Online Resource 1: Supplementary methods

Article: When Does Supervision Beat Reconstruction? Magnification Shift and the Robustness of AI Quality Control in Digital Pathology

Journal: Journal of Medical Systems

Authors: Konstantinos Moutselos, Ilias Maglogiannis

Corresponding author: Konstantinos Moutselos, Department of Digital Systems, University of Piraeus, Piraeus, Greece; kmouts@unipi.gr

Reference numbers in square brackets refer to the reference list of the article. Sections, tables and figures without the prefix S are those of the article.

S1 Data

S1.1 Development data: clean slides and ink-augmented reference

- **Clean slides in cross-validation.** Each fold holds out one slide without artifacts. A slide counted as clean if at most 0.01% of its pixels were annotated. All four qualify. One of them (TCGA-55-A492) carries a small fold annotation: 2,003 pixels, 510 cells at the 10× read.
- **Clean slides on the benchmark.** A slide is clean when none of its annotation lies within the scored coverage (11 slides). One of them (TCGA-56-8628) has 21 annotated air-bubble pixels, all outside it.
- **Clean false positives** are counted on clean-slide tissue outside the annotation. Annotated cells on a clean slide count toward sensitivity.
- The released AIRAQc annotations cover tissue only, by design, so marker ink lying off tissue is never annotated. The saturation-based scored region (SAT) nevertheless keeps much of it, because ink is strongly saturated.
- Teal ink was detected on each slide thumbnail at the map grid, in HSV space: hue 150–210°, saturation ≥ 0.30, value ≤ 0.90. Detection was followed by a 3 × 3 opening and a 3 × 3 closing.
- For the secondary (augmented) reference, teal inside SAT was added to the pen-marking class, following the ink-augmented reference of [15].
- Teal also occurs on 4 of the 16 training slides, where about three quarters of it lies outside the annotations. For the probe, teal cells were excluded from both positive and negative labels (S4). They entered training only as image content of tokens labelled by their other cells.

S1.2 SlideGuard cohorts

- **TCGA:** 25 slides, five each of breast, colon, lung, prostate and lymphatic tissue, with no overlap with our TCGA slides at slide or patient level.
- **Dartmouth-Hitchcock Medical Center:** 10 slides. The benchmark record describes them as breast; the files are those of the Dartmouth Lung Cancer Histology Dataset [17] (lung

adenocarcinoma). They were obtained under that dataset's research-use agreement. The dataset owner approved, in writing, uploading predictions, reporting scores and showing up to three slides.

- **Maria Skłodowska-Curie National Research Institute of Oncology:** 6 slides of lymphatic tissue (40×, 0.12 µm per pixel).
- **Mouse skin:** 12 slides (20×, 0.39 µm per pixel).
- The record gives only the cohort sizes. We assigned the non-TCGA files to cohorts from file names, group sizes and scanner metadata.
- Six DHMC slides are designated high quality by the organizers (DHMC_0017, 0044, 0055, 0063, 0071, 0102).

S1.3 Fırat dataset

- 2,127 RGB microscope images (3840 × 2160 pixels, 10× objective, 0.49 µm per pixel) from 302 slides, using the corrected slide mapping of Version 2 of the record [19].
- Two fold images have empty masks and were excluded from the fold endpoints.
- Each image carries a burned-in 100 µm scale bar at a fixed position. A fixed rectangle around the bar and its label (0.39% of the image) was treated as outside tissue for every detector (S6).

S1.4 Backbone exposure

- The generative backbone [8] and the second encoder [13] were pretrained on corpora that include TCGA. TCGA slides are therefore not unseen at the image level, although no artifact labels entered pretraining.
- TCGA slides of external sets are reported as a separate stratum.

S2 Generative detector

Adaptation.

- PixCell-1024 [8] is adapted with LoRA [20]: rank 16, α 16, no dropout.
- The adapters sit on the query, key, value and output projections of every self- and cross-attention layer, 224 modules in total.

Recipe K2 [9]:

- Training pool: 2,969 artifact-free 1,024 × 1,024 patches at 10×, selected from the training slides by entropy-based tissue detection, with patches containing marker ink removed.
- Training: 1,000 steps, batch size 4, learning rate 1×10^{-5}.
- Training monitor:
  - Every 50 steps, the mean reconstruction error on 96 clean patches from the two held-out AIRAQc slides (TCGA-34-5232, TCGA-BH-A0H5; outside the 16 training and 24 benchmark slides) was logged.
  - The checkpoint was fixed at step 1,000 in advance and was not selected on this value.

