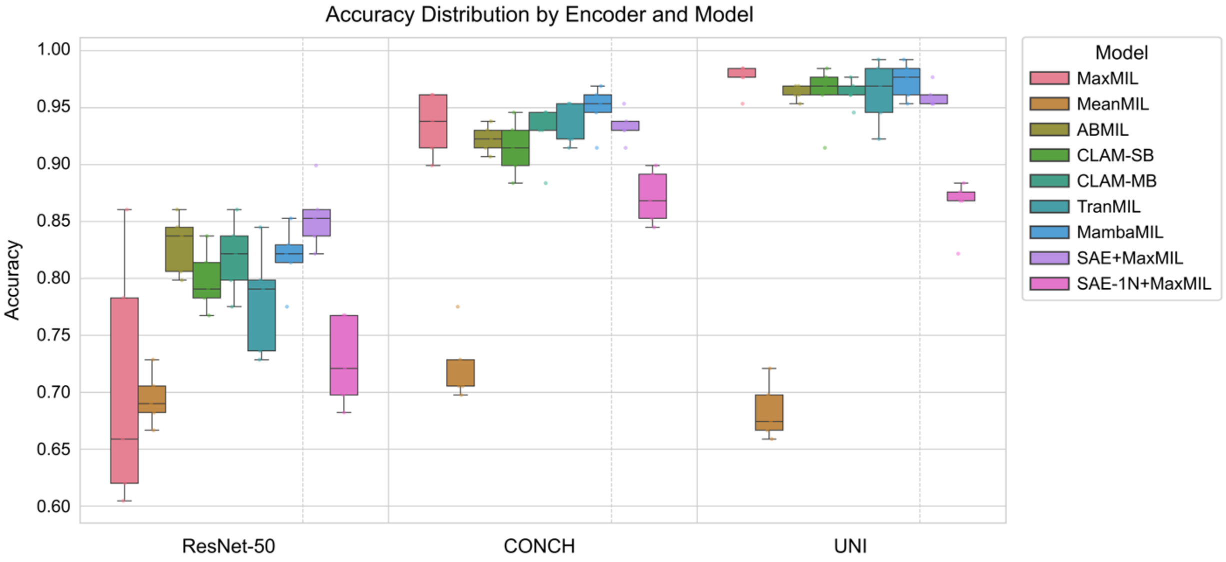
Supplementary Information

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

Supplementary Figure 1: Whole-slide level performance on Camelyon16

Whole-slide level performance of different MIL models using various pretrained embeddings (ResNet-50, CONCH, and UNI) on the Camelyon16 dataset. Accuracy results are reported here.



Supplementary Table 1: Whole-slide level performance on Camelyon16

	ResNet-50		CONCH		UNI	
Model	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
MaxMIL	0.7054 (± 0.1114)	0.7408 (± 0.1153)	0.9349 (± 0.0277)	0.9662 (± 0.0111)	0.9752 (± 0.0127)	0.9988 (± 0.0013)
MeanMIL	0.6946 (± 0.0236)	0.6209 (± 0.0417)	0.7225 (± 0.0317)	0.7559 (± 0.0455)	0.6837 (± 0.0254)	0.6194 (± 0.0498)
ABMIL	0.8295 (± 0.0263)	0.8502 (± 0.0314)	0.9225 (± 0.0123)	0.9557 (± 0.0036)	0.9628 (± 0.0065)	0.9985 (± 0.0018)
CLAM-SB	0.7984 (± 0.0274)	0.8208 (± 0.0453)	0.9147 (± 0.0245)	0.9493 (± 0.0061)	0.9612 (± 0.0274)	0.9970 (± 0.0046)
CLAM-MB	0.8186 (± 0.0332)	0.8345 (± 0.0750)	0.9271 (± 0.0255)	0.9583 (± 0.0058)	0.9628 (± 0.0115)	0.9972 (± 0.0024)
TransMIL	0.7798 (± 0.0435)	0.8334 (± 0.0480)	0.9333 (± 0.0187)	0.9568 (± 0.0190)	0.9628 (± 0.0287)	0.9879 (± 0.0137)
MambaMIL	0.8186 (± 0.0283)	0.8348 (± 0.0752)	0.9488 (± 0.0209)	0.9534 (± 0.0115)	0.9736 (± 0.0161)	0.9990 (± 0.0010)
SAE+MaxMIL	0.8543 (± 0.0261)	0.8703 (± 0.0266)	0.9333 (± 0.0126)	0.9727 (± 0.0103)	0.9597 (± 0.0090)	0.9966 (± 0.0013)
SAE-1N+MaxMIL	0.7271 (± 0.0351)	0.7659 (± 0.0499)	0.8713 (± 0.0211)	0.8858 (± 0.0215)	0.8636 (± 0.0217)	0.8921 (± 0.0123)

Supplementary Table 2: Whole-slide level performance on PANDA

	Radboud & Karolinska		Radboud		Karolinska	
Model	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
MaxMIL	0.9453	0.9837	0.9128	0.9667	0.9468	0.9905
MeanMIL	0.9406	0.9822	0.9360	0.9707	0.9505	0.9878
ABMIL	0.9548	0.9895	0.9341	0.9789	0.9560	0.9928
CLAM-SB	0.9444	0.9889	0.9380	0.9763	0.9596	0.9918
CLAM-MB	0.9576	0.9910	0.9264	0.9731	0.9560	0.9934
TransMIL	0.9510	0.9886	0.9205	0.9705	0.9560	0.9871
MambaMIL	0.9453	0.9889	0.9322	0.9787	0.9486	0.9899
SAE+MaxMIL	0.9331	0.9812	0.9302	0.9724	0.9394	0.9784
SAE-1N+MaxMIL	0.8756	0.9351	0.8430	0.9020	0.8037	0.8673

Supplementary Methods

BatchTopK Sparse Autoencoder (BatchTopK SAE) Settings

We trained a BatchTopK sparse autoencoder (BatchTopKSAE) on patch embeddings derived from each encoder in order to disentangle polysemantic features into monosemantic, human-interpretable latent concepts. The architecture and training procedure were adapted based on the output dimensionality of each encoder. Hyperparameters were optimized to minimize reconstruction loss while ensuring low dead unit rates—i.e., latent neurons that remained consistently inactive during training.

Encoder-specific Configurations

We systematically evaluated the following parameter ranges:

- Sparsity levels: k ∈ {8, 16, 32}
- Expansion factors: {2, 4, 8}

Supplementary Table 1 summarizes the optimal configuration identified for each encoder.

Supplementary Table 3: Sparse autoencoder configurations for each encoder

Encoder	Input Dim	Hidden Dim	Expansion	K (Sparsity)	Avg. Density (%)
ResNet-50	1024	2048	2	32	1.56
CONCH	512	2048	4	16	0.78
UNI	1024	4096	4	32	0.78