Paired-Acquisition Neural Factorization for Computational Pathology

Matthew Vaishnav
Independent Researcher, Kitchener-Waterloo, Canada
https://github.com/matthewvaishnav/computational-pathology-research

Abstract

I trained a neural factorization model to separate biological tissue signal from scanner and acquisition signal using biology-preserving paired observations. The principal experiment uses 2,400 real histopathology images from 480 aligned regions on 48 human H&E slides, each scanned on five devices. A frozen Paired-Acquisition Neural Factorization objective reduced biological scanner-probe accuracy from 0.7825 to 0.3989 on DINOv2 features while improving paired agreement and preserving near-perfect cross-scanner tissue retrieval. Without backbone-specific tuning, the same objective transferred to pathology-native Phikon and ImageNet ResNet50. Across the two unseen feature families, the slide-blocked scanner-probe difference was -0.4019 with 95% bootstrap interval $[-0.4145, -0.3895]$, favorable on all 48 slides. Scanner identity remained strongly recoverable from a compact acquisition branch, indicating factor separation rather than representational erasure.

The method follows from two earlier failures. A locked adversarial study reduced site leakage but did not improve held-out tumor accuracy, and unconditional removal of a learned site component erased tumor-predictive information. Controlled latent-factor experiments then showed that ordinary observations and paired acquisitions are distinct resources. Under matched compute, doubling paired exposure improved recovery by $+0.037357$, while increasing unique anchors from 50 to 100 helped substantially more than increasing them from 100 to 200. Downstream PANDA, CAMELYON17, and PCam studies test whole-slide aggregation, institutional influence, external-center generalization, and systems constraints. The evidence supports a narrow conclusion: acquisition-associated information must be separated conditionally, with explicit protection of biology, rather than removed merely because it predicts site.

1 Introduction

Digital pathology representations contain more than morphology. Scanner optics, color response, compression, staining, preparation, local workflow, and institutional policy can all be encoded alongside tumor-relevant tissue structure. A representation may retrieve the correct specimen and support a downstream classifier while still making scanner or center identity easy to recover. Conversely, a method that suppresses every site-predictive direction can delete useful biology.

The central problem is therefore not ordinary domain invariance. It is *conditional factorization*: acquisition-sensitive variation should become less recoverable from the biological representation, but tissue and diagnostic information must remain. This is difficult from ordinary observations alone because site and biology may be correlated. If a site-associated direction predicts tumor status, observational data cannot determine whether that direction is nuisance, biology, or an entangled mixture.

Biology-preserving acquisition pairs change the information structure. The same tissue region is observed under different scanners or acquisition states. What changes across a pair is candidate acquisition variation; what remains shared is candidate biology. Paired-Acquisition Neural Factorization uses this intervention structure to learn separate biological and acquisition branches and evaluates both what is removed and where it goes.

The specific contributions of this paper are as follows. First, I provide a locked negative-result sequence showing that lower site leakage alone does not imply better tumor prediction and that unconditional subtraction can erase biology. Second, I introduce a paired-acquisition factorization objective with explicit biological retention, acquisition prediction, reconstruction, variance control, and cross-factor decorrelation. Third, I evaluate the frozen objective on real paired scanner data with original-slide blocking and transfer it without retuning across three substantially different feature families. Fourth, I measure the separate roles of ordinary observations, unique paired anchors, and repeated paired exposure in controlled latent-factor experiments. Finally, I connect representation failure to later whole-slide and federated studies without using those studies to inflate the central paired-scanner claim. Focused companion PDFs carry exact protocols, frozen result tables, and reproduction scripts for study-specific claims.

2 Linked Study Packages

The public PDF is the research-program overview, not the complete protocol archive for every bounded claim. Study-specific repositories carry the exact preprocessing boundary, frozen split design, result tables, and reproduction scripts. I use logical method names in the text and link verified child-package repositories and PDFs from this table.

Table 1: Focused study packages linked from the main research PDF.

Package	What it contains	Repository status
Paired-Acquisition Neural Factorization SCORPION core study	Primary paired-scanner method study on 48 original human H&E slides, 480 aligned tissue regions, five scanners, DINOv2/Phikon/ResNet50 transfer, and pair-integrity falsification controls.	repository; PDF
Paired-Acquisition Neural Factorization external canine SCC validation	Independent five-scanner canine SCC paired-acquisition validation on 805 geometry-qualified five-view regions from 44 biological samples. The locked projection reduced sample-blocked scanner-probe accuracy by 0.380 absolute while preserving same-region retrieval.	repository; PDF
Paired-Acquisition Neural Factorization allocation study	Resource-allocation study testing biological pair diversity versus anchor repetition under matched pair-presentation budgets. The strongest allocation contrast was 200 versus 50 pairs at budget 6,400, with positive budget-doubling effects across allocations.	repository; PDF

Verified study-package repositories and PDFs are linked from the public paper once the public URLs are live.

3 The Data

3.1 SCORPION paired-scanner tissue

The primary real-data study uses the public SCORPION benchmark [19]. It contains 48 original human H&E slides, ten aligned tissue regions per slide, and one image from each of five scanners for every region: AT2, GT450, DP200, P1000, and B300. The resulting dataset contains 480 biological regions and 2,400 images.

The original slide is the statistical unit. All regions and all scanner views from one slide remain in the same fold. Optimization seeds are averaged within slide before bootstrap and sign-flip inference. Patches, scanner views, regions, and repeated backbones are not treated as independent biological samples.

3.2 Controlled latent-factor data

Synthetic experiments generate known biological and acquisition factors, then mix them under controlled overlap and linear or nonlinear observation functions. Because the true factors are available, recovery can be measured using canonical correlation, Procrustes alignment, cross-factor leakage, paired invariance, counterfactual transport, and task retention. These experiments do not imitate the full complexity of histopathology; they isolate when the proposed supervision is sufficient or insufficient.

3.3 Downstream pathology datasets

PANDA supplies 10,611 verified slide-level Phikon feature bags for prostate grading and federated stress experiments [2]. CAMELYON17/WILDS supplies 455,954 examples from five centers for natural external-center validation [1, 9]. PCam supplies a public patch-classification benchmark used for full-test, bootstrap, threshold, and simulated-site validation. These datasets test later levels of the learning pipeline; they are not substitutes for the paired-acquisition experiment.

4 The Paired-Acquisition Neural Factorization Model

4.1 Paired acquisition supervision

Let $x_{r,s} \in \mathbb{R}^D$ be the frozen feature for tissue region r acquired by scanner s. A biology-preserving set contains several views of the same region:

$$\mathcal{V}_r = \{x_{r,s_1}, \dots, x_{r,s_m}\}.$$

The biological branch should be stable across members of $\mathcal{V}_r$, whereas the acquisition branch should retain information that distinguishes scanner states.

Under a simple latent model,

$$x_{r,s} = Bz_{b,r} + Az_{a,s} + \varepsilon_{r,s},$$

a paired difference cancels the shared biological term:

$$x_{r,s'} - x_{r,s} = A(z_{a,s'} - z_{a,s}) + (\varepsilon_{r,s'} - \varepsilon_{r,s}).$$

The neural model does not assume the observation map is exactly linear, but the cancellation argument explains why paired observations contain information that ordinary site labels do not.

4.2 Architecture

For a standardized frozen feature $x \in \mathbb{R}^D$, Paired-Acquisition Neural Factorization uses two factor encoders, a joint decoder, and two scanner heads:

$$z_b = f_b(x) \in \mathbb{R}^{d_b}, \qquad d_b = 256, \tag{1}$$

$$z_a = f_a(x) \in \mathbb{R}^{d_a}, \qquad d_a = 64, \tag{2}$$

$$\widehat{x} = g([z_b; z_a]) \in \mathbb{R}^D, \tag{3}$$

$$\widehat{y}_b = q_b(\mathrm{GRL}_\gamma(z_b)), \tag{4}$$

$$\widehat{y}_a = q_a(z_a), \qquad \gamma = 1. \tag{5}$$

Both encoders are MLPs with hidden width 512:

$$h_c(x) = \mathrm{LN}\{\mathrm{GELU}(W_{c,1}x + b_{c,1})\}, \tag{6}$$

$$f_c(x) = W_{c,2}h_c(x) + b_{c,2}, \qquad c \in \{b, a\}. \tag{7}$$

The decoder has widths $(d_b + d_a) \to 512 \to D$. The biological scanner head has widths $256 \to 128 \to C$, the acquisition scanner head has widths $64 \to 64 \to C$, and $C = 5$ scanners. Gradient reversal is the identity in the forward pass and negates the encoder gradient:

$$\mathrm{GRL}_\gamma(z) = z, \qquad \frac{\partial\, \mathrm{GRL}_\gamma}{\partial z} = -\gamma I. \tag{8}$$

Thus q_b predicts scanner identity while f_b receives the opposite gradient; q_a and f_a minimize ordinary scanner cross-entropy. The architecture is therefore the factorization

$$x \mapsto (z_b, z_a) = (f_b(x), f_a(x)), \qquad \widehat{x} = g([z_b; z_a]), \quad (9)$$

where z_b is optimized for same-region agreement and scanner suppression, and z_a is optimized for scanner prediction and acquisition retention.