  - For some runs, the step-1,000 value was an acceptance check: at most 0.2754, i.e. 1.2 times the value of the fold-1 seed-7 K2 run. These were the k-fold A and H2 runs and H2all (each with its seed-7 repeat), and K2all_s7. All passed (0.193–0.221).
  - The other K2 runs (k-fold K2 and K2s7, K2all) and freshcond were not gated. Their values (0.218–0.237) were also below the threshold.
- Conditioning [21]: at every step the cross-attention input receives a fresh draw $e = \mu + \sigma \cdot \varepsilon$, $\varepsilon \sim N(0, I)$. Here μ and σ are the per-token-position mean and standard deviation of UNI2-h embeddings of the clean training patches. At inference the null embedding is used, so scoring is unchanged.

Scoring.

- Each tissue-bearing patch is encoded by the VAE into a 16-channel, 128 × 128 latent (one posterior sample).
- The latent is noised to t = 800, and the noise prediction $\hat{\varepsilon}$ is computed on the unconditional path.
- The score is $s = \text{mean}_c (\hat{\varepsilon} - \varepsilon)^2$ per latent position, i.e. per 8 × 8-pixel cell (≈ 8 µm).
- Patches tile the slide with stride 1,024 and are kept if ≥ 10% is tissue. The noise seed (42) is reset per slide.
- Patch errors are placed on the slide grid and averaged where patches overlap, which happens only at clamped borders.

S3 Error-space hinge (H2)

H2 adds two terms to the K2 denoising loss at every step. Both act on the unconditional scoring path, with t drawn uniformly from 750–850.

Anchor. Plain denoising on four clean patches from the pool: loss mean $(\hat{\varepsilon} - \varepsilon)^2$, weight 1.0.

Hinge.

- The hinge acts on the statistic the decision rule thresholds: the per-position error map, filled outside tissue with the running 65th percentile of clean error and smoothed with a Gaussian (σ = 2 cells).
- For one annotated patch per artifact type with ≥ 25 interior tissue cells of that type, the loss is the mean over those cells of relu(log Q99 − log s). It is then averaged over types.
- **Q99:** the 99th percentile of the same smoothed statistic on clean patches, over interior tissue cells ≥ 4 cells from the patch border. It is tracked as an exponential moving average (decay 0.95) and excluded from the gradient.
- The hinge positions, Q99 and the fill are restricted to entropy-detected tissue excluding teal ink.
- Weight $\lambda = 0.1$. An anchor-only arm (A, $\lambda = 0$) isolates the hinge's contribution.

Labels. Artifact positions come from the training-fold annotations only, through a per-fold catalogue. The all-slide variant (H2all) used all 16 slides.

Design history.

- H2 is the second hinge design. In the first (H), the hinge acted on the scored path with no anchor on that path.
- During H's training, the clean error on the two held-out slides rose from 0.43–0.49 at step 100 to 0.70–0.92 (seed 42) and 3.0–5.3 (seed 7) at step 1,000.
- H2 adds the anchor. Its settings and the acceptance check of S2 were fixed after these training logs and before H was evaluated on the test folds.
- On the test folds H detected almost nothing: its sensitivity was at most 0.004 for every artifact type, against 0.10–0.67 for K2. It was not pursued and is not reported further.
- H2's design was therefore informed by the held-out slides, not by test-fold results. No H2 setting changed afterwards. The only later change was in training data: H2all was trained on all 16 slides, before the development benchmark and the external sets were evaluated.

S4 Supervised probes

Features.

- Each 1,024-pixel patch is split into 4 × 4 tokens of 256 pixels.
- Each token is resized to 224 pixels (bilinear) and encoded with UNI2-h (revision d517a8dd) [10, 12]; the class token (1,536 dimensions) is kept.

Labels.

- **Region:** the coverage of the generative heatmap ∧ the entropy tissue mask ∧ ¬teal. A token enters only if ≥ 25% of its 1,024 cells lie in the region.
- **Positive:** the union of the four type masks covers ≥ 25% of the token's region cells.
- **Negative:** no annotation in the token's region cells.
- **Excluded:** all other tokens.
- Training census for probe_all: 13,902 positive, 17,445 negative and 8,269 excluded tokens.

Classifier. Standardization followed by L2-regularized logistic regression. The regularization strength was chosen by cross-validation grouped by slide (four groups).

Second encoder.

- Phikon-v2 (revision 2ae989a9) [13] replaces UNI2-h; its class token has 1,024 dimensions.
- Tokens were resized as for UNI2-h (bilinear rather than the bicubic resize of the model card), so the encoders differ only in the network.
- Phikon-v2 is distributed under the Owkin non-commercial licence with a ShareAlike term. The Phikon-v2 probe parameters and the evaluation outputs obtained with Phikon-v2 are in our deposit under that licence; the embeddings are not released.

Scale-augmented probe.

