Histoscope: Expert-Grounded Inspection of Sparse Autoencoder Features in Histopathology Foundation Models

Mirza Nasir Hossain [1] **Sarah L Bell** [2] **Gareth Bryson** [2] **David Harris-Birtill** [1]

Abstract

Computational pathology lacks open tools for inspecting sparse autoencoder (SAE) features, and few studies test whether such features align with expert judgement. We present Histoscope, an open-source tool for inspecting SAE features derived from histopathology foundation-model embeddings, and evaluate it through a blinded two-rater pathologist study. Using UNI embeddings from the SPIDER colorectal dataset, we train a TopK SAE and use Histoscope to rank features by diagnostic-class selectivity, measured with per-class AUPRC. Pathologists judged 82% of feature panels monosemantic in a confidence-stratified sample comprising the 50 highest- and 50 lowest-ranked features by Histoscope's AUPRC selectivity rule. All 50 Histoscope-mono features were confirmed, while 32 of the 38 resolved Histoscope-poly features were also judged monosemantic, mostly because they captured morphology spanning multiple diagnostic classes. Inspection tools of this kind can help make pathology foundation models more auditable by domain experts by connecting internal features to morphological evidence rather than treating embeddings as opaque vectors. We release Histoscope, the trained SAE, and the evaluation protocol to support expert-grounded SAE inspection in computational pathology.

1. Introduction

Sparse autoencoders (SAEs) have become a standard method for probing internal representations in large language and vision models (Bricken et al., 2023; Cunningham et al., 2023; Olson et al., 2025), supported by open libraries and dashboards (Lieberum et al., 2024; Decode Research, 2024; McDougall, 2024). However, computational pathology still lacks open tools for inspecting SAE features, and to our knowledge no prior study has tested whether SAE-derived features from a histopathology foundation model align with direct expert morphological judgement.

Computational pathology increasingly relies on pretrained foundation models such as UNI (Chen et al., 2024) and UNI2-h (Mahmood Lab, 2025), which are used as feature extractors for downstream tasks including classification, segmentation, and prognosis prediction (Azad et al., 2023). UNI is a vision-transformer pathology encoder pretrained with DINOv2 self-supervision on over 100 million patches from more than 100,000 diagnostic haematoxylin and eosin (H&E) whole-slide images spanning 20 major tissue types (Chen et al., 2024). Although these models achieve strong benchmark performance across medical imaging tasks (Hossain & Harris-Birtill, 2026), their internal representations remain difficult to interpret (Le et al., 2024). Training an SAE on pathology embeddings may reveal candidate features, but visual similarity alone is not enough to establish whether a feature corresponds to a single meaningful morphological concept. In practice, this creates a gap between researchers who can build interpretability tools and pathologists who can assess whether the features those tools surface are actually coherent.

Prior work has begun to explore SAE-style decomposition in medical imaging. PLUTO-SAE (Le et al., 2025) trained an SAE on a proprietary pathology foundation model and evaluated selected features by correlating SAE dimensions with automatically derived quantitative features such as cell-type counts. However, this work did not directly evaluate SAE feature coherence against pathologist judgement or release inspection tooling or SAE weights. In a related setting, CytoSAE (Dasdelen et al., 2025) trained an SAE on embeddings from a haematology foundation model and evaluated the resulting concepts against expert annotations.

We address these gaps by decomposing an open-weights histopathology foundation model (UNI) with an SAE, building open tooling for inspecting the resulting features, and directly validating those features against pathologist judgement. The tool, Histoscope, surfaces activation statistics,

[1]School of Computer Science, University of St Andrews [2]NHS Greater Glasgow and Clyde. Correspondence to: Mirza Nasir Hossain <mnh3@st-andrews.ac.uk>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

top-activating patches, and AUPRC-based labels derived from diagnostic-class selectivity. Unlike general-purpose SAE dashboards such as Neuronpedia (Neuronpedia, 2026) and SAEDashboard (Bloom & McDougall, 2024), Histoscope focuses on diagnostic-class structure through per-class AUPRC, recall, and activation heatmaps.

In a blinded two-rater study, pathologists confirmed all 50 features labelled monosemantic by Histoscope, but also judged 32 of 38 features labelled polysemantic by Histoscope as monosemantic, showing that the AUPRC rule misses coherent morphology when it spans multiple diagnostic classes. We release Histoscope, the trained SAE, and the evaluation protocol to support replication and extension.

