Interpretability analysis on a pathology foundation model reveals biologically relevant embeddings across modalities

Nhat Le [1] **Ciyue Shen** [1] **Chintan Shah** [1] **Blake Martin** [1] **Daniel Shenker** [1] **Harshith Padigela** [1] **Jennifer Hipp** [1] **Sean Grullon** [1] **John Abel** [1] **Harsha Vardhan Pokkalla** [1 2] **Dinkar Juyal** [1]

Abstract

Mechanistic interpretability has been explored in detail for large language models (LLMs). For the first time, we provide a preliminary investigation with similar interpretability methods for medical imaging. Specifically, we analyze the features from a ViT-Small encoder obtained from a pathology Foundation Model via application to two datasets: one dataset of pathology images, and one dataset of pathology images paired with spatial transcriptomics. We discover an interpretable representation of cell and tissue morphology, along with gene expression within the model embedding space. 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

Mechanisitic interpretability (MI) aims to study neural networks by reverse-engineering them (Olah, 2022; Cammarata et al., 2020a; Elhage et al., 2021; Bereska & Gavves, 2024). Under this paradigm, "features" are defined as the fundamental units of neural networks, and "circuits" are formed by connecting features via weights (Cammarata et al., 2020a). 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. 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, with the aim to associate features with individual neurons that represent disentangled concepts in these sparse networks. 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). In LLMs, MI has been used to understand phenomenon such as in-context learning (Olsson et al., 2022), grokking (Nanda et al., 2023a), uncovering biases and deceptive behavior (Templeton et al., 2024). While the Universality Hypothesis (Olah et al., 2020) states that similar features and circuits are learned across different models tasks, other studies (Chughtai et al., 2023) found mixed evidence for this claim.

1.2. Interpretability in pathology

Histopathology (interchangeably used with pathology) is the diagnosis and study of diseases, involving microscopic examination of cells and tissues, and plays a critical role in disease diagnosis and grading, treatment decision-making, and drug development (Walk, 2009; Madabhushi & Lee, 2016). Digitized whole-slide images (WSIs) of pathology samples can be gigapixel-sized, contain millions of areas of interest and contain biologically relevant entities across a wide range of characteristic length scales. 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).

Spatial gene expression provides rich information on how cells behave and biological processes happen in their spatial contexts. In cancer biology, they can be useful to study tumor heterogeneity, reflect disease mechanisms, and derive prognostic markers for cancer (Arora et al., 2023; Denisenko et al., 2024). Studying interpretability in the context of spatial gene expression can establish the connections between different biological modalities, (histopathology and

[1]PathAI, Boston, Massachusetts, USA [2]Employee of PathAI at the time of study. Correspondence to: Nhat Le <nhat.le@pathai.com>.

transcriptomics in our case), and therefore increase our confidence in the utility of a foundation model in multimodal settings.

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

- Most current image datasets are object-centric, and interpretability analysis is often restricted on these datasets. In contrast, a single pathology image patch can contain 10^6 regions of interest (i.e, cell nuclei). The number of active concepts are bounded by the underlying biological structures, and identifying every concept can be critical depending on downstream applications.
- A significant challenge in such datasets is "batch effects," where models can latch on to spurious features instead of learning relevant morphology-related features. This problem can be particularly insidious in pathology images due to reasons such as high-frequency artifacts and systematic confounders from image acquisition, etc. (Howard et al., 2020). While previous attempts have been made to disentangle biological content from incidental attributes (Nguyen et al., 2024), having a better understanding of which circuits correspond to each of these categories can lead to the design of more robust models for real-world applications.
- Having a bottom-up understanding of which features contribute to certain predictions for an image will enable us to model useful interventions at increasing levels of complexity, going from activation-based methods (Vig et al., 2020; Chan et al., 2022) to text-based interventions, e.g. "How will this image of a tissue change if we administer a drug that has been shown to demonstrate a particular biological effect?"
- Medicine is inherently multimodal (Topol, 2023). Recent advances in the field of spatial biology, i.e. spatially resolved technologies to extract molecular information in native tissue locations, provide ample opportunities to draw connections and learn shared patterns across modalities such as histopathology, genomics, and transcriptomics. (Bressan et al., 2023).

