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

Nhat M. Le[*]
PathAI
Boston, Massachusetts, USA
nhat.le@pathai.com

Neel Patel[*]
PathAI
Boston, Massachusetts, USA
neel.patel@pathai.com

Ciyue Shen
PathAI
Boston, Massachusetts, USA
judy.shen@pathai.com

Blake Martin
PathAI
Boston, Massachusetts, USA
blake.martin@pathai.com

Alfred Eng
PathAI
Boston, Massachusetts, USA
alfred.eng@pathai.com

Chintan Shah
PathAI
Boston, Massachusetts, USA
chintan.shah@pathai.com

Sean Grullon
PathAI
Boston, Massachusetts, USA
sean.grullon@pathai.com

Dinkar Juyal
PathAI
Boston, Massachusetts, USA
dinkar.juyal@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 in 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 model activations. In this work, we train a Sparse Autoencoder on the embeddings of a pathology pretrained foundation model. We discover an interpretable sparse representation of biological concepts within the model embedding space. We perform an investigation into how these representations are associated with quantitative human-interpretable features. Our work paves 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, to neural networks used in pathology.

Mechanistic interpretability aims to study neural networks by reverse-engineering them, providing insights into their internal workings Olah [2022], Cammarata et al. [2020a], Elhage et al. [2021],

[*] These authors contributed equally to this work

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Bereska and Gavves [2024]. 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 Cammarata et al. [2020a]. This conceptualization allows researchers to dissect complex neural networks and understand how they process and represent information.

According to the Superposition Hypothesis Elhage et al. [2022], Olah et al. [2020], 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.

Bricken *et al.* Bricken et al. [2023] use Sparse Autoencoders – a form of dictionary learning – to decompose multilayer perceptron (MLP) activations into a number of features greater than the number of neurons. The aim is to associate features with individual neurons that represent disentangled concepts in these sparse networks. This approach holds promise for improving the interpretability of AI systems in pathology, potentially allowing for more precise identification of diagnostic features.

Nanda *et al.* Nanda et al. [2023b] provide evidence that these features are linear combinations of neurons for OthelloGPT, in line with the linear representation hypothesis proposed by Mikolov et al. [2013]. This finding suggests that complex concepts in neural networks, including those used in pathology applications, may be represented as linear combinations of simpler features.

In Large Language Models (LLMs), MI has been used to understand phenomena such as in-context learning Olsson et al. [2022], grokking Nanda et al. [2023a], and uncovering biases and deceptive behavior Templeton et al. [2024]. While these studies primarily focus on language models, their insights may have implications for image-based AI systems used in pathology. The Universality Hypothesis Olah et al. [2020] posits that similar features and circuits are learned across different models and tasks. However, other studies Chughtai et al. [2023] 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.

Sparse autoencoders have emerged as an important tool for extracting monosemantic features from the embeddings of complex models Bricken et al. [2023], Cunningham et al. [2023], Rajamanoharan et al. [2024a], Makhzani and Frey [2014]. This paper aims to explore the application of sparse autoencoders in disentangling neural representations in pathology-focused self-supervised models, investigate the presence and implications of polysemantic neurons in these systems, and examine the potential of mechanistic interpretability techniques to improve the transparency and reliability of AI-assisted pathology diagnostics. By advancing our understanding of these areas, we seek to contribute to the development of more interpretable and trustworthy AI systems in pathology, ultimately enhancing their utility and acceptance in clinical practice.

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 Walk [2009], Madabhushi and Lee [2016]. 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 Bulten et al. [2020], Campanella et al. [2019], Wang et al. [2016]. Work on interpretability in pathology has focused on assigning spatial credit to WSI-level predictions Javed et al. [2022], Lu et al. [2020], computing human-interpretable features from model output heatmaps Diao et al. [2021], and visualization of multi-head self-attention values on image patches Chen et al. [2024].

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 Kang et al. [2023], Dippel et al. [2024], Vorontsov et al. [2023], Filiot et al. [2023], Chen et al. [2024]. 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 Howard et al. [2020]. 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 Vig et al. [2020], Chan et al. [2022] to text-based interventions, such as predicting tissue changes in response to drug administration.
- **Multimodal Integration:** Medicine is inherently multimodal Topol [2023]. Recent advances in spatial biology provide opportunities to draw connections and learn shared patterns across modalities like histopathology, genomics, and transcriptomics Bressan et al. [2023]. MI can help in understanding these cross-modal relationships.
- **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.