2. Methods

2.1. Sparse autoencoder

We extracted 1024-dimensional patch embeddings from the CLS-token representation of frozen UNI using 224 × 224 SPIDER (Nechaev et al., 2025) colorectal patches. We trained a TopK SAE (Gao et al., 2025) with $k = 250$, 48× expansion, tied encoder–decoder weights, and a pre-encoder bias. The SAE was trained for 2 epochs with batch size 32, learning rate 10^{-4}, and MSE reconstruction loss. On the held-out test split, it explained 83.4% of variance (normalised reconstruction MSE = 0.166). All 49,152 dictionary features fired on at least one held-out test patch above their feature-wise 95th-percentile train-activation threshold, so the dead-feature count under this criterion was zero.

Hyperparameters were selected during tool development before the pathologist study. We first explored L1/ReLU SAEs across expansion factors and sparsity penalties, then switched to TopK SAEs after qualitative Histoscope inspection showed more stable feature panels. We compared TopK variants with and without tied weights and pre-encoder bias, and swept expansion factors up to 64× and sparsity values around the final regime ($k = 150$–300). The pathologist study evaluated the resulting frozen SAE.

2.2. Histoscope and the rating interface

Figure 1 shows a cropped view of the Histoscope interface. Histoscope presents SAE features through an interactive activation heatmap, with features on the x-axis and diagnostic classes on the y-axis. Features can be sorted by AUPRC or feature ID; auxiliary heuristic scores, including purity and class-activation margin, are also shown for inspection. Selecting a feature opens a panel showing top-activating patches, activation distribution across classes, and the AUPRC-based label used in this study.

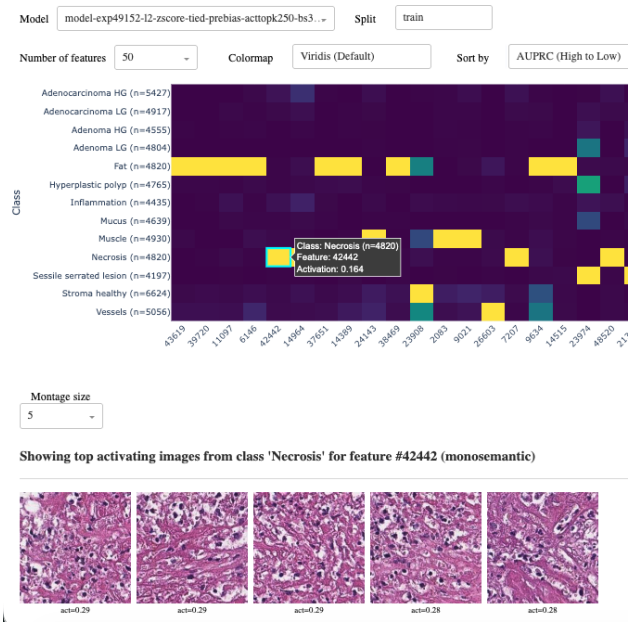

Figure 1. Cropped view of the Histoscope interface, showing the feature heatmap, sorting controls, selected feature, and top-activating patches. The full-width interface, including the activation-distribution and metric panels, is shown in Appendix A.

2.3. AUPRC-based feature stratification

Histoscope stratifies SAE features using per-class one-vs-rest area under the precision–recall curve (AUPRC) (Davis & Goadrich, 2006). The rule was developed before the pathologist study as a triage heuristic for selecting candidate feature panels. Initial experiments with class-mean and top-k purity heuristics were sensitive to outliers, so we instead treated each feature as a retrieval score over patches and asked whether high activation consistently retrieved one diagnostic class.

For each feature, we computed per-class AUPRC on the train and test splits and combined the two profiles by taking, for each class, the lower value across splits. The dominant class is the class with the highest combined AUPRC; the AUPRC gap is the difference between the highest and second-highest combined AUPRC, reducing cases where broad multi-class features are treated as monosemantic.

Final groups used fixed AUPRC thresholds, gap, and a split-specific recall floor. At the 95th-percentile train-activation threshold, the dominant class had to satisfy recall $\geq \max(0.01, \min(0.045, p_c))$ on both splits, where p_c is its prevalence on that split. This adaptive floor replaced an earlier fixed 10% floor that was too strict for rare classes while still excluding features with isolated activations.

