# PATH-DISSECT: Interpreting Concept Representation and Classifier Reliance in Pathology Foundation Models with Sparse Autoencoders

Mayur Parvatikar[1], Xingjian Li[2*], and Min Xu[2*]

[1] Department of Information Science and Engineering, RV College of Engineering, Bengaluru, India
mayurgp.is21@rvce.edu.in

[2] Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, USA
xingjia2@andrew.cmu.edu, mxu1@cs.cmu.edu

**Abstract.** Pathology foundation models achieve strong diagnostic performance, yet the morphological basis of their predictions remains opaque, limiting trust for safe clinical deployment. Sparse autoencoders (SAEs) have emerged as a promising tool for recovering interpretable sparse features that align with tissue concepts from such models, but existing analyses are largely descriptive, characterising what a model encodes without establishing whether a downstream classifier depends on it. We introduce PATH-DISSECT, a depth-resolved analysis of concept representation and classifier reliance in frozen pathology foundation models. By training sparse autoencoders at multiple vision transformer blocks, PATH-DISSECT identifies class-associated features, validates their class selectivity through alignment, purity, and specificity, and measures how far a frozen linear probe depends on them through causal intervention on the representation during the forward pass. Applied to UNI and Virchow2 encoders on colorectal and breast tissue patches, PATH-DISSECT provides a layer-wise account of where validated concepts first become cleanly detectable and where the probe becomes most sensitive to removing them, separating concept representation from use. We further show that a discovered feature set can be used to localise an injected spurious cue and mitigate the classifier's reliance on it, making PATH-DISSECT a practical tool for interpreting and auditing pathology foundation models.

**Keywords:** Sparse autoencoders · Pathology foundation models · Interpretability · Causal intervention.

## 1 Introduction

Computational pathology foundation models are large vision transformers pretrained on vast histopathology datasets by self-supervised learning [3,32,30,17].

* These authors are co-corresponding authors.

These models are popular for tasks such as tumor classification, biomarker discovery, and prognosis prediction across organs, and they can be adapted through transfer, few-shot, or zero-shot learning [15]. Although their benchmark performance is strong, they remain black boxes where the internal representations driving the classifications are opaque [21], and the contribution of morphological concepts to a prediction is unknown. This matters for clinical adoption, where a model that reaches the right answer from the wrong morphology is indistinguishable from one that does not until someone looks inside.

Efforts to interpret pathology models fall into two families. Post-hoc attribution highlights influential regions through class-activation or attention maps [31,1] while intrinsically interpretable models reason over a predefined vocabulary of histological patterns, such as prototype- and handcrafted-dictionary models for atypical breast lesions [23,22]. Both give spatial or example-based explanations of where a model looked, rather than an account of what the model internally encodes. Sparse dictionary learning has emerged as a promising route to improving the interpretability of neural networks [7,11,2,27,29,25,8]. A sparse autoencoder (SAE) is a two-layer network that implements this without supervision, re-expressing a high-dimensional representation as a sparse code over a learned dictionary. SAEs were developed for language models [18,16] and have since proven effective for vision models [28,12]. They now extend to medical imaging, where they dissect chest-X-ray representations [26], brain MRI foundation models [20], and hematology cell embeddings [4]. Applied to pathology foundation models, SAEs recover sparse features that align with tissue concepts and trace how they sharpen across encoder depth [14], while recent work assembles them into a large concept atlas used to audit and steer downstream predictions [13].

Our overarching goal is to understand not only which concepts a pathology foundation model encodes, but how much a downstream classifier depends on them across depth. PATH-DISSECT addresses three questions: (RQ1) Which concepts does the model represent, and how reliably? (RQ2) At what depth does each concept first become cleanly detectable? (RQ3) Does a downstream classifier depend on it, and where? By discovering, validating, and causally intervening on concepts inside the frozen encoder, PATH-DISSECT moves beyond a model's visual vocabulary to how that vocabulary drives a downstream decision, and we apply it to mitigate reliance on an injected shortcut. We study UNI [3] and Virchow2 [32] across colorectal and breast patches, giving a depth-resolved account of concept representation and classifier reliance in pathology foundation models.

