Patch-level phenotype identification via weakly supervised neuron selection in sparse autoencoders for CLIP-derived pathology embeddings

Keita Tamura[1,†], Yao-zhong Zhang[2,†,*], Yohei Okubo[2] and Seiya Imoto[2,*]

1. School of Medicine, Hiroshima University, Hiroshima, 734-8553, Japan

2. Division of Health Medical Intelligence, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, 108-8639, Japan

[*] *Corresponding E-mail: yaozhong@ims.u-tokyo.ac.jp and imoto@hgc.jp*

[†] *Equal contribution*

Computer-aided analysis of whole slide images (WSIs) has advanced rapidly with the emergence of multi-modal pathology foundation models. In this study, we propose a weakly supervised neuron selection approach to extract disentangled representations from CLIP-derived pathology foundation models, leveraging the interpretability of sparse autoencoders. Specifically, neurons are ordered and selected using whole-slide level labels within a multiple instance learning (MIL) framework. We investigate the impact of different pre-trained image embeddings derived from general and pathology images and demonstrate that a selected single neuron can effectively enable patch-level phenotype identification. Experiments on the Camelyon16 and PANDA datasets demonstrate both the effectiveness and explainability of the proposed method, as well as its generalization ability for tumor patch identification.

Keywords: Whole-slide image; Sparse autoencoder; Explainable AI

1. Introduction

Whole-slide images (WSIs)—gigapixel digital scans of histopathological slides—have become central to the digitization of diagnostic pathology. They are now routinely used to assist pathologists with tumor grading, biomarker quantification, and prognostic risk stratification. Owing to their cellular-level resolution, WSIs have the potential to reduce inter-observer variability and reveal subtle spatial patterns that are difficult to discern by eye. However, these advantages come with significant technical challenges. A single WSI can exceed 100,000 × 100,000 pixels, making naïve end-to-end model training computationally infeasible. Additionally, tissue appearance varies widely across scanners, laboratories, and staining protocols, while diagnostically relevant structures are often sparse and heterogeneous. Although slide-level labels (e.g., "tumor present") are relatively easy to acquire, generating detailed patch-wise or pixel-wise annotations is labor-intensive and costly, resulting in a severe supervision bottleneck.

To address the challenges of weak supervision and extreme resolution, most modern computational pathology pipelines adopt a multiple-instance learning (MIL)[1] framework. In this paradigm, each slide is divided into hundreds or thousands of tiles (also referred to as patches

throughout the text). Each tile is encoded into a feature vector using a representation network, and the resulting embeddings are aggregated via a pooling operator to produce a slide-level prediction trained using only global labels. For tile-level representation, early MIL approaches used generic convolutional neural networks pretrained on natural image datasets such as ImageNet[2,3] (e.g., ResNet[4]). More recent methods have shifted toward domain-specific pathology foundation models, such as UNI,[5] a self-supervised vision transformer trained on over 100 million tiles from more than 100,000 WSIs, and GigaPath,[6] which scales pretraining to over one billion tiles to capture gigapixel context. In parallel, CLIP-based vision-language models (e.g., CONCH[7]) align pathology images with textual descriptions to produce semantically rich embeddings that support zero-shot phenotype querying via natural language. For bag-level aggregation, a variety of strategies have been developed, each influencing slide-level performance differently. These include simple pooling methods (e.g., max and mean), attention-based pooling, gated attention mechanisms, and more expressive set-based functions such as Deep Sets or Transformer-style pooling. Both the choice of representation and the method of aggregation play critical roles in the effectiveness and generalizability of MIL-based WSI models.

While these methods have demonstrated strong slide-level performance, patch-level interpretability is increasingly important for clinical translation. MIL inherently provides coarse localization of discriminative regions by associating predictions with patch-level features. Among existing approaches, the type of method using attention-based aggregator with MIL, such as ABMIL[8] and CLAM,[9] has become a popular method due to its ability to assign learnable weights to individual tiles, highlighting regions most influential to the final prediction. This enables the generation of attention heatmaps without requiring dense annotations. However, a key limitation of attention-based aggregator is that attention weights are normalized within each slide, making them non-comparable across different WSIs. This restricts their utility in standardized patch-level phenotype identification and limits population-scale interpretability. Moreover, the attention mechanism typically operates over entangled high-dimensional representations, which are difficult to interpret biologically.

To overcome these limitations, we draw inspiration from recent advances in model interpretability using sparse autoencoders (SAEs).[10] Recently introduced for dissecting representations in large language models (LLMs)[11] and protein language models (PLMs),[12] SAEs project dense embeddings into sparse, overcomplete spaces by enforcing sparsity constraints during training. Remarkably, many individual neurons in the SAE bottleneck correspond to monosemantic units, each selectively activating for specific, semantically coherent concepts—such as a functional motif in a protein sequence or a syntactic role in text. These properties make SAEs powerful tools for understanding and manipulating the internal representations of large foundation models.

Building on SAE for embeddings of the pathology foundation model, we propose a novel framework for patch-level phenotype identification via weakly supervised neuron selection. Specifically, we train an SAE on CLIP-like pathology embeddings and identify single neurons whose activation patterns correlate with slide-level labels within the MIL paradigm. These neurons represent disentangled, semantically meaningful features that are both predictive of slide-level outcomes and spatially localizable at the patch level, thus enabling interpretable

phenotype annotation without requiring dense supervision. We conducted experiments on the Camelyon16[13] and PANDA[14] datasets, and the results demonstrate that the proposed approach achieves competitive performance in tumor patch identification while offering interpretable and generalizable spatial phenotyping.