Features with dominant-class AUPRC ≥ 0.90, AUPRC-gap ≥ 0.20, and recall above this floor on both splits were placed in the monosemantic group. Features meeting the

same AUPRC-gap and recall criteria with AUPRC 0.65–0.90 were placed in the nearly-mono group; all others were placed in the polysemantic group. Thresholds were fixed pre-evaluation and used only for sampling.

2.4. Feature sampling

For the study, Histoscope's AUPRC-based labels were collapsed into a binary sampling scheme, with nearly-mono treated as monosemantic. We selected the 50 highest-ranked monosemantic features and the 50 lowest-ranked polysemantic features. The study used top-activating patches rather than random patches, matching the intended inspection setting: users inspect the strongest visual evidence for each feature. The feature sample was confidence-stratified rather than representative of the full SAE dictionary.

2.5. Evaluation protocol

Two pathologists completed the study in a dedicated rating app. Each feature rating included: (a) morphology descriptor selection from a controlled vocabulary, with Mixed/Other escape hatches; (b) a consistency Likert score from 1–5 (1=very mixed/diffuse, 3=moderately consistent, 5=very consistent); and (c) a diagnostic-relevance Likert score from 1–5. Diagnostic relevance is not analysed here and will be revisited with a refined protocol in future work. Each feature was presented as a panel of its 12 top-activating patches, giving 1,200 patches across the 100-feature study sample. The study was split into two sessions of 50 unique features, with 10 Session A features repeated in Session B to assess intra-rater consistency. Raters were blind to all labels, thresholds, and feature-sampling groups.

We treated a feature as pathologist-confirmed monosemantic when both raters assigned consistency $\geq 4/5$. In this setting, consistency operationalises monosemanticity as a coherent morphological pattern or tightly bounded set, rather than a single diagnostic class. A stricter endpoint additionally required descriptor-set overlap ≥ 0.5, computed as the Jaccard similarity between the two raters' selected morphology descriptors.

The controlled vocabulary was initially generated by Gemini 3 Pro (Google, 2025) from non-study SPIDER patches and refined with both pathologists before the study. Inter-rater reliability was assessed using ICC(2,k) (Shrout & Fleiss, 1979), exact agreement, within-1 agreement, and morphology Jaccard. ICC(2,k) is a two-way random-effects, absolute-agreement statistic for the mean of k raters; here, $k = 2$. We compared Histoscope's AUPRC-based labels against resolved pathologist consensus; intra-rater test–retest is reported descriptively because only 10 features were repeated.

3. Results

3.1. Primary endpoint

Within the 100-feature set, both pathologists independently rated 82 features as monosemantic ($\geq 4/5$), giving $P(\text{feature-mono}) =$ **0.82** with a 95% Wilson CI of [0.73, 0.88]. Under the strict endpoint, which additionally required morphology-descriptor Jaccard ≥ 0.5, the rate dropped to 0.43 with a bootstrap CI of [0.34, 0.53]. Within this strict subset, pathologists achieved perfect descriptor-set agreement on 13 features; for 12 of these, both raters selected a single identical morphology descriptor.

The gap between the primary and strict endpoints suggests many disagreements were vocabulary-level: pathologists often agreed that a feature depicted coherent morphology but did not always choose the same descriptors for it.

3.2. Inter-rater reliability

On the monosemanticity rating, inter-rater reliability was ICC(2,k) = 0.77, 95% CI $[0.65, 0.84]$, with 62% exact agreement and 96% within-1 agreement between raters. Morphology agreement, measured as mean Jaccard similarity between the raters' selected descriptors, was 0.48, 95% CI $[0.43, 0.53]$.

Seven top-level taxonomy families were fixed during vocabulary construction before ratings were collected (Section 2.5): Architecture, Cellular, Nuclear, Cytoplasmic, Stroma, Inflammation, and Other. Collapsing descriptors to these families raised mean Jaccard from 0.48 to 0.72, a mean gain of 0.25. Of the 87 features where raters did not select identical descriptor sets, 71 showed improved overlap after family collapse, 27 became exact family matches, and 3 had no family overlap.

Reliability was stable across sessions: morphology Jaccard 0.50 vs 0.46, monosemanticity exact agreement 62% in both, and monosemanticity ICC 0.81 vs 0.72, indicating no clear exposure effect.