4.3 Objective

Let a minibatch be $\{(x_i, r_i, s_i)\}_{i=1}^n$, where r_i is tissue-region identity and $s_i \in \{1, \dots, C\}$ is scanner identity. Write $Z_b \in \mathbb{R}^{n \times d_b}$ and $Z_a \in \mathbb{R}^{n \times d_a}$ for the stacked factors, and let $u_i = z_{b,i}/\|z_{b,i}\|_2$. For anchor i, define $P(i) = \{p \neq i : r_p = r_i\}$ and

$$S_i = \sum_{j \neq i} \exp(u_i^\top u_j/\tau), \quad (10)$$

$$\ell_i = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(u_i^\top u_p/\tau)}{S_i}, \quad (11)$$

$$\mathcal{L}_{\text{pair}} = \frac{1}{n} \sum_{i=1}^n \ell_i, \qquad \tau = 0.1. \quad (12)$$

The reconstruction loss is the mean squared error

$$\mathcal{L}_{\text{recon}} = \frac{1}{nD} \|\widehat{X} - X\|_F^2. \quad (13)$$

For $Z \in \mathbb{R}^{n \times d}$, define

$$v_j(Z) = \frac{1}{n} \sum_{i=1}^n (Z_{ij} - \bar{Z}_j)^2, \quad (14)$$

$$\ell_{\text{var}}(Z) = \frac{1}{d} \sum_{j=1}^d \left[1 - \sqrt{v_j(Z) + \varepsilon}\right]_+, \qquad \varepsilon = 10^{-4}, \quad (15)$$

$$\mathcal{L}_{\text{var},b} = \ell_{\text{var}}(Z_b), \qquad \mathcal{L}_{\text{var},a} = \ell_{\text{var}}(Z_a). \quad (16)$$

Let $Z_b^c = Z_b - \mathbf{1}\bar{z}_b^\top$ and

$$C_b = \frac{1}{n-1} (Z_b^c)^\top Z_b^c. \quad (17)$$

The within-biological covariance penalty is

$$\mathcal{L}_{\text{cov},b} = \frac{1}{d_b} \left\| C_b - \text{diag}(\text{diag}\, C_b) \right\|_F^2. \quad (18)$$

The scanner losses are forward cross-entropies:

$$\mathcal{L}_{\text{scan},b} = -\frac{1}{n} \sum_i \log[\text{softmax}\{q_b(\text{GRL}_\gamma(z_{b,i}))\}]_{s_i}, \quad (19)$$

$$\mathcal{L}_{\text{scan},a} = -\frac{1}{n} \sum_i \log[\text{softmax}\{q_a(z_{a,i})\}]_{s_i}. \quad (20)$$

Only the gradient entering f_b is reversed.
Let $Y_s \in \mathbb{R}^{n \times C}$ be the one-hot scanner matrix. Column-standardize Z_b and Y_s using the same $\varepsilon = 10^{-4}$ variance floor, producing $\widetilde{Z}_b$ and $\widetilde{Y}_s$. The direct scanner-dependence penalty is

$$\mathcal{L}_{\text{dep}} = \frac{1}{d_b C} \left\| \frac{1}{n-1} \widetilde{Z}_b^\top \widetilde{Y}_s \right\|_F^2. \quad (21)$$

For $Z_a^c = Z_a - \mathbf{1}\bar{z}_a^\top$, the cross-factor penalty is

$$\mathcal{L}_{\text{xcov}} = \frac{1}{d_b d_a} \left\| \frac{1}{n-1} (Z_b^c)^\top Z_a^c \right\|_F^2. \quad (22)$$

The frozen Paired-Acquisition Neural Factorization objective used in the five-fold and cross-backbone experiments is

$$\mathcal{L} = \mathcal{L}_{\text{pair}} + \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{var},b} + 0.25\mathcal{L}_{\text{var},a} \quad (23)$$

$$+ 0.01\mathcal{L}_{\text{cov},b} + 0.5\mathcal{L}_{\text{scan},b} + 0.5\mathcal{L}_{\text{scan},a} \quad (24)$$

$$+ 20\mathcal{L}_{\text{dep}} + 0.05\mathcal{L}_{\text{xcov}}. \quad (25)$$

The paired-consistency baseline is

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{pair}} + \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{var},b} + 0.01\mathcal{L}_{\text{cov},b}. \quad (26)$$

It shares the main retention and collapse controls but has no acquisition branch, scanner classifiers, direct scanner-dependence penalty, or cross-factor penalty.

5 Preventing Representation Failure

5.1 Why site invariance was not enough

The first Paired-Acquisition Neural Factorization study used a shared feature encoder, tumor head, and gradient-reversal site head. A locked confirmation reached approximately 0.9264 held-out tumor accuracy versus 0.9270 for the tumor-only baseline. The paired accuracy difference was approximately -0.000625 with an interval crossing zero. Site leakage decreased slightly, but the result was not a diagnostic-performance improvement.
A later factorized model separated diagnostic and site branches with reconstruction and covariance control. Validation selected the diagnostic-only baseline. An unconditional subtraction experiment was more decisive: the learned site-associated component remained strongly tumor-predictive, so subtracting it deleted useful biological signal. These failures define the design requirement for Paired-Acquisition Neural Factorization: removed information must be accounted for, and biology must be protected directly.

5.2 Collapse and factor-retention controls

A low scanner-probe score is meaningless if the biological branch is constant. Paired-Acquisition Neural Factorization therefore reports effective rank, per-dimension variance, tissue-region retrieval, paired cosine similarity, reconstruction, acquisition-branch scanner prediction, and acquisition-branch tissue retrieval. The desired pattern is not zero information everywhere. It is high tissue retention and reduced scanner recovery in biology, paired with high scanner recovery and substantially lower tissue retrieval in acquisition.

5.3 Ordinary observations and paired resources

The controlled program treats three quantities separately: ordinary observations N_u, unique paired anchors N_p, and paired repetitions R_p. Ordinary observations learn the task manifold. Unique anchors expose more biological locations under acquisition changes. Repetition supplies additional optimization signal for already observed interventions.
The pair-repeat experiments hold total paired presentations fixed while changing the allocation between N_p and R_p. This distinguishes a true anchor-diversity effect from the simpler explanation that a model merely received more pair-loss updates.

Table 2: Frozen representation audit before Paired-Acquisition Neural Factorization training.

Encoder	Cosine	Retrieval	Scanner
ResNet50	0.8849	0.9550	0.7560
DINOv2-B	0.9851	0.9995	0.8400
Phikon	0.8787	0.9825	0.9640

6 Details of Learning

The frozen DINOv2 objective was selected through bounded calibration on development data, followed by independent-seed confirmation and final five-fold evaluation. The direct dependence weight of 20 was fixed before the final test stage. Cross-backbone transfer to Phikon and ResNet50 was preregistered before inspecting those outcomes. The transfer experiments used new shared seeds 701–705, the same rotating slide folds, and no backbone-specific objective or schedule changes.

Each SCORPION model used 75 epochs, region batch size 32, AdamW learning rate 3×10^{-4}, and weight decay 10^{-4}. Every contrastive minibatch included all five scanner views of selected tissue regions. Inference used slide-block bootstrap intervals and sign-flip tests after averaging optimization seeds within each slide.

The pair-repeat synthetic study used overlap 0.975, nonlinear mixing, 320 optimizer steps, observational batch size 128, paired batch size 64, $N_u \in \{750, 1500\}$, pair-consistency and hybrid-curriculum methods, and ten matched seeds. The two paired-presentation budgets were 6,400 and 12,800. Within each budget, the tested allocations were 50, 100, and 200 unique anchors with compensating repetition counts.

7 Results

7.1 Frozen paired-scanner factorization

Before Paired-Acquisition Neural Factorization training, scanner identity was strongly recoverable from all three feature families despite high tissue retention. The pathology-native Phikon representation had the highest baseline scanner-probe score.

On DINOv2, the frozen five-fold comparison reduced scanner-probe accuracy from 0.7825 to 0.3989, increased mean paired cosine from 0.8476 to 0.8789, and preserved mean tissue retrieval at approximately 0.9998.

