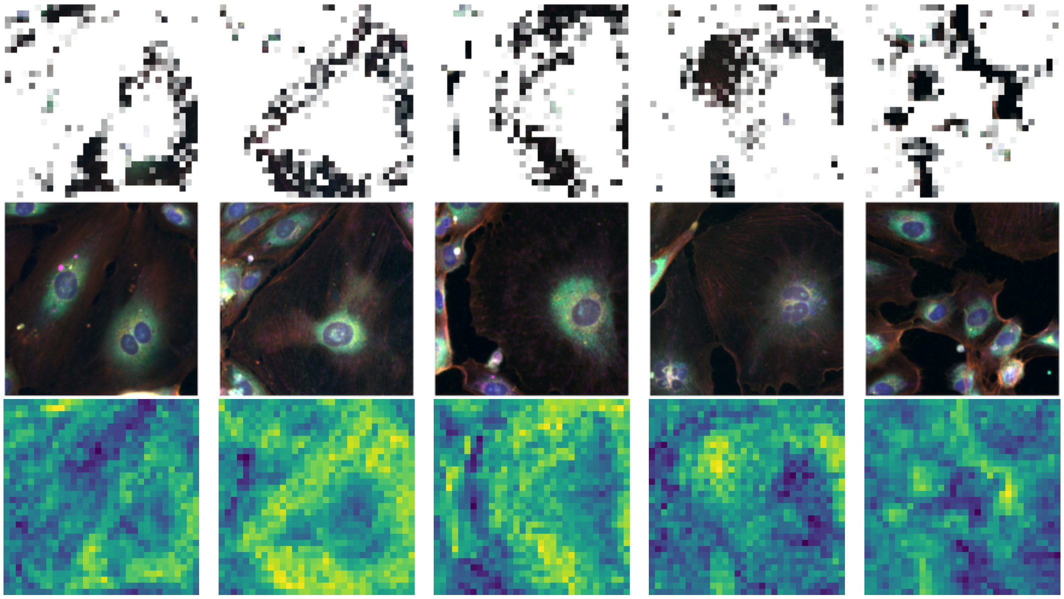


Figure 6: This feature appears to be firing for cells that are unusually large with spread out actin. Note that the feature focuses on the actin channel (red) surrounding the cell.

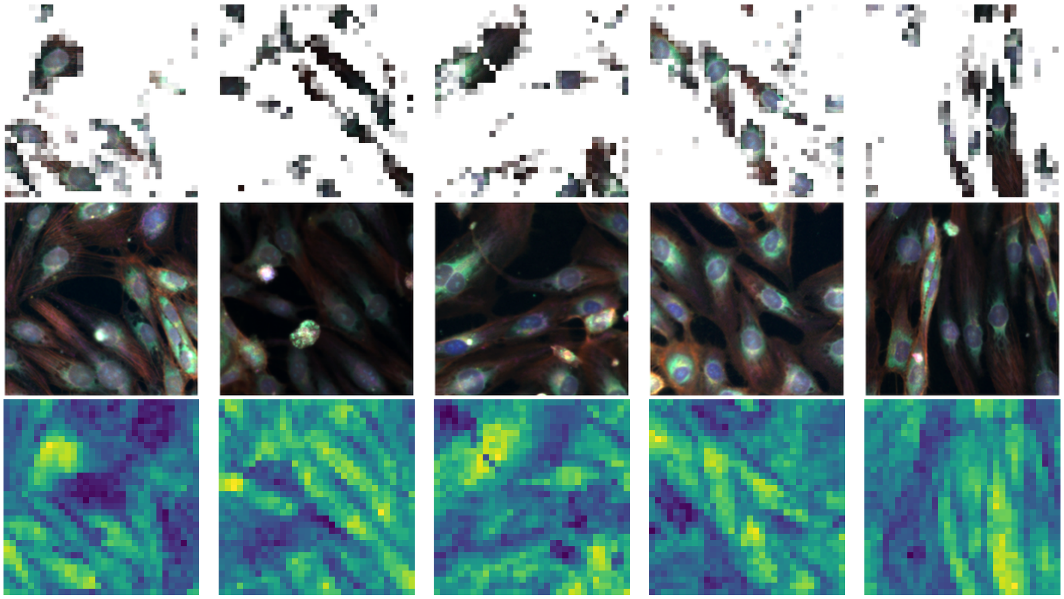


Figure 7: This feature appears to be active for long spindly cells, with the features are most aligned for the long "stretched out" section of the cells.