3.3. Intra-rater test-retest

To test whether raters were consistent with themselves across sessions, we re-presented 10 features from Session A in Session B, shuffled in with new features so they did not appear in a block. Exact agreement on the monosemanticity rating was 90% for one rater and 70% for the other; both raters reached 100% within-1 agreement, and morphology Jaccard was 0.64 and 0.61 respectively. At the ≥ 4 cut-off that drives the primary endpoint, neither rater's binary judgement flipped on re-exposure.

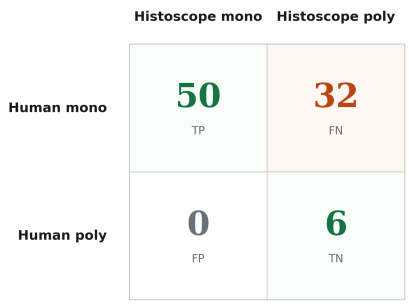

Figure 2. Comparison between Histoscope's AUPRC-based labels and dual-pathologist consensus in the study sample. Only resolved cases are included (n=88; 12 unresolved features excluded).

3.4. Tool validation

We define resolved features as those where both raters agreed on the binary monosemantic/polysemantic judgement; 88 of 100 features were resolved, with all 12 unresolved cases coming from Histoscope's polysemantic group. Restricting to the resolved set, Figure 2 compares Histoscope's AUPRC-based labels against pathologist consensus. All 50 Histoscope-mono features were confirmed as monosemantic by consensus, while 32 of 38 Histoscope-poly features were also considered monosemantic, mostly in cases where morphology spanned diagnostic classes.

Of the 50 monosemantic-labelled features, 50/50 met the primary endpoint and 36/50 met the strict endpoint; of the 50 polysemantic-labelled features, 32/50 met the primary endpoint and 7/50 met the strict endpoint.

4. Discussion

Pathologists judged 82% of the features surfaced by Histoscope to be monosemantic. Agreement was strongest where Histoscope itself had made a confident positive call: all 50 features labelled monosemantic by the tool were also judged monosemantic by both pathologists. Its polysemantic calls were less reliable, with 32 of 38 resolved cases instead judged monosemantic. These misses were concentrated in features whose morphology spans multiple diagnostic classes, which the one-vs-rest AUPRC rule cannot reward.

The mono–poly difference has both calibration and structural components. Recalibrating the AUPRC threshold against expert ratings would likely recover some missed features. However, the missed features had a different profile. Most AUPRC-positive confirmed features were class-concentrated: 45 of 50 came from a single diagnostic class. By contrast, most AUPRC-negative features that pathologists still judged monosemantic were class-distributed: 29 of 32 spanned two or more classes. Because the AUPRC rule is defined against diagnostic classes, it rewards class-concentrated features and pushes coherent cross-class morphology into the polysemantic group. Expert evaluation remains necessary to identify these features and distinguish them from genuinely polysemantic activations.

Inter-rater reliability was strong for monosemanticity, with $ICC(2, k) = 0.77$ and 96% within-1 agreement. This suggests that the study's core judgement—whether a feature reflects a single coherent morphological pattern—is reproducible across expert raters, despite the ambiguity and overlap that often characterise histopathology.

Raw descriptor agreement was lower, but family-level analysis suggested that much of the disagreement reflected vocabulary granularity. Collapsing descriptors to seven top-level morphology families raised mean Jaccard from 0.48 to 0.72, and only 3 features did not show family-level overlap.

Our evaluation differs from prior pathology SAE work in its ground truth: PLUTO-SAE (Le et al., 2025) validated features against automated cell-count proxies, which scales but does not directly test pathologist recognition. Direct expert ratings also let us separate concept coherence from descriptor matching: a coherent morphological pattern is the better primary endpoint; descriptor agreement is a stricter secondary measure.

5. Limitations

Preprocessing inconsistency. After the study, we found that the post-hoc export pipeline normalised test embeddings using test-split rather than train-split statistics. This was isolated to the export script and did not affect SAE training or feature identities, but could alter displayed patch rankings. An audit of the displayed stimuli against the corrected pipeline found near-identical embeddings (cosine similarity 0.9984) and 97.6% mean overlap across the 100 unique study panels ($k = 12$). No panel changed dominant tissue class, and no analysed feature changed its AUPRC-based group assignment under the corrected pipeline. Ratings were nevertheless collected on stimuli generated by the uncorrected pipeline, which we report as a limitation.