## 2 The PATH-DISSECT Framework

PATH-DISSECT (Fig. 1) consists of three stages: feature discovery, concept validation, and depth-resolved causal intervention, which we further apply to mitigate shortcut reliance.

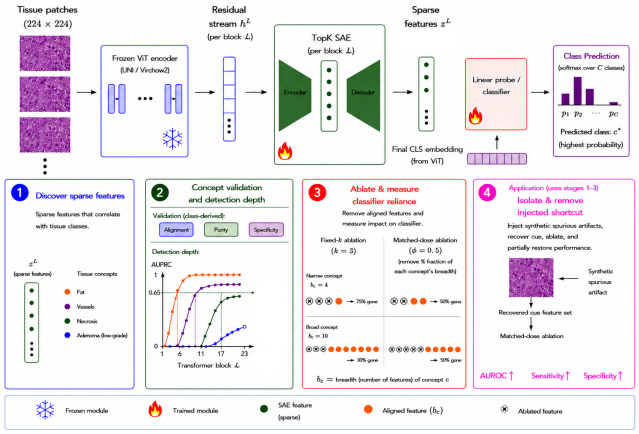


**Fig. 1.** The PATH-DISSECT framework pipeline, with per-block SAEs on a frozen encoder, validated then ablated across depth.

**Preliminaries.** A frozen ViT encoder $f$ (UNI or Virchow2) maps a $224 \times 224$ H&E patch $x$ to a class-token (CLS) embedding. We write $h^L = f^L(x) \in \mathbb{R}^d$ for its residual-stream CLS activation at block $L$. We work with labelled patches $\{(x_i, y_i)\}$ where $y_i \in \{1, \ldots, C\}$ indexes tissue classes. A multinomial logistic-regression probe is fitted on the final CLS embedding of the training split and then frozen. All reliance we report is reliance of this probe: a different classifier family could exploit different interactions and redundancies, so our conclusions are scoped to a linear readout. For the causal intervention experiments we measure the drop in true-class log-probability, which is more sensitive than accuracy because it captures loss of confidence even when the true class remains the top prediction.

### 2.1 Learning sparse features (Stage 1)

The residual-stream activation $h^L$ superposes many tissue concepts in one vector [6], so at each block we train a TopK SAE [8] that decomposes it into a sparse code,

$$h \approx \mathrm{Dec}(z) = b + \sum_{i=1}^{m} z_i \, d_i, \quad z = \mathrm{Enc}(h) = \mathrm{TopK}_{\kappa}\big(\mathrm{ReLU}(W(h-b)+b_e)\big), \quad (1)$$

with encoder weight $W$, pre-encoder bias $b$, encoder bias $b_e$, and a code $z \in \mathbb{R}^m$ of $m$ latents with $\kappa$ nonzero entries per patch, trained to minimise reconstruction MSE. Each latent $z_i$ is a feature, and its decoder direction $d_i$ (the $i$-th column of Dec) is the fixed vector it writes into the residual stream when active. A feature is class-associated when $z_i$ correlates with a tissue class. A concept is a tissue class, and since a class is usually carried by several features we write $\mathcal{S}_c$ for its class-selective feature set, the top-$k$ aligned features. All our selection

and validation criteria are class-derived, so they establish that a feature tracks a tissue label, not that it encodes a single coherent morphology; the top-activating patches (Fig. 3A) show the morphology involved but are illustrative rather than confirmatory.

### 2.2 Concept validation and detection depth (Stage 2)

A class-associated feature is useful only if it consistently tracks its class rather than an incidental correlation. For each class $c$ we take its rank-1 aligned feature on train and score it on held-out test under three class-derived properties: Alignment $a_c = |r_c|$ is the absolute point-biserial correlation with the class indicator [14], used rather than mean class activation because normalisation by the off-class baseline rewards a feature that fires only for its class, not one that merely fires strongly. Purity is the fraction of its top-activating patches truly of class $c$, against the class-prevalence chance floor. Specificity, from the feature's cross-class profile $|r_{cj}|$, is its own-class over mean off-class alignment, its largest off-class term naming the confusable class.