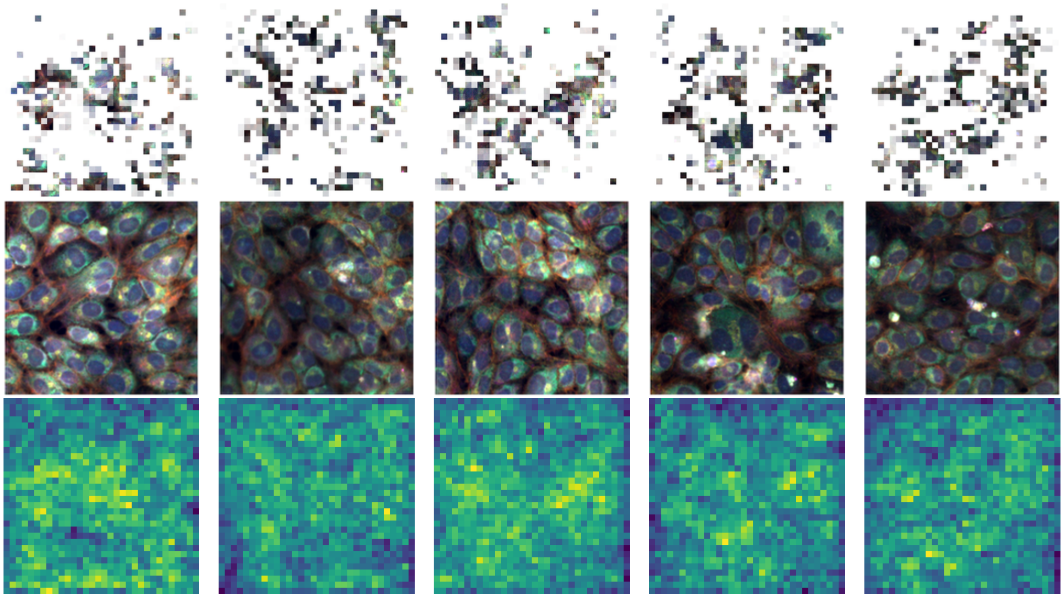


Figure 8: This feature is active for tightly clumped cells. The heatmaps are less clearly interpretable for these images, but appear to be active when neighboring nuclei are touching.