Both unseen backbones passed all seven preregistered criteria without retuning. Averaging Phikon and ResNet50 method differences within the same 48 slides gave scanner-probe difference -0.4019 with interval $[-0.4145, -0.3895]$, mean paired-cosine difference $+0.0578$ with interval $[0.0539, 0.0615]$, and worst paired-cosine difference $+0.0628$ with interval $[0.0585, 0.0672]$.

The acquisition branch retained scanner-probe accuracies of 0.8602, 0.9711, and 0.7845 for DINOv2, Phikon, and ResNet50, respectively. Its tissue retrieval was much lower than the biological branch: 0.1042, 0.0739, and 0.1705. Every biological dimension retained nonzero test-fold variance. The result is therefore consistent with separation, not a constant representation or simple information deletion.

Pair-integrity falsification further separated scanner suppression from tissue preservation. Broken-pair controls reduced scanner-probe accuracy similarly or more than true same-region pairs, but degraded paired-tissue consistency and same-region retrieval (Tables 4 and 5). This indicates that the useful scanner-suppression/tissue-preservation tradeoff depends on true same-tissue pairing, rather than scanner suppression alone.

The same falsification pattern reproduced on the unseen Phikon and ResNet50 feature families. True same-tissue pairs preserved tissue identity substantially better than shuffled-pair controls in both backbones, even when shuffled controls achieved comparable or stronger scanner-probe suppression.

7.2 External paired-scanner validation package

A companion external validation study applies the locked Paired-Acquisition Neural Factorization protocol to a public five-scanner canine squamous-cell-carcinoma dataset after geometry qualification. On 805 complete five-view regions from 44 biological samples, Paired-Acquisition Neural Factorization reduced sample-blocked scanner-probe accuracy from 0.7529 to 0.3614. The scanner-probe contrast was -0.3803 with bootstrap interval $[-0.3996, -0.3611]$, favorable in all 44 sample blocks, while same-region retrieval remained non-inferior. A later representation audit tested the branch-separation claim directly across folds 0–4 and seeds 911–915: scanner prediction fell from 0.878 in raw DINOv2 to 0.300 in the Paired-Acquisition Neural Factorization biological branch while same-region retrieval rose from 0.873 to 0.959; the acquisition branch retained scanner prediction at 0.968 while same-region retrieval fell to 0.031. Stronger linear, random-forest, kNN, and nonlinear MLP probes reproduced the same separation pattern. The exact dataset audit, P1000 orientation-normalization boundary, frozen encoder triage, five-fold result tables, and reproduction scripts are assigned to the companion repository and PDF.

External pair-integrity falsification repeated the same control on canine SCC. True same-region pairs preserved tissue identity better than both shuffled controls (Table 6). Region-shuffled pairs reduced scanner-probe accuracy more than true pairs, but degraded paired cosine and same-region retrieval. This separates scanner suppression from useful tissue preservation outside the primary SCORPION benchmark.

7.3 Simple scanner-removal baseline controls

A peer-review objection is that ordinary post-hoc operations on frozen embeddings might achieve comparable scanner suppression without paired-acquisition training. To test this, linear scanner-subspace projection (top-k logistic scanner-discriminative directions removed after fold-fit standardization) and PCA component removal (top-k PCA directions removed) were evaluated on frozen DINOv2 features for both SCORPION and external canine SCC across the same five folds and five seeds.

Linear scanner-subspace projection preserved tissue structure (cosine 0.881 SCORPION, 0.739 canine) but saturated at the learned scanner-subspace rank ($k \geq 4$), leaving scanner probe at 0.724 and 0.707—far above the neural references of 0.399 and 0.361. PCA component removal suppressed scanner further (0.560 SCORPION, 0.641 canine) but caused substantial tissue damage: cosine fell from 0.987 to 0.806 on

Table 3: Paired-Acquisition Neural Factorization minus paired-consistency slide-blocked contrasts. Positive cosine and retrieval values and negative scanner-probe values are favorable.

Backbone	Scanner probe	Mean cosine	Worst cosine	Mean retrieval	Worst retrieval
DINOv2-Base	-0.3854	$+0.0310$	$+0.0311$	-0.00008	-0.00042
Phikon	-0.4353	$+0.0900$	$+0.0963$	$+0.00060$	$+0.00479$
ResNet50	-0.3685	$+0.0255$	$+0.0294$	$+0.00808$	$+0.01688$

Table 4: SCORPION DINOv2 pair-integrity falsification across five folds and five seeds. Lower scanner probe is not sufficient: higher cosine and retrieval indicate retained tissue identity.

Condition	Scanner	Mean cos.	Worst cos.	Mean ret.	Worst ret.
True pairs	0.3998	0.8796	0.8502	0.9999	0.9991
Region-shuffled	0.3736	0.8089	0.7632	0.9954	0.9839
Sample-shuffled	0.3588	0.7668	0.7168	0.9792	0.9494

SCORPION and from 0.919 to 0.598 on canine SCC. No simple baseline came within 0.03 scanner probe of neural factorization on either dataset.

The same pattern—linear projection preserves tissue but undersuppresses scanner; PCA suppresses scanner more but damages tissue; neural factorization achieves the best tradeoff—held across both the primary SCORPION benchmark and an external canine SCC dataset with different scanners and tissue types. This cross-dataset replication separates the baseline objection from one-dataset artifacts.

7.4 Resource allocation under matched compute

At 6,400 paired presentations, the seed-blocked 200-versus-50 anchor contrast was $+0.017865$, with 95% interval $[0.010560, 0.026079]$, positive in all ten seed blocks, and exact two-sided sign-flip $p = 0.001953$. The 200-versus-100 contrast was only $+0.001578$. At 12,800 presentations, the corresponding contrasts were $+0.013050$ and $+0.002422$. Doubling paired exposure increased recovery by $+0.037357$ overall, with interval $[0.030315, 0.043657]$, positive in all ten seed blocks, and exact $p = 0.001953$. This establishes separate exposure and allocation effects. It does not establish a universal optimum or a recovered phase boundary.

7.5 Whole-slide modeling on PANDA

The whole-slide studies compare simple and learned aggregation under the same broad feature-bag setting. Mean-pooled Phikon plus an MLP reached validation QWK 0.7274. Gated AttentionMIL reached 0.8100. Three listed tuned TransnnMIL runs reached 0.8155, 0.8225, and 0.8086. An 18-run grid across six learning rates and three seeds reached a best mean validation QWK of 0.8257 ± 0.0169 at learning rate 10^{-4}.

These results show a large improvement from mean pooling to attention-based MIL and establish a stable TransnnMIL training recipe. They do not yet prove architectural superiority over AttentionMIL, TransMIL, or nnMIL under fully matched compute.

7.6 Federated site-signal auditing

In the full-PANDA label-noise study, a tuned observable detector-switch policy kept clean performance effectively unchanged while improving global QWK by $+0.008009$ at 25% dominant-site corruption and $+0.007482$ at 35% corruption. A stronger transfer experiment froze a detector calibrated under label noise and applied it without retuning to systematic conservative ordinal threshold shift.

At 35% and 45% shift, the bootstrap intervals for all three headline metrics were positive. Removing the most frequent diagnostic family reduced trigger rate but retained positive 35/45-shift means. In a 36-setting calibration-neighborhood sweep, 29 settings preserved low clean switching and positive 35%/45% gains across global QWK, macro-F1, and worst-site QWK.

CAMELYON17 then supplied natural center shift. With frozen ImageNet features, held-out-center accuracy was 0.8312 for FedAvg-style sample weighting, 0.9132 for equal-client weighting, and 0.9094 for dominant-source downweighting. With source-trained CAMELYON17 features, the corresponding accuracies were 0.9052, 0.9318, and 0.9322. FedAvg-style source-training accuracy reached 0.9991 in the latter study, showing that stronger source fit did not imply stronger unseen-center performance.

7.7 Patch and systems studies

On the full 32,768-example PCam public test set, the patch classifier reached 0.8526 accuracy and 0.9394 AUC; uncertainty was evaluated with 1,000 bootstrap resamples. Threshold analysis reduced missed tumor predictions by 61.7% relative to the documented default operating point. Simulated-site PCam studies verified that equal, volume-based, and contribution-aware weighting implementations changed institutional influence on real image data, although patch-level performance was comparatively insensitive to those weight changes.

For the CAMELYON17 feature/head regime, 100 fp32 rounds across three clients were estimated at 24.98 GB for a full ResNet18 exchange versus 2.35 MB for a 512-to-2 head, a ratio of approximately 10,894. A coefficient-noise probe preserved the equal-client and downweighted advantages through noise standard deviation 0.20. A separate infrastructure simulation measured heterogeneous update times, dropout, failed rounds, and synchronous straggler burden. These are bounded systems studies, not formal privacy or hospital-deployment validation.