- Every training token was also embedded after the ×0.5/×2 resampling of its patch (S6) and added with the same label and slide group.
- Cross-validation folds therefore contain both copies, so its cross-validation scores are not comparable with the unaugmented probe's.

Probability map. Token probabilities are painted onto the 8 µm grid and averaged where patches overlap.

S5 Decision rules

R0.

- Two levels, the 65th and 75th percentiles of the error map, are computed without labels, pooled over the slides evaluated and within the tissue region.
- Each map is filled outside tissue with the lower level, smoothed ($\sigma = 2$ cells) and clipped to the two levels.
- It is then thresholded per slide with Otsu's method and cleaned by a 5×5 closing and opening.

Benchmark menu. For comparability with [15], the generative baseline also used leave-one-slide-out selection over a 5×5 menu of percentile pairs (the pair with the highest pooled F1 on the other 23 slides).

Probe.

- The probability map, zero-filled outside tissue and smoothed ($\sigma = 2$ cells), is thresholded at $\tau = 0.5$ with the same morphology.
- A leave-one-slide-out choice over $\tau \in \{0.3, \ldots, 0.9\}$ is a secondary rule.

Combination rules. All were selected on the development benchmark by leave-one-slide-out on pooled F1 (released reference) and then frozen.

- **C-ref:** thresholds the generative error at the 99th percentile of its smoothed values over positions the probe calls clean (probability < 0.1), with a pooled fallback below 2,048 such cells.
- **C-both:** keeps C-ref detections within 32 cells of a probe detection (probability ≥ 0.5).
- **PC:**
  - Per slide, the generative error is thresholded at the q-quantile of its smoothed values over SAT cells the probe calls clean (probability < p_c).
  - Grid: $q \in \{0.995, 0.99, 0.975, 0.95\}$, $p_c \in \{0.05, 0.1, 0.2\}$; frozen at $q = 0.95$, $p_c = 0.05$.
  - Falls back to R0 (65/75) when the reference set has fewer than $10/(1 - q)$ cells.
- **PR:** the probe mask (before morphology) is dilated by $m \in \{0, 8, 16\}$ cells. Otsu is applied to the generative error inside each connected component, then the morphology, and the result is intersected with the dilated mask. Frozen at $m = 16$.
- **BF:**

  - **Prior:** the probe map, zero-filled outside SAT and smoothed (σ = 2), clipped to $[10^{-4}, 1 - 10^{-4}]$ and converted to log-odds.
  - **Error statistic:** z = log(E_s / median_SAT E_s), from the smoothed error E_s.
  - **Likelihood ratio:** estimated once per generative recipe from the pooled out-of-fold k-fold maps of the matching arm. Artifact cells are SAT cells in the annotation union; clean cells are SAT cells without annotation. z uses 100 equal-width bins between its 0.1st and 99.9th percentiles, with values outside clipped. LLR = log f_art − log f_clean, after adding one to every count.
  - **Decision:** a cell is flagged when log-odds(prior) + α · LLR(z) ≥ 0, followed by the morphology. α was selected over {0.5, 1, 2} and frozen at 0.5.
- **Rejected on the development benchmark:** a local-window PC, an expected-error residual, seeded region growing, and fusion of two diffusion steps. None met the selection criterion that BF met (Results, Section 3.6), so none was carried forward to the external sets; their values are not tabulated.

S6 Magnification handling

Development slides. Four slides had been read at about 2.0 µm per pixel: two training slides (TCGA-HU-8610, and TCGA-W5-AA30, the clean slide of fold 4) and two clean benchmark slides (TCGA-HU-A4GC, TCGA-HU-A4GF).

- They were re-read through route 2 of the guard below.
- Region masks that exist only on the old grid were brought to the new grid by repeating each cell 2 × 2, replicating the last row or column where the new grid is one cell larger.
- Training was not repeated. The generative training pool, the hinge positions and the probe's training tokens from the two training slides remain at the old read.

Guard (SlideGuard, Fırat and the re-read slides), in order:

1. a pyramid level within 0.9–1.1 µm per pixel, read directly;
2. a level at about 0.5 µm per pixel, downsampled by 2 (box filter);
3. the nearest finer level, box-resampled by f ∈ (1, 4];
4. otherwise, rejection.

Where file metadata were unreliable, the pixel size came from the dataset's metadata and was checked against the evaluation server's mask dimensions.

Fırat.

- The images were read through route 2: 1920 × 1080 pixels at 0.98 µm per pixel, at most four overlapping patches per image.
- The pixel size, given by the dataset authors, was reproduced from the scale bar: 204 pixels per 100 µm, median over the 2,074 images in which the automatic measurement found the bar at its usual position. In the other 53 it picked up another dark structure. A template check showed the bar at its usual position in 49 of them and shifted by 9 pixels in four.