1.3. Summary of contributions

In this work, we performed a preliminary interpretability analysis on the embedding dimensions extracted from a vision foundation model trained on histopathology images. We described for the first time the image characteristics represented in specific embedding dimensions of a pathology foundation model, and performed a linear regression analysis of how biologically-relevant concepts such as cell nuclear characteristics are represented in the embedding space (for the purpose of this work, we use "embedding" to refer to the output vector of the foundation model, and interpretable "features" are derived from these embeddings). We finally demonstrated the utility of the interpretability analysis in analyzing morphological changes associated with spatial gene expression.

The contributions of our work are as follows:

- We find that single dimensions in the embedding space capture complex higher-order concepts involving polysemantic combination of atomic characteristics including cell appearance and nuclear morphology.
- Linear combinations of these embedding dimensions predict nuclear characteristics including size, shape, color and orientation.
- Regression weights for predicting the nuclear color and orientation are invariant across organs, supporting zero-shot decoding in these characteristics in unseen domains.
- Foundation model embeddings predict spatial gene expression, providing evidence for multimodal behavior. The interpretation of these embeddings aligns with the biological mechanisms reported in the literature.
- Training a sparse autoencoder allows further dissection of polysemantic embedding dimensions via sparse dictionaries of interpretable features, with each feature representing characteristics such as cell and tissue features, geometric structures and image artifacts.

2. Method

2.1. Datasets

We used three publicly available TCGA (The Cancer Genome Atlas) (Weinstein et al., 2013) datasets consisting of images from three organs: breast (TCGA-BRCA), lung (TCGA-LUAD), and prostate (TCGA-PRAD). We selected 362, 130 and 324 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, 100 cancer cells and 100 fibroblast cells were randomly sampled from the model predictions, and image patches (224 x 224 pixels at a high resolution, 0.25 microns per pixel) were created centered on the selected cells. For the spatial gene expression analysis, we used the public dataset from (Barkley et al., 2022).

2.2. Foundation model and embedding extraction

Image patches were passed through a frozen ViT-Small encoder taken from 'PLUTO' - a pathology pretrained foundation model (Juyal et al., 2024). Each image patch outputs a 384-dimensional embedding vector corresponding to the average embedding of the four center 16x16 patch tokens.

We used an instance segmentation model (Abel et al., 2024) to extract biological characteristics of the cell at the center of each patch. Extracted characteristics include area, major and minor axis lengths (characterizing shape), orientation in degrees with respect to the horizontal axis, and nuclear stain color measured by the saturation, grayscale, red/green, blue/yellow channels in the LAB color space.

All samples in the spatial transcriptomics dataset contain hundreds to thousands of spots locations with corresponding gene expression information. For each of the spot locations in the 3 samples used in this study, we extract the 384-dimensional embedding vector corresponding to the average embedding of all the patch tokens.

3. Identification of interpretable feature dimensions in PLUTO embedding space

3.1. Embedding dimensions encode biologically-relevant features

We first inspected each of the 384 dimensions of the PLUTO embedding space to determine if they represent singular features of the image. For each dimension, we randomly sampled 3 patches that have the lowest 5% and the highest 5% activation values across the TCGA-BRCA dataset (Figure 1).

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).

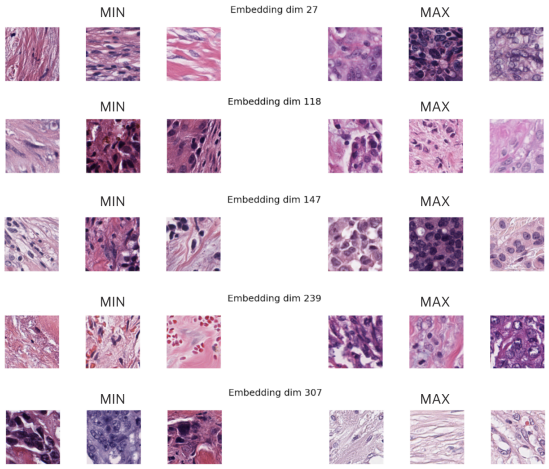

Figure 1. Visualization of features activating each embedding dimension. In each dimension, 3 example patches in the lowest 5% and highest 5% respectively of that dimension's activation are visualized. Inspection of each these patches reveals that multiple atomic features vary within each embedding dimension, including background stain color, cell size, shapes or morphologies. Some dimensions correspond to complex concepts that are relevant to pathology.