8 Discussion

The main result is not merely that a scanner classifier became worse. Scanner information remained strongly available in the acquisition branch, tissue identity remained almost saturated in biology, and the frozen objective transferred without retuning across a general self-supervised encoder, a pathology-native encoder, and a conventional ImageNet encoder. Cross-backbone pair-integrity falsification further

Table 5: SCORPION cross-backbone pair-integrity falsification. Lower scanner probe alone is not sufficient; higher cosine and retrieval indicate retained tissue identity.

Backbone	Condition	Scanner	Mean cos.	Worst cos.	Mean ret.	Worst ret.
Phikon	True pairs	0.5200	0.8645	0.8300	0.9997	0.9974
Phikon	Region-shuffled	0.4341	0.6913	0.6022	0.9603	0.8992
Phikon	Sample-shuffled	0.4085	0.6158	0.4741	0.8763	0.6757
ResNet50	True pairs	0.3145	0.6544	0.5978	0.9726	0.9455
ResNet50	Region-shuffled	0.3045	0.5188	0.4487	0.8839	0.7980
ResNet50	Sample-shuffled	0.3067	0.5083	0.4439	0.8775	0.8003

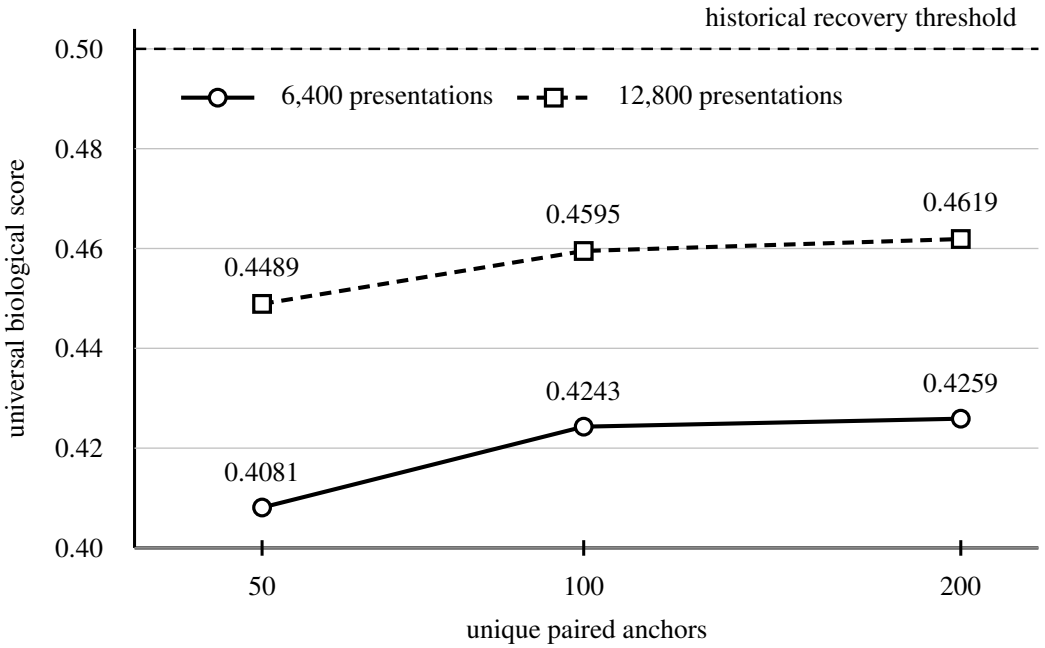

Figure 1: Pair-repeat allocation of unique paired anchors and repeated presentations. More total exposure produced the largest gain. Increasing unique-anchor diversity helped, with diminishing returns beyond approximately 100 anchors in the tested regime. No allocation crossed the historical 0.50 recovery threshold.

Table 6: External canine SCC DINOv2 pair-integrity falsification across five sample-blocked folds and five seeds. Lower scanner probe alone is not sufficient; higher cosine and retrieval indicate retained tissue identity.

Condition	Scanner	Mean cos.	Worst cos.	Mean ret.	Worst ret.
True pairs	0.3614	0.7300	0.6567	0.9334	0.8844
Region-shuffled	0.3057	0.5422	0.4211	0.7293	0.5158
Sample-shuffled	0.4093	0.5849	0.4971	0.7183	0.5650

showed that broken-pair controls can suppress scanner probes while damaging tissue preservation on Phikon and ResNet50. Simple scanner-removal baselines—linear scanner-subspace projection and PCA component removal on frozen features—could not reproduce the neural tradeoff on SCORPION or external canine SCC. This pattern is difficult to explain as one-backbone overfitting, wholesale representational collapse, or scanner suppression alone.

The external canine SCC audit strengthens the same interpretation outside the primary human SCORPION benchmark. Across five folds, five seeds, and stronger probe families, the biological branch remained weakly scanner-decodable while retaining biological identity, and the acquisition branch remained strongly scanner-decodable while losing biological retrieval behavior. This is representation-identifiability evidence, not diagnostic validation, but it directly tests the central separation claim.

The negative studies are equally important. Gradient reversal changed leakage without improving tumor prediction. A factorized model did not beat its diagnostic-only baseline. Unconditional subtraction removed tumor-predictive information. Together, these findings reject a common shortcut: site predictability is not sufficient evidence that a direction is disposable.

The controlled resource experiments add a second result. Unique paired specimens and repeated pair presentations are not interchangeable, but neither is pair diversity the only resource. Over the tested range, doubling total exposure produced the largest improvement, moving from 50 to 100 anchors produced a meaningful additional gain, and moving from 100 to 200 produced only a small gain. This suggests that future real paired-acquisition collections should report both unique biological anchors and total paired exposure rather than calling both simply “more paired data.”

The later PANDA, CAMELYON17, and PCam studies expose related authority problems at different scales. A pooling rule decides which patches dominate a slide. FedAvg decides that sample count controls institutional influence. Source-training accuracy can become nearly perfect while external-center accuracy remains worse. These studies do not validate one complete clinical system, but they show why representation,

Table 7: Paired-Acquisition Neural Factorization separates biological identity from acquisition identity. On an external multi-scanner canine SCC benchmark, raw DINOv2 features encode both scanner and biological identity. Paired-Acquisition Neural Factorization produces a biological branch with low scanner decodability and high biological recoverability, and an acquisition branch with high scanner decodability and low biological recoverability. Projected representations are summarized over folds 0–4 and seeds 911–915.

Representation	Runs	Scanner probe ↓	Region R@1 ↑	Sample R@1 ↑	Eff. rank	Zero-var
Raw DINOv2	1	0.878	0.873	0.986	442.802	0.000
Paired reference	25	0.746	0.940	0.999	186.481	0.000
Paired-Acquisition Neural Factorization biological	25	**0.300**	**0.959**	**0.999**	180.615	0.000
Paired-Acquisition Neural Factorization acquisition	25	**0.968**	0.031	0.421	47.355	0.000

Probe battery	Target	Raw DINOv2	Paired ref.	Paired-Acquisition Neural Factorization bio.	Paired-Acquisition Neural Factorization
Linear/RF/kNN	scanner_id ↓	0.662	0.573	**0.255**	**0.**
Linear/RF/kNN	sample_id ↑	0.942	0.983	**0.989**	0.
MLP	scanner_id ↓	0.832	0.740	**0.207**	**0.**
MLP	sample_id ↑	0.940	0.977	**0.986**	0.

Table 8: Scanner-removal baseline controls. Simple post-hoc operations on frozen DINOv2 features cannot match the scanner-suppression/tissue-preservation tradeoff of Paired-Acquisition Neural Factorization on either dataset.

Dataset	Condition	Scanner	Mean cos.	Worst cos.	Mean ret.
SCORPION	Neural Factorization	0.399	0.879	0.850	0.9998
SCORPION	Best linear ($k\geq4$)	0.724	0.881	0.850	0.9999
SCORPION	Best PCA ($k=32$)	0.560	0.806	0.765	1.0000
SCORPION	Original frozen	0.865	0.987	0.982	0.9986
Canine SCC	Neural Factorization	0.361	0.730	0.657	0.9334
Canine SCC	Best linear ($k\geq4$)	0.707	0.739	0.665	0.8738
Canine SCC	Best PCA ($k=32$)	0.641	0.598	0.481	0.9347
Canine SCC	Original frozen	0.863	0.919	0.891	0.8329

slide aggregation, institutional influence, and operational constraints must be audited separately.

