Learning biologically relevant features in a pathology foundation model using sparse autoencoders

Nhat Minh Le[1] Ciyue Shen[1] Neel Patel[1] Chintan Shah[1] Darpan Sanghavi[1]
Blake Martin[1] Alfred Eng[1] Daniel Shenker[1] Harshith Padigela[1] Raymond Biju[1]
Syed Ashar Javed[1] Jennifer Hipp[1] John Abel[1] Harsha Pokkalla[1] Sean Grullon[1] Dinkar Juyal[1]
[1]PathAI Inc, Boston, USA
nhat.le@pathai.com

Abstract

Pathology plays an important role in disease diagnosis, treatment decision-making and drug development. Previous works on interpretability for machine learning models on pathology images have revolved around methods such as attention value visualization and deriving human-interpretable features from model heatmaps. Mechanistic interpretability is an emerging area of model interpretability that focuses on reverse-engineering neural networks. Sparse Autoencoders (SAEs) have emerged as a promising direction in terms of extracting monosemantic features from polysemantic model activations. In this work, we trained a Sparse Autoencoder on the embeddings of a pathology pretrained foundation model. We found that Sparse Autoencoder features represent interpretable and monosemantic biological concepts. In particular, individual SAE dimensions showed strong correlations with cell type counts such as plasma cells and lymphocytes. These biological representations were unique to the pathology pretrained model and were not found in a self-supervised model pretrained on natural images. We demonstrated that such biologically-grounded monosemantic representations evolved across the model's depth, and the pathology foundation model eventually gained robustness to non-biological factors such as scanner type. The emergence of biologically relevant SAE features was generalizable to an out-of-domain dataset. Our work paved the way for further exploration around interpretable feature dimensions and their utility for medical and clinical applications.

1. Introduction

1.1. Mechanistic Interpretability

Artificial Intelligence (AI) has made significant strides in various domains, including healthcare and pathology. As these systems become more complex and widely adopted, understanding their internal mechanisms becomes crucial for ensuring reliability, addressing biases, and fostering trust. This paper focuses on the application of mechanistic interpretability (MI) techniques, particularly sparse autoencoders (SAEs), to neural networks used in pathology.

Mechanistic interpretability aims to study neural networks by reverse-engineering them, providing insights into their internal workings [2, 5, 17, 32]. This approach is particularly relevant in pathology, where understanding the decision-making process of AI systems can have significant implications for patient care and diagnostic accuracy. In the MI paradigm, "features" are defined as the fundamental units of neural networks, and "circuits" are formed by connecting features via weights [5]. This conceptualization allows researchers to dissect complex neural networks and understand how they process and represent information.

According to the Superposition Hypothesis [18, 33], a neuron can be polysemantic, i.e., it can store multiple unrelated concepts. Consequently, a neural network can encode more features than its number of neurons. This concept is particularly intriguing in the context of pathology, where complex visual patterns and subtle tissue variations must be recognized and interpreted.

SAEs have been used in NLP [3, 13] to achieve a more monosemantic unit of analysis compared to the model neurons. In vision datasets, SAEs trained on layers of convolutional neural nets have uncovered interpretable features such as curve detectors [6, 22]. Various improvements to SAEs have been suggested, including k-sparse [29] and gated sparse [35] autoencoders, and using JumpReLU [36] instead of ReLU as the activation function.

In Large Language Models (LLMs), MI has been used to understand phenomena such as in-context learning [34], grokking [31], and uncovering biases and deceptive behavior [37]. While these studies primarily focus on language models, their insights may have implications for image-based AI systems used in pathology. The Universality Hy-

pothesis [33] posits that similar features and circuits are learned across different models and tasks. However, other studies [12] have found mixed evidence for this claim. Understanding the extent of universality in neural networks could have significant implications for the transferability and generalizability of AI systems in pathology across different types of analyses or tissue samples.

1.2. Interpretability in Pathology

Histopathology, often used interchangeably with pathology, is the diagnosis and study of diseases through microscopic examination of cells and tissues. It plays a critical role in disease diagnosis and grading, treatment decision-making, and drug development [28, 40]. Digitized whole-slide images (WSIs) of pathology samples can be gigapixel-sized, containing millions of areas of interest and biologically relevant entities across a wide range of characteristic length scales.

Machine learning (ML) has been applied to pathology images for tasks such as segmentation of biological entities, classification of these entities, and end-to-end weakly supervised prediction at a WSI level [4, 7, 41]. Work on interpretability in pathology has focused on assigning spatial credit to WSI-level predictions [24, 27], computing human-interpretable features from model output heatmaps [14], and visualization of multi-head self-attention values on image patches [11].

Foundation Models (FMs) are promising for pathology as they can take advantage of large amounts of unlabeled data to build rich representations which can be easily adapted for downstream tasks in a data-efficient manner [11, 15, 19, 26, 39]. The diversity of pre-training data powers these models to generate robust representations, enabling them to generalize better than individual task-specific models trained on smaller datasets. Additionally, these models can be used as a universal backbone across different tasks, reducing the development and maintenance overhead associated with bespoke task-specific models.

We believe that histopathology data is a promising area for Mechanistic Interpretability (MI)-based analysis, for the following reasons:

- **Rich and Complex Data:** Unlike object-centric image datasets, a single pathology image patch can contain up to 10^6 regions of interest (e.g., cell nuclei). The number of active concepts is bounded by underlying biological structures, and identifying every concept can be critical for downstream applications.
- **Addressing Batch Effects:** Pathology images are susceptible to "batch effects", where models may learn spurious features instead of relevant morphology-related features. This issue arises from high-frequency artifacts and systematic confounders in image acquisition [23]. MI can help disentangle biological content from incidental attributes, leading to more robust models for real-world applications.
- **Enabling Precise Interventions:** A bottom-up understanding of feature contributions to predictions can enable modeling of useful interventions at increasing levels of complexity. This ranges from activation-based methods [10, 38] to text-based interventions, such as predicting tissue changes in response to drug administration.
- **Enhancing Model Transparency:** MI can provide insights into the decision-making process of AI systems in pathology, potentially improving their interpretability and trustworthiness in clinical settings.
- **Facilitating Novel Discoveries:** By uncovering the internal mechanisms of AI models trained on pathology data, MI may lead to new biological insights or hypotheses that were not apparent through traditional analysis methods.

