This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material: AUC Is Not Enough

Mirza Nasir Hossain, Sarah L. Bell, Gareth Bryson, and David Harris-Birtill

This supplement documents the corrected, matched-Z experiment used throughout the main manuscript. All empirical tables derive from the recorded runs listed here; no new runs were performed solely to create this supplement.

S1.1 Preprocessing, Matched-Z, and Panel Sensitivity

For raw UNI embeddings $X_{\mathrm{tr}}, X_{\mathrm{te}}$, per-coordinate statistics are estimated on training data only:

$$(\mu, \sigma) = \mathrm{stats}(X_{\mathrm{tr}}), \quad \widetilde{X}_s = (X_s - \mu)/\sigma, \quad Z_s = C(\widetilde{X}_s), \quad s \in \{\mathrm{tr}, \mathrm{te}\}.$$

The compressor C is fitted or trained without test data. Logistic regression is fitted on Z_{tr} and evaluated on Z_{te}; the evidence analysis reconstructs the identical Z_{te} and applies the unchanged TopK SAE. The historical expert panels used test-split statistics for test embeddings, whereas the primary analysis uses train-fitted statistics throughout. Differences between references are at most 0.0033 and do not alter the result.

Table 1. Sensitivity to corrected versus historical reference panels. AUC always uses the corrected matched-Z protocol.

		Top-12 overlap		Retained@100	
PCA	AUC ratio	Corrected	Historical	Corrected	Historical
5	0.9921	0.0325	0.0308	0.1992	0.1958
9	1.0003	0.1008	0.0975	0.3842	0.3850
48	1.0041	0.5042	0.5008	0.9592	0.9600

S1.2 Compression, Reconstruction, and Classifier Configuration

PCA is centered, train-fitted, and inverted with its learned inverse. GRP uses Gaussian projections (seeds 42–46) and a Moore–Penrose pseudoinverse. UMAP-AE uses encoder widths $1024 \to 512 \to 256 \to k$ and a symmetric decoder, with LayerNorm and GELU. Its objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{MSE}} + 0.25\mathcal{L}_{\mathrm{cosine}} + 0.05(\mathcal{L}_{\mathrm{positive}} + \mathcal{L}_{\mathrm{negative}}).$$

UMAP-AE uses a 15-nearest-neighbor graph, 20 AdamW epochs (10^{-3} learning rate, 10^{-5} weight decay), 20,000 training rows, 12,000 graph rows, negative-sample rate 2, and seeds 42–44.

The dimension grid was 1–50, then 64, 96, 128, 192, 256, 384, 512, 768, and 1024. AUC used cuML logistic regression after scaling fitted on Z_{tr}, macro one-vs-rest ROC-AUC, dimension-indexed L_2 settings, and at most 10,000 iterations, without SPIDER test-label calibration. Evidence metrics used the fixed tied-weight TopK checkpoint (250 active features per patch) and the 100 pre-existing reviewed features.

S1.3 Metrics, Thresholds, and Sanity Checks

For feature j, $a_j, a'_j \in \mathbb{R}^{13,193}$ are original and reconstructed activation vectors, $T_j^q, T_j'^q$ are their top-q patch sets, w_j is the tied SAE direction, and R is a linear round trip:

$$d_j = \cos(w_j, Rw_j), \qquad r_j = \mathrm{corr}(a_j, a'_j),$$
$$o_j = |T_j^{12} \cap T_j'^{12}|/12, \quad q_j = |T_j^{12} \cap T_j'^{100}|/12.$$

Direction cosine is undefined for nonlinear UMAP-AE. A reviewed feature is alive when TopK-active on at least 50 test patches; dead/awakened rates use all 100 reviewed features. AUPRC-profile stability is Pearson correlation between the original and reconstructed 13-class one-vs-rest AUPRC vectors. Dominant-class stability preserves their argmax; the margin-aware result uses the 82 features with original top-vs-second margin above 0.05.