The main limitations are clear. SCORPION contains 48 original slides and five scanners from one benchmark. Tissue retrieval is not a clinical endpoint. Pair-integrity falsification strengthens the same-tissue pairing interpretation but does not prove universal robustness to all pairing errors or acquisition shifts. Linear probes cannot prove that every form of acquisition information has been removed; the canine stronger-probe battery reduces this concern but still cannot prove absolute independence. Synthetic recovery laws are local to the tested generator, metrics, and budgets. PANDA and CAMELYON17 results are retrospective and largely feature-level. Privacy-noise and infrastructure experiments are probes, not deployment guarantees. Companion PDFs provide exact study-specific boundaries so that external, canine, synthetic, whole-slide, and federated results are not collapsed into one clinical claim. No experiment establishes treatment benefit, prospective workflow safety, regulatory readiness, or universal biological–acquisition identifiability.

9 Conclusion

Biology-preserving paired acquisitions can supervise a neural factorization in which tissue identity is retained while scanner identity becomes substantially less recoverable from the biological branch. The same frozen objective transferred across DINOv2, Phikon, and ResNet50 on real human histopathology, while a compact acquisition branch retained scanner information and cross-backbone pair-integrity controls reproduced the true-pair tissue-preservation advantage. Simple post-hoc scanner-removal baselines on frozen features could not match the neural tradeoff on either SCORPION or external canine SCC. The external canine audit reproduced the branch-separation pattern across folds, seeds, and stronger probes. Earlier failures show why unconditional site removal is unsafe, and pair-repeat experiments show that unique anchors and repeated exposure are separate resources.

The broader research record extends this principle to slide aggregation, federated influence, external-center validation, and systems constraints. Focused companion PDFs now carry the detailed protocols and exact tables for bounded studies, including the SCORPION core package, external canine SCC validation, and matched-budget allocation study. The supported claim remains bounded: acquisition-associated variation should be separated conditionally and audited explicitly, not erased merely because it predicts site.

A Complete Empirical Study Record

The main text is deliberately narrow and result-first. The following appendix preserves the complete study-by-study research record, including methods, secondary experiments, negative results, systems probes, and claim boundaries.

B Broader computational-pathology study record

The detailed SCORPION paired-scanner study above is the newest and strongest locked Paired-Acquisition Neural Factorization result, but it is not the whole research record. Earlier public PDF versions treated the repository as a three-level platform—representation, whole-slide aggregation, and institutional aggregation—and reported several distinct studies inside those levels. This section restores that study-by-study structure and adds the newer evidence without using one experiment to overstate another.

B.1 Study map

Table 9: Fixed detector transfer to conservative ordinal threshold shift. Deltas are detector-switch minus the clean-strategy baseline.

Shift	Trigger	Global QWK Δ	Macro-F1 Δ	Worst-site QWK Δ
0%	13.3%	−0.00025	+0.00009	+0.00151
25%	33.3%	+0.00129	+0.00290	+0.00353
35%	60.0%	+0.00542	+0.00838	+0.00991
45%	73.3%	+0.01053	+0.01512	+0.01290

Table 10: Empirical studies represented in the public research record. Each row has its own unit of analysis and claim boundary.

Study	Data or experimental substrate	Primary question
Adversarial Paired-Acquisition Neural Factorization confirmation	Pathology-derived representations	Does lower site leakage improve tumor prediction?
Factorized and subtraction falsification	Pathology-derived representations	Can a learned site component be removed without deleting biology?
Synthetic two-resource phase maps	Controlled latent-factor generators	How do ordinary observations and paired interventions jointly control recovery?
Pair-repeat anchor allocation	Controlled nonlinear high-overlap regime	Under fixed paired exposure, how much does unique-anchor diversity matter relative to repetition?
PANDA baseline hierarchy	10,611 verified Phikon feature bags	How much do mean pooling, gated attention, and TransnnMIL differ for whole-slide grading?
TransnnMIL stability grid	PANDA feature bags	Is the custom MIL model stable across optimization settings and seeds?
Dominant-site label-noise stress	PANDA-derived federations	When does sample-volume dominance become unsafe under controlled corruption?
Fixed detector transfer	PANDA-derived federations	Does a label-noise-calibrated detector transfer to systematic ordinal threshold shift?
Detector ablation and sensitivity	PANDA-derived federations	Is the detector dependent on one diagnostic family or one exact calibration?
CAMELYON17 external-center validation	455,954 examples from five centers	Does sample-proportional source weighting generalize best to an unseen center?
CAMELYON17 center-subspace projection	CAMELYON17-trained features	Can source-center leakage be attenuated without collapsing tumor signal?
Communication accounting	CAMELYON17 feature/head regime	What communication burden is avoided by feature/head federation?
Privacy-noise stress	CAMELYON17-trained features	Do aggregation differences survive coefficient-noise perturbation?
Infrastructure-friction simulation	Three heterogeneous source-client profiles	How do speed, dropout, and stragglers affect practical federation?
PCam patch validation	Full public PCam test set	How well does the patch classifier perform, and how stable are its estimates?
PCam federated strategy probes	Real PCam patches split into simulated sites	Do institutional weighting strategies behave differently under balanced and heterogeneous partitions?

C Earlier Paired-Acquisition Neural Factorization studies

C.1 Locked adversarial confirmation

The first neural Paired-Acquisition Neural Factorization study used a shared feature encoder with a tumor head and a gradient-reversal site head. Development experiments suggested a small task improvement and reduced site leakage. A locked confirmation over new seeds did not reproduce the task advantage: the adversarial model reached approximately 0.9264 held-out tumor accuracy versus 0.9270 for the tumor-only baseline, a paired difference of approximately -0.000625 with an interval crossing zero. The site-probe difference was approximately -0.003258, with the paired bootstrap interval excluding zero.

The result therefore established a representation change, not a diagnostic-performance improvement. Lower recoverable site information was achievable, but it was not sufficient evidence that the representation had become more useful for tumor prediction.

C.2 Factorized model selection and unconditional-subtraction falsification

A later model used a shared stem with diagnostic and acquisition branches, reconstruction pressure, and cross-branch covariance control. Validation-based selection preferred the diagnostic-only baseline rather than the more elaborate factorized model. A subsequent subtraction experiment was more decisive: the learned acquisition-associated component remained strongly tumor-predictive, so removing it also removed useful biological information.

This falsified the simple rule that a direction should be deleted merely because it predicts site. The negative result directly motivated biology-preserving paired supervision and the later SCORPION study.

C.3 Synthetic two-resource phase maps

Controlled generators made the latent biological and acquisition factors known, allowing recovery to be measured rather than inferred indirectly. The experiments varied observational sample size N_u, unique paired-anchor count N_p, biological–acquisition overlap, linear versus nonlinear mixing, neural method, and seed. Pair-consistency, learned-operator, and hybrid-curriculum variants were evaluated using task retention, site leakage, cross-factor leakage, paired biological invariance, counterfactual transport, canonical-correlation recovery, and Procrustes recovery.

Under the compute-controlled high-overlap nonlinear protocol, sustained recovery boundaries were:

Table 11: Compute-controlled paired-anchor boundaries in the frozen synthetic follow-up.

Method	N_u	Sustained paired-anchor crossing
Hybrid curriculum	750	175
Hybrid curriculum	1500	150
Pair consistency	750	225
Pair consistency	1500	125

The threshold analysis used interval- and right-censored fits because some cells did not cross the historical recovery threshold within the tested intervention budget. The resulting relationship is a local, threshold-dependent empirical family: more observational data can reduce the paired-anchor requirement, while stronger overlap and nonlinear mixing increase it. It is not presented as a universal law.

C.4 Pair-repeat anchor diversity versus repetition

A separate allocation study fixed observational presentations, optimizer steps, pair-batch size, and total paired presentations while redistributing the paired budget between unique anchors and repetitions. The high-overlap nonlinear setting used $N_u \in \{750, 1500\}$, pair-consistency and hybrid-curriculum methods, and ten matched seeds.

Table 12: Mean universal biological score across four method-by-N_u strata under matched paired-presentation budgets.

Paired presentations	50 anchors	100 anchors	200 anchors
6,400	0.4081	0.4243	0.4259
12,800	0.4489	0.4595	0.4619

At 6,400 presentations, the seed-blocked 200-versus-50 contrast was $+0.017865$, with 95% bootstrap interval $[0.010560, 0.026079]$, a positive difference in all ten seed blocks, and exact two-sided sign-flip $p = 0.001953$. The 200-versus-100 contrast was only $+0.001578$. At 12,800 presentations, the corresponding contrasts were $+0.013050$ and $+0.002422$. Doubling total paired exposure increased recovery by $+0.037357$ overall, with interval $[0.030315, 0.043657]$, positive effects in all ten seed blocks, and exact $p = 0.001953$.

The supported interpretation is that total paired exposure was the dominant resource in this regime, unique-anchor diversity provided an additional benefit, and gains diminished beyond approximately 100 anchors. No allocation crossed the historical 0.50 recovery threshold, so this is an allocation and exposure result rather than a recovered phase boundary.