1.3. Summary of Contributions

This work presents an interpretability analysis of the embedding dimensions derived from a vision foundation model trained on histopathology images. We employ SAEs on the model embeddings to uncover monosemantic representations of biologically-relevant and human interpretable features. Our study provides the first detailed characterization of the image attributes represented within embedding dimensions of a pathology foundation model and how they evolve through the model's depth.

The main contributions of our work are as follows:

- We demonstrate that individual dimensions in the embedding space encapsulate complex, higher-order concepts through polysemantic combinations of fundamental characteristics like cell appearance and nuclear morphology.
- We train a sparse auto-encoder (SAE) to disentangle polysemantic embedding dimensions, revealing a sparse dictionary of interpretable features that represent cell and tissue characteristics, geometric structures, and image artifacts.
- We investigate the features identified by the SAEs and show that specific SAE dimensions are correlated with counts of key cell types in pathology (plasma cells, cancer cells, lymphocytes, fibroblasts, macrophages).
- These cell-type specific dimensions are found in the last layers, while SAE dimensions from the earlier layers mainly correlate with color features, providing evidence that the model gains invariance to spurious factors for downstream pathology tasks.
- We demonstrate that the SAE representations are robust and can be generalized across metadata such as disease, stain and scanner type.

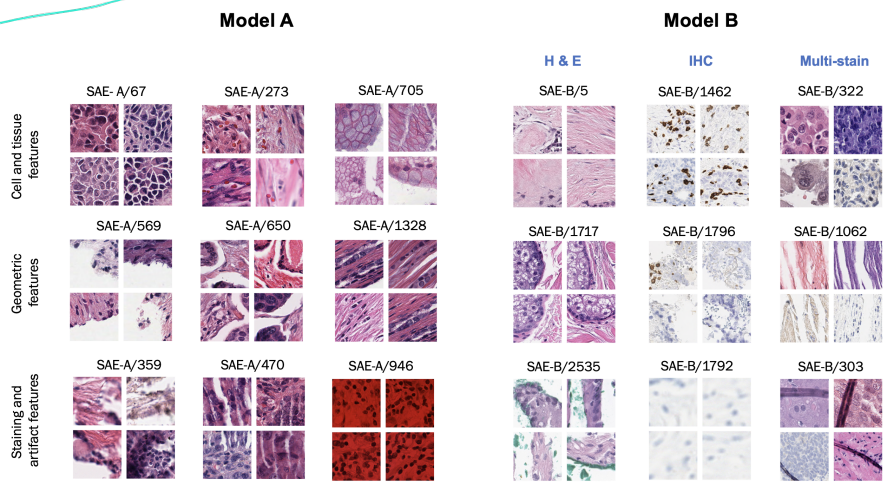


Figure 1. Feature visualization of SAE hidden dimensions revealed interpretable dictionary of pathology features. For each SAE hidden dimension of model A (trained on the TCGA dataset) and model B (trained on the 1M dataset), 4 out of the top 16 images that activated that dimension were visualized. Manual examination revealed interpretable features represented by these dimensions. For model A, these include cell and tissue features specific to H & E stain (top row: poorly differentiated carcinoma with distinct cell separation, red blood cells, mucin); geometric features (middle row: edge of tissue, clefting between cancer and stroma, diagonal fibers); staining and artifact features (bottom row: blur, sectioning artifact, red stain). For model B, some SAE dimensions were specific to H & E stain (first column: collagen-enriched fibroblasts, circular clusters of tumor cells, surgical ink), some were specific to IHC stain (second column: stained lymphocytes, edge of tissue, blur), and others generalized across stains (third column: large cancer cells, vertical structures, tissue folds).

2. Method

2.1. Datasets

We used three datasets for experimentations. For the training dataset, we used 1.1 million image patches (1M dataset) (224 x 224 pixels at a high resolution, 0.25 microns per pixel) including both haematoxylin & eosin (H & E) and immunohistochemistry (IHC) stains, sampled from the train set of 'PLUTO' - a pathology pretrained foundation model [25], covering oncology, IBD (inflammatory bowel disease) and MASH (metabolic dysfunction-associated steatohepatitis).

Two different datasets were mainly used for evaluation. The first (TCGA dataset) comprised three publicly available TCGA (The Cancer Genome Atlas) [42] datasets containing H & E-stained histology images from three organs: breast (TCGA-BRCA), lung (TCGA-LUAD), and prostate (TCGA-PRAD). We selected 951, 493 and 488 WSIs from these datasets respectively for the analysis. This dataset was used in the early experimentation of SAE training. The second dataset (CPTAC) comprised two publicly available CPTAC (Clinical Proteomic Tumor Analysis Consortium) datasets containing H & E-stained histology images from two cancer types: cutaneous melanoma (CPTAC-CM, 256 WSIs), and head and neck cancer (CPTAC-HNSCC, 228 WSIs). The two evaluation datasets were chosen from different data sources and purposely from different organ types to maximize the data diversity.

2.2. Embedding extraction

All the images for the train and evaluation datasets were passed through a frozen ViT-S encoder taken from 'PLUTO'. For each image patch, we extracted 384-dimensional embedding vectors corresponding to the CLS token residual stream in layers 1-12 with 12 as the output layer. For baseline comparison, we used a self-supervised vision transformer DINO [8].

2.3. Biological and color feature extraction

To investigate the representation of cellular features in SAE dimensions, we deployed PathExplore, a machine-learning model (PathExplore is for research use only. Not for use in diagnostic procedures) [1, 30], on the evaluation datasets. On each slide, the model detected and classified cell types including cancer cells, lymphocytes, macrophages, fibroblasts, plasma cells. The count and density of each cell class on each image were computed. In addition, color features, including gray-scale intensity, LAB colorspace, and saturation, were extracted in each image patch by taking the average or standard deviation of feature values across all the

pixels in the image at its original resolution.