Vocabulary–rater dependence. The controlled vocabulary was refined with the same two pathologists who then completed the study, so descriptor-level agreement is not independent of the raters who shaped the vocabulary. This dependence bears most directly on the strict Jaccard-based endpoint; the primary consistency endpoint does not require shared descriptor choices. An extended study is planned to include an additional independent rater who did not participate in vocabulary construction.

Scope and baseline. The study involved two pathologists rating top-activating patch panels from a single SAE trained

on UNI activations from the SPIDER colorectal dataset. It therefore evaluates the coherence of Histoscope's displayed feature evidence rather than the full activation distribution of each SAE feature, and does not by itself establish generalisation across more raters, different encoders, or full-dictionary feature behaviour. Future work can also compare SAE decomposition against simpler clustering baselines to isolate the specific utility of sparse coding in this domain.

6. Conclusion

We presented Histoscope, an open-source tool for inspecting SAE features in histopathology foundation models, with a blinded two-pathologist evaluation. In this 100-feature study, 82% of features were judged monosemantic; all 50 Histoscope-mono features were confirmed by both pathologists, while 32 of 38 resolved Histoscope-poly features were also judged monosemantic. Within this confidence-stratified sample, the AUPRC rule identified a high-confidence class-concentrated subset but missed coherent cross-class morphology. AUPRC should therefore be used to rank and prioritise features for inspection rather than to filter low-selectivity features from view; expert review remains necessary to identify coherent morphology spanning diagnostic classes. Histoscope and the evaluation protocol provide a foundation for expert-grounded SAE inspection in computational pathology, with clear extensions to other encoders, datasets, and expert groups.

Reproducibility

We release the Histoscope source code, trained SAE checkpoint and configuration, and corrected preprocessing pipeline. We also release the evaluation protocol: the controlled vocabulary, rater instructions, rating-app code, prespecified thresholds, feature-sampling procedure, and example patch panels. All materials are available at https://github.com/mnhcorp/histoscope.

Impact Statement

This paper presents work whose goal is to advance mechanistic interpretability for computational pathology. Histoscope is a research tool and is not intended for clinical diagnosis or decision-making without further validation.

References

Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., and Merhof, D. Foundational models in medical imaging: A comprehensive survey and future vision, 2023. URL https://arxiv.org/abs/2310.18689.

Bloom, J. and McDougall, C. Saedashboard. https://github.com/jbloomAus/SAEDashboard, 2024. Software tool for visualising and analysing sparse autoencoder dashboards.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N. L., 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. https://transformer-circuits.pub/2023/monosemantic-features, 2023. Transformer Circuits Thread.

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A. H., Shaban, M., Williams, M., Vaidya, A., Sahai, S., Oldenburg, L., Weishaupt, L. L., Wang, J. J., Williams, W., Le, L. P., Gerber, G., and Mahmood, F. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30:850–862, 2024. doi: 10.1038/s41591-024-02857-3. URL https://doi.org/10.1038/s41591-024-02857-3.

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.

Dasdelen, M. F., Lim, H., Buck, M., Götze, K. S., Marr, C., and Schneider, S. CytoSAE: Interpretable cell embeddings for hematology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 77–86. Springer, 2025. URL https://papers.miccai.org/miccai-2025/0206-Paper4885.html.

Davis, J. and Goadrich, M. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 233–240, Pittsburgh, Pennsylvania, USA, 2006. ACM. doi: 10.1145/1143844.1143874. URL https://doi.org/10.1145/1143844.1143874.

Decode Research. SAE Lens. https://github.com/decoderesearch/SAELens, 2024. Software library for training and analysing sparse autoencoders.

Gao, L., Dupré la Tour, T., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Google. Gemini 3 Pro. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/3-pro, 2025. Model documentation; accessed 2026-04-26.

Hossain, M. N. and Harris-Birtill, D. BioFuse: An embedding fusion framework for biomedical foundation models. *PLOS ONE*, 21(3):e0320989, 2026. doi: 10.1371/journal.pone.0320989. URL https://doi.org/10.1371/journal.pone.0320989.