Separately, we measure how class-concentrated the SAE features are, via the Shannon entropy $H_c$ of each feature's class-association distribution $p_{cj} = |r_{cj}|/\sum_k |r_{ck}|$, low when a feature is class-selective, read in bits against the raw-embedding baseline [26]. A per-block probe on $h^L$ gives a bulk-decodability reference.

These properties show that a feature is reliably associated with its class, but not when its class first acquires a clean detector. Following the class-derived AUPRC criterion in [10] (imbalance-robust [5], scored as the smaller of train and test), a class has a clean detector once its best feature clears a per-class bar ($> 0.90$, or $[0.65, 0.90)$ with a top-vs-second margin $\geq 0.20$) as that class's dominant feature, and its detection depth is the first block where this holds. Each block has its own dictionary and the best feature is re-selected at every block, so detection depth records the first block at which some feature meets the criterion for that class. We establish no correspondence between features at different blocks, so this indexes when a class becomes cleanly detectable in the residual stream rather than the formation or propagation of a single feature.

### 2.3 Causal intervention (Stage 3)

Stage 2's properties are correlational, so Stage 3 tests necessity directly by intervening on the sparse features that carry a concept. At block $L$ we mean-ablate the class-selective feature set $\mathcal{S}_c$: writing $z = \mathrm{Enc}(h)$ for the SAE codes of the CLS activation $h$, we reset each feature in $\mathcal{S}_c$ to its training-set mean $\bar{z}_j$ and subtract only its decoder contribution, leaving the SAE residual intact,

$$\hat{h} = h - \sum_{j \in \mathcal{S}_c} (z_j - \bar{z}_j)\, d_j, \qquad \bar{z}_j = \frac{1}{N} \sum_{n=1}^{N} z_j^{(n)}, \tag{2}$$

then run the rest of the network on $\hat{h}$ untouched. The edit acts on SAE features, not pixels, so it targets the concept representation inside the model, and averaging rather than zeroing keeps it on-distribution [9]. The causal effect is the drop in true-class log-probability, averaged over class-$c$ patches with at least one active feature in $\mathcal{S}_c$,

$$\Delta_c(L) = \left\langle \log p(y{=}c \mid h) - \log p(y{=}c \mid \hat{h}) \right\rangle, \tag{3}$$

a graded measure of the probe's reliance. Repeating it at every block localises where the probe is most sensitive to a concept. Comparing $\Delta_c$ across concepts at a fixed $k$ misleads, because concepts differ in breadth $b_c$, the number of features aligned with a concept above a threshold (measured independently of any ablation), so the same $k$ removes a different fraction of each. We therefore ablate at matched dose, setting $k_c = \lceil \phi\, b_c \rceil$ for a fraction $\phi \in (0, 1]$ of each concept's breadth, so every concept loses the same share of its features. Any remaining difference in $\Delta_c$ then reflects reliance rather than dose.

## 3 Results

**Setup.** We evaluate PATH-DISSECT on SPIDER [19], using the official train and test split for colorectal (13 classes, 63,989/13,193 patches) and breast (18 classes, 80,858/12,034). From frozen UNI (ViT-L) and Virchow2 (ViT-H) we train one TopK SAE per block at six evenly spaced blocks. A forward hook reads the CLS token (index 0) of each block's raw output, centred by the training mean and divided by one scalar to unit mean per-dimension variance. Adam $3{\times}10^{-4}$, 5% warmup then cosine decay, 4000 steps, batch 1024, 8× expansion ($m$=8192 for UNI, 10240 for Virchow2), $\kappa$=150, decoder re-projected to unit norm every step, dead latents resampled, final checkpoint, seed 42. SAEs, the probe, feature rankings and breadth all use train; test scores alignment and purity, supplies the intervention subsample (up to 500 patches per class), and enters the detection criterion only via $\min(\text{train}, \text{test})$ AUPRC. The frozen logistic-regression probe reaches 0.875 to 0.892 accuracy and 0.84 to 0.88 macro-F1.