2.4. Sparse autoencoder training

The SAEs were trained using a loss function given by $\frac{1}{k}(\sum_{i=1}^{k}||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2 + \lambda \sum_{i=1}^{k}||\mathbf{f}_i||_1)$, where k is the batch size, $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$ are the raw and reconstructed embeddings, and $\mathbf{f}_i$ are the learned features of image i [3, 20]. Dead neuron resampling was implemented to reduce the fraction of dead neurons [3, 20]. We used Adam optimizer with a learning rate of 0.001, expansion factors of 1, 8, 16, 32; and L1-penalty weight in 0.001, 0.004, 0.006, 0.008, 0.01. Results were reported for models with an expansion factor of 8 and L1-penalty weight of 0.004.

3. Training a sparse autoencoder on PLUTO embeddings reveals interpretable features

3.1. Interpretability analysis of PLUTO embeddings

We first manually inspected each of the 384 dimensions of the PLUTO embedding space to determine if they represent singular features of the images. For each dimension, we randomly sampled 5 patches that have the lowest 5% and the highest 5% activation values across the TCGA-BRCA dataset (see Supplementary section).

The embedding dimensions tended to encode multiple image characteristics. For example, dimension 27 was more activated for larger cells (compared to smaller cells), purple background (compared to red background), and non-elongated cell shapes. Dimension 118 was more activated for mucinuous and round structure and less activated for fibrous structures.

By visual inspection, most embedding dimensions encoded a combination of these cellular, tissue and background-stain related characteristics, suggesting a polysemantic representation of atomic properties. Certain combinations of the atomic properties corresponded to complex concepts that were relevant to pathology, such as the distinction between cancer epithelium and stroma tissue (captured in dimension 27 and 147), or the presence of red blood cells (captured in dimension 239). However, the polysemantic features represented in these dimensions prevented interpretability analysis of these dimensions.

Sparse autoencoder models were fit to the CLS token embedding from the output layer (layer 12) of the training dataset (1M dataset). Training the SAE on this dataset (including multiple organs, stains and cell types) leads to generalizable representation of useful features in the embedding dimensions of the model. We visualized the images that have the highest activation value for a given SAE dimension. This revealed highly interpretable features, as shown in Figure 1, including cell and tissue features such as poorly differentiated carcinoma, geometric structures such as vertical fibers, and staining and artifact features.

With the use of the diverse training set, individual dimensions of the trained SAE model exhibited robust representations, where single dimensions represented the same features regardless of stain type. Consistent with this, 247/3072 dimensions (8.0%) had representations of both H & E and IHC stains in the top 100 activating patches, and some of these dimensions represented interpretable concepts across stain types (Figure 1, rightmost column). 374/3072 dimensions (12.2%) were H & E-specific while 1451/3072 dimensions (47.2%) were IHC-specific. This result showed that when trained with diverse datasets, SAE dimensions can represent stain-specific features and exhibit cross-stain generalization.

Training on the diverse dataset reduced the fraction of dead neurons in the SAE intermediate layer compared to a model that was trained only on the TCGA dataset. Similar to previous work for natural language [3], we identified a cluster of "ultra-sparse" features that activated for very few images ($< 0.1\%$ of the dataset). The fraction of these ultra-sparse features were reduced with the incorporation of more diverse training data for the model trained on the 1M dataset (20%) compared to the model trained on TCGA dataset (88%) (see Supplementary section). For subsequent analyses, we used the SAEs trained on the 1M dataset, and used the TCGA dataset as a held-out evaluation dataset.

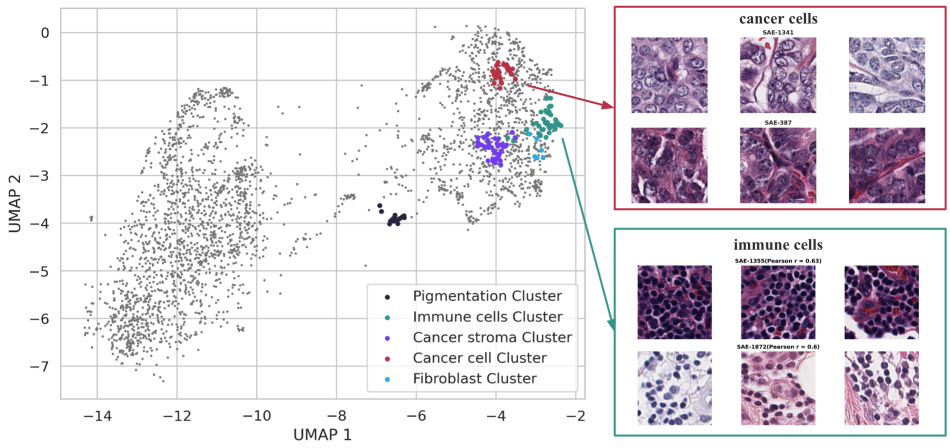
3.2. PLUTO SAE dimensions represent interpretable pathology-relevant concepts

Using the TCGA evaluation dataset, we performed unsupervised clustering on the UMAP representations of the SAE dimensions using HDBSCAN, following the analysis strategy of [3] (Figure 2). To understand the meanings of some of the clusters, we manually examined image patches activating the SAE dimensions within each cluster.

Of the 139 clusters obtained using HDBSCAN, we found clusters (Figure 2) containing SAE features correlated with unique histological concepts such as immune cell presence (Cluster 27), cancer stroma (Cluster 33), fibroblast cells (Cluster 37) and circular cancer cells (Cluster 41). Notably, cluster 0 features were associated with abnormal pigmentation, such as carbon accumulating black anthracotic macrophages (SAE-1745) as well as artifactual pigmentations from residual brown stain (SAE-2034) and from marker ink (SAE-2842) (see Supplementary section).