These factors highlight the potential of MI to significantly advance our understanding and application of AI in pathology, ultimately improving diagnostic accuracy and treatment decisions in healthcare.

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. Our study provides the first detailed characterization of the image attributes represented within specific embedding dimensions of a pathology foundation model. To move towards monosemantic representations, we employ sparse autoencoders (SAEs) on the embedding outputs, aiming to identify interpretable features within the SAE's hidden dimensions. Further interpretability analysis of these hidden dimensions revealed clusters of related histopathology concepts, and correlation between single SAE dimensions with human-interpretable features characterizing cell densities.

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 autoencoder to enable the disentanglement of polysemantic embedding dimensions, revealing a sparse dictionary of interpretable features that represent cell and tissue characteristics, geometric structures, and image artifacts.
- We examine the effect of training SAEs on complex datasets consisting of multiple stain types, uncovering lower fraction of dead neurons and ultra-sparse features, and identifying features that generalize across multiple staining techniques.

- We perform a clustering analysis on the SAE dimensions, identify groups of related features that encode for related histopathology concepts.
- We conduct quantitative comparisons between human-interpretable features and distinct SAE dimensions, finding varying degrees of correlation across different cell types.

2 Polysemanticity in pathology foundation model embeddings

2.1 Datasets and embedding extraction

We use 2 datasets for experimentation, which we term as dataset A and B. For dataset A, we used three publicly available TCGA (The Cancer Genome Atlas) Weinstein et al. [2013] datasets consisting of H & E (haematoxylin & eosin)-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. A machine-learning model, PathExplore (PathExplore is for research use only. Not for use in diagnostic procedures.) Markey et al. [2023], Abel et al. [2024], was deployed on these images to detect and classify cell types from the WSIs. On each slide, we sampled 100 cells from each cell type (cancer cells, lymphocytes, macrophages, fibroblasts, plasma cells, and indication-specific cell types). Image patches (224 x 224 pixels at a high resolution, 0.25 microns per pixel) were created centered on the selected cells.

For dataset B, we used 1.1 million image patches, including both H & E and IHC (immunohistochemistry) stains, sampled from the train set of 'PLUTO' - a pathology pretrained foundation model Juyal et al. [2024], covering oncology, IBD (inflammatory bowel disease) and MASH (metabolic dysfunction-associated steatohepatitis). All the images for dataset A and B were passed through a frozen ViT-Small encoder taken from 'PLUTO'. Each image patch outputs a 384-dimensional embedding vector corresponding to the CLS token.

2.2 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 (Figure 1).

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

By visual inspection, most embedding dimensions similarly encode a combination of these cellular, tissue and background-stain related characteristics, suggesting a polysemantic representation of these atomic properties. *Certain combinations of the atomic properties correspond to complex concepts that are 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 multiple features represented in these dimensions prevented interpretability analysis of these dimensions.

3 Training a sparse autoencoder on PLUTO embeddings reveals interpretable features

Sparse autoencoders (SAEs) have been used in NLP Bricken et al. [2023], Cunningham et al. [2023] 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 Gorton [2024], Cammarata et al. [2020b]. Various improvements to SAEs have been suggested, including k-sparse Makhzani and Frey [2014] and gated sparse Rajamanoharan et al. [2024a] autoencoders, and using JumpReLU Rajamanoharan et al. [2024b] instead of ReLU as the activation function. Inspired by previous work, we investigate training SAEs on top of PLUTO's embeddings and analyzing the sparse features for interpretable dimensions.