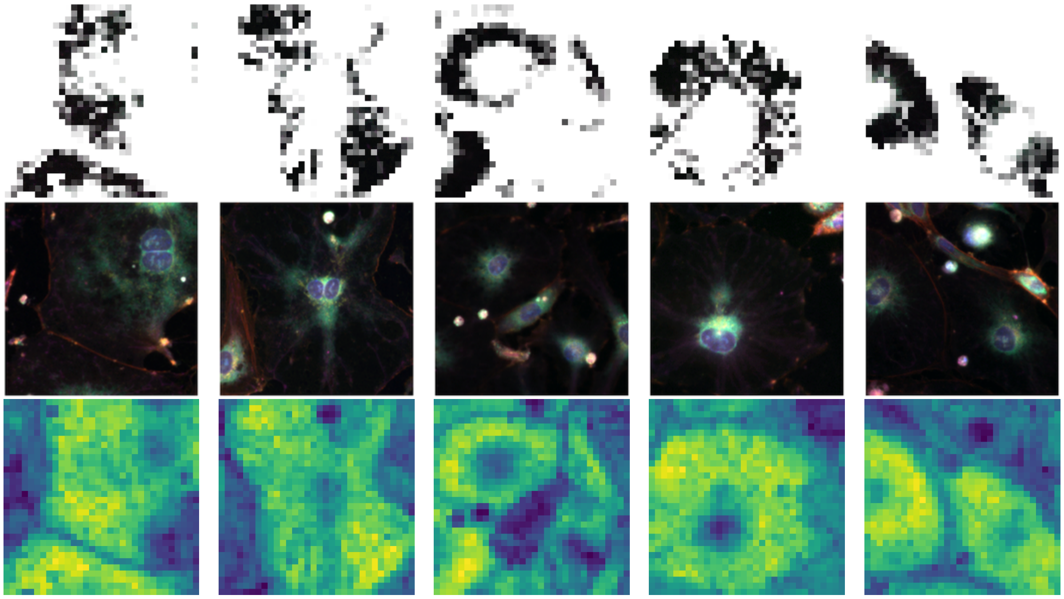
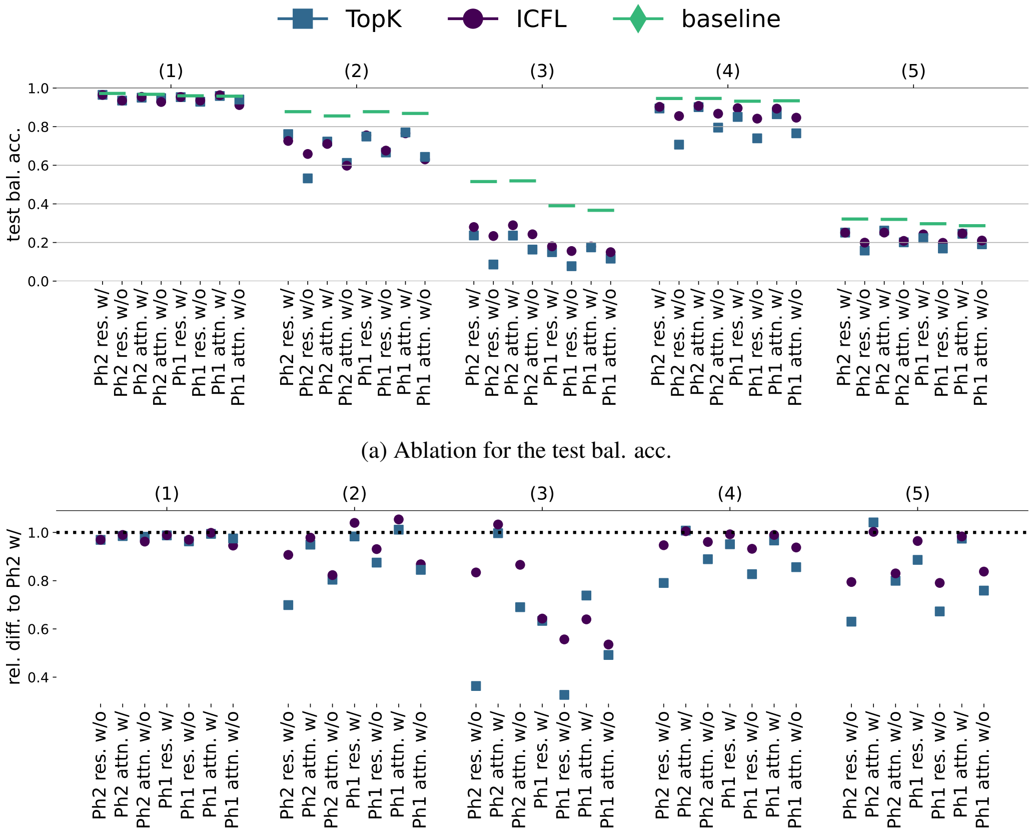


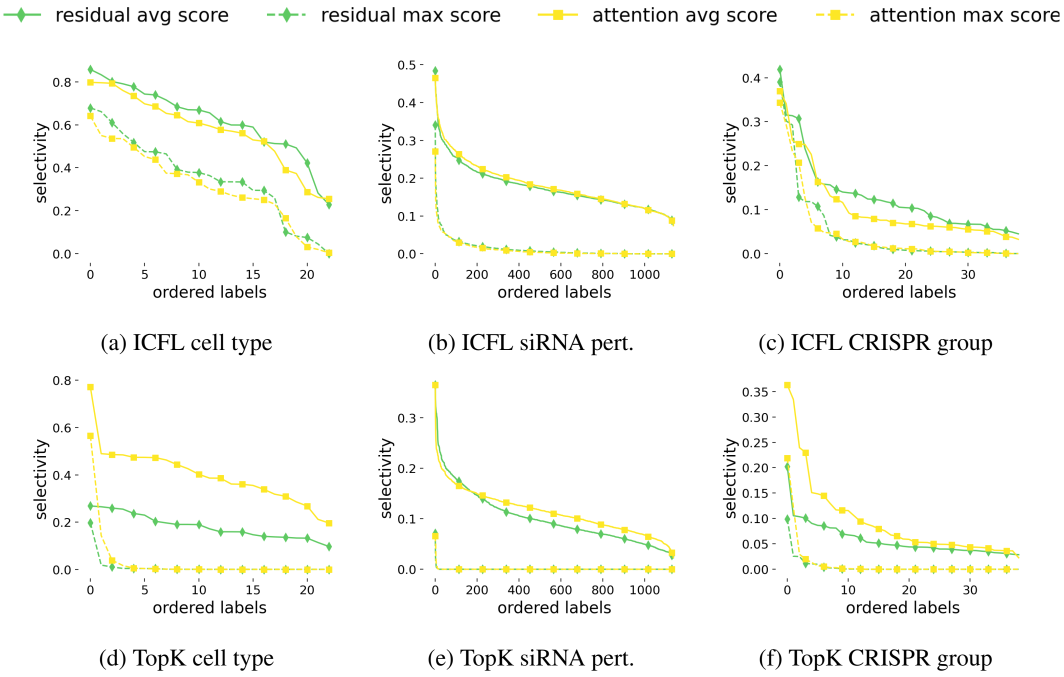
Figure 9: This feature show a similar behavior to the feature in Figure 6