2. Methods

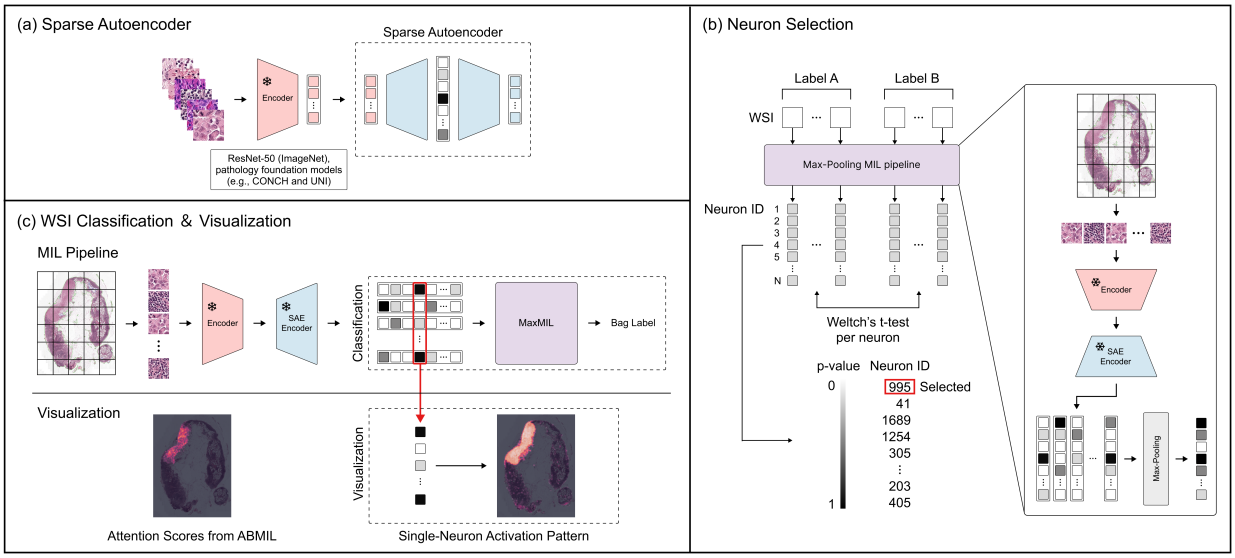

Figure 1. Overall pipeline of the proposed SAE-1N method. (a) SAE applied for different embeddings of patch images. (b) Neuron selection method for SAE disentangled neurons with weak supervision from the whole-slide labels. (c) MIL pipelines for the whole-slide level classification and patch-level classification with a selected neuron.

An overview of our proposed method SAE-1N is illustrated in Figure 1. SAE-1N denotes the single-neuron selection strategy based on disentangled neurons derived from a sparse autoencoder (SAE). We begin by applying an SAE to disentangle latent representations from patch-level embeddings extracted using a pretrained pathology foundation model. Each latent neuron is then evaluated for its capacity to support whole-slide phenotype classification within a multiple instance learning (MIL) framework. To identify the most informative neurons, we assess their discriminative power between phenotype classes using statistical testing, enabling principled neuron ranking and selection. The top-ranked neurons are subsequently employed for downstream tasks, such as slide- and patch-level phenotype prediction.

2.1. *Sparse Autoencoder for Disentangled Representation*

Sparse autoencoders (SAEs) have emerged as a powerful framework for disentangling superposed representations in high-dimensional neural embeddings. By learning sparse and overcomplete latent codes, SAEs promote the emergence of monosemantic units—individual neurons that consistently correspond to human-interpretable concepts. This property has been

successfully leveraged in mechanistic interpretability studies of large language models,[11] protein language models,[12] and structure–function models such as Evo2,[15] where SAEs have been shown to recover latent dimensions aligned with semantic, biochemical, or structural factors from otherwise entangled activations.

The model architecture of SAE follows the conventional encoder–decoder design of autoencoders but employs an overcomplete latent space, where the number of latent units substantially exceeds the input dimensionality. For learning sparse and overcomplete latent codes, we adopt the BatchTopK Sparse Autoencoder (SAE-BatchTopK),[16] an extension of SAE-TopK.[17] SAE-TopK constrains each input to activate exactly k latent units, which improves interpretability but imposes a fixed sparsity level per sample. SAE-BatchTopK generalizes this idea by enforcing the sparsity constraint across a mini-batch, thereby encouraging feature reuse, allowing per-sample variability, and maintaining consistent average sparsity during training. In practice, this enables simpler inputs to be represented with fewer active neurons, while more complex inputs naturally recruit a larger subset, yielding more flexible and efficient representations.