### 3.1 Features are class-concentrated and concepts become detectable at concept-dependent depths

We select each class's best feature on train and score it on held-out test. Alignment sharpens with depth in all four settings (Fig. 2). Best-feature $|r|$ reaches 0.72 (UNI) and 0.75 (Virchow2) on colorectal and 0.51 and 0.62 on breast, and top-100 purity climbs from 0.59 at block 3 to 0.90 at block 19 for UNI, and from 0.62 at block 4 to 0.94 at block 26 for Virchow2. A 50-permutation label-shuffle null puts the best of 8192 features at only 0.02 (95th percentile), and across both encoders the SAEs stay well-conditioned, with held-out variance explained of 0.82 to 0.99 for UNI and 0.90 to 0.99 for Virchow2 and dead units confined to early blocks, so neither selection bias nor a loss of SAE quality explains

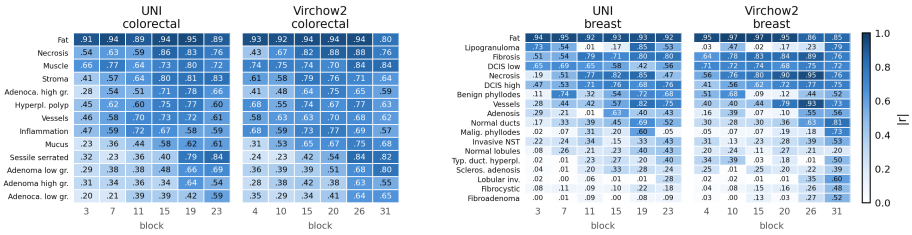


**Fig. 2.** Stage 2 validation: per-class best-feature alignment $|r|$ across depth for UNI and Virchow2 on colorectal (left two panels) and breast (right two). Darker = stronger held-out alignment, cleaner in later blocks.

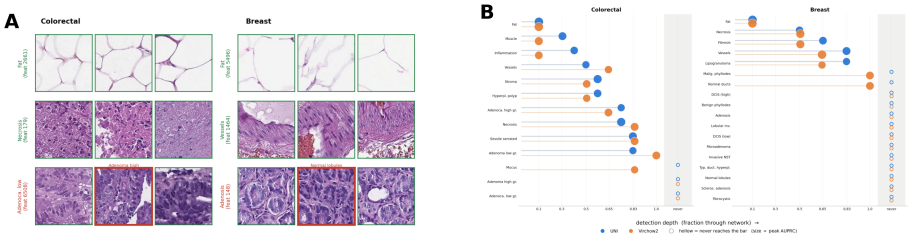


**Fig. 3.** Stage 2. **(A)** Top-3 patches for two class-pure and one cross-class feature (red = look-alike intruder from another class); qualitative illustration. **(B)** Detection depth, the first block at which that class has a clean detector (dot $\propto$ peak AUPRC, hollow = never).

the trend. Individual directions are seed-dependent [24] (median $|\cos|$=0.35), while per-class alignment and ablation ordering reproduce ($\rho \geq 0.83$). Features are class-concentrated, with class-association entropy below the raw-embedding baseline at every block in both encoders, and they are specific, correlating 9.3 to 12.7 times more with their own class than with the rest. The exceptions are informative and occur in both organs. UNI's adenocarcinoma low-grade feature aligns more with adenoma high-grade (0.64 vs 0.42), a confusion that also shows up among its top-activating patches (Fig. 3A), and Virchow2's fibroadenoma feature aligns more with benign phyllodes (0.67 vs 0.52). Bulk decodability rises only gradually (0.74→0.87), but detection depth is strongly concept-dependent and agrees across encoders (Fig. 3B). Fat has a clean detector by block 3, followed by muscle, inflammation and vessels, while adenoma low-grade, hyperplastic polyp and sessile serrated lesion clear the bar only near the output. Under this conservative bar 10/13 and 11/13 colorectal and 5/18 and 7/18 breast concepts qualify for UNI and Virchow2 respectively, a lower bound since it discards features that span several diagnostic classes [10], and those that never clear it are the cross-class-confusable ones rather than absent representations. The UNI colorectal count moves from 11/13 to 8/13 as the bar spans $[0.60, 0.70]$, but the detection ordering is unchanged, so it is not an artifact of the 0.65 threshold.