D PANDA whole-slide studies

D.1 Mean pooling, AttentionMIL, and TransnnMIL comparison

The PANDA work studies weakly supervised prostate-cancer grading from whole-slide Phikon feature bags on the PANDA challenge data [2]. After HDF5 readability verification, 10,611 slide-level feature files were retained. Three principal model families were evaluated: a mean-pooled Phikon MLP, gated AttentionMIL [6], and TransnnMIL, a custom direction combining TransMIL-style correlated self-attention [20] with nnMIL-style gated aggregation [15].

Table 13: PANDA whole-slide validation evidence. QWK denotes quadratic weighted kappa.

Model or run	Validation QWK
Mean-pooled Phikon + MLP	0.7274
Gated AttentionMIL	0.8100
Tuned TransnnMIL, seed 42	0.8155
Tuned TransnnMIL, seed 123	0.8225
Tuned TransnnMIL, seed 2025	0.8086

The model hierarchy establishes that attention-based slide aggregation substantially improved on simple mean pooling in this feature-bag setting. TransnnMIL was favorable on two of three listed tuned seeds, but the margin over AttentionMIL remained small.

D.2 TransnnMIL optimization-stability grid

The stabilization study used the same 10,611 readable PANDA bags, three seeds, and six learning rates. The recipe used AdamW, warmup-cosine scheduling, two warmup epochs, gradient clipping at norm 1.0, and early stopping.

Table 14: TransnnMIL stability across three seeds per learning rate.

Learning rate	Mean best val QWK	Std.	Min.	Max.
1×10^{-4}	**0.8257**	0.0169	0.8087	0.8425
2×10^{-4}	0.8245	0.0192	0.8077	0.8455
3×10^{-4}	0.8238	0.0160	0.8127	0.8422
5×10^{-4}	0.8158	0.0144	0.8042	0.8319
7×10^{-4}	0.8117	0.0163	0.7994	0.8301
1×10^{-3}	0.8160	0.0170	0.7998	0.8337

This study supports optimizer stabilization and a wider usable learning-rate range. It does not yet establish that the complete TransnnMIL architecture is superior to simpler MIL methods under matched compute. Optional hierarchical, topology, and pruning branches remain separate implementation directions until controlled ablations validate them.

E PANDA federated site-signal studies

E.1 Dominant-site label-noise stress

PathologyFL partitions PANDA-derived Phikon slide representations into simulated clients and applies label corruption only to the largest client. The study asks whether FedAvg's implicit rule—more samples means more influence—remains safe when sample volume and task-specific label signal diverge.

In the full-cache 15-seed study, a tuned observable switch kept FedAvg in the clean regime and moved to a 50% cross-site blend when clean-calibrated diagnostics crossed a trigger threshold. Selected paired results were:

Table 15: Selected full-PANDA label-noise stress results for the tuned observable switch.

Dominant-site corruption	Metric	Delta versus FedAvg	
0%	Global QWK	$+0.000126$	$[-0.000$
25%	Global QWK	$+0.008009$	$[0.001$
35%	Global QWK	$+0.007482$	$[0.003$
45%	Worst-site QWK	$+0.009713$	$[-0.000$

The clean false-switch rate was 6.7%, while trigger rates increased strongly under 25–45% corruption. The result

supports a conditional stress-regime mechanism, not a universal replacement for FedAvg.

E.2 Fixed detector transfer to conservative ordinal threshold shift

A stronger transfer study froze one detector calibrated under dominant-site label noise and applied it without retuning to a systematic conservative ordinal threshold shift. The fixed rule used the 0.10 lower quantile, 0.80 upper quantile, at least three triggered diagnostic families, and no entropy diagnostic.

The conservative ordinal threshold-shift detector-transfer results are reported in Table 9. We do not duplicate the table here. The key pattern is that clean-regime triggering remains low, while trigger rates increase under stronger conservative shift and the detector-transfer deltas become positive for global QWK, macro-F1, and worst-site QWK in shifted regimes.

At 25%, 35%, and 45% conservative shift, the fixed detector increasingly switches away from FedAvg and yields positive deltas across the reported summary metrics, while the 0% clean condition remains near neutral.

At 35% and 45% shift, the bootstrap intervals for all three headline deltas were positive. The 25% regime was directionally positive but inconclusive. The scientific value is transfer across stress mechanisms: the detector was calibrated under random label-noise stress and evaluated as frozen under systematic ordinal bias.

E.3 Diagnostic ablation and calibration sensitivity

The detector uses global QWK, worst-site QWK, site-QWK spread, mean absolute ordinal error, and severe-error rate. In the conservative-shift study, mean absolute error was the most frequent trigger, followed by worst-site and global-QWK degradation. Site spread alone was the weakest single-family variant.

Removing the strongest diagnostic family, `mean_abs_error_high`, reduced the mean 35/45 trigger rate to 50.0% but retained positive mean deltas of $+0.00701$ global QWK, $+0.01003$ macro-F1, and $+0.00856$ worst-site QWK. A 36-setting neighborhood sweep varied lower quantile, upper quantile, and minimum trigger count; 29 of 36 configurations preserved a clean trigger rate at or below 20% and positive 35%/45% gains across all three headline metrics. This study weakens the explanation that the transfer result is a one-feature shortcut or a single hand-picked calibration.

F CAMELYON17 external-center studies

F.1 Natural center-shift design

CAMELYON17/WILDS provides natural hospital-center labels rather than only simulated client partitions [1, 9]. The audited feature-level dataset contains 455,954 examples across five centers. Centers 0, 3, and 4 are source-training centers, center 1 is OOD validation, and center 2 is the held-out OOD test center.

F.2 Frozen ImageNet ResNet18 feature study

With frozen ImageNet ResNet18 features, FedAvg-style equal-patch weighting reached 0.8312 held-out test accuracy. Equal-client weighting reached 0.9132, an improvement of 8.20 percentage points, while downweighting the dominant source improved accuracy by 7.82 points. A validation-aware detector switch improved held-out accuracy by 6.58 points, and 43 of 112 detector settings were robust-positive in the threshold sweep.

F.3 CAMELYON17-trained ResNet18 feature study

A second study repeated the aggregation comparison using features from a source-trained CAMELYON17 ResNet18 checkpoint.

Table 16: Held-out center-2 accuracy under two feature families and three aggregation policies.

Feature extractor	FedAvg-style	Equal-client	Downweig
Frozen ImageNet ResNet18	0.8312	0.9132	
CAMELYON17-trained ResNet18	0.9052	0.9318	

Under the source-trained representation, FedAvg-style weighting reached 0.9991 source-training accuracy while generalizing worse to center 2 than equal-client or dominant-source-downweighted weighting. The result supports the narrow observation that stronger source fit and sample-proportional influence do not automatically imply better external-center generalization.

F.4 Paired-Acquisition Neural Factorization center-subspace projection branch

A separate CAMELYON17 mechanism branch asked whether source-center information in the CAMELYON17-trained feature substrate could be attenuated without deleting tumor signal. Learned adversarial and subtractive variants did not materially reduce post-hoc center leakage. An explicit supervised center-subspace projection did: fitting a linear center classifier on an internal calibration split, removing the top right-singular directions of the centered classifier weight matrix, and probing on a separate internal split reduced center decodability while leaving tumor AUC essentially unchanged.

Table 17: CAMELYON17 center-subspace projection rank sweep. The effective center-discriminant rank is four after centering the five one-vs-rest class weight vectors.

Removed rank	Center accuracy	Tumor AUC	Tumor accuracy
0	0.8946	0.9903	0.9550
1	0.8530	0.9903	0.9564
2	0.8256	0.9903	0.9564
3	0.8092	0.9901	0.9562
4	0.7636	0.9903	0.9560

This is not a clinical-performance improvement claim and not complete source-center invariance. Its value is mechanistic: the adversarial variants failed, but a partially removable center-discriminant subspace exists and can be attenuated without collapsing the tumor probe.

G Federated systems stress studies

G.1 Communication accounting

Communication was bounded for the CAMELYON17 feature/head regime. Under an explicit comparison with three source clients, 100 fp32 rounds, and a binary ResNet18 full-model proxy, full-model exchange was estimated at 24.98 GB. A 512-to-2 feature/head learner required approximately

2.35 MB, making the full-model proxy about 10,894 times larger.

Equal-client and dominant-source-downweighted policies used the same feature/head communication budget as FedAvg-style weighting, so their held-out-center gains were not purchased with additional traffic. A detector-style decision after 5, 10, or 25 rounds corresponds to estimated full-model communication reductions of 95%, 90%, or 75% relative to 100 rounds. This is a bounded accounting result, not a measured wide-area hospital-network deployment.