3.2. Identification of axes encoding quantitative cell features

We investigated whether linear combinations of the embedding dimensions might represent atomic characteristics of the center nucleus of each patch. These features include nuclear size, shape, color and orientation, which are important in differentiating high-grade from low-grade cancer.

We fit a linear regression model with L1 regularization ($\alpha = 0.1$) to the embedding dimensions to predict each of these nuclear characteristics. L1 regularization was used to determine if a sparse basis in the embedding space can sufficiently represent these characteristics (Bricken et al., 2023). On the test set (20% of the patches), these models achieved Pearson correlation of 0.51 to 0.91 with actual features of the cells (Table 1), with nuclear color having the highest performance, followed by size/shape and orientation. The sparsity level under the hyperparameter configuration is characterized in the Supplementary section.

4. Generalizability of encoding dimensions

It is hypothesized that invariance of image representations across domains in the foundation model embedding space supports cross-domain generalization. To test this hypothesis using our linear regression approach, we determined if the same axes (embedding weights) encode the nuclear size,

Category	Characteristic	BRCA	LUAD	PRAD
Size	Area	0.68	0.69	0.59
Shape	Major axis	0.57	0.56	0.51
Shape	Minor axis	0.73	0.70	0.66
Color	Saturation	0.88	0.91	0.84
Color	Grayscale	0.88	0.89	0.81
Color	Green/Red	0.82	0.75	0.82
Color	Blue/Yellow	0.77	0.74	0.81
Orientation	Cos-orientation	0.62	0.59	0.60
Orientation	Sin-orientation	0.62	0.60	0.62

Table 1. Performance of linear regression models (L1-regularized) in predicting cell nuclear characteristics related to size, shape, color and orientation based on the 384 embedding activations. The table reports the Pearson correlation between the actual and predicted values of the features.

shape, orientation and color characteristics across different datasets.

Category	Characteristic	BRCA→PRAD
Size	Area	0.62
Shape	Major axis	0.50
Shape	Minor axis	0.68
Color	Saturation	0.84
Color	Grayscale	0.83
Color	Green/Red	0.83
Color	Blue/Yellow	0.79
Orientation	Cos-orientation	0.59
Orientation	Sin-orientation	0.61

Table 2. Out-of-domain performance of linear regression model. L1-regularized linear regression model was fit on the embedding dimensions to predict each of the 9 nuclear characteristics in BRCA. The model was evaluated on the PRAD dataset, using the Pearson correlation between the predicted and actual characteristics.

For each of the 9 nuclear characteristics, we fit an L1-regularized regression model ($\alpha = 0.1$) to predict the characteristic from the embedding in one dataset (for e.g. BRCA), and tested the performance of that model on another dataset (for e.g., PRAD), quantified by the Pearson correlation between the actual and predicted characteristic. Performance of the model is equivalent to the in-domain PRAD model (Table 2), demonstrating cross-domain generalization.

5. Embedding dimensions capture morphological changes associated with gene expression

We fit linear regression models (L1-regularized) using PLUTO embeddings to predict spatial expression of the most variable genes shared across samples. To facilitate the interpretation of the embeddings, we focused on two genes with known and different spatial gene patterns in the literature - COL1A2, which encodes for collagen and is usually expressed in stromal cells (Retief et al., 1985), and WFDC2, a malignancy marker expressed in ovarian and endometrial tumor cells (Schummer et al., 1999; Barkley et al., 2022). For COL1A2, we trained the models on a breast cancer sample (BRCA) and predicted on an ovarian cancer sample (OVCA). For WFDC2, we trained the models on the OVCA sample and predicted on a uterine corpus endometrial cancer sample (UCEC). Models achieved a predictive accuracy for of Pearson r = 0.831 for COL1A2 and Pearson r = 0.528 for WFDC2, Fig. 3 in Supplementary section, suggesting that these embeddings are generalizable across different samples and cancer types.