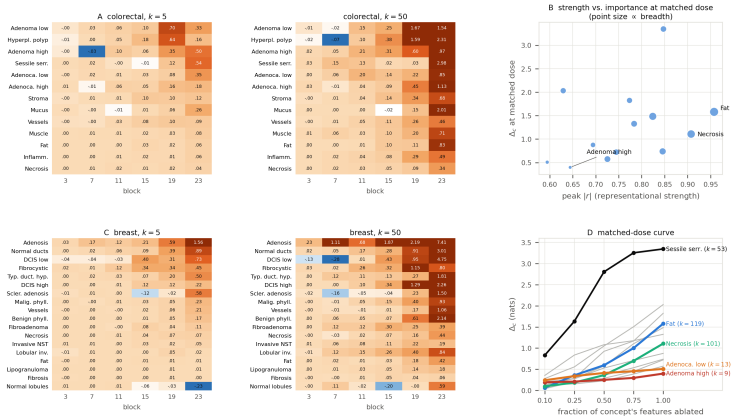


**Fig. 4.** Stage 3 (UNI). Fixed-$k$ ablation effect by depth at $k$=5 and $k$=50 (A, C), alignment versus matched-dose probe reliance (B), and the matched-dose curve (D). All effects are true-class log-probability drops in nats.

### 3.2 Probe sensitivity concentrates near the readout, and dose confounds fixed-$k$ comparison

A concept can be represented without the probe using it, so we test necessity for the probe directly by mean-ablating each class's feature set inside the frozen forward pass and reading the true-class log-probability drop. In all four encoder–organ settings this drop far exceeds matched random-feature ablation, which averages under 0.001 nats for UNI colorectal, so the probe depends on these representations and not merely on their decodability, and the effect peaks late, in the final two blocks (colorectal 13/13 and breast 17/18 for UNI, 10/13 and 16/18 for Virchow2, Fig. 4A,C). A concept ablated at an early block is largely rebuilt by the output, with up to 96% of its on-class feature set returning downstream for a representative concept. The late peak thus reflects where an edit persists, i.e. depth-dependent sensitivity rather than late formation. Negative effects, where ablation instead raises the probability, stay at early blocks, the exception being Normal lobules, whose $k$=5 feature set remains negative to the readout and works against its own class (Fig. 4C). Cross-concept comparison needs care, as concepts differ in breadth and a fixed feature count removes a different share of each. Matching the dose to each concept's breadth, applied at the final block (block 23 for UNI), removes any significant breadth dependence in every setting, unchanged at thresholds $|r|$>0.2 or 0.4 (Fig. 4D). With dose controlled, a concept's representational strength no longer predicts how far the probe relies on it (Fig. 4B): in no encoder–organ setting are better-aligned concepts more important to the probe, consistent with representation and use being distinct axes. Breadth rankings are preserved across sparsity budgets $\kappa \in \{32, 64, 150, 300\}$ (Spearman $\rho \geq 0.94$), so breadth reflects the concept rather than the dictionary configuration.

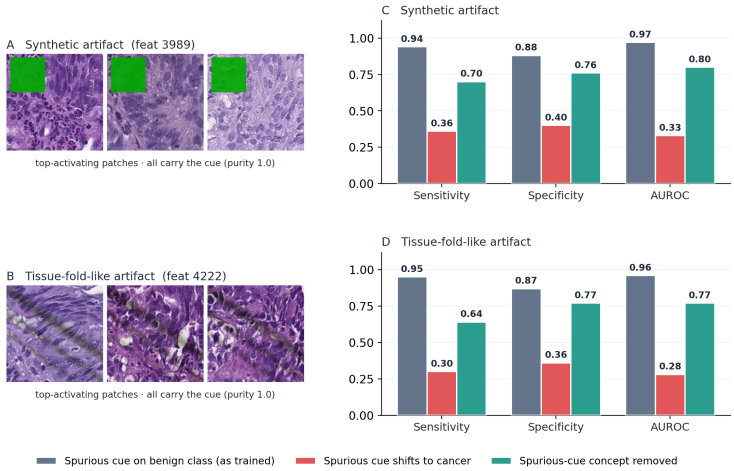


**Fig. 5.** Mitigating an injected shortcut: PATH-DISSECT isolates the spurious cue (A,B) and matched-dose ablation partially restores tumour sensitivity (C,D).