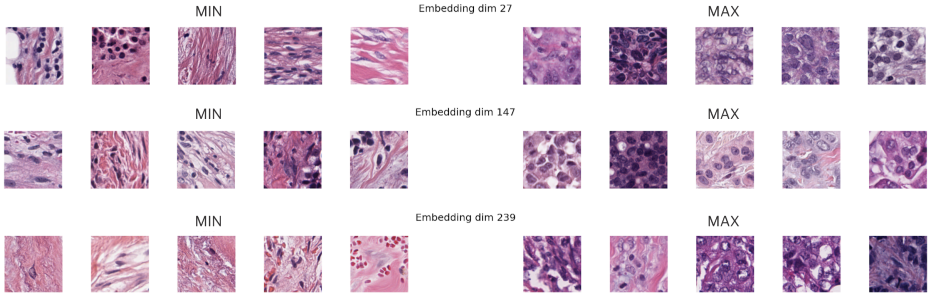

Figure 1: Visualization of features activating each embedding dimension. In each dimension, 5 example patches in the lowest 5% and highest 5% respectively of that dimension's activation are visualized. Inspection of each these patches reveals that multiple atomic features vary within each embedding dimension, including background stain color, cell size, shapes or morphologies. Some dimensions correspond to complex concepts that are relevant to pathology.

Two sparse autoencoder models were fit separately to the CLS token embedding of datasets A and B. Our hypothesis is that training SAEs on a more diverse dataset (including multiple organs, stains and cell types) leads to more generalizable representation of useful features in the embedding dimensions of the model. For simplicity, we will refer the first model as "model A" and the second model as "model B".

The two SAEs use an expansion factor of 8 and 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 Bricken et al. [2023], Foundation [2024]. Dead neuron resampling was implemented to reduce the fraction of dead neurons Bricken et al. [2023], Foundation [2024]. We tried 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. A single training run took approximately 30 minutes on a Quadro RTX 8000 GPU. The fraction of dead neurons remains lower than 4% for different values of hyperparameters.

3.1 Visualization of learned SAE features

We visualized the images that have the highest activation value for a given SAE dimension. This revealed highly interpretable features, as shown in Figure 2. These include cell and tissue features such as poorly differentiated carcinoma, geometric structures such as vertical fibers, and staining and artifact features.

With the incorporation of diverse training data in model B, SAE dimensions of model B exhibited multimodal representations, where single SAE dimensions represent 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 represent interpretable concepts across stain types (Figure 2, rightmost column). 374/3072 dimensions (12.2%) were H & E-specific while 1451/3072 dimensions (47.2%) were IHC-specific. This result shows that when trained with diverse datasets, SAE dimensions can represent both stain-specific features and exhibit cross-stain generalization.

Training on the diverse dataset (dataset B) reduced the fraction of dead neurons in the SAE intermediate layer. Similar to previous work for natural language Bricken et al. [2023], 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 are reduced with the incorporation of more diverse training data for model B (20%) compared to model A (88%).