SAE-BatchTopK was trained to minimize a loss consisting of a reconstruction term and an auxiliary penalty that discourages persistent neuron inactivity. Given a batch of B inputs $\{x_i\}_{i=1}^{B}$ of size D, the total loss is defined as $\mathcal{L} = L_{\text{recon}} + L_{\text{aux}}$, where $L_{\text{recon}} = \frac{1}{BD}\sum_{i=1}^{B} \|x_i - \hat{x}_i\|_2^2$ is the reconstruction loss, and $L_{\text{aux}} = \frac{\lambda_{\text{aux}}}{BD}\sum_{i=1}^{B} \| \sum_{j\in\mathcal{D}} w_j h_{i,j} - (x_i - \hat{x}_i)\|_2^2$ is the auxiliary loss to avoid dead latents. Here, $h_i = W_{\text{enc}} x_i + b_{\text{enc}}$ is the latent activation, $\tilde{h}_i = \text{BatchTopK}(h_i)$ the sparsified vector, $\hat{x}_i = W_{\text{dec}}\tilde{h}_i + b_{\text{dec}}$ the reconstruction, w_j the j-th column from the decoding weights W_{dec}, and $\mathcal{D}$ indexes latent units inactive for $n_{\text{inactive}} = 5$ consecutive batches. As the default setting, training was conducted per dataset for 100 epochs using the Adam optimizer ($\text{lr} = 10^{-3}, \beta_1 = 0.9, \beta_2 = 0.99$), with a batch size of 4096 and gradient norm clipping at 10^5. Decoder weights were re-normalized after each update, and the checkpoint with the lowest validation loss was selected for downstream analysis. Complete configuration details are available in the supplementary material. In the following sections, SAE and SAE-BatchTopK are used interchangeably.

2.2. *Pretrained Pathology Models for Patch Embedding*

The emergence of large-scale foundation models has significantly reshaped the landscape of computational pathology. Early approaches commonly relied on convolutional neural networks (CNNs) pretrained on natural image datasets such as ImageNet,[2] with architectures like ResNet-50[4] serving as generic feature extractors for histopathology patches. While these models provided a strong starting point for weakly supervised learning frameworks such as multiple-instance learning (MIL), their limited domain specificity constrained performance on complex pathology tasks. Recent advances, such as CONCH (CONtrastive learning from Captions for Histopathology)[7] and UNI,[18] have demonstrated that representations learned directly from large-scale pathology data can generalize more effectively across diverse diagnostic settings. These models, typically built upon Vision Transformer backbones and trained using multimodal contrastive or self-supervised objectives, yield domain-specialized embeddings that substantially improve downstream WSI-level classification. However, the result-

ing feature spaces are increasingly high-dimensional and entangled. Despite their predictive strength, these embeddings remain difficult to interpret, presenting a growing performance–interpretability trade-off that limits clinical transparency.

To explore the internal structure and interpretability of high-dimensional embeddings from pathology foundation models, we analyze patch-level features extracted from three representative pretrained encoders. Whole-slide images (WSIs) are first segmented into non-overlapping patches of size 256×256 at a resolution of $0.5\,\mu$m/pixel using the CLAM framework.[9] The following models are used to generate patch-level embeddings: (1) ResNet-50,[4] pretrained on ImageNet,[2] as a conventional convolutional baseline, producing 1024-dimensional features; (2) UNI,[18] a self-supervised Vision Transformer trained on over 100 million image patches from diagnostic slides using a DINOv2-style distillation framework,[19] yielding 1024-dimensional embeddings optimized for general-purpose tissue representation; and (3) CONCH,[7] a CLIP-style foundation model built on a CoCa architecture,[20] trained on 1.17 million image–caption pairs to align histopathology patches with free-text descriptions, producing 512-dimensional semantically enriched embeddings. These diverse encoders reflect the spectrum of representation learning strategies in computational pathology, ranging from natural image transfer learning to large-scale self-supervised and multimodal pretraining. We applied SAE-BatchTopK to each set of patch embeddings to assess whether sparse, monosemantic structure could be recovered from these dense representations. We paid particular attention to CONCH, as its multimodal alignment provides a compelling test case for evaluating interpretability through disentanglement.

2.3. *Disentangled Neuron Explainability and Neuron-phenotype Association*

For identifying the potential meaning of the neurons disentangled by SAEs, prior work has explored various strategies across domains such as language modeling and computational pathology. In the context of large language models (LLMs), mechanistic interpretability studies utilize sparse autoencoders in combination with prompt-based querying to associate individual neurons with semantic functions, often by inspecting example activations.[11] In pathology, ProtoMIL[21] introduces an end-to-end explainable multiple-instance learning (MIL) framework that first trains a sparse autoencoder to extract high-level, human-interpretable concepts from patch-level embeddings. These concepts are then used for slide-level classification. Crucially, ProtoMIL supports *human-in-the-loop* interpretability by enabling domain experts to manually inspect and disable irrelevant or spurious concepts, thereby aligning model predictions more closely with clinical reasoning.

In this work, we propose a weakly supervised strategy to associate whole-slide phenotype labels with individual patch-level neurons disentangled by the sparse autoencoder. Analogous to how neuron activations are profiled in large language models (LLMs) to capture shared input semantics, our approach treats the slide-level supervision as an indirect indicator of neuron selectivity. However, unlike LLM-based methods that rely on coarse approximations of neuron meaning or manual annotation, our method directly links each neuron to a specific whole-slide label. This precise supervision enables the identification of neurons that are both

semantically aligned and highly discriminative, supporting downstream applications such as patch-level phenotype annotation with improved interpretability.