Le, N., Shen, C., Shah, C., Martin, B., Shenker, D., Padigela, H., Hipp, J. A., Grullon, S., Abel, J., Pokkalla, H. V., and Juyal, D. Interpretability analysis on a pathology foundation model reveals biologically relevant embeddings across modalities. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024. URL https://openreview.net/forum?id=briEoJFKof.

Le, N. M., Shen, C., Patel, N., Shah, C., Sanghavi, D., Martin, B., Eng, A., Shenker, D., Padigela, H., Biju, R., Javed, S. A., Hipp, J., Abel, J., Pokkalla, H., Grullon, S., and Juyal, D. Evaluating the utility of sparse autoencoders for interpreting a pathology foundation model. In *Proceedings of the 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI at NeurIPS*, 2025. URL https://openreview.net/pdf/0443a698c1fb99df7a57c5b6dd60b02cf7173fea.pdf.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramár, J., Dragan, A., Shah, R., and 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. 277–300. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

Mahmood Lab. UNI2-h: Pathology foundation model. https://huggingface.co/MahmoodLab/UNI2-h, 2025. Model card; accessed 2026-04-26.

McDougall, C. SAE Visualizer. https://github.com/callummcdougall/sae_vis, 2024. Software library for creating sparse autoencoder feature dashboards.

Nechaev, D., Pchelnikov, A., and Ivanova, E. SPIDER: A comprehensive multi-organ supervised pathology dataset and baseline models, 2025. URL https://arxiv.org/abs/2503.02876.

Neuronpedia. Neuronpedia. https://www.neuronpedia.org, 2026. Open mechanistic interpretability platform.

Olson, M. L., Hinck, M., Ratzlaff, N., Li, C., Howard, P., Lal, V., and Tseng, S.-Y. Probing the representational power of sparse autoencoders in vision models, 2025. URL https://arxiv.org/abs/2508.11277.

Shrout, P. E. and Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86 (2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420.

A. Histoscope Interface Examples

Figure 3. Histoscope main interface before feature selection, showing the feature heatmap, dataset/model controls, sorting controls, model-level summary, and AUPRC-based classification counts. Selecting a heatmap cell opens the feature-specific summary, activation distribution, and top-activating patch panel shown in the following examples. Additional tabs, including cluster analysis, patch browser, feature analysis, flipper analysis, and analysis report, provide secondary inspection views not used in this study.

A.1. Additional Inspection Views

Cluster analysis. Displays precomputed feature clusters, their two-dimensional organisation, representative patches, and cluster-level summaries, allowing related SAE features to be inspected together.

Patch browser. Allows patches to be filtered by tissue class and inspected through their SAE activation profiles, including the features most strongly activated by each patch.

Feature analysis. Provides galleries of top-activating patch panels grouped by dataset split and dominant tissue class, supporting rapid comparison across features.

Flipper analysis. Surfaces features whose dominant class association changes between the training and test splits, highlighting potentially unstable feature interpretations.

Analysis report. Displays the saved model-level analysis report, including configuration details, reconstruction and activation statistics, and generated evaluation summaries.

A.2. Reviewed Feature Examples

Figures 4 and 5 illustrate the two principal evaluation outcomes. Feature 14964 shows class-concentrated morphology identified as Histoscope-mono and confirmed by both pathologists. Feature 25534 shows coherent morphology spanning diagnostic classes: Histoscope labelled it poly because of its low class selectivity, whereas both pathologists judged its patch panel monosemantic. These examples are illustrative rather than representative of either feature group.