(a) Ablation for the test bal. acc.

(b) Relative difference to Ph2 w/

Figure 10: a) The test bal. acc. of linear probes trained on the original representation (solid lines) and reconstructions from ICFL features and TopK SAEs for representations taken from the residual stream and attention output of Ph2 (larger model) and Ph1 (smaller model), as well as with PCA whitening and without. b) Same as a) but depicting the relative difference in linear probing accuracy compared to Ph2 residual stream using PCA



(a) ICFL cell type

(b) ICFL siRNA pert.

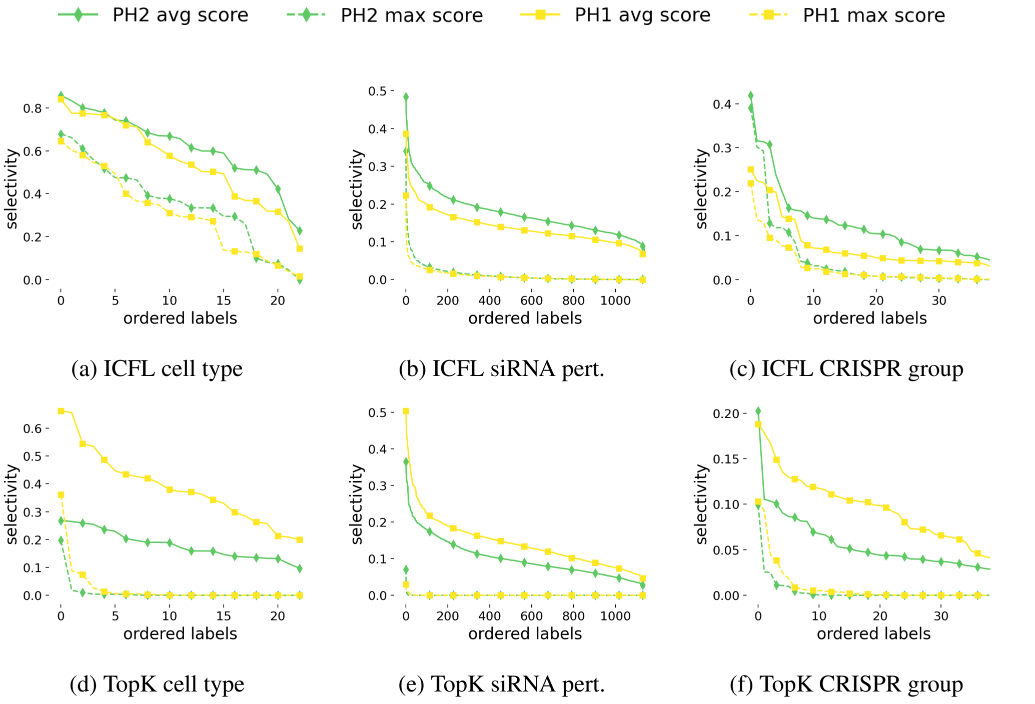
(c) ICFL CRISPR group

(d) TopK cell type

(e) TopK siRNA pert.

(f) TopK CRISPR group

Figure 11: The selectivity scores as in Figure 2 for ICFL (first row) and TopK (second row) when using representations from the residual stream (green) and the attention block (yellow).



(a) ICFL cell type

(b) ICFL siRNA pert.

(c) ICFL CRISPR group

(d) TopK cell type

(e) TopK siRNA pert.

(f) TopK CRISPR group

Figure 12: The selectivity scores as in Figure 2 for ICFL (first row) and TopK (second row) when using representations from the residual stream from Ph2 (green) and Ph1 (yellow) using PCA whitening.