Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Konstantinos Moutselos

kmouts@unipi.gr

University of Piraeus

Ilias Maglogiannis

University of Piraeus

Research Article

Keywords: digital pathology, quality control, artifact detection, foundation models, distribution shift, diffusion models

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11447191/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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

Konstantinos Moutselos (ORCID 0000-0002-6759-8540) and **Ilias Maglogiannis** (ORCID 0000-0003-2860-399X)

Department of Digital Systems, University of Piraeus, Piraeus, Greece

Corresponding author: Konstantinos Moutselos, kmouts@unipi.gr

Abstract

Purpose: Automated quality control of whole-slide images must stay reliable when scanning conditions change. We compared supervised probes on foundation-model features with generative reconstruction for artifact detection.

Methods: A linear probe on UNI2-h features and diffusion-reconstruction detectors adapted from PixCell used the same 16 annotated training slides. We evaluated them in cross-validation, on a 24-slide benchmark, on a blind 53-slide multi-cohort benchmark scored by its organizers, and on two external sets (289 cases; 2,127 images). Resampling to half resolution and back mimicked half magnification on the same tissue. We also assessed calibration and monitoring of the probe on known-clean slides.

Results: At matching magnification the probe was the better detector. It flagged no tissue on clean cross-validation slides, and on the blind benchmark its precision was 0.29 against 0.10 at similar Dice. Under resampling its false positives on clean tissue rose 227-fold on the benchmark and in every organ of the external images, while the reconstruction detectors' did not; at matched false positives the ranking reversed. Its calibration error rose from 0.05 to 0.30. A second encoder, Phikon-v2, showed a 3.8-fold rise. Training the UNI2-h probe with resampled copies removed most of its rise but lowered accuracy on the original images. Rules combining both detectors inherited the probe's failure. On known-clean slides the probe's flagged-tissue fraction separated the 5× external set and resampled images (AUROC 0.94 and 0.89).

Conclusions: Laboratories using foundation-model probes for quality control should verify acquisition magnification, monitor the probe on clean tissue, and keep a reconstruction detector as a safeguard.

Keywords: digital pathology; quality control; artifact detection; foundation models; distribution shift; diffusion models

1 Introduction

Whole-slide images carry artifacts from slide preparation and scanning: tissue folds, out-of-focus regions, air bubbles and marker ink, among others. Left undetected, they mislead both the pathologist and the downstream algorithm [1, 2]. Manual checking does not scale to a digital service or a research cohort, so automated quality control (QC) now precedes most analyses [3, 4].

Two approaches are in use. Supervised detectors learn each artifact type from annotations [4]. They are accurate on the types and data they were trained on, but only on those; a recent multi-center benchmark found that their accuracy varies with center, tissue and species [5]. Reconstruction detectors instead model artifact-free tissue and flag what they cannot reproduce [6, 7]. Diffusion models can be trained for this on clean tissue [6] or adapted from a generative foundation model for pathology [8]. They need no artifact labels, though their false positives depend on how clean tissue is defined [9].

Pathology foundation models have shifted the balance. Their frozen features turn a few annotated slides into a strong supervised detector, a linear probe [10], and with a light decoder they segment tissue structures from few labels [11]. If such a probe can be trained from the same annotations, it is fair to ask what reconstruction still adds.

The answer matters for deployment, where scanners, protocols and magnifications vary. QC software usually selects a pyramid level by its nominal magnification. When a file lacks the expected level or carries wrong metadata, the model silently receives tissue at another scale. This happened to four slides of our own benchmark (Section 2.4). We are not aware of a study that compares the two approaches on the same annotations, at matched false-positive rates, both at the training magnification and away from it.

We compared a linear probe on UNI2-h features [10, 12] with diffusion-reconstruction detectors built on PixCell [8] in three settings (Fig. 1): (i) in distribution (cross-validation and a benchmark with clean control slides); (ii) on a blind multi-cohort benchmark acquired at matching magnification [5]; and (iii) under acquisition shift (two external sets and resampling experiments that mimic a lower magnification on the same tissue). We repeated the resampling experiments with a second encoder [13] and with a probe also trained on resampled copies of its training data (scale augmentation). We also measured the probe's calibration, tested rules that combine the probe with a reconstruction detector, and asked whether these shifts can be detected from the probe's behavior on clean slides.