We develop a neuron selection framework that identifies SAE neurons associated with whole-slide phenotype labels. Taking Camelyon16 as an example, where WSIs are labeled as tumor-positive or tumor-negative at the whole-slide level, we leverage patch-level embeddings from pretrained encoders and link them to slide-level signals through statistical inference. This enables explanation at the neuron level without requiring dense annotations or human-in-the-loop labeling. After training SAE-BatchTopK, we compute each neuron's slide-level activations by max-pooling across all patches P_i from slide i: $a_{i,j} = \max_{p \in P_i} h_{p,j}$, where $h_{p,j}$ is the activation of neuron j for patch p. Slides are grouped by phenotype into tumor-positive ($\mathcal{P}^+$) and tumor-negative ($\mathcal{P}^-$) sets based on slide-level labels. To assess neuron–phenotype association, we apply Welch's t-test to compare activation distributions:

$$t_j = \frac{\bar{a}_j^+ - \bar{a}_j^-}{\sqrt{\frac{s_{+,j}^2}{|\mathcal{P}^+|} + \frac{s_{-,j}^2}{|\mathcal{P}^-|}}}, \tag{1}$$

where $\bar{a}_j^{\pm}$ and $s_{\pm,j}^2$ denote the mean and variance of slide-level activations across slides in each phenotype. Neurons are then ranked by statistical significance (i.e., p-values), and top-ranked units are selected as interpretable detectors for downstream patch-level phenotype identification. In SAE-1N, only the top-1 neuron is selected for use.

This approach links sparse neuron activations to weak supervisory signals, aligning tumor-associated neurons with localized high-activation patterns. By isolating individual units associated with disease states, this method provides a transparent mechanism for tracing WSI-level predictions back to single, interpretable features, thereby enhancing explainability while incurring only a reduced compromise in predictive performance.

3. Results

We evaluated the proposed method on two benchmark whole-slide image datasets: Camelyon16[13] and PANDA.[14] The Camelyon16 dataset is a widely used benchmark for evaluating algorithms in the automated detection of breast cancer metastases in lymph node whole-slide images. It comprises a total of 399 WSIs, including 270 training slides sourced from two pathology centers and 129 test slides. The dataset provides both slide-level labels and detailed pixel-level annotations of tumor regions. During training, only whole-slide labels were used in a weakly supervised manner. For evaluation, however, we additionally incorporated patch-level and pixel-level annotations to assess model performance. We conducted five-fold stratified cross-validation on the Camelyon16 dataset for slide-level metastasis classification.

The PANDA (Prostate cANcer graDe Assessment) dataset is a large-scale benchmark comprising more than 10,000 prostate WSIs collected from two institutions: Radboud University Medical Center (Radboud) and Karolinska Institutet (Karolinska). The Radboud data contain fine-grained, gland-centric annotations that distinguish stroma from epithelium, whereas the Karolinska data provide region-level masks in which both benign and cancer regions include stromal and epithelial tissue. For slide-level cancer detection, we reformulated the ISUP grades such that Grade Group 0 was considered normal, whereas Grade Groups 1–5 were considered

tumors. For the patch-level segmentation task, categories 3–5 in Radboud data and category 2 in Karolinska data were treated as tumor, while categories 0–2 in Radboud data and categories 0–1 in Karolinska data were treated as non-tumor. Owing to its larger scale relative to Camelyon16, we performed experiments under both mixed-institution and independent-institution splits. In each setting, the data were randomly partitioned into training, validation, and test sets following an 80/10/10 ratio.

3.1. *SAE enables disentanglement of superposed features in neural representations*

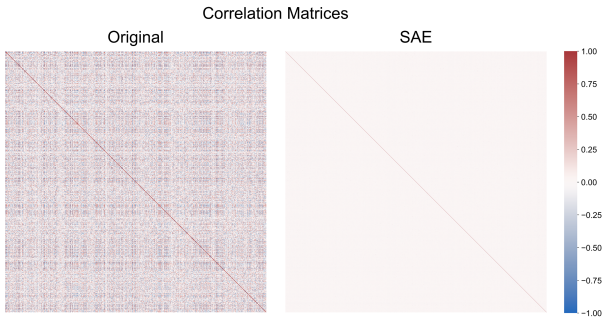

Figure 2. Correlation matrices of latent neuron activations before and after applying the SAE to CONCH embeddings on the Camelyon16 validation set. The left panel shows the correlation matrix (512×512) of the individual units in the original CONCH embeddings. The right panel displays the correlation matrix (2048×2048) of the neurons in the SAE-transformed representations of the same embeddings.

Figure 2 presents the Pearson correlation matrices of neuron activations before and after applying the SAE to CONCH embeddings on the Camelyon16 validation set. (In the left panel, each unit in the original CONCH embedding is treated as a neuron.) Each neuron is represented by its activation values across all samples. After applying the SAE, the correlation values at off-diagonal positions generally decrease, indicating the effectiveness of the disentanglement achieved by the SAE.

3.2. *Benchmarking Multiple Instance Learning Models with Diverse Embeddings and Aggregation Methods*

To benchmark performance, we evaluated multiple MIL models using various pretrained embeddings and aggregation strategies on the Camelyon16 data, including MaxMIL, MeanMIL, ABMIL,[8] CLAM,[9] TransMIL,[22] and MambaMIL.[23] MaxMIL and MeanMIL are used as baselines. ABMIL introduces a trainable soft attention mechanism that assigns adaptive weights to individual instances, enabling the model to focus on diverse discriminative regions while maintaining interpretability. CLAM (single- or multi-branch) extends ABMIL with class-specific attention and feature-space clustering regularization.TransMIL incorporates transformers to model long-range spatial dependencies among patches, and MambaMIL leverages a state-

space sequence model to efficiently capture global context across large bags of instances. For patch-level image embeddings, we evaluate three widely used pretrained encoders: ResNet-50, trained on natural images from ImageNet; CONCH, a CLIP-style model pretrained on paired pathology images and text captions; and UNI, a self-supervised Vision Transformer trained on a large-scale corpus of diagnostic whole-slide pathology images.