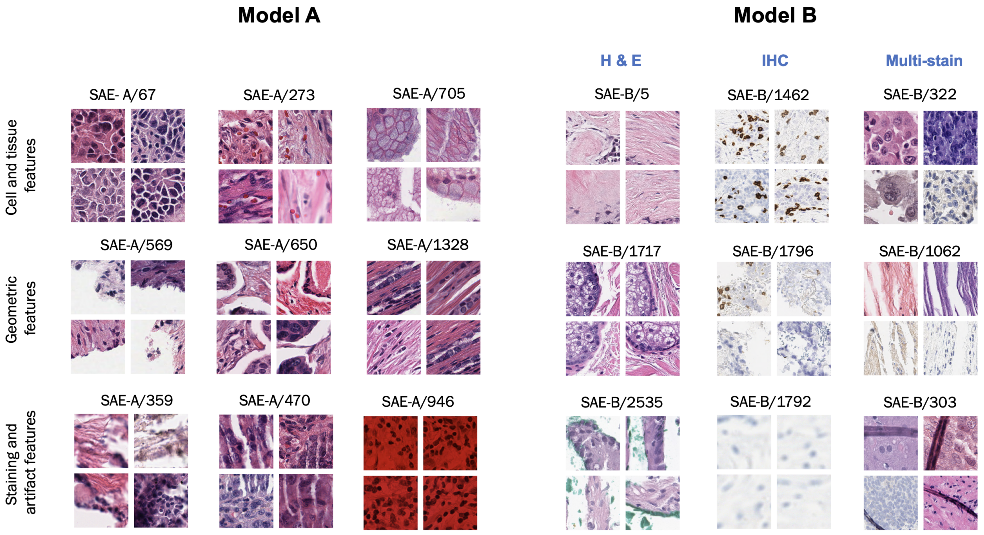

Figure 2: Feature visualization of SAE hidden dimensions reveals interpretable dictionary of pathology features. For each SAE hidden dimension of model A and model B, 4 out of the top 16 images that activated that dimension are visualized. Manual examination revealed interpretable features represented 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 are specific to H & E stain (first column: collagen-enriched fibroblasts, circular clusters of tumor cells, surgical ink), some are specific to IHC stain (second column: stained lymphocytes, edge of tissue, blur), and others generalize across stains (third column: large cancer cells, vertical structures, tissue folds).

3.2 Unsupervised clustering of SAE dimensions reveal distinct clusters of histological concepts

Pathology domain presents continuous, quantifiable and clinically relevant features, such as cell type density and area of tissue regions. We perform experiments to determine whether these features can be captured within single SAE dimensions.

Using dataset A as a held-out set for model B, we performed unsupervised clustering on the UMAP representations of the SAE dimensions using HDBSCAN, following the analysis strategy of Bricken et al. [2023] (Figure 3). 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, shown in Figure 4, containing SAE features correlated with unique histogical concepts such as immune cell presence (Cluster 27), cancer stroma (Cluster 33), fibroblast cells (Cluster 37) and circular cancer cells (Cluster 41) (Table 1). 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) (Figure 4).

3.3 Biological interpretability of SAE dimensions

In order to further understand individual SAE dimensions, we calculated the Pearson's correlation (ρ) of the activation values with human-interpretable features (HIFs) Diao et al. [2021] quantifying tumor microenvironment characteristics such as counts of cancer cells, plasma cells, lymphocytes,

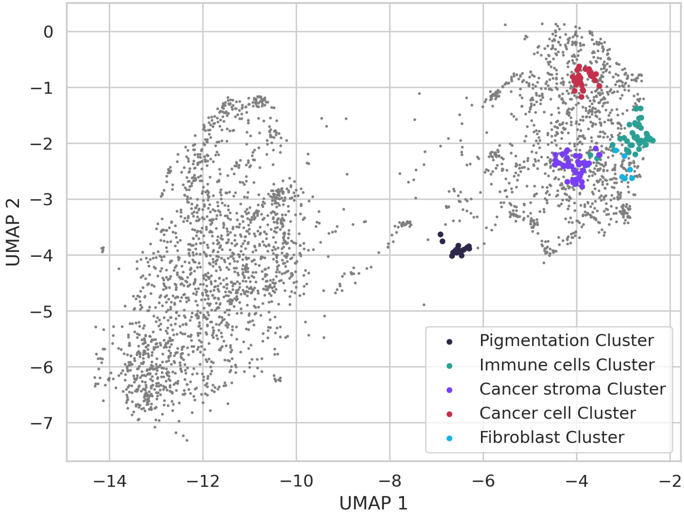

Figure 3: UMAP of 3072 SAE features from model B. Several clusters clearly associated with histological concepts are highlighted.