Model-predicted patches with high COL1A2 expression have visible differences in morphological features (pink background with elongated cell shapes) compared to patches with low COL1A2 expression (red background). On the other hand, model-predicted patches with high WFDC2 expression show purple background with round cells, which are indicative of tumor cells. These observations align with prior literature that COL1A2 is highly expressed in stromal cells (Retief et al., 1985) and WFDC2 is expressed in malignant cancer cells (Schummer et al., 1999; Barkley et al., 2022). Noticeably, one of the embedding dimensions (embedding dim 147), with high linear coefficient in predicting WFDC2 expression, contributes negatively to the prediction of COL1A2. When visually examining this embedding dimension in the TCGA dataset (section 3.1), we found that it encodes morphological features that distinguish cancer cells and stromal cells, confirming the generalizability of the embeddings across different datasets and modalities.

6. 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). Inspired by previous work, we investigate training SAEs on top of PLUTO's embeddings and analyzing the sparse features for interpretable dimensions.

A sparse autoencoder was fit to the CLS token embedding. The SAE uses an expansion factor of 8 and with loss function given by $\frac{1}{|X|}\sum_{i=1}^{|X|}||\mathbf{x}_i-\hat{\mathbf{x}}_i||^2+\frac{\lambda}{|X|}\sum_{i=1}^{|X|}||\mathbf{f}_i||_1$, where $|X|$ 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). In order to better capture the diversity of pathology images, we trained the SAE using an expanded dataset consisting of 665,090 images from the three TCGA organs, but consisting of diverse cell types in-

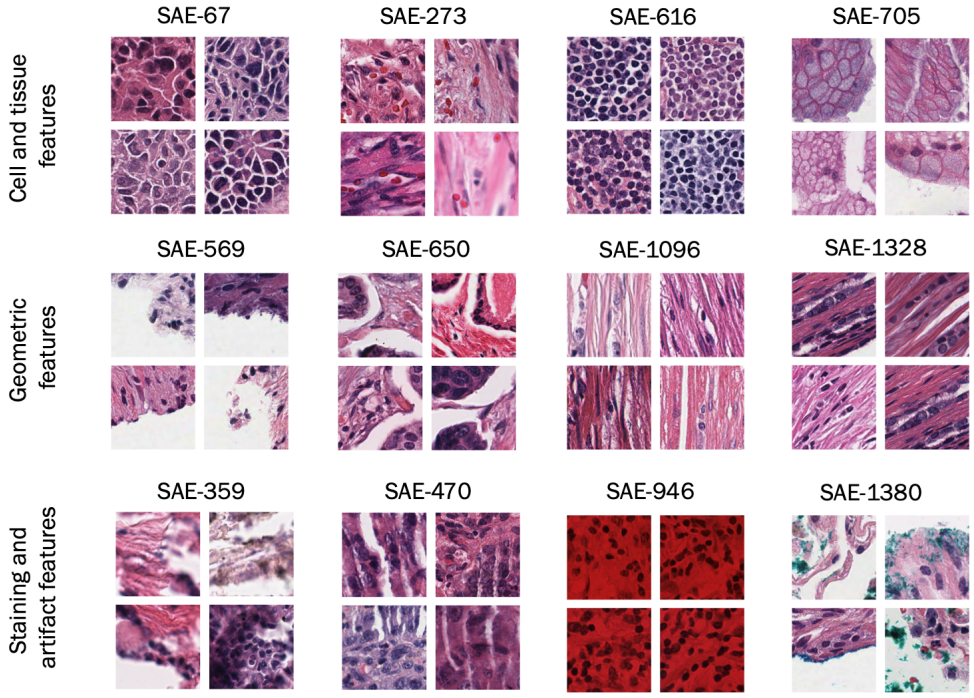

Figure 2. Feature visualization of SAE hidden dimensions reveals interpretable dictionary of pathology features. For each SAE hidden dimension, 4 out of the top 16 images that activated that dimension are visualized. Manual examination revealed highly interpretable features represented by these dimensions. These include cell and tissue features (top row: poorly differentiated carcinoma with distinct cell separation, red blood cells, dense lymphoid cells, mucin); geometric features (middle row: edge of tissue, clefting between cancer and stroma, vertical fibers, diagonal fibers); staining and artifact features (bottom row: blur, sectioning artifact, red stain, surgical ink).

cluding cancer cells, lymphocytes, macrophages, fibroblasts, as well as indication-specific cell types. The supplementary section contains more details about the change in training behavior with variation in the sparsity penalty. The fraction of dead neurons remains lower than 4% for different values of hyperparameters.