Thresholds are AUC ≥ 0.99, activation correlation ≥ 0.90, overlap ≥ 0.50, retained@100 ≥ 0.90, AUPRC-profile correlation ≥ 0.99, and dominant-class stability ≥ 0.90. Composite feature recovery requires the four continuous feature criteria (activation, overlap, retained@100, AUPRC) simultaneously; unreached criteria remain NR. Random expectations are $12/13{,}193 = 0.000910$ for strict overlap and $100/13{,}193 = 0.007580$ for retained@100. The direct 1024-D AUC is 0.990902. Full-rank PCA returns overlap, retained@100, and activation correlation of 1.000000, confirming identity of the evaluation path when no dimension is removed.

S1.4 Complete Threshold Crossings and Sensitivity

Table 2 gives all crossings behind the main threshold figure. GRP and UMAP-AE are means over five and three seeds; NR means the threshold was not reached through 1024 dimensions.

Table 2. First dimension crossing each threshold. Parentheses show the seedwise min–max when applicable.

Metric	PCA	GRP (5 seeds)	UMAP-AE (3 seeds)
AUC recovery ≥ 0.99	4	37.4 (31–40)	2.3 (2–3)
Activation correlation ≥ 0.90	64	768	NR
Top-12 overlap ≥ 0.50	47	409.6 (384–512)	NR
Retained@100 ≥ 0.90	34	307.2 (256–384)	78.3 (43–128)
AUPRC profile ≥ 0.99	50	972.8 (768–1024)	NR
Dominant class ≥ 0.90	26	460.8 (384–512)	6.0 (5–7)

The PCA sensitivity grids vary utility and evidence thresholds independently; each cell reports AUC crossing / evidence crossing.

Table 3. PCA threshold sensitivity. Left: strict top-12 overlap; right: retained@100. Cells give AUC dimension / evidence dimension.

Strict overlap					Retained@100				
Evidence \ AUC	.95	.97	.99	.999	Evidence \ AUC	.95	.97	.99	.999
.25	2/20	3/20	4/20	8/20	.50	2/12	3/12	4/12	8/12
.33	2/27	3/27	4/27	8/27	.75	2/21	3/21	4/21	8/21
.50	2/47	3/47	4/47	8/47	.90	2/34	3/34	4/34	8/34
.67	2/128	3/128	4/128	8/128	.95	2/43	3/43	4/43	8/43

The strict utility-before-evidence ordering persists throughout. The lenient corner (AUC ≥ 0.999, retained@100 ≥ 0.50) narrows to 8/12, but only requires six original patches to appear anywhere among the reconstructed top 100.

S1.5 SAE Fine-Tuning on PCA-47 Reconstructions

To distinguish irrecoverable compression loss from a fixed SAE merely being misaligned with reconstructed inputs, the SAE was fine-tuned on corrected PCA-47 training reconstructions. Both arms used AdamW without weight decay and updated all SAE parameters. The conservative arm used one epoch at learning rate 10^{-5}; the strong arm used three epochs at 10^{-4}. Final-epoch checkpoints were evaluated on PCA-47 test reconstructions.

Table 4. PCA-47 SAE fine-tuning sanity arm. Values are means across the 100 reviewed features, except reconstruction loss.

Condition	Recon. loss	Activation corr.	Top-12 overlap	Retained@100
Unchanged SAE	5.142×10^{-5}	0.884	0.502	0.953
1 epoch, 10^{-5}	4.074×10^{-5}	0.881	0.497	0.951
3 epochs, 10^{-4}	1.817×10^{-6}	0.769	0.403	0.878

Lower reconstruction loss did not improve reviewed-feature preservation. The stronger update also rotated feature directions relative to the original SAE (mean direction cosine 0.873), so its lower loss does not recover the same reviewed feature identities.

S1.6 Exploratory Feature-Level Heterogeneity

Composite recovery has median dimension 64, with 43/100 reviewed features recovering by PCA-47. The source-study mono and poly groups have medians 39 and 192, respectively. This association is descriptive and partly metric-coupled: the groups are AUPRC-defined and the composite endpoint includes AUPRC-profile stability. The full six-page supplement, heterogeneity figure, optimization diagnostics, and machine-readable receipts remain available with the code.