3.3. Comparison to non-pathology ViT model

To investigate whether the pathology-relevant SAE features emerged due to the pre-training method of PLUTO on pathology images, we evaluated our SAE methodology on a baseline ViT-S that matched PLUTO's design. This ViT-S was trained on ImageNet-1k using the self-supervised



Cluster ID	Cluster name	Histological concepts represented in cluster
0	Abnormal pigmentation	Carbon accumulating black anthracotic macrophages, artifactual pigmentations from residual brown stain, and from marker inks
27	Immune cells	Immune cells such as lymphocytes, plasma cells and macrophages.
33	Cancer stroma	Cancer-associated stroma
37	Fibroblast cells	Fibroblast cells
41	Cancer cells	Circular cancer cells

Figure 2. UMAP of 3072 SAE dimensions from model trained on the 1M dataset. Feature clusters were identified by HDBSCAN and were interpreted by manual inspection. Several clusters clearly associated with histological concepts were highlighted. For cancer and immune cell clusters, visualizations of top 3 patches that maximally activate the SAE dimension were shown.

DINO method, and was obtained from the timm library [9, 16, 43]. We chose the same input patch size as PLUTO and SAE training methodology as in section 2 to ensure fair comparison.

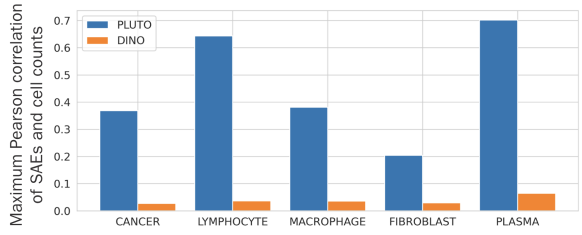


Figure 3. Pearson correlations of SAE dimensions of PLUTO and DINO models with counts of pathology-relevant cell types, showing much higher correlations of the PLUTO SAE dimensions with the cell count features.

To compare the SAE representations of PLUTO and DINO, we used human-interpretable features (HIFs) [14] quantifying tumor microenvironment characteristics such as counts of cancer cells, plasma cells, lymphocytes, macrophages and fibroblasts (see Section 2.3). We first calculated the Pearson's correlation (ρ) of the SAE activation values of the PLUTO with five human-interpretable features representing cell counts. We identified SAE dimensions from PLUTO with the highest correlation with each cell count HIF: plasma cells (ρ = 0.7), lymphocytes (ρ) = 0.63), cancer cells (ρ) = 0.37), macrophages (ρ) = 0.38), and fibroblasts (ρ) = 0.21).

In contrast to the SAE dimensions of PLUTO, SAE dimensions of DINO showed weak association with pathology-relevant concepts. SAE dimensions of this model showed poor correlations with the counts of different cell types (Figure 3). These differences in the representations of the two models demonstrated the value of training on pathology-specific data.

3.4. Evaluation of SAE monosemanticity using pathology-relevant cellular features

To evaluate the monosemanticity of the SAE dimensions, we investigated whether the SAE with highest correlation with the counts of specific cell types also associated with other cell types. SAE-1736, which exhibited a strong correlation with plasma cell counts, showed minimal correlation ($\rho < 0.1$) with other cell types. Images with the highest activation values for SAE-1736 consistently demonstrated

a high presence of plasma cells and captured specific histological features, such as eccentric nuclei surrounded by pale blue cytoplasm. The linear relationship between SAE-1736 activation and plasma cell counts was further illustrated in Figure 4A,C. As the average SAE-1736 activation increased, plasma cell counts rose linearly, while the counts of other cell types remained constant or decreased.

In contrast, no such monosemantic feature was found in the PLUTO embedding space. The strongest plasma cell-associated PLUTO dimension, 148, exhibited only a moderate correlation with plasma cell counts (ρ = 0.29) and was also correlated with the presence of other cell types, as shown in Figure 4B and D. To quantify monosemanticity with respect to these five cellular concepts, we defined a probability distribution $p_i = \frac{|\rho_i|}{\sum_j |\rho_j|}$ where $\rho_1, ...\rho_5$ are the Pearson correlations with the individual cell class densities. Monosemantic features would result in a low entropy S of the distribution. We found that SAE-1736 has a low entropy ($S = 0.42$), while PLUTO dimension 148 has higher entropy ($S = 0.65$, close to the maximum possible entropy $S_{max} = 0.70$). This suggested that SAE dimensions both show stronger correlations with biological concepts (plasma cells), and a higher degree of monosemanticity than original PLUTO embedding dimensions.

4. Emergence of monosemanticity across PLUTO layers

Previous work on convolutional networks [44] and vision transformers [21] suggested emergence of complex features in downstream layers of deep networks. Via SAEs, we investigated how the representation of pathology-relevant concepts evolve across layers of PLUTO.

For this investigation, we extracted the embeddings of the CLS token in every layer of PLUTO and trained separate SAEs on these embeddings. ρ for maximally correlated SAEs per cell type for each layer was shown in Figure5A. At earlier layers (L1 to L6), SAE dimensions correlated with low-level color features such as intensity, hue and saturation (Figure 5B). Correlations of these SAE dimensions with cellular features were low ($\rho < 0.5$ for lymphocytes and $\rho < 0.3$ for the other four cell types).

At later layers (L7 to L12), association of SAE dimensions with color features decreased, while association with cell features increased, suggesting that the feature representations in these layers were biologically meaningful while being invariant to lower level features. More specifically, the SAE representation of plasma cells did not emerge until layer 10 (Figure 5A).

Investigating how plasma cell monosemanticity emerged across layers, we observed that the maximally correlated plasma cell SAEs for earlier layers had higher correlation with lymphocyte counts (e.g. Layer 5 SAE-32 ρ = 0.45, Layer 9 SAE-100 ρ = 0.36, Figure 5C). In other words, embeddings from earlier PLUTO layers could not produce monosemantic SAEs related to plasma cell presence, but might be capturing some simpler and generic characteristics shared between lymphocytes and plasma cells, such as darkly stained nuclei. This was consistent with our observations on correlation of these SAE dimensions with color features. We found a decrease in the entropy of the SAE dimension that showed the strongest association with plasma cell counts (Figure 5D), suggesting increased monosemanticity in later layers.