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.

7. Conclusion and Future Work

We performed a preliminary investigation of the features represented in the embedding space of a pathology foundation model. Single embedding dimensions were found to represent higher-order pathology-related concepts composed of atomic characteristics of cellular and tissue properties. Future work can be done to further decompose these concepts into their atomic properties and understand their joint representation in the embedding space.

We demonstrated the existence of axes encoding interpretable nuclear characteristics (size, shape, color and orientation) in the embedding space of a pathology foundation model. These axes generalize across datasets involving three different organs (lung, breast and prostate), potentially supporting zero-shot decoding of feature values on unseen data.

The investigation of embeddings revealed association of morphological properties and gene expression in a separate spatial transcriptomics dataset. Training a sparse autoencoder enables the extraction of relatively interpretable features corresponding to distinct biological characteristics, geometric features and image acquisition artifacts. Investigation of these interpretable axes motivates further work in discovering explainable, multimodal features of large pathology foundation models.

Acknowledgement

The authors would like to thank Jacqueline Brosnan-Cashman for her help with the illustrations and proofreading the manuscript.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abel, J., Jain, S., Rajan, D., Padigela, H., Leidal, K., Prakash, A., Conway, J., Nercessian, M., Kirkup, C., Javed, S. A., Biju, R., Harguindeguy, N., Shenker, D., Indorf, N., Sanghavi, D., Egger, R., Trotter, B., Gerardin, Y., Brosnan-Cashman, J. A., Dhoot, A., Montalto, M. C., Parmar, C., Wapinski, I., Khosla, A., Drage, M. G., Yu, L., and Taylor-Weiner, A. 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.

Arora, R., Cao, C., Kumar, M., Sinha, S., Chanda, A., McNeil, R., Samuel, D., Arora, R. K., Matthews, T. W., Chandarana, S., Hart, R., Dort, J. C., Biernaskie, J., Neri, P., Hyrcza, M. D., and Bose, P. Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. *Nature Communications*, 14 (1), August 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-40271-4. URL http://dx.doi.org/10.1038/s41467-023-40271-4.

Barkley, D., Moncada, R., Pour, M., Liberman, D. A., Dryg, I., Werba, G., Wang, W., Baron, M., Rao, A., Xia, B., França, G. S., Weil, A., Delair, D. F., Hajdu, C., Lund, A. W., Osman, I., and Yanai, I. Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nat. Genet.*, 54(8):1192–1201, August 2022.

Bereska, L. and Gavves, E. Mechanistic interpretability for ai safety – a review, 2024.

Bressan, D., Battistoni, G., and Hannon, G. J. 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.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., and Litjens, G. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology*, 21(2):233–241, 2020.

Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., Schubert, L., Voss, C., Egan, B., and Lim, S. K. Thread: Circuits. *Distill*, 2020a. doi: 10.23915/distill.00024. https://distill.pub/2020/circuits.

Cammarata, N., Goh, G., Carter, S., Schubert, L., Petrov, M., and Olah, C. Curve detectors. *Distill*, 2020b. doi: 10.23915/distill.00024.003. https://distill.pub/2020/circuits/curve-detectors.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., and Fuchs, T. J. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8): 1301–1309, 2019.

Caruana, R., Lundberg, S., Ribeiro, M. T., Nori, H., and Jenkins, S. Intelligible and explainable machine learning: Best practices and practical challenges. KDD '20, pp. 3511–3512, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3406707. URL https://doi.org/10.1145/3394486.3406707.

Castro, D. C., Walker, I., and Glocker, B. Causality matters in medical imaging. *Nature Communications*, 11(1), July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17478-w. URL http://dx.doi.org/10.1038/s41467-020-17478-w.

Chan, L., Garriga-Alonso, A., Goldwosky-Dill, N., Greenblatt, R., Nitishinskaya, J., Radhakrishnan, A., Shlegeris, B., and Thomas, N. Causal scrubbing, a method for rigorously testing interpretability hypotheses. *AI Alignment Forum*, 2022.