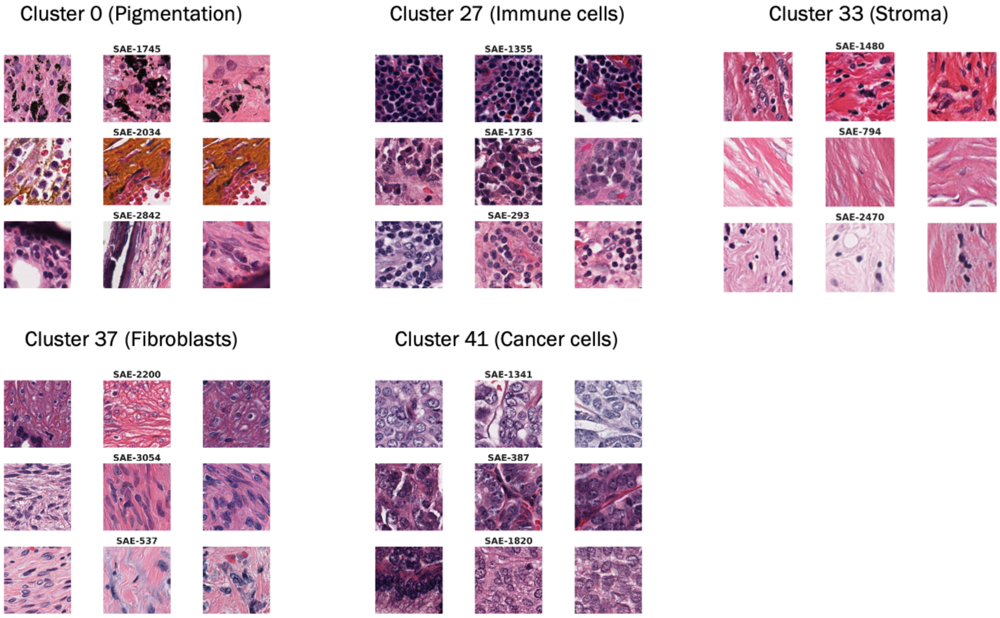

Figure 4: Visualization of features within key clusters identified by the UMAP analysis. For each cluster, each row represents an SAE dimension from that cluster, and shows 3 patches that maximally activate that dimension.

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

Table 1: Characterization of SAE feature clusters identified by the UMAP analysis. Feature clusters were identified by HDBSCAN and were interpreted by manual inspection.

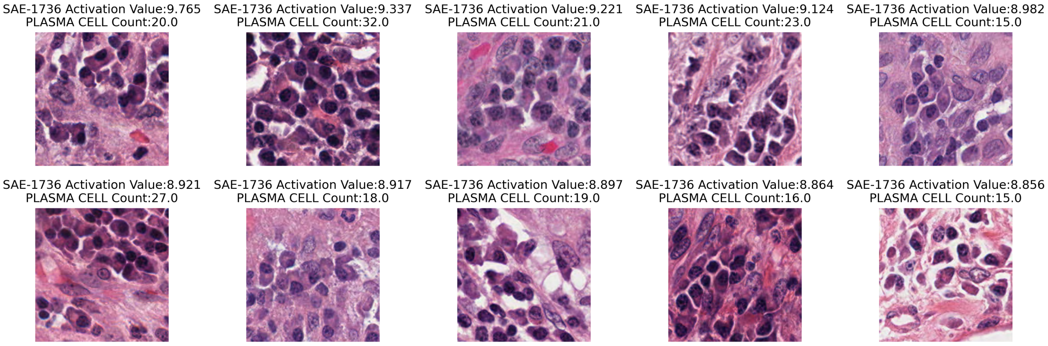
SAE-1736 Activation Value:9.765
PLASMA CELL Count:20.0
SAE-1736 Activation Value:9.337
PLASMA CELL Count:32.0
SAE-1736 Activation Value:9.221
PLASMA CELL Count:21.0
SAE-1736 Activation Value:9.124
PLASMA CELL Count:23.0
SAE-1736 Activation Value:8.982
PLASMA CELL Count:15.0
SAE-1736 Activation Value:8.921
PLASMA CELL Count:27.0
SAE-1736 Activation Value:8.917
PLASMA CELL Count:18.0
SAE-1736 Activation Value:8.897
PLASMA CELL Count:19.0
SAE-1736 Activation Value:8.864
PLASMA CELL Count:16.0
SAE-1736 Activation Value:8.856
PLASMA CELL Count:15.0

Figure 5: SAE-1736 captures plasma cell histology. Top-10 images with the highest SAE-1736 activation values and the corresponding plasma cell counts are shown.

macrophages and fibroblasts. These cell types possess distinct morphological characteristics, that may be captured by monosemantic SAE dimensions. To that end, we identified the following dimensions with the highest correlation with each cell count HIF: SAE-1736 with plasma cells (ρ = 0.7), SAE-1355 with lymphocytes (ρ = 0.63), SAE-1341 with cancer cells (ρ = 0.37), SAE-293 with macrophages (ρ = 0.31), and SAE-825 with fibroblasts (ρ = 0.21). The immune cell Cluster 27, identified in the previous section, contained SAE-1355, SAE-1736 and SAE-293 and the cancer cell Cluster 41 contained SAE-1341. SAE-825, although unclustered, was very close to other fibroblast features in Cluster 37 in the UMAP embeddings space.