5. Robustness of SAE representations across domains

5.1. SAE correlations on an out-of-domain evaluation dataset

Finally, we verified that our previous results extended across other domains of pathology images. We extracted embeddings and deployed SAEs on the CPTAC dataset, which covers two different oncology indications. We performed similar correlation analysis between SAE dimensions and cell features. We confirmed the following discoveries (Figure 6): 1. representations of biologically relevant cellular features by SAE evolved across PLUTO layers, and 2. the monosemantic representation of plasma cell by SAE-1736 was robust across different domains of pathology images.

5.2. Robustness of PLUTO to scanners and stains

The low correlation of the SAE dimensions with color features suggested that PLUTO learned a representation of the pathology-relevant concepts that was robust to sources of variation such as stain types and scanner type which can have a large impact on model performance of downstream tasks.

To investigate whether the cell-type specific SAE dimensions as identified in sections 3.4 and 4 were robust to scanners, we computed differences in these SAE activations between the images scanned with GT450 scanners and all other images (quantified by Cohen's d, Figure 7A). The Cohen's d metric was moderate in the earlier layers and decreased to close to zero at the last layer, showing that the cell-type specific dimensions identified above did not show systematic scanner-specific variations. A similar analysis was done on H&E and other stains, showing reduced stain-specific variation in the SAE dimensions trained on the last layers of PLUTO (Figure 7B).

5.3. Feature universality of SAE dimensions

Previous work suggested that SAE dimensions are more likely to be useful (representing true monosemantic features

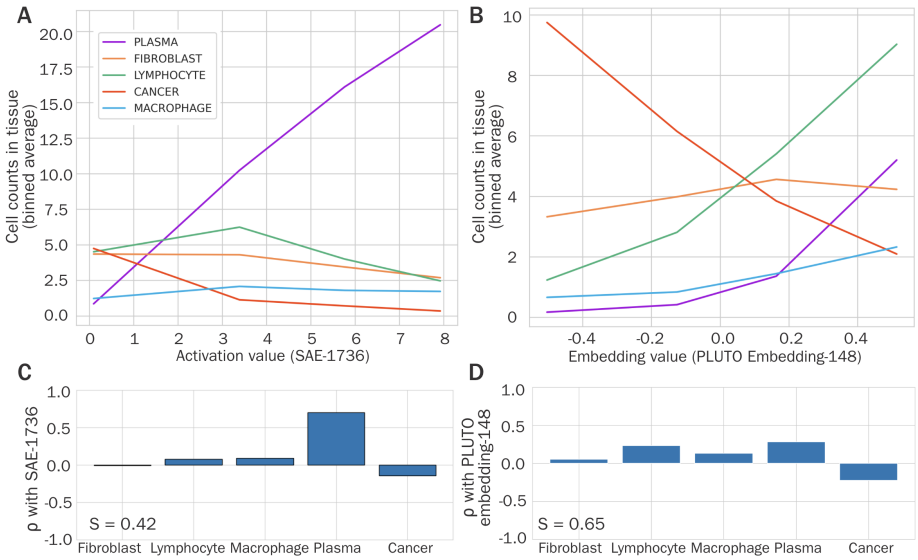


Figure 4. SAE-1736 monosemantically encoded plasma cell-specific information. Top panels show the average cell counts across bins of (A) SAE-1736 activation values, and (B) PLUTO dimension 148. Average plasma cell counts (shown in purple) increased linearly with increasing SAE-1736 activation values, while counts of other cell types decreased or remained constant. In contrast, counts of lymphocytes, macrophages, and plasma cells all increased monotonically with increasing PLUTO-148 feature values. C) Correlation between SAE-1736 activation and counts of five cell types, showing monosemantic correlation with only the plasma cell counts. D) Same as C, but for PLUTO dimension 148