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A. H., Shaban, M., et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

Chughtai, B., Chan, L., and Nanda, N. A toy model of universality: Reverse engineering how networks learn group operations, 2023.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Denisenko, E., de Kock, L., Tan, A., Beasley, A. B., Beilin, M., Jones, M. E., Hou, R., Muirí, D. Ó., Bilic, S., Mohan, G. R. K. A., Salfinger, S., Fox, S., Hmon, K. P. W., Yeow, Y., Kim, Y., John, R., Gilderman, T. S., Killingbeck, E., Gray, E. S., Cohen, P. A., Yu, Y., and Forrest, A. R. R. Spatial transcriptomics reveals discrete tumour microenvironments and autocrine loops within ovarian cancer subclones. *Nat. Commun.*, 15(1):2860, April 2024.

Diao, J. A., Wang, J. K., Chui, W. F., Mountain, V., Gullapally, S. C., Srinivasan, R., Mitchell, R. N., Glass, B., Hoffman, S., Rao, S. K., et al. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nature communications*, 12(1):1–15, 2021.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy$_m odel/index.html$.

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

Howard, F. M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O. I., Kather, J. N., Cipriani, N., Grossman, R., and Pearson, A. T. 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.

Ilse, M., Tomczak, J. M., and Welling, M. Attention-based deep multiple instance learning, 2018.

Javed, S. A., Juyal, D., Padigela, H., Taylor-Weiner, A., Yu, L., and Prakash, A. Additive mil: Intrinsically interpretable multiple instance learning for pathology, 2022.

Juyal, D., Padigela, H., Shah, C., Shenker, D., Harguindeguy, N., Liu, Y., Martin, B., Zhang, Y., Nercessian, M., Markey, M., Finberg, I., Luu, K., Borders, D., Javed, S. A., Krause, E., Biju, R., Sood, A., Ma, A., Nyman, J., Shamshoian, J., Chhor, G., Sanghavi, D., Thibault, M., Yu, L., Najdawi, F., Hipp, J. A., Fahy, D., Glass, B., Walk, E., Abel, J., Pokkalla, H., Beck, A. H., and Grullon, S. Pluto: Pathology-universal transformer, 2024.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F. Data efficient and weakly supervised computational pathology on whole slide images, 2020.

Lundberg, S. and Lee, S.-I. A unified approach to interpreting model predictions, 2017.

Madabhushi, A. and Lee, G. 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).

Markey, M., Kim, J., Goldstein, Z., Gerardin, Y., Brosnan-Cashman, J., Javed, S. A., Juyal, D., Padigela, H., Yu, L., Rahsepar, B., 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.

Mikolov, T., Yih, W.-t., and Zweig, G. Linguistic regularities in continuous space word representations. In Vanderwende, L., Daumé III, H., and Kirchhoff, K. (eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability, 2023a.

Nanda, N., Lee, A., and Wattenberg, M. Emergent linear representations in world models of self-supervised sequence models, 2023b.

Nguyen, T. H., Juyal, D., Li, J., Prakash, A., Nofallah, S., Shah, C., Gullapally, S. C., Yu, L., Griffin, M., Sampat, A.,

Abel, J., Lee, J., and Taylor-Weiner, A. Contrimix: Scalable stain color augmentation for domain generalization without domain labels in digital pathology, 2024.

Olah, C. Mechanistic interpretability, variables, and the importance of interpretable bases, 2022. https://transformer-circuits.pub/2022/mech-interp-essay/index.html.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits, 2020. https://distill.pub/2020/circuits/zoom-in.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Retief, E., Parker, M. I., and Retief, A. E. Regional chromosome mapping of human collagen genes alpha 2(i) and alpha 1(i) (colia2 and colia1). *Human Genetics*, 69(4): 304–308, April 1985. ISSN 1432-1203. doi: 10.1007/bf00291646. URL http://dx.doi.org/10.1007/BF00291646.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should i trust you?": Explaining the predictions of any classifier, 2016.

Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., and Sarkar, R. The shapley value in machine learning, 2022.

Schummer, M., Ng, W. V., Bumgarner, R. E., Nelson, P. S., Schummer, B., Bednarski, D. W., Hassell, L., Baldwin, R. L., Karlan, B. Y., and Hood, L. Comparative hybridization of an array of 21 500 ovarian cdnas for the discovery of genes overexpressed in ovarian carcinomas. *Gene*, 238(2):375–385, October 1999. ISSN 0378-1119. doi: 10.1016/s0378-1119(99)00342-x. URL http://dx.doi.org/10.1016/S0378-1119(99)00342-X.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2): 336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Topol, E. J. 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.