- In every image the bar and its label lie inside the masked rectangle.
- The scale-bar rectangle was removed from the tissue region before the decision rule and was not scored.

Controlled resampling.

- Each 1,024-pixel RGB patch was downsampled to 512 pixels and upsampled back to 1,024, bilinear in both steps (Pillow 10.2).
- Resampling precedes the token split and resize. Tissue gating, coverage and regions were kept from the original patches.
- The same function served the diagnostics, the scale-augmented probe and the second-encoder probe.

S7 Evaluation details

Regions.

- Sensitivity: inside SAT.
- Clean-slide false positives:
  - inside entropy-detected tissue for cross-validation, as in [9];
  - inside SAT on the development benchmark;
  - inside the provided tissue annotation (ANN) on SlideInspect.
- **Region sensitivity:** SAT keeps 66% of the out-of-focus annotation in cross-validation and 73% on the benchmark, against 95% and 77% for the entropy region (true 10× read; shares of the annotation within the generative coverage). Over all annotated cells, SAT keeps 84% and 82%, and the entropy region 94% and 85%. All conclusions were re-scored in the entropy regions with every setting fixed (Table S2).

FP-matched comparison. The probe's sensitivity was traced against clean FP along R0 percentile pairs from 45/55 to 95/99 and along τ from 0.5 to 0.99. Each generative arm was compared with the probe's sensitivity interpolated at the arm's own clean FP.

Fırat.

- **Regions:** SAT-F, SAT minus the scale-bar rectangle (primary), and COV-F, the whole field minus the rectangle (secondary).
- **Rules:** R0 levels were pooled over all images, with Otsu applied per image; the probe used its native rule.
- **Endpoints:** fold sensitivity on fold images; clean FP on clean images.
- **Intervals:** bootstrap over slides, stratified by clean and fold slides.
- **Reference masks:** transferred to the 8 µm grid by nearest-pixel sampling.

SlideGuard.

- Binary masks (background, tissue, artifact) at 2 µm per pixel were submitted once per arm through the server's web interface and kept off the public leaderboard.
- The server returns per-slide artifact Dice, precision, recall and artifact-free Dice within the organizers' tissue annotation.
- Contrasts use a paired bootstrap over slides by stratum.
- Per-cohort means for the comparison with the published methods use the slide set of the benchmark paper's Table 2 [5].

S8 Shift monitor M3

- A token is eligible if ≥ 512 of its 1,024 cells lie in the tissue region.
- **Reference:** the 31,410 eligible tokens of the 16 training slides, in the entropy region.
- **Test tokens:** each set's scored region (SAT; SAT-F for Fırat).
- **Covariance:** Ledoit–Wolf shrinkage [22].
- **Unit value:** the median distance over eligible tokens. For Fırat, a slide's value is the mean of its images' medians.
- Because the reference and test regions differ, absolute M3 values are not interpreted.

S9 Calibration

- **Unit:** one probe token, at the native output before painting, smoothing or thresholding. Probabilities were recomputed from the stored model and embeddings.
- **Labels:** the training rule of S4 on the development benchmark at the 10× read. Coverage comes from the K2all heatmaps, the entropy mask is the stored one (2 × 2 repeat for the two re-read slides), and teal is formed from the slide thumbnail.
- **Settings:** original read and resampled (Δ-B), for probe_all and the scale-augmented probe.
- **Metrics:** reliability tables (15 equal-width bins), ECE (token-weighted), Brier score, and the mean predicted probability on clean slides.
- **Intervals:** stratified slide bootstrap (20,000 replicates, seed 0).
- **Checks before measurement:**
  - the label rule reproduced the training census;
  - the recomputed probabilities, passed through the decision rule, reproduced the stored per-slide counts for all 96 slide × probe × read combinations.

S10 Analyses added after the SlideInspect results (post hoc)

- **(i)** Re-embedding the development slides after resampling (Δ-B).
- **(ii)** Per-case decomposition of H2's out-of-focus loss on SlideInspect.
- **(iii)** Re-scoring every conclusion in the entropy regions.
- **(iv)** Teal-ink census and the fraction of unannotated teal flagged per arm.
- **(v)** Three probe variants with teal labels:

  - T1: annotated teal admitted as positive;
  - T2: teal exclusion removed;
  - T3: all training teal labelled as pen marking.
- **(vi)** Fırat resampling (Δ-F), with generative R0 levels recomputed label-free and the probe at τ 0.5. The primary contrast was the difference-in-differences of the change in clean false positives (probe − H2).
- **(vii)** The per-organ and per-slide breakdown of (vi).
- **(viii)** The second encoder.
- **(ix)** The scale-augmented probe.
- **(x)** Shift monitors.
- **(xi)** The four re-read slides at both reads.
- **(xii)** Calibration.