Notably, 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, as shown in Figure 5. The linear relationship between SAE-1736 activation and plasma cell counts is further illustrated in Figure 6. As the average SAE-1736 activation increases, plasma cell counts rise steeply and linearly, while the counts of other cell types remain constant or decrease.

In contrast, a similar monosemantic feature was not 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 6. This highlights the unique monosemantic nature of the SAE-1736 dimension, which encodes plasma cell-specific characteristics that were not captured by the PLUTO embeddings.

3.4 Feature universality of SAE dimensions

We then examine the feature universality of the SAE dimensions by comparing the SAE activations from model A to those from model B. We found that models trained on different datasets are able to uncover SAE dimensions that capture the same histological concepts. For example, SAE-1736 from model B and SAE-2541 from model A are 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

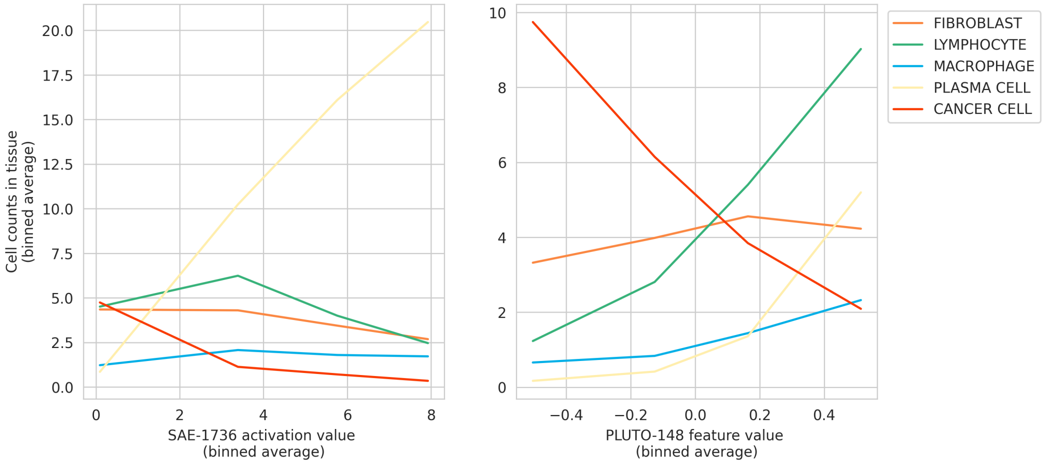

Figure 6: SAE-1736 monosemantically encodes plasma cell-specific information. The left plot shows average cell counts across bins of SAE-1736 activation values, while the right plot shows the same across bins of PLUTO dimension 148. Average plasma cell counts (shown in yellow) increase linearly with increasing SAE-1736 activation values, while counts of other cell types decrease or remain constant. In contrast, counts of lymphocytes, macrophages, and plasma cells all increase monotonically with increasing PLUTO-148 feature values.

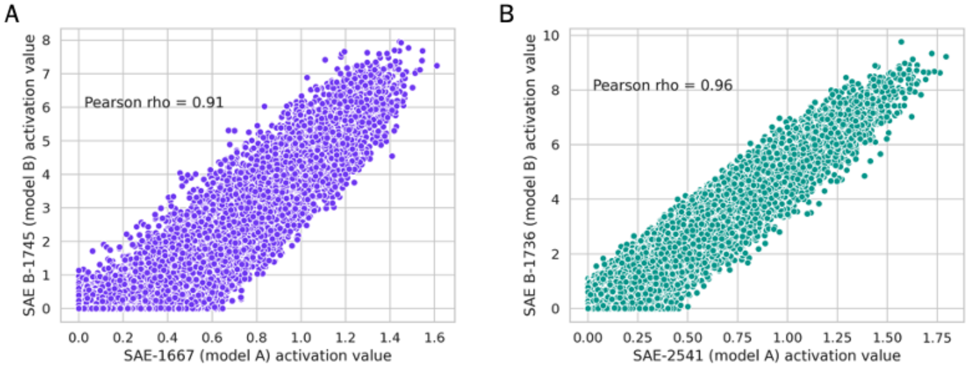

Figure 7: A) Anthracotic macrophage SAE feature comparison between model A and B. B) Plasma cell SAE feature comparison between model A and B. The high correlation values demonstrate that models trained on different datasets are able to uncover SAE dimensions that capture the same histological concepts

anthracotic macrophages ($\rho = 0.91$) (Figure 7). These findings demonstrate the universality of the learned SAE features and suggests generalizability of the SAEs.