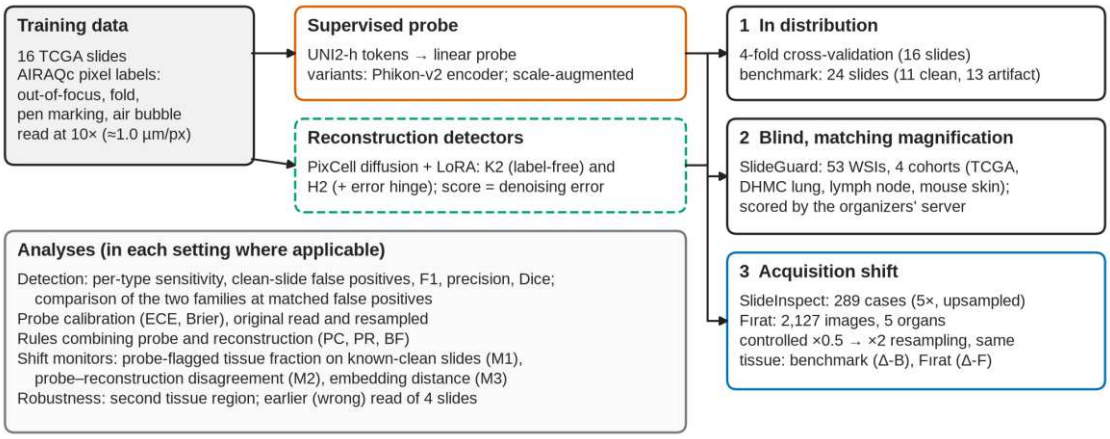


Fig. 1 Study design. Both detector families were trained on the same 16 annotated TCGA slides read at 10×: a linear probe on UNI2-h features (solid orange border) and diffusion-reconstruction detectors adapted from PixCell (dashed green border). They were evaluated in three settings: in distribution (cross-validation and a 24-slide benchmark); on a blind benchmark at matching magnification, scored by its organizers; and under acquisition shift (two external sets, and resampling to half resolution and back, on the same tissue). The analyses listed in the grey box were applied in each setting where the required labels or outputs existed. Δ-B and Δ-F denote the resampling experiments on the benchmark and on the Fırat images

Our contributions are:

1. a matched comparison of a foundation-model probe and diffusion reconstruction for artifact detection, in and out of distribution, including a blind third-party evaluation;
2. evidence that acquisition magnification reverses their ranking, shown on the same tissue, with its dependence on the encoder, its effect on calibration, and the cost of a training remedy;
3. a simple monitor, the fraction of tissue that the probe flags on known-clean slides, and the negative result that combining the probe with a reconstruction detector does not remove the failure;
4. a practical account of silent pyramid-level selection, with results reported at both the erroneous and the corrected resolution.

2 Methods

2.1 Data, splits and labels

Development data. AIRAQc provides pixel annotations for four artifact types (out-of-focus, tissue fold, pen marking, air bubble) on 50 diagnostic slides of The Cancer Genome Atlas (TCGA) from 50 patients, 25 with and 25 without artifacts [14]. Following [15], eight slides were excluded (no usable patches or too little tissue), leaving 16 for training, 24 for the benchmark and 2 held out. The held-out slides only monitored the training of the reconstruction detectors (an acceptance check; Online Resource 1) and entered no evaluation. Each patient contributes one slide, so slide-level splits and resampling are

patient-level. For in-distribution testing, the 16 training slides were split into four folds of four slides, and each fold was held out once. The 24 benchmark slides (11 without annotated artifacts, 13 with artifacts) followed the split and protocol of [15].

Detections and annotations are compared on a common grid of 8 × 8 µm squares ("grid cells"): the diffusion detectors score each square, and the probe's token predictions are painted onto the same grid. The reference for pooled metrics is every annotated pixel of the released annotation images. Of the 8.79 million scored annotated grid cells on the benchmark, 0.30% carry no artifact-type colour (anti-aliased annotation edges) and 0.07% (one slide) carry the dataset's no-tissue label; both enter pooled metrics but not per-type sensitivities. The annotations cover tissue only, so marker ink off tissue is never annotated; a secondary reference adds teal ink to the pen-marking class (Online Resource 1).