G.2 Privacy-noise robustness probe

A coefficient-noise stress test used the CAMELYON17-trained ResNet18 features and added Gaussian perturbations to logistic-head coefficients at standard deviations 0.0, 0.01, 0.03, 0.05, 0.10, and 0.20, with five repeats per level. Equal-client and dominant-source-downweighted policies remained positive versus FedAvg-style weighting at every tested level. At noise standard deviation 0.20, the gains were approximately $+1.98$ and $+1.90$ percentage points.

This is not differential privacy, privacy accounting, membership-inference testing, or an update-inversion audit. It is a robustness probe under coefficient perturbation.

G.3 Infrastructure-friction simulation

A separate systems simulation modeled heterogeneous client speed, dropout, synchronous stragglers, an asynchronous proxy, communication cost, and detector-style reduced-round operation. The illustrative clients required 4, 7, and 13 minutes per local update. The outputs quantify wall-clock burden, failed rounds, straggler sensitivity, active-client counts, and traffic.

This study provides reproducible infrastructure accounting but is not a real hospital deployment benchmark. Containerized orchestration and measured network experiments remain future work.

H PCam validation studies

H.1 Patch-level classifier and uncertainty analysis

The PCam study provides the repository's patch-level cancer-classification validation layer. The reported validation AUC was 0.9537. On the full 32,768-example public test set, the model reached 0.8526 accuracy and 0.9394 AUC. One thousand bootstrap resamples were used to report uncertainty.

A threshold and false-negative analysis examined screening-style operating trade-offs. In the documented analysis, threshold adjustment reduced missed tumor predictions by 61.7% relative to the default operating point. This is retrospective benchmark analysis, not a prospective screening claim.

H.2 Federated strategy probes on PCam

Real PCam patches were also divided into simulated sites to exercise equal, volume-based, prestige-style legacy, and FAIR-WEIGHTS-H aggregation. The balanced benchmark completed without performance degradation from FAIR-WEIGHTS-H. Under heterogeneous simulated sites, the strategies produced different weighting trajectories, but patch-level performance was comparatively insensitive to those differences.

The value of this study is methodological: it verifies that the aggregation implementations run on real image data and that distinct policies actually alter institutional influence. It does not establish that FAIR-WEIGHTS-H improves clinical fairness or predictive performance.

I How the studies fit together

The study sequence follows the learning pipeline:

- **PCam** establishes patch-level classification, uncertainty, and threshold-analysis infrastructure.
- **PANDA MIL** tests slide-level aggregation from pathology foundation-model feature bags.
- **Earlier Paired-Acquisition Neural Factorization** falsifies unconditional site removal and motivates paired supervision.
- **Synthetic Paired-Acquisition Neural Factorization** measures recovery when the ground-truth latent factors are known.
- **SCORPION Paired-Acquisition Neural Factorization** tests paired-acquisition separation on real human tissue across multiple representation families.
- **PANDA federated stress** audits sample-volume dominance under controlled label-process misalignment.
- **CAMELYON17** tests related aggregation hypotheses under natural held-out-center shift and tests whether explicit center-subspace projection can attenuate center leakage without collapsing tumor signal.
- **Communication, privacy-noise, and infrastructure studies** bound systems constraints without claiming that deployment is solved.

The integrated hypothesis is not that one architecture solves all of computational pathology. It is that representation formation, whole-slide aggregation, institutional influence, and systems constraints are separate failure surfaces that require separate experiments and explicit claim boundaries.

J Supporting Identifiability Calculations

K Paired-intervention identifiability calculations

This section is a theoretical and synthetic extension motivating the next Paired-Acquisition Neural Factorization experiments. It is separate from the detector-switch evidence above and should not be interpreted as a clinical validation result.

K.1 Why pairing changes the information structure

Assume a pathology representation is generated by biological and acquisition factors:

$$x = Bz_b + Az_a + \varepsilon. \tag{27}$$

For a paired acquisition intervention preserving biology,

$$x_i = Bz_{b,i} + Az_{a,i} + \varepsilon_i, \tag{28}$$

$$x'_i = Bz_{b,i} + Az'_{a,i} + \varepsilon'_i. \tag{29}$$

Subtracting gives

$$\Delta x_i = x'_i - x_i = A(z'_{a,i} - z_{a,i}) + (\varepsilon'_i - \varepsilon_i), \tag{30}$$

so the shared biological term cancels. If acquisition latents have covariance Σ_a and the noise is isotropic,

$$\mathrm{Cov}(\Delta x) = 2A\Sigma_a A^\top + 2\sigma^2 I. \tag{31}$$

Table 18: Geometric difficulty proxy.

ρ	$1-\rho^2$	$(1-\rho^2)^{-2}$
0.50	0.7500	1.78
0.75	0.4375	5.22
0.90	0.1900	27.70
0.95	0.0975	105.19

Thus the leading eigenspace of paired differences estimates the acquisition subspace. A first-order biological projection is

$$P_{\perp A} = I - AA^\top, \qquad x_{\text{bio}} = P_{\perp A}x. \tag{32}$$

K.2 Geometric overlap penalty

Let B and A be orthonormal bases with canonical overlap ρ. Removing the acquisition subspace retains approximately

$$E_{\text{bio,retained}} \approx 1-\rho^2. \tag{33}$$

Hence $\rho = 0.90$ leaves only 0.19 of the biological energy outside the acquisition subspace, while $\rho = 0.95$ leaves 0.0975. Any component in $\text{span}(B) \cap \text{span}(A)$ is not uniquely assignable without additional assumptions.

K.3 Finite-pair sample complexity

Let $m = pn$ be the number of paired examples, D the observed dimension, and γ the eigengap separating acquisition directions from noise. A Davis–Kahan-style perturbation argument suggests

$$\sin\Theta(\widehat{A}, A) \lesssim \frac{C}{\gamma}\sqrt{\frac{D+\log(1/\delta)}{pn}}. \tag{34}$$

Requiring this estimation error to be smaller than the surviving biological margin gives the provisional scaling prediction

$$p_{\min} \gtrsim \frac{C^2(D+\log(1/\delta))}{n\gamma^2(1-\rho^2)^2}. \tag{35}$$

This predicts diminishing returns in paired coverage, error decay proportional to $(pn)^{-1/2}$, and a sharp rise in required pairing as $\rho \to 1$. The exact exponent and overlap dependence remain hypotheses rather than established theorems for the learned nonlinear models.

K.4 Synthetic two-resource recovery boundary

A follow-up synthetic phase-map study treated unpaired observations and paired interventions as distinct resources. It varied observational sample size
$N_u \in \{500, 1000, 3000, 10000\}$, high biological–acquisition overlap $\rho \in \{0.90, 0.95\}$, linear versus nonlinear mixing, a grid of paired-anchor counts up to 300, five seeds, and pairing-aware methods. The endpoint N_p^* was the smallest paired-anchor count at which the mean universal biological score crossed a predefined threshold and remained above it for every larger tested pair count.

At the primary sustained-recovery threshold $\tau = 0.50$, 43 of 47 available conditions had observed boundaries and four were right-censored above the tested pair-count range. A model using $\log N_u$, overlap, and nonlinear mixing achieved the best leave-condition-out prediction and slightly outperformed a larger model with method-specific offsets.

The uncensored empirical fit was

$$\log N_p^* = -1.568 - 0.659\log N_u \tag{36}$$
$$+ 10.8792\rho + 0.5345V, \tag{37}$$

Table 19: Recovery-boundary model comparison at $\tau = 0.50$. Lower leave-condition-out log RMSE is better.

Model	Log RMSE
Power + overlap + nonlinearity	0.455
Full model with method offsets	0.459
Power only	0.599
Constant boundary	0.881

where $V = 1$ denotes nonlinear mixing and $V = 0$ denotes linear mixing. Equivalently,

$$N_p^* \approx 0.208\, N_u^{-0.659} \tag{38}$$
$$\times \exp(10.8792\rho)\exp(0.5345V). \tag{39}$$

An interval/right-censored likelihood fit used all 47 conditions and treated the four missing primary boundaries as exceeding the maximum tested pair count. It yielded

$$\log N_p^* = -5.077 - 0.8086\log N_u \tag{40}$$
$$+ 15.8347\rho + 0.5277V, \tag{41}$$

and therefore

$$N_p^* \approx 0.00624\, N_u^{-0.8086} \tag{42}$$
$$\times \exp(15.8347\rho)\exp(0.5277V). \tag{43}$$

The nonlinear multiplier is $\exp(0.5277) \approx 1.695$. Within this tested synthetic regime, more observational data was associated with fewer required paired anchors, whereas overlap and nonlinear mixing increased the paired-anchor requirement.