Varoquaux, G. and Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digital Medicine*, 5 (1), April 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00592-y. URL http://dx.doi.org/10.1038/s41746-022-00592-y.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Sakenis, S., Huang, J., Singer, Y., and Shieber, S. Causal mediation analysis for interpreting neural nlp: The case of gender bias, 2020.

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

Wang, D., Khosla, A., Gargeya, R., Irshad, H., and Beck, A. H. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

A. Interpretability in medical imaging

Interpretability is crucial for Machine Learning (ML) in medical imaging; it builds decision-makers' trust, enables model developers to debug silent failure modes and shortcut-learning, and reduces the chances of catastrophic model failures in real-world deployments (Castro et al., 2020; Varoquaux & Cheplygina, 2022; Caruana et al., 2020). Existing interpretability methods on medical imaging datasets include gradient-based methods (Selvaraju et al., 2019; Simonyan et al., 2014), model-agnostic (Rozemberczki et al., 2022; Lundberg & Lee, 2017; Ribeiro et al., 2016) and intrinsically interpretable techniques (Ilse et al., 2018; Javed et al., 2022). In comparison, mechanistic interpretability methods have been relatively underexplored.

B. Results of model-fitting using L2-regularization

Nuclear Characteristics	Breast	Lung	Prostate
Area	0.78	0.77	0.71
Major axis length	0.67	0.65	0.63
Minor axis length	0.82	0.78	0.76
Saturation	0.94	0.95	0.90
Grayscale	0.94	0.94	0.88
Green/Red	0.92	0.90	0.91
Blue/Yellow	0.90	0.86	0.89
Cos-orientation	0.69	0.65	0.65
Sin-orientation	0.68	0.65	0.66

Table 3. Performance of linear regression models (L2-regularized) in predicting cell nuclear characteristics related to size, shape, color and orientation based on the 384 embedding activations. The table reports the Pearson correlation between the actual and predicted values of the features.

C. Sparsity of L1 regression model

Nuclear Characteristics	Breast	Lung	Prostate
Area	42	45	23
Major axis length	28	29	25
Minor axis length	33	34	30
Saturation	21	19	16
Grayscale	18	19	19
Green/Red	16	15	16
Blue/Yellow	10	12	14
Cos-orientation	9	9	10
Sin-orientation	15	12	17

Table 4. Sparsity in the embedding regression of nuclear characteristics. The table shows the number of non-zero regression weights (out of 384) in the L1-regularized regression model for each nuclear feature.

D. Generalization performance of L1-regularized linear regression models

Category	Characteristic	PRAD→LUAD	LUAD→BRCA
Size	Area	0.64	0.68
Shape	Major axis	0.51	0.52
Shape	Minor axis	0.68	0.73
Color	Saturation	0.85	0.82
Color	Grayscale	0.85	0.83
Color	Green/Red	0.78	0.78
Color	Blue/Yellow	0.80	0.79
Orientation	Cos-orientation	0.58	0.61
Orientation	Sin-orientation	0.60	0.62

Table 5. Out-of-domain performance of linear regression model on PRAD→LUAD and LUAD→BRCA generalization experiments. L1-regularized linear regression model was fit on the embedding dimensions to predict each of the 9 nuclear characteristics in one dataset, and tested on the out-of-domain dataset. Values reported are the Pearson correlations between the predicted and actual characteristics.