External data. No external slide overlaps the development slides at slide or patient level, and apart from SlideGuard's TCGA cohort the external sources come from other institutions, scanners or cameras. SlideInspect [16] contains 289 hematoxylin and eosin (H&E) cases from several organs, with pixel annotations for out-of-focus, fold and mounting artifacts; 55 cases without artifacts serve as clean controls. Its images are distributed at 5× and were upsampled ×2 (bilinear) to the working resolution. SlideGuard [5] comprises 53 whole-slide images in four cohorts (TCGA; Dartmouth-Hitchcock lung [17]; lymphatic tissue from the Maria Skłodowska-Curie National Research Institute of Oncology; mouse skin). Its annotations are held by the organizers and were used only through their evaluation server. TCGA slides of external sets are reported separately because the generative backbone and the second encoder were pretrained on TCGA images, so external claims rest on the non-TCGA cohorts. The Fırat set [18, 19] contains 2,127 microscope images of veterinary H&E teaching slides from five organs and 302 slides: 899 images (150 slides) are free of artifacts and 1,228 (152 slides) contain tissue folds with pixel masks, and no slide contributes both. Datasets, magnification routes and inclusion criteria are listed in Table S1 (Online Resource 3).

2.2 Detectors

Generative detectors. The detectors adapt PixCell-1024 [8], a latent diffusion model for histopathology, to artifact-free tissue with low-rank adaptation [20]. K2 follows the recipe of [9]: it is trained on 2,969 artifact-free patches of the training slides with the conditioning regularizer of [21]. For scoring, each 1,024-pixel patch at 10× is noised to step t = 800, and the artifact score is the squared noise-prediction error per latent position (one 8 × 8 µm grid cell); the noise seed is reset per slide. H2 adds to K2 an anchor term on clean patches and a hinge that raises the error at annotated artifact positions above the 99th percentile of clean error, with weight 0.1 (Online Resource 1). Unlike K2, H2 therefore uses artifact labels. H2 is the second hinge design; the first, without the anchor, inflated clean error during training and detected almost no artifacts (Online Resource 1, S3). K2 and H2 were trained per fold for in-distribution testing and on all 16 slides (K2all, H2all) for the benchmark and external sets, each with two seeds (42, 7).

Supervised probes. Each patch was divided into 4 × 4 tokens of 256 pixels and encoded with UNI2-h [10, 12] (class token, 1,536 dimensions). A standardized, L2-regularized logistic regression predicts whether a token contains an artifact, with the regularization strength chosen by cross-validation grouped by slide. A token is positive if annotated artifacts cover ≥ 25% of its tissue grid cells and negative if it contains no annotation; other tokens and teal-ink grid cells are excluded. Per-fold probes were used in distribution, and one probe trained on all 16 slides (probe_all) elsewhere. Two variants were trained: the same pipeline with Phikon-v2 [13] as a second encoder, and a scale-augmented probe, trained with an added resampled copy of every token (Section 2.4).

2.3 Decision rules

Each detector outputs a score map: reconstruction error for the diffusion detectors, artifact probability for the probe. A decision rule turns this map into a binary artifact mask, on which every endpoint is computed (Section 2.6).

- **R0 (generative maps):** two label-free levels, the 65th and 75th percentiles of the pooled error within tissue, clip a smoothed map (σ 2 grid cells), which is thresholded per slide with Otsu's method and cleaned by a 5 × 5 closing and opening.
- **Probe:** the smoothed probability map is thresholded at $\tau = 0.5$ with the same morphology.
- **Combination rules:** three rules combine the probe with a reconstruction detector. PC thresholds the generative error at a quantile over probe-clean tissue; PR applies Otsu within the dilated probe mask; BF combines the probe probability as prior with an error likelihood ratio. Each was selected on the development benchmark by leave-one-slide-out and then frozen (Online Resource 1).

2.4 Magnification handling

All inputs are processed at about 1.0 µm per pixel (10×). Nearest-pyramid-level selection had silently read four of our development slides at about 2.0 µm per pixel; they were re-read at 10×, all primary results use that read, and the earlier read is a sensitivity analysis (Table S5). External slides were read through an explicit guard: a level within 0.9–1.1 µm per pixel; otherwise a 0.5 µm level downsampled by 2; otherwise box-filter resampling by a factor up to 4; otherwise rejection.

Controlled resampling. To mimic a lower magnification on the same tissue, every 1,024-pixel patch was downsampled to 512 pixels and upsampled back (bilinear, both steps). Tissue gating and regions were kept from the original patches. On the development benchmark only the probe input was resampled (Δ-B); on Fırat both detectors' inputs were resampled (Δ-F).

2.5 Shift monitors

Three annotation-free statistics were computed per slide:

- **M1:** the