The qualitative sign pattern was stable in ordinary least-squares sensitivity analyses at thresholds 0.45, 0.50, 0.55, and 0.60: the coefficient of $\log N_u$ remained negative, while the coefficients of overlap and nonlinear mixing remained positive. The number of observed boundaries decreased from 47 at $\tau = 0.45$ to 27 at $\tau = 0.60$, so stricter-threshold estimates are increasingly influenced by right censoring. Censored fits converged at $\tau = 0.50$ and $\tau = 0.55$; non-converged fits are not used as headline estimates.

These results support a threshold-dependent empirical family

$$N_p^*(\tau) = C_\tau N_u^{-\alpha_\tau}\exp(\gamma_\tau\rho)\exp(\eta_\tau V), \tag{44}$$

not a universal coefficient set. The raw-overlap term is a local approximation over the tested high-overlap range and does not by itself confirm the theoretical $(1-\rho^2)^{-2}$ dependence.

K.5 Compute-controlled follow-up

The fixed-epoch phase map coupled observational sample size to optimization exposure because larger datasets produced more minibatches per epoch. A follow-up protocol therefore fixed optimizer steps at $S = 320$, observational batch size at 128, and one separate paired-supervision batch per optimizer step. It preserved held-out pairs for diagnostics and evaluated the harder unseen setting $\rho = 0.975$ with nonlinear mixing across 20 seeds.

Under the predefined mean sustained-recovery rule, hybrid curriculum required 175 anchors at $N_u = 750$ and 150 at $N_u = 1500$. Pair consistency required 225 and 125 anchors, respectively. These point crossings were close to the threshold and should not be read as sharply identified constants.

A seed-cluster bootstrap produced broad, right-censored intervals. The median reduction was 25 anchors for hybrid

Table 20: Universal biological score at $N_p = 225$ and $R_p = 128$.

Method	$N_u = 750$	$N_u = 1500$
Hybrid curriculum	0.503	0.524
Pair consistency	0.495	0.505

Table 21: Mean recovery across four method-by-N_u strata under matched paired-supervision allocations.

Budget	$N_p \times R_p$	Biological score	Factor separation
6400	50×128	0.4081	0.7487
6400	100×64	0.4243	0.7626
6400	200×32	0.4259	0.7661
12800	50×256	0.4489	0.7874
12800	100×128	0.4595	0.7922
12800	200×64	0.4619	0.7987

curriculum and 50 for pair consistency, but both intervals crossed zero. The probability of a positive reduction was 0.564 and 0.670, respectively. These results are suggestive, not confirmatory.

The fixed-epoch relation remains a description of that training protocol, but recovery also depends on optimization exposure. A more complete resource description is

$$\text{Recovery} = f(N_u, N_p, R_p, S, \rho, V), \tag{45}$$

where R_p denotes paired-anchor repetition and S denotes optimizer steps.

K.6 Exact anchor diversity–repetition pilot

A subsequent exact-schedule pilot separated unique paired anchors N_p from presentations per anchor R_p. Every selected anchor appeared exactly R_p times in a deterministic shuffled schedule. The pilot crossed $N_u \in \{750, 1500\}$, $N_p \in \{50, 125, 225\}$, $R_p \in \{32, 64, 128\}$, two methods, and five new seeds, for 180 neural-network fits.

Increasing repetition improved the universal biological score in all 12 fixed-$(N_u, N_p, \text{method})$ slices. Increasing unique-anchor diversity also improved it in all 12 fixed-$(N_u, R_p, \text{method})$ slices.

Task AUC remained approximately 0.96–0.98 while biological recovery, factor separation, and paired-invariance error changed substantially. Strong task prediction therefore did not certify a disentangled biological representation.

K.7 Matched paired-supervision budgets

To separate unique intervention diversity from repeated exposure, two exact-schedule experiments fixed total paired presentations within each budget while changing the allocation between N_p and R_p. The first compared

$$50 \times 128 = 100 \times 64 = 200 \times 32 = 6400, \tag{46}$$

with 100 pair-loss-active steps. The second doubled repetition at every anchor count,

$$50 \times 256 = 100 \times 128 = 200 \times 64 = 12800, \tag{47}$$

with 200 pair-loss-active steps. Both experiments used the same 10 matched seeds, $N_u \in \{750, 1500\}$, hybrid curriculum and pair consistency, 320 optimizer steps, and 40,960 observational presentations. Each budget therefore comprised 120 neural-network fits.

For global allocation contrasts, the four method-by-N_u cells were first averaged within each seed, leaving 10 independent matched seed blocks. At 6400 presentations, 200 anchors exceeded 50 anchors by 0.0179 (95% seed-bootstrap interval $[0.0106, 0.0261]$), with a positive difference in all 10 seed blocks and an exact two-sided sign-flip $p = 0.00195$. The 100-versus-50 contrast was also positive on average, but its interval included zero (0.0163, interval $[-0.0017, 0.0328]$, $p = 0.113$). At 12800 presentations, the 200-versus-50 contrast remained positive (0.0131, interval $[0.0015, 0.0251]$), although the exact test was less conclusive ($p = 0.0762$). The 100-versus-50 contrast was 0.0106 ($p = 0.195$).

Differences between 200 and 100 anchors were near zero at both budgets: 0.0016 at 6400 presentations and 0.0024 at 12800 presentations, with exact p-values of 0.822 and 0.607. The allocation result therefore indicates that concentrating all supervision on only 50 unique anchors is inefficient in this regime, while the incremental benefit beyond approximately 100 anchors is small and uncertain.

Doubling total paired exposure produced the larger and more reproducible effect. Averaged within each seed across all 12 matched cells, 12800 presentations improved the universal biological score over 6400 by 0.0374 (95% seed-bootstrap interval $[0.0303, 0.0437]$). All 10 seed blocks improved and the exact two-sided sign-flip test gave $p = 0.00195$. The improvement was positive for each allocation separately: 0.0408 at 50 anchors, 0.0352 at 100 anchors, and 0.0360 at 200 anchors.

These controlled results identify total paired exposure as the dominant resource over this range, with a smaller diversity-allocation effect and diminishing returns between 100 and 200 unique anchors. No allocation mean at either budget crossed the predefined recovery threshold of 0.50, so the experiments characterize resource efficiency and response to supervision rather than a confirmed recovery boundary.

K.8 Objective conflict in transport models

For an operator trained with task, environment, reconstruction, transport, and consistency losses,

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_e \mathcal{L}_{\text{env}} + \lambda_r \mathcal{L}_{\text{recon}} + \lambda_t \mathcal{L}_{\text{transport}} + \lambda_c \mathcal{L}_{\text{consistency}}, \tag{48}$$

the transport and consistency objectives may conflict. Their gradient cosine is

$$G_{ct} = \frac{\nabla \mathcal{L}_c^\top \nabla \mathcal{L}_t}{\|\nabla \mathcal{L}_c\| \, \|\nabla \mathcal{L}_t\|}. \tag{49}$$

A negative G_{ct} indicates that reconstruction of acquisition-specific detail is opposing biological invariance.

References

[1] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019.

[2] Wouter Bulten et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine*, 28:154–163, 2022.

[3] Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David P. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25: 1301–1309, 2019.

[4] Alexandre Filiot, Roux Ghermi, Amandine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[6] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2127–2136, 2018.

[7] Peter Kairouz, H. Brendan McMahan, Brendan Avent, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2): 1–210, 2021.

[8] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5132–5143, 2020.

[9] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard L. Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Brian A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[10] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-IID data silos: An experimental study. *IEEE International Conference on Data Engineering*, pages 965–978, 2022.

[11] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020.

[12] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-IID features via local batch normalization. In *International Conference on Learning Representations*, 2021.

[13] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5:555–570, 2021.

[14] Ming Y. Lu, Dehan Kong, Jana Lipkova, Richard J. Chen, Rajendra Singh, Drew F. K. Williamson, Tiffany Y. Chen, and Faisal Mahmood. Federated learning for computational pathology on gigapixel whole slide images. *Medical Image Analysis*, 76:102298, 2022.

[15] Xiangde Luo, Jinxi Xiang, Yuanfeng Ji, and Ruijiang Li. nnMIL: A generalizable multiple instance learning framework for computational pathology. *arXiv preprint arXiv:2511.14907*, 2025.

[16] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

[17] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[18] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, et al. The future of digital health with federated learning. *npj Digital Medicine*, 3:119, 2020.

[19] Jeongun Ryu, Heon Song, Seungeun Lee, Soo Ick Cho, Jiwon Shin, Kyunghyun Paeng, and Sérgio Pereira. SCORPION: Addressing scanner-induced variability in histopathology. *arXiv preprint arXiv:2507.20907*, 2025.

[20] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer-based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 2136–2147, 2021.