E. Embedding interpretability in the context of spatial gene expression

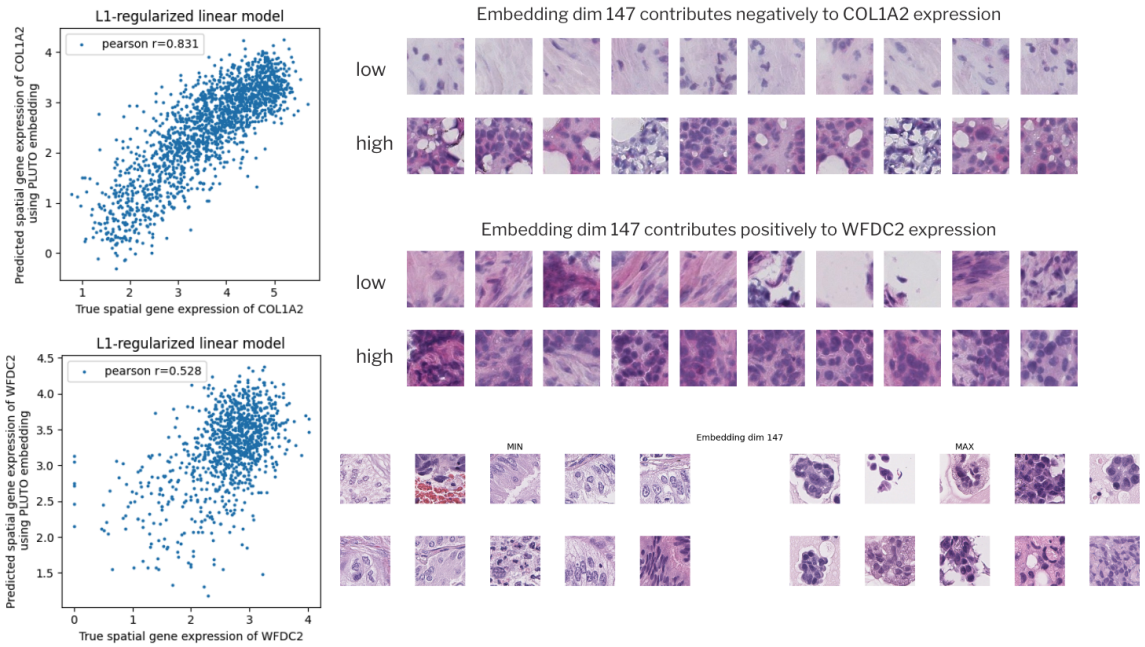

Figure 3. Embedding interpretability in the context of spatial gene expression. PLUTO embeddings can predict spatial gene expression with good accuracy. One of the embedding dimensions (embedding dim 147) that contributes the most to both gene expressions shows morphological features that distinguish cancer cells from stromal cells. The observations align with known biology of the genes. Similar morphological features corresponding to embedding dimension 147 are also observed in the TCGA dataset.

F. Sparse autoencoder training behavior

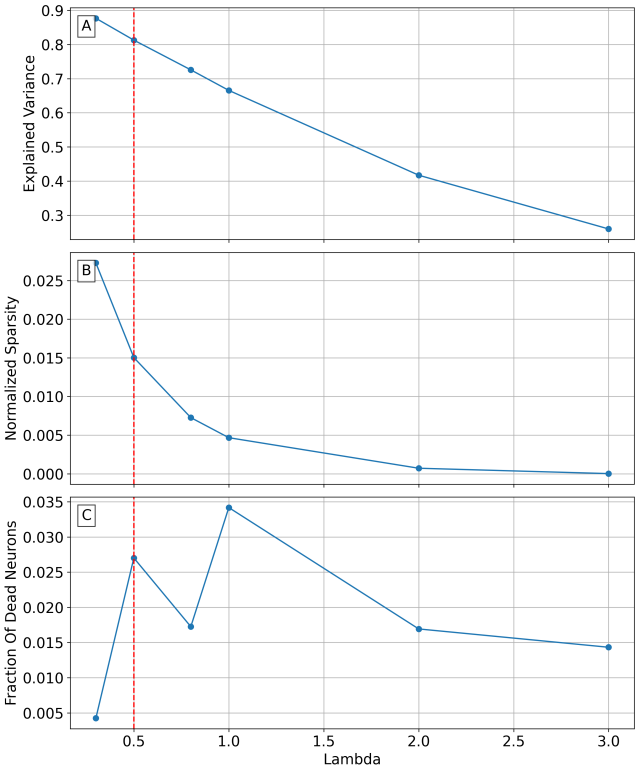

Figure 4. SAE training behavior. As the regularization parameter (lambda) increases, explained variance decreases (A) and average sparsity increases (B). The fraction of dead neurons (C) remained less than 4%, showing that most neurons in the hidden layer were activated by at least one input image and demonstrating that the network is utilized at almost full capacity. Dashed line represents the chosen parameter ($\lambda = 0.5$) for feature visualization.