### 3.3 Ablating the discovered cue feature set mitigates reliance on an injected shortcut

On a challenging class pair of adenoma high grade (benign) against adenocarcinoma low grade (malignant), we add an artificial cue to the benign class [13] at cue–label correlation 0.90: a green corner mark in one run, a synthetic tissue fold in another. Splits are slide-disjoint (3,000 biased train, 1,000 validation, 1,152 cue-decorrelated probe (correlation 0.01) for identifying the cue feature, 1,944 each reversed and clean test). Moving the cue onto cancer collapses the ranking (AUROC 0.97→0.33 mark, 0.96→0.28 fold) and drops tumour sensitivity from 0.94 to 0.36. An SAE trained on the same biased data isolates the cue in one clean feature set (purity 1.0, Fig. 5A,B); ablating it at a dose matched to its breadth, at the final block, lifts sensitivity (0.36→0.70), diagnostic specificity (benign correctly passed, 0.40→0.76) and AUROC (0.33→0.80), while size-matched random ablations leave AUROC at the collapsed baseline. Recovery is partial and the cue is injected by construction, so this isolates and removes a known artifact rather than discovering an unknown shortcut.

## 4 Conclusion

PATH-DISSECT provides a depth-resolved view of how pathology foundation models represent tissue concepts and how far a linear classifier depends on them. By combining sparse feature discovery, concept validation, and causal intervention, it distinguishes concepts that are merely represented from those a downstream linear classifier depends on. Across UNI and Virchow2 on colorectal and breast tissue, we show that concept detectability and probe sensitivity follow

different trajectories across depth, and that concept breadth is important for interpreting intervention effects, so representation analysis alone is insufficient for understanding model behaviour. Beyond explaining model decisions, PATH-DISSECT offers a practical approach for auditing pathology foundation models and identifying shortcut representations. Our conclusions are scoped to a linear readout, to class-associated feature sets rather than verified morphological units, to per-block dictionaries without cross-block feature correspondence, and to a shortcut we injected rather than discovered. Future work will extend it to token-level representations, circuit tracing, and automated concept interpretation.

**Reproducibility.** `https://github.com/MayurGP25/PATH-DISSECT` hosts the scripts and artifacts.

**Acknowledgments.** This work was supported in part by U.S. NSF grants DBI-2238093, DBI-2422619, IIS-2211597, and MCB-2205148. This work was also supported in part by an Amazon Research Award (Fall 2025 CFP).

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to this article.

## References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4190–4197 (2020)
2. Bricken, T., Templeton, A., Batson, J., et al.: Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread (2023)
3. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
4. Dasdelen, M.F., Lim, H., Buck, M., Götze, K.S., Marr, C., Schneider, S.: CytoSAE: Interpretable cell embeddings for hematology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 77–86 (2025)
5. Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning (ICML). pp. 233–240 (2006)
6. Elhage, N., Hume, T., Olsson, C., et al.: Toy models of superposition. arXiv preprint arXiv:2209.10652 (2022)
7. Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al.: A mathematical framework for transformer circuits. Transformer Circuits Thread (2021)
8. Gao, L., Dupré la Tour, T., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., Wu, J.: Scaling and evaluating sparse autoencoders. In: International Conference on Learning Representations. pp. 26721–26754 (2025)
9. Heimersheim, S., Nanda, N.: How to use and interpret activation patching. arXiv preprint arXiv:2404.15255 (2024)