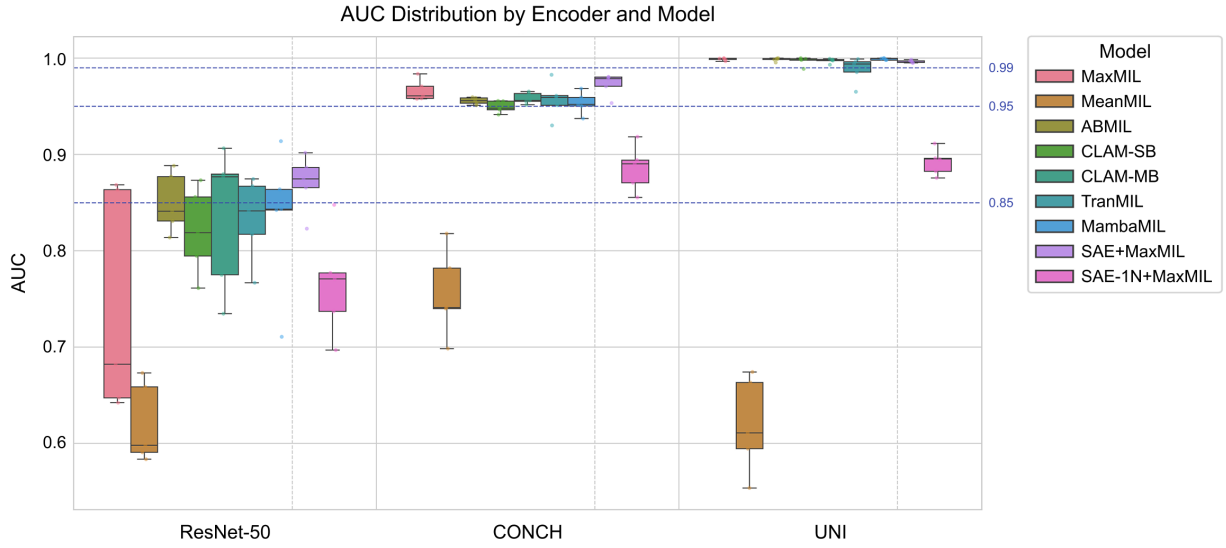

Figure 3. Whole-slide level performance of different MIL models using various pretrained embeddings (ResNet-50, CONCH, and UNI) on the Camelyon16 dataset. Area Under the Curve (AUC) scores are reported as the primary performance metric. Accuracy results are provided in the supplementary material.

As illustrated in Figure 3, pathology-specific pretrained encoders (CONCH and UNI) consistently outperform the general-purpose ResNet-50 encoder across all multiple instance learning (MIL) models on the Camelyon16 dataset. For instance, using ABMIL, the UNI encoder achieves the highest mean AUC, followed closely by CONCH. In contrast, ResNet-50 yields the lowest performance with the widest interquartile ranges, indicating greater variability and less robust feature extraction. These results underscore the necessity of domain-specific pretraining: encoders trained on large-scale histopathology datasets capture tissue-relevant cues that are absent in models pretrained on natural images like ImageNet. This observation also aligns with findings reported for pathology-based embeddings in survival analysis.[24]

Across all three encoders, MeanMIL consistently produced the weakest performance. This limitation is particularly pronounced in the Camelyon16 dataset, where tumor regions occupy only a small fraction of each WSI—causing average pooling to dilute the discriminative signal from informative patches. In contrast, MaxMIL emerges as a strong baseline, particularly when paired with a high-quality encoder. For example, the UNI encoder combined with MaxMIL achieved median AUC and accuracy scores of 0.9988 and 0.9752, respectively. Attention-based models such as ABMIL, CLAM-SB, and CLAM-MB, as well as transformer-based models like

TransMIL and MambaMIL, generally matched or slightly outperformed MaxMIL approaches. Among them, MambaMIL exhibited the most consistent performance across encoders, suggesting that global sequence modeling further enhances feature aggregation. Therefore, the quality of pretrained embeddings is a dominant factor in WSI-level classification performance, and while architectural innovations in MIL can offer marginal gains, better encoders can elevate simple pooling strategies to near-optimal performance.

3.3. *Phenotype Discrimination with SAE-1N on different Embeddings*

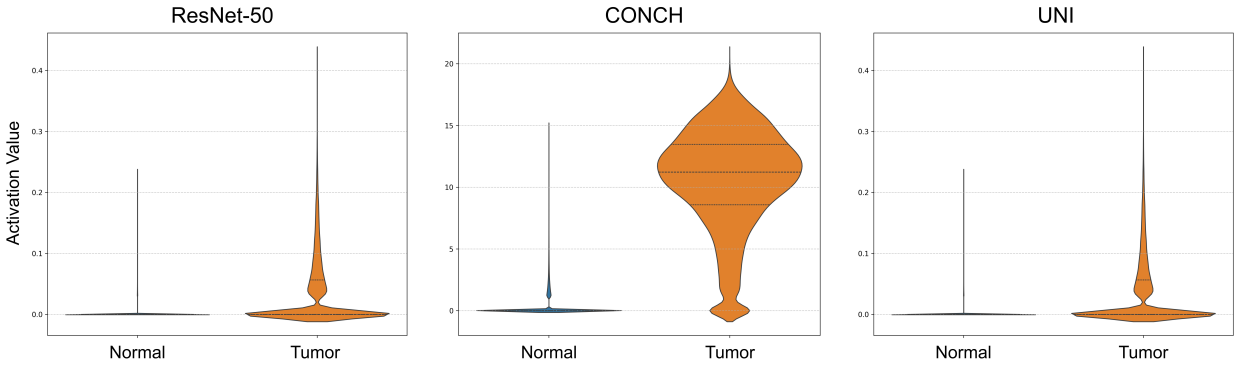

Figure 4. Violin plots depicting activation distributions of the top-selected SAE neurons for each encoder (ResNet-50, CONCH, and UNI). Activations are grouped according to patch-labels ("Normal" and "Tumor"), derived from pixel-level annotations in Camelyon16, where patches containing $\geq 20\%$ annotated tumor regions are labeled as "Tumor".