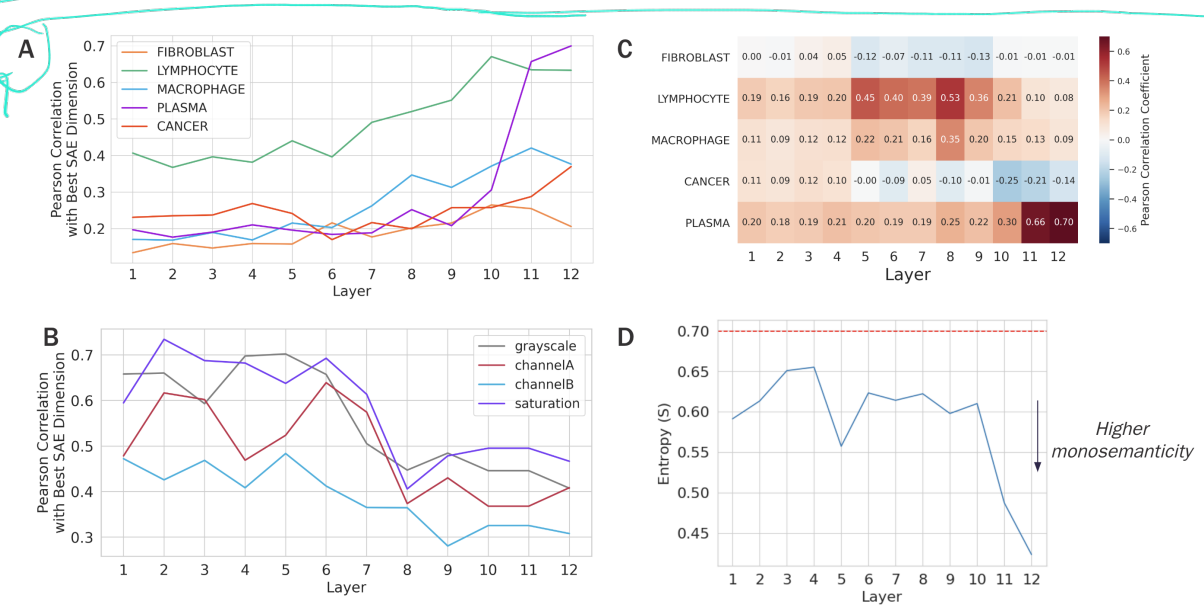


Figure 5. Monosemanticity emerged in later layers of PLUTO A) Correlation of cell count features with dimensions of SAE models trained on the embeddings of PLUTO across layers. In each layer, the five SAE dimensions with the highest correlation with the counts of each of the five cell classes were plotted. B) Correlation of color features with dimensions of SAE models trained on the embeddings of different layers of PLUTO C) For each layer, we found the SAE dimension with the highest correlation with count of plasma cells. We then measured monosemanticity of that dimension by calculating the correlation with other cell type counts. D) Entropy of the best plasma-cell SAE for each layer with respect to the five cell types (lower entropy implies higher monosemanticity). Red dotted line represents maximum possible entropy ($S_{max} = 0.70$)

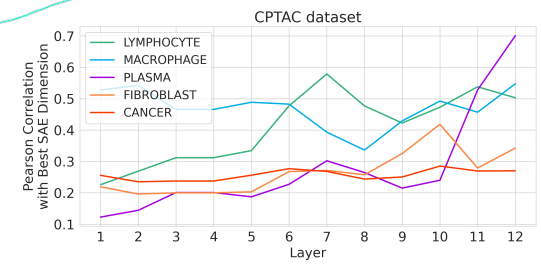


Figure 6. Maximum correlations of SAE dimensions with cellular features in CPTAC dataset. Method for computing the Pearson correlation at each layer is the same as Figure 5A

in the real world) if they display *universality* (the same feature is discovered across independently trained SAE models [3]). We examined feature universality of the SAE dimensions of PLUTO by comparing the SAE activations from two models, one trained on the 1M dataset, and one trained on the TCGA dataset. We found that these two models were able to uncover SAE dimensions that captured the same histological concepts. For example, SAE-1736 from the model trained on the 1M dataset and SAE-2541 from the model trained on the TCGA dataset were highly correlated (ρ = 0.96) and both represent abundance of plasma cells; SAE-1745 from model B and SAE-1667 from model A both represent abundance of anthracotic macrophages (ρ = 0.91) (see Supplementary section). These findings demonstrate the universality of the learned SAE features and suggests generalizability of the SAEs.

6. Limitations and future work

In this work, we restricted our analysis to a vanilla SAE. We left the application of newer variations such as gated SAE and k-sparse SAE in pathology to future work. Similarly, analysis around how these findings translate to SAEs trained on other pathology foundation models can be the subject of further studies. Another area of exploration can be around performing intervention in the SAE latent space and characterizing its impact on robustness to spurious features.

7. Conclusion

We performed an investigation of the features represented in the embedding space of a pathology foundation model. Single embedding dimensions were found to demonstrate polysemanticity in terms of representing higher-order pathology-related concepts composed of atomic characteristics of cellular and tissue properties. Training a SAE enabled the extraction of relatively monosemantic and interpretable features corresponding to distinct biological characteristics, geometric features and image acquisition artifacts. Analysis with human-interpretable features revealed

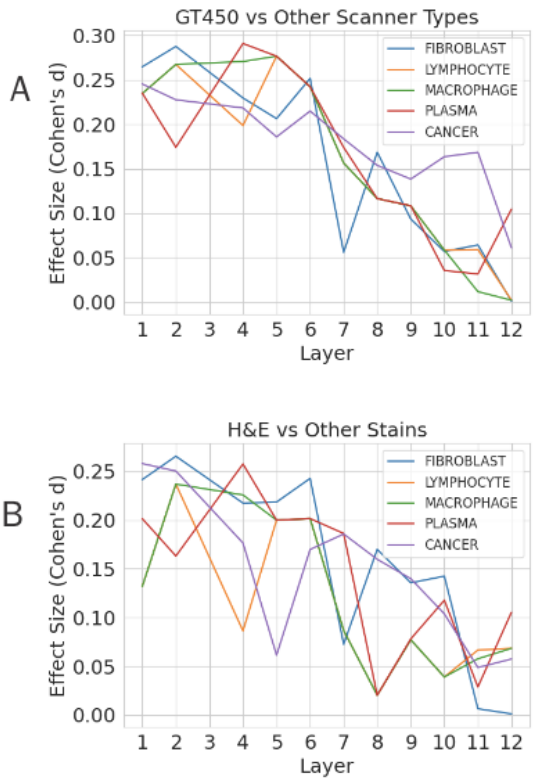


Figure 7. Robustness of cell-type specific SAE dimensions to scanners and stains. In each layer, we identified five SAE dimensions that showed the highest Pearson correlation with counts of fibroblasts, lymphocytes, macrophages, plasma cells and cancer cells. The difference in activation of those SAE dimensions were computed between patches from (A) GT450 versus other scanner types, or (B) H&E versus other stains. Effect sizes were quantified by Cohen's d metric.

correlations of SAE activations with counts of different cell types. Clustering of SAE dimensions revealed distinct groups corresponding to related and interpretable concepts such as anomalous pigmentation, malignant regions and inflammation. These features demonstrated generalization across multiple stains.

An in-depth investigation of feature representations of individual layers from the pathology foundation model provided insight on the evolution of these feature across model depth. Generic features of an image, such as colors, were learned early on by the model, while biologically relevant features, such as abundance of individual cell types, emerged at later layers of the model. Consequently, we demonstrated that the model gains invariance to biologically irrelevant features, such as scanner types. This study provided concrete evidence that embeddings extracted from this pathology foundation model were biologically-grounded, facilitating downstream models that

are built upon these embeddings for solving pathology-relevant tasks. Overall, investigation of sparse features is a promising direction and motivates further work in discovering explainable, generalizable features of pathology foundation models.