4 Conclusion

We performed a preliminary 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 sparse autoencoder enables the extraction of relatively monosemantic and interpretable features corresponding to distinct biological characteristics, geometric features and image acquisition artifacts. These features demonstrate generalization across multiple stains.

Analysis with human-interpretable features reveals correlations of SAE activations with counts of different cell types. Clustering of SAE dimensions reveals distinct groups corresponding to related and interpretable concepts such as anomalous pigmentation, malignant regions and inflammation.

Our work is one of the first investigations of sparse features of pathology foundation models. To address some limitations of this study, future directions will include comparative analysis using other interpretability techniques and baseline models, and investigating the generalizability of the results using diverse datasets. Overall, investigation of sparse features is a promising direction and motivates further work in discovering explainable, generalizable features of pathology foundation models.

References

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, Jun 2024. ISSN 2397-768X. doi: 10.1038/s41698-024-00623-9. URL https://doi.org/10.1038/s41698-024-00623-9.

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

Dario Bressan, Giorgia Battistoni, and Gregory J. Hannon. The dawn of spatial omics. *Science*, 381(6657):eabq4964, 2023. doi: 10.1126/science.abq4964. URL https://www.science.org/doi/abs/10.1126/science.abq4964.

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.

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.

Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. Thread: Circuits. *Distill*, 2020a. doi: 10.23915/distill.00024. https://distill.pub/2020/circuits.

Nick Cammarata, Gabriel Goh, Shan Carter, Ludwig Schubert, Michael Petrov, and Chris Olah. Curve detectors. *Distill*, 2020b. doi: 10.23915/distill.00024.003. https://distill.pub/2020/circuits/curve-detectors.

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.

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.

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.

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

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

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.

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.

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.

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.

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. doi: 10.1101/2023.07.21.23292757. URL https://www.medrxiv.org/content/early/2023/07/26/2023.07.21.23292757.

AI Safety Foundation. Sparse autoencoder, 2024. URL https://github.com/ai-safety-foundation/sparse_autoencoder. Accessed: 2024-08-29.

Liv Gorton. The missing curve detectors of inceptionv1: Applying sparse autoencoders to inceptionv1 early vision, 2024. URL https://arxiv.org/abs/2406.03662.

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. doi: 10.1101/2020.12.03.410845. URL https://www.biorxiv.org/content/early/2020/12/05/2020.12.03.410845.

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.

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.

Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio 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, June 2023.

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.

Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2016.06.037. URL https://www.sciencedirect.com/

science/article/pii/S1361841516301141. 20th anniversary of the Medical Image Analysis journal (MedIA).

Alireza Makhzani and Brendan Frey. k-sparse autoencoders, 2014. URL https://arxiv.org/abs/1312.5663.

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.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090.

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

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models, 2023b.

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.

Christopher Olah. Mechanistic interpretability, variables, and the importance of interpretable bases, 2022.

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.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders, 2024a. URL https://arxiv.org/abs/2404.16014.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders, 2024b. URL https://arxiv.org/abs/2407.14435.

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. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Eric J. Topol. As artificial intelligence goes multimodal, medical applications multiply. *Science*, 381(6663):eadk6139, 2023. doi: 10.1126/science.adk6139. URL https://www.science.org/doi/abs/10.1126/science.adk6139.

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.

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.

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

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.

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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The experiments in the main body of the paper support the claims in the abstract and the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed limitations in the Conclusion section and future directions to address the limitations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not have any novel theoretical result.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We discuss the dataset composition and SAE training hyperparameters.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: While we cannot share the code, we cite the open-source SparseAutoencoder library we use in the main text.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details are shared in the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: While we do not report error bars, results are presented as descriptive statistics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have discussed training time and compute resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms to the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This is a preliminary investigation, we leave a thorough assessment of societal impact to future work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Asset creators are credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets are released.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing is used.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.