We evaluated the effectiveness of the proposed method across different embedding types using pixel-level annotations. Patch-level labels ("Normal" and "Tumor") were derived from the Camelyon16 training set, where patches containing $\geq 20\%$ annotated tumor regions were labeled as "Tumor". To assess the discriminative power of the selected neuron, we grouped patches by their assigned labels and examined the activation values of the selected neuron within each group. As shown in Figure 4, SAE-1N using the CONCH embedding exhibits a clear distinction between tumor and normal patches, whereas SAE-1N applied to ResNet-50 and UNI embeddings does not demonstrate such separation. Since UNI embeddings are also pre-trained on pathology images, one possible explanation for the stronger discriminative capacity of CONCH embeddings is their use of contrastive learning between pathology images and textual descriptions. This cross-modal alignment may coarsely encode phenotype-relevant information into the embeddings, enabling SAE-1N to isolate disentangled neurons that are more effective for phenotype discrimination.

3.4. *Comparison of SAE-1N and Attention-Based Methods for Patch-Level Phenotype Identification*

We further compared the performance of SAE-1N with attention-based MIL models, including ABMIL, CLAM-SB, CLAM-MB, and TransMIL, where attention scores were used empirically

to infer patch-level predictions. For the original CONCH embeddings, we leveraged the model's zero-shot capability within a contrastive learning framework. Text embeddings were generated using CONCH's text encoder with the prompt "An image of metastatic breast carcinoma in a lymph node" for Camelyon16 and "An image of prostate cancer" for the PANDA dataset. Cosine similarity between this text embedding and each patch image embedding was then computed to produce a heatmap. Following prior work,[25] prediction thresholds for the above methods were determined using the validation set, guided by patch-level annotations to optimize classification performance.

Table 1. Comparison of spatial overlap performance on the Camelyon16 and PANDA datasets. For Camelyon16, results are reported as mean Dice coefficients with standard deviations obtained from five-fold cross-validation. For PANDA, Dice coefficients are provided for the combined Radboud and Karolinska (R&K) set, as well as for the Radboud and Karolinska subsets individually.

Method	**Camelyon16**	**PANDA**		
		R&K	**Radboud**	**Karolinska**
Zero-shot (CONCH)	0.1612 (± 0.0107)	0.5361	0.5969	0.4990
ABMIL	0.5146 (± 0.0482)	0.5606	0.6198	0.5017
CLAM-SB	0.4785 (± 0.0630)	0.5563	0.5822	0.4767
CLAM-MB	0.5551 (± 0.0208)	0.5781	0.5596	0.4945
TransMIL	0.0950 (± 0.0269)	0.4566	0.4443	0.4549
SAE-1N	0.6135 (± 0.0309)	0.5705	0.6375	0.5129

The performance results are presented in Table 1. On Camelyon16, SAE-1N achieves the mean Dice score of 0.6135, outperforming attention-based MIL methods such as ABMIL (0.5146), CLAM-SB (0.4785), and CLAM-MB (0.5551). On the PANDA dataset, SAE-1N also performs competitively: it achieves 0.5705 on the combined Radboud & Karolinska (R&K) split, surpassing ABMIL (0.5606) and CLAM-SB (0.5563), and achieves the best score on Radboud (0.6375), while obtaining 0.5129 on Karolinska. The zero-shot method exhibits strong sensitivity to text prompt design. Its performance is relatively lower on Camelyon16, and it further decreases to 0.0942 when the prompt is changed to "An image of tumor". Overall, SAE-1N demonstrates the effectiveness of leveraging a single selected neuron to target epithelial regions in WSIs for patch-level phenotype identification.

3.5. *Interpretable Phenotype Visualization Using SAE-1N*

We visualized the activation values of the selected neuron from SAE-1N to highlight phenotype-relevant regions. Three representative whole-slide images from the Camelyon16 test set were selected, each containing tumors of different sizes (small, medium, and large). For the PANDA dataset, one example was randomly selected from Radboud and one from Karolinska. The activation maps from SAE-1N were compared with the attention scores produced by the ABMIL model.