References

[1] John Abel, Suyog Jain, Deepta Rajan, Harshith Padigela, Kenneth Leidal, Aaditya Prakash, Jake Conway, Michael Nercessian, Christian Kirkup, Syed Ashar Javed, Raymond Biju, Natalia Harguindeguy, Daniel Shenker, Nicholas Indorf, Darpan Sanghavi, Robert Egger, Benjamin Trotter, Ylaine Gerardin, Jacqueline A. Brosnan-Cashman, Aditya Dhoot, Michael C. Montalto, Chintan Parmar, Ilan Wapinski, Archit Khosla, Michael G. Drage, Limin Yu, and Amaro Taylor-Weiner. Ai powered quantification of nuclear morphology in cancers enables prediction of genome instability and prognosis. *npj Precision Oncology*, 8(1):134, 2024. 3

[2] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety – a review, 2024. 1

[3] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. 1, 4, 8

[4] Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, and Geert Litjens. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology*, 21(2):233–241, 2020. 2

[5] Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. Thread: Circuits. *Distill*, 2020. https://distill.pub/2020/circuits. 1

[6] Nick Cammarata, Gabriel Goh, Shan Carter, Ludwig Schubert, Michael Petrov, and Chris Olah. Curve detectors. *Distill*, 2020. https://distill.pub/2020/circuits/curve-detectors. 1

[7] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 2

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *CoRR*, abs/2104.14294, 2021. 3

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5

[10] Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldwosky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing, a method for rigorously testing interpretability hypotheses. *AI Alignment Forum*, 2022. 2

[11] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024. 2

[12] Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations, 2023. 2

[13] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. 1

[14] James A Diao, Jason K Wang, Wan Fung Chui, Victoria Mountain, Sai Chowdary Gullapally, Ramprakash Srinivasan, Richard N Mitchell, Benjamin Glass, Sara Hoffman, Sudha K Rao, et al. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nature communications*, 12 (1):1–15, 2021. 2, 5

[15] Jonas Dippel, Barbara Feulner, Tobias Winterhoff, Simon Schallenberg, Andreas Kunft Gabriel Dernbach, Stephan Tietz, Timo Milbich, Simon Heinke, Marie-Lisa Eich, Julika Ribbat-Idel, Rosemarie Krupar, Philipp Jurmeister, David Horst, Lukas Ruff, Klaus-Robert Müller, Frederick Klauschen, and Maximilian Alber. RudolfV: A Foundation Model by Pathologists for Pathologists, 2024. 2

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 5

[17] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. 1

[18] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. 1

[19] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling Self-Supervised Learning

for Histopathology with Masked Image Modeling. *medRxiv*, 2023. 2

[20] AI Safety Foundation. Sparse autoencoder, 2024. Accessed: 2024-08-29. 4

[21] Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. What do vision transformers learn? a visual exploration, 2022. 6

[22] Liv Gorton. The missing curve detectors of inceptionv1: Applying sparse autoencoders to inceptionv1 early vision, 2024. 1

[23] Frederick M. Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I. Olopade, Jakob N. Kather, Nicole Cipriani, Robert Grossman, and Alexander T. Pearson. The impact of digital histopathology batch effect on deep learning model accuracy and bias. *bioRxiv*, 2020. 2

[24] Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive mil: Intrinsically interpretable multiple instance learning for pathology, 2022. 2

[25] Dinkar Juyal, Harshith Padigela, Chintan Shah, Daniel Shenker, Natalia Harguindeguy, Yi Liu, Blake Martin, Yibo Zhang, Michael Nercessian, Miles Markey, Isaac Finberg, Kelsey Luu, Daniel Borders, Syed Ashar Javed, Emma Krause, Raymond Biju, Aashish Sood, Allen Ma, Jackson Nyman, John Shamshoian, Guillaume Chhor, Darpan Sanghavi, Marc Thibault, Limin Yu, Fedaa Najdawi, Jennifer A. Hipp, Darren Fahy, Benjamin Glass, Eric Walk, John Abel, Harsha Pokkalla, Andrew H. Beck, and Sean Grullon. Pluto: Pathology-universal transformer, 2024. 3

[26] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sergio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3344–3354, 2023. 2

[27] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data efficient and weakly supervised computational pathology on whole slide images, 2020. 2

[28] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016. 20th anniversary of the Medical Image Analysis journal (MedIA). 2

[29] Alireza Makhzani and Brendan Frey. k-sparse autoencoders, 2014. 1

[30] Miles Markey, Juhyun Kim, Zvi Goldstein, Ylaine Gerardin, Jacqueline Brosnan-Cashman, Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Limin Yu, Bahar Rahsepar, et al. Abstract b010: Spatially-resolved prediction of gene expression signatures in h&e whole slide images using additive multiple instance learning models. *Molecular Cancer Therapeutics*, 22(12_Supplement):B010–B010, 2023. 3

[31] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023. 1

[32] Christopher Olah. Mechanistic interpretability, variables, and the importance of interpretable bases, 2022. 1

[33] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits, 2020. https://distill.pub/2020/circuits/zoom-in. 1, 2

[34] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. 1

[35] Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders, 2024. 1

[36] Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders, 2024. 1

[37] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. 1

[38] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias, 2020. 2

[39] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Juan Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David Klimstra, Brandon Rothrock, and Thomas J. Fuchs. Virchow: A Million-Slide Digital Pathology Foundation Model, 2023. 2

[40] Eric E Walk. The role of pathologists in the era of personalized medicine. *Archives of pathology & laboratory medicine*, 133(4):605–610, 2009. 2

[41] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016. 2

[42] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer

genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013. 3

[43] Ross Wightman. Pytorch image models. https://github.com/huggingface/pytorch-image-models, 2019. 5

[44] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks, 2013. 6

Learning biologically relevant features in a pathology foundation model using sparse autoencoders

Supplementary Material

S8. Quantification of dead and ultra-low density neurons across SAE models

For each SAE dimension, we computed the number of dead neurons (neurons that showed no activation across the entire dataset), and the number of ultra-low density features (features where the fraction of active neurons is below 0.1%).