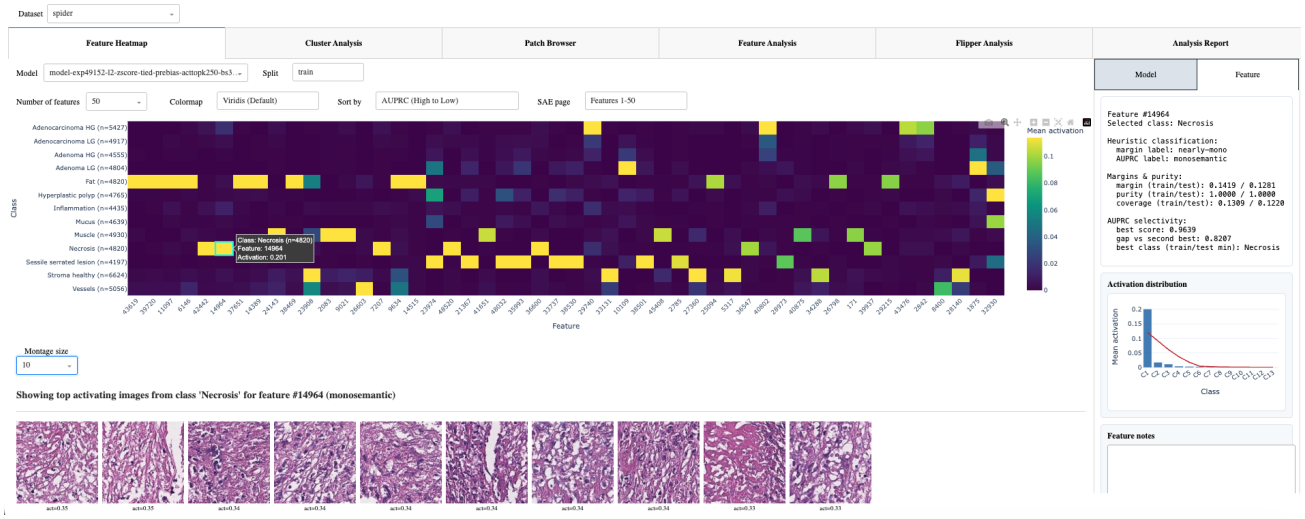

Figure 4. Selected-feature view for a Histoscope-mono feature confirmed by both pathologists. Feature 14964 was assigned to the monosemantic group by the AUPRC-based rule (AUPRC = 0.964; top-patch class purity = 1.00; dominant class = necrosis), and both raters scored it 5/5 for monosemanticity, identifying coherent necrosis and nuclear debris morphology.

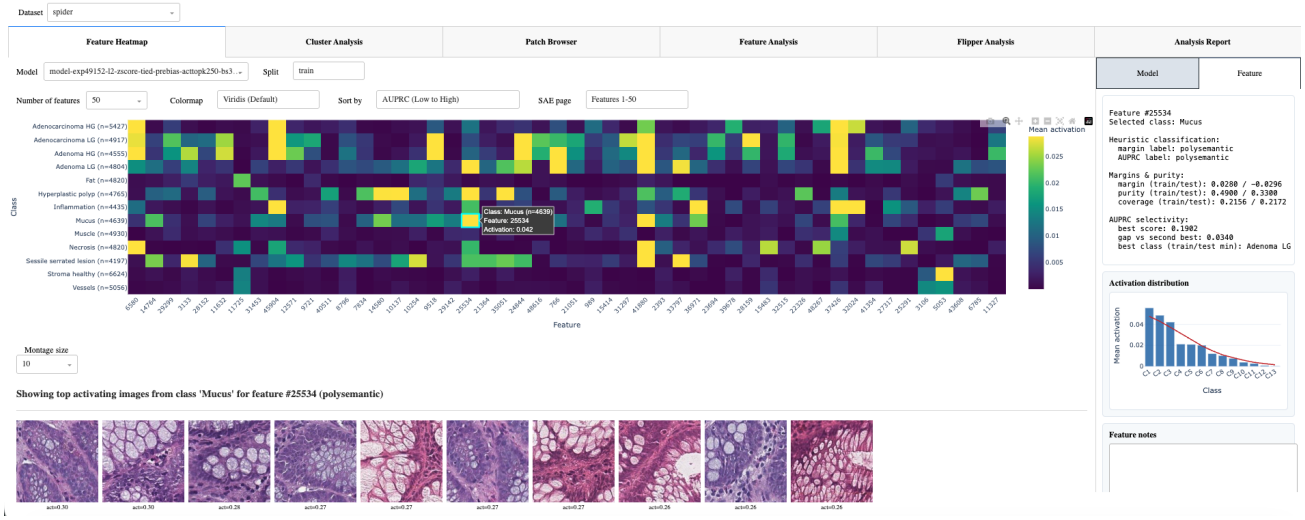

Figure 5. Selected-feature view for a Histoscope-poly feature judged monosemantic by both pathologists. Feature 25534 was assigned to the polysemantic group by the AUPRC-based rule (AUPRC = 0.196; top-patch class purity = 0.33), but both raters scored it as monosemantic (5/4), identifying coherent glandular and goblet-cell morphology across the displayed patches.