10. Hossain, M.N., Bell, S.L., Bryson, G., Harris-Birtill, D.: Histoscope: Expert-grounded inspection of sparse autoencoder features in histopathology foundation models. In: Mechanistic Interpretability Workshop at ICML 2026 (2026)
11. Huben, R., Cunningham, H., Smith, L., Ewart, A., Sharkey, L.: Sparse autoencoders find highly interpretable features in language models. In: International Conference on Learning Representations. pp. 7827–7845 (2024)
12. Joseph, S., Suresh, P., Hufe, L., Stevinson, E., Graham, R., Vadi, Y., Bzdok, D., Lapuschkin, S., Sharkey, L., Richards, B.A.: Prisma: An open source toolkit for mechanistic interpretability in vision and video. arXiv preprint arXiv:2504.19475 (2025)
13. Kim, C., Kaczmarzyk, J., Savant, D., Zhao, Z., Koo, P.K., Lee, S.I.: Dissecting and directing pathology foundation models. bioRxiv (2026), doi:10.64898/2026.06.12.731496
14. Le, N.M., Shen, C., Patel, N., Shah, C., Sanghavi, D., Martin, B., Eng, A., Shenker, D., et al.: Learning biologically relevant features in a pathology foundation model using sparse autoencoders. arXiv preprint arXiv:2407.10785 (2024)
15. Li, D., Wan, G., Wu, X., Wu, X., He, Y., Chen, Z., Hao, N., Zhao, C.: A survey on computational pathology foundation models. In: NeurIPS 2025 Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI (2025)
16. Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramár, J., Dragan, A., Shah, R., Nanda, N.: Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. In: Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP. pp. 278–300 (2024)
17. Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024)
18. Marks, S., Rager, C., Michaud, E.J., Belinkov, Y., Bau, D., Mueller, A.: Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In: International Conference on Learning Representations. pp. 23888–23923 (2025)
19. Nechaev, D., Pchelnikov, A., Ivanova, E.: SPIDER: A comprehensive multi-organ supervised pathology dataset and baseline models. arXiv preprint arXiv:2503.02876 (2025)
20. Nerrise, F., Yin, L., Abbasi, M.H., Pohl, K.M., Adeli, E.: GeoSAE: Geometric prior-guided layer-wise sparse autoencoder annotation of brain MRI foundation models. arXiv preprint arXiv:2605.01829 (2026)
21. Orlov, A.V., Makus, Y.V., Ashniev, G.A., Orlova, N.N., Nikitin, P.I.: What do biological foundation models compute? Sparse autoencoders from feature recovery to mechanistic interpretability. bioRxiv (2026), doi:10.64898/2026.03.04.709491
22. Parvatikar, A., Choudhary, O., Ramanathan, A., Jenkins, R., Navolotskaia, O., Carter, G., Tosun, A.B., Fine, J.L., Chennubhotla, S.C.: Prototypical models for classifying high-risk atypical breast lesions. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 143–152 (2021)
23. Parvatikar, A., Choudhary, O., Ramanathan, A., Navolotskaia, O., Carter, G., Tosun, A.B., Fine, J.L., Chennubhotla, S.C.: Modeling histological patterns for differential diagnosis of atypical breast lesions. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 550–560 (2020)
24. Paulo, G., Belrose, N.: Sparse autoencoders trained on the same data learn different features. arXiv preprint arXiv:2501.16615 (2025)

25. Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., Nanda, N.: Improving dictionary learning with gated sparse autoencoders. arXiv preprint arXiv:2404.16014 (2024)
26. Renzulli, R., Lepoutre, C., Cassano, E., Grangetto, M.: MedSAE: Dissecting MedCLIP representations with sparse autoencoders. arXiv preprint arXiv:2510.26411 (2025)
27. Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., et al.: Open problems in mechanistic interpretability. arXiv preprint arXiv:2501.16496 (2025)
28. Stevens, S., Chao, W.L., Berger-Wolf, T., Su, Y.: Interpretable and testable vision features via sparse autoencoders. arXiv preprint arXiv:2502.06755 (2025)
29. Templeton, A., Conerly, T., Marcus, J., et al.: Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread (2024)
30. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**(8015), 181–188 (2024)
31. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2921–2929 (2016)
32. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., et al.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint arXiv:2408.00738 (2024)