Dataset	Expansion factor	Common	Dead	Ultra-low density
TCGA	8	9.1%	2.7%	88.2%
1M	1	99.7%	0%	0.3%
1M	8	80%	0.2%	20%
1M	16	62%	0.7%	37%
1M	32	46%	1.4%	53%

Table S1. Fraction of features belonging to the common, dead or ultra-low density feature groups across models.

S9. Visualization of individual SAE features within key UMAP clusters

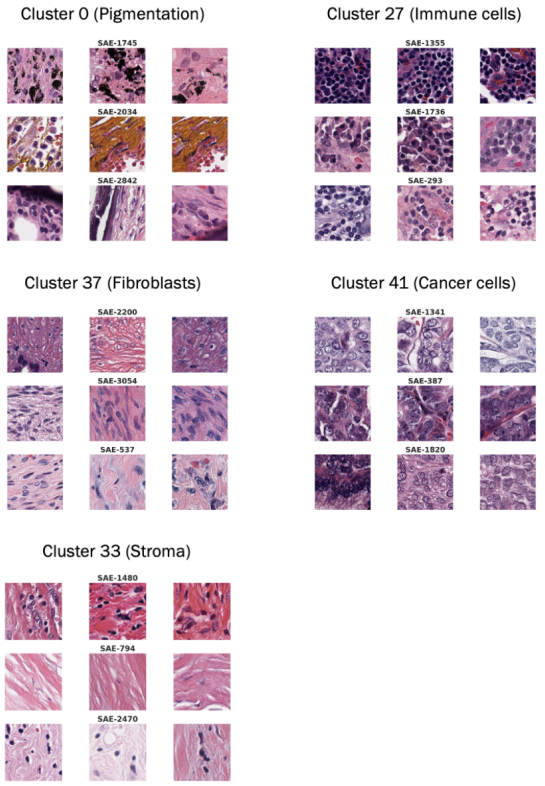


Figure S8. Visualization of features within key clusters identified by the UMAP analysis. For each cluster, each row represents a SAE dimension from that cluster, and shows 3 patches that maximally activate that dimension.

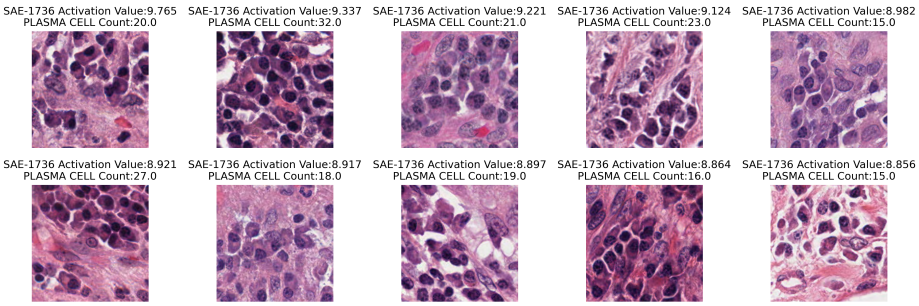


Figure S9. SAE-1736 captures plasma cell histology. Top-10 images with the highest SAE-1736 activation values and the corresponding plasma cell counts are shown.

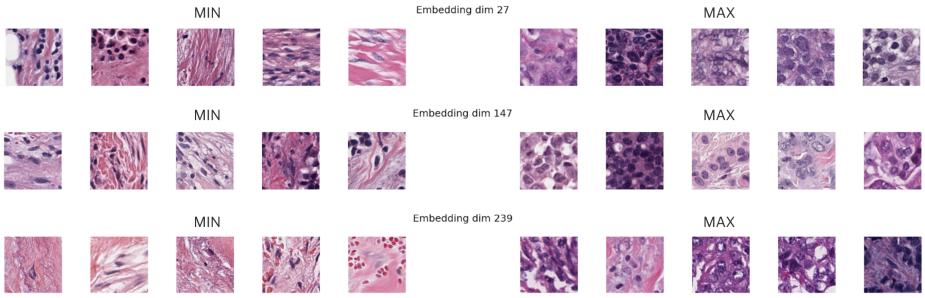


Figure S10. Visualization of features activating each embedding dimension. In each dimension, 5 example patches in the lowest 5% and highest 5% respectively of that dimension's activation are visualized. Inspection of each these patches reveals that multiple atomic features vary within each embedding dimension, including background stain color, cell size, shapes or morphologies. Some dimensions correspond to complex concepts that are relevant to pathology.

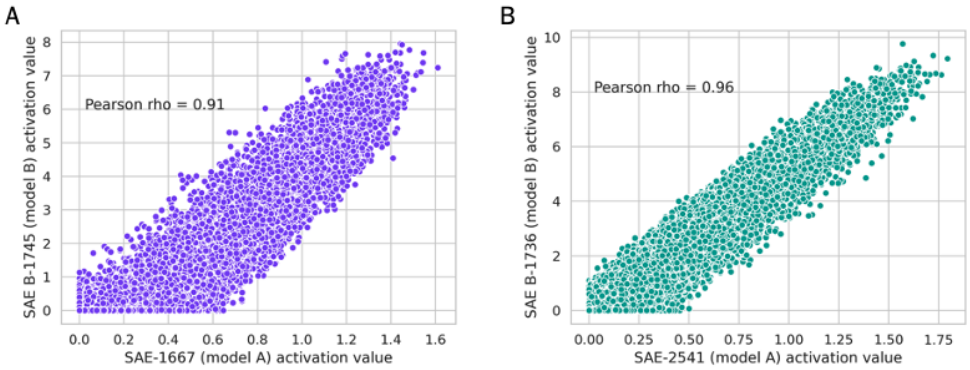


Figure S11. A) Anthracotic macrophage SAE feature comparison between model A and B. B) Plasma cell SAE feature comparison between model A and B. The high correlation values demonstrate that models trained on different datasets are able to uncover SAE dimensions that capture the same histological concepts.