As shown in Figure 5, attention-based methods such as ABMIL often highlight only a subset of the tumor region, even in slides where the tumor occupies a large area. This limitation

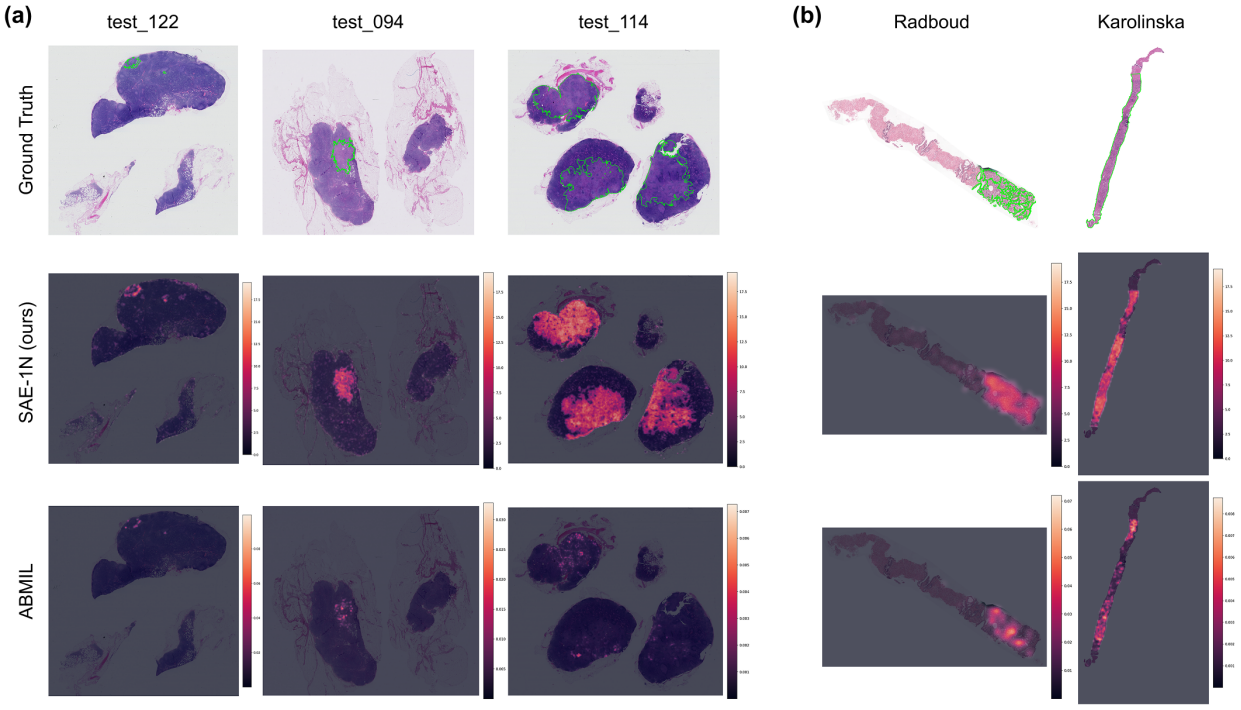

Figure 5. Visualization results from SAE-1N and ABMIL. From top to bottom: the ground-truth annotation, the activation heatmap of the selected SAE neuron (SAE-1N), and the attention map generated by ABMIL. (a) Camelyon16: From left to right, three cases with different tumor sizes (small: test_122, medium: test_094, and large: test_114). (b) PANDA: From left to right, two cases from the PANDA dataset (Radboud: 3424fa4daa40e0bb944a1f579ca895ca, and Karolinska: 4fc84c3c665c970865284b352e2134e3). The SAE-1N visualizations for PANDA were obtained from the model trained on the combined Radboud and Karolinska (R&K) splits.

arises because, in slide-level classification tasks, a small number of high-attention patches is often sufficient for accurate prediction, even if these patches do not comprehensively represent the entire tumor area. Consequently, conventional MIL classification metrics may overlook such incomplete tumor coverage. Moreover, attention scores reflect regions that contribute most to the model's prediction but do not necessarily correspond to the actual presence of tumor tissue. In contrast, SAE-1N activations show a direct and statistically validated association with tumor regions, achieved through the weakly supervised selection of a neuron linked to the whole-slide–level phenotype.

Unlike attention scores computed via Softmax, which are relative within each slide, the activation value of the selected SAE neuron is defined on an absolute scale based on the training set. This enables consistent evaluation across samples and supports a more accessible visualization strategy for whole-slide image analysis. By relying on a single neuron, SAE-1N offers a simple yet transparent approach for tumor localization in computational pathology.

4. Discussion

Our work diverges from both Le et al.[26] and ProtoMIL[21] in three key respects: supervision, interpretability granularity, and the usage of sparse autoencoders (SAEs). Le et al. demonstrated that unsupervised SAE training on pathology embeddings can uncover monosemantic latent units correlated with cell types. However, their work does not connect latent features to explicit diagnostic tasks. In contrast, our method introduces weak slide-level phenotype supervision within an MIL framework to select individual SAE neurons that are directly predictive of tumor presence. This links each neuron to patch-level phenotype localization, moving beyond unsupervised discovery into task-specific interpretability. ProtoMIL applies SAE to extract human-interpretable "concepts" from embeddings, which are then used in an inherently interpretable MIL classification pipeline. It further allows pathologists to intervene by disabling spurious concepts. While similar in its use of SAE-derived concepts, our work focuses on interpretable neuron selection rather than prototype aggregation, and critically evaluates performance across different pretrained embeddings—especially CLIP-style models like CONCH. We target single-neuron explainability and patch-level phenotype detection, rather than slide-level prototype-based interpretability. Thus, our method complements these works by emphasizing neuron-level transparency, enabling more precise, localized explanations tied to clinical phenotypes.

The advantage of using a single-neuron sparse autoencoder (SAE) as the input feature lies in its high interpretability: a single neuron is directly linked to the final prediction. For instance, in the Camelyon16 classification task that distinguishes tumor from normal tissue, the selected neuron can be clearly attributed to either tumor or normal regions. This makes it a highly intuitive and interpretable marker for visualization. Although this study primarily evaluates the proposed method using the Camelyon16 and PANDA datasets, the approach is not limited to phenotypes such as tumor versus normal. The proposed method can be extended to other phenotype types as well. On the other hand, the current method has primarily been developed for identifying two phenotypes. For multiple phenotypes, the neuron selection strategy based on whole-slide supervision could be extended, for example, by employing Welch's ANOVA. We plan to explore these extensions as part of future work.

5. Conclusion

In this work, we propose a novel approach that leverages a sparse autoencoder to learn disentangled representations from CLIP-based pathology foundation models. We further introduce a statistically grounded weakly supervised method for selecting phenotype-associated neurons for downstream tasks. Our results demonstrate that the proposed method serves as a simple and interpretable indicator for identifying patch-level phenotypes. Experiments on the Camelyon16 and PANDA datasets also show that the proposed SAE-1N model achieves competitive performance on the patch-level annotation task.

Competing interests

The authors declare no competing interests.

Acknowledgment

This work was carried out as a part of the Genome research in CAncers and Rare Diseases(G-CARD) project, supported by the Japan Agency for Medical Research and Development (AMED) under Grant Number 23ck0106873h0002, 23ck0106873h0002, 24ck0106873h0003 and 25ck0106873h0004. The computing resources in this work were provided by Human Genome Center, the Institute of Medical Science, the University of Tokyo.

Code availability

The source code is available at `https://github.com/tamukei/sae-1n`

References

1. T. G. Dietterich, R. H. Lathrop and T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, *Artificial Intelligence* **89**, 31–71 (1997).
2. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, Imagenet: A large-scale hierarchical image database, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
3. M. Talo, Convolutional neural networks for multi-class histopathology image classification, *arXiv preprint arXiv:1903.10035* (2019).
4. K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
5. R. J. Chen, T. Ding, M. Y. Lu, D. F. K. Williamson, G. Jaume, B. Chen, A. Zhang, D. Shao, A. H. Song, M. Shaban *et al.*, Towards a general-purpose foundation model for computational pathology, *Nature Medicine* **30**, 850–862 (2024).
6. H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu *et al.*, A whole-slide foundation model for digital pathology from real-world data, *Nature* **630**, 181–188 (2024).
7. M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber *et al.*, A visual-language foundation model for computational pathology, *Nature Medicine* **30**, 863–874 (2024).
8. M. Ilse, J. M. Tomczak and M. Welling, Attention-based deep multiple instance learning, *International Conference on Machine Learning (ICML)*, 2018.
9. M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri and F. Mahmood, Data-efficient and weakly supervised computational pathology on whole-slide images, *Nature Biomedical Engineering* **5**, 555–570 (2021).
10. A. Makhzani and B. Frey, k-sparse autoencoders, *arXiv preprint arXiv:1312.5663* (2013).
11. R. Huben, H. Cunningham, L. R. Smith, A. Ewart and L. Sharkey, Sparse autoencoders find highly interpretable features in language models, *International Conference on Learning Representations (ICLR) 2024*, 2024.
12. E. Simon and J. Zou, Interplm: discovering interpretable features in protein language models via sparse autoencoders, *Nature Methods* (2025).
13. B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol *et al.*, Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer, *Jama* **318**, 2199–2210 (2017).
14. W. Bulten, K. Kartasalo, P.-H. C. Chen, P. Ström, H. Pinckaers, K. Nagpal, Y. Cai, D. F. Steiner, H. Van Boven, R. Vink *et al.*, Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge, *Nature medicine* **28**, 154–163 (2022).

15. G. Brixi, M. G. Durrant, J. Ku, M. Poli, G. Brockman, D. Chang, G. A. Gonzalez, S. H. King, D. B. Li, A. T. Merchant *et al.*, Genome modeling and design across all domains of life with evo 2, *bioRxiv* (2025).
16. B. Bussmann, P. Leask and N. Nanda, Batchtopk sparse autoencoders, *arXiv preprint arXiv:2412.06410* (2024).
17. L. Gao, T. D. la Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike and J. Wu, Scaling and evaluating sparse autoencoders, *arXiv preprint arXiv:2406.04093* (2024).
18. R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban *et al.*, Towards a general-purpose foundation model for computational pathology, *Nature Medicine* **30**, 850–862 (2024).
19. M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, Dinov2: Learning robust visual features without supervision, *arXiv preprint arXiv:2304.07193* (2023).
20. J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini and Y. Wu, Coca: Contrastive captioners are image-text foundation models, *arXiv preprint arXiv:2205.01917* (2022).
21. S. Sun, D. van Midden, G. Litjens and C. F. Baumgartner, Prototype-based multiple instance learning for gigapixel whole slide image classification, *arXiv preprint arXiv:2503.08384* (2025).
22. Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji *et al.*, Transmil: Transformer based correlated multiple instance learning for whole slide image classification, *Advances in Neural Information Processing Systems* **34**, 2136–2147 (2021).
23. S. Yang, Y. Wang and H. Chen, Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024.
24. K. M. Papadopoulos and T. Stathaki, Comparing imagenet pre-training with digital pathology foundation models for whole slide image-based survival analysis, *arXiv preprint arXiv:2405.17446* (2024).
25. W. Zhang, J. Chen and C. Kanan, Insight: Explainable weakly-supervised medical image analysis, *arXiv preprint arXiv:2412.02012* (2024).
26. N. M. Le, C. Shen, N. Patel, C. Shah, D. Sanghavi, B. Martin, ... and D. Juyal, Learning biologically relevant features in a pathology foundation model using sparse autoencoders, *arXiv preprint arXiv:2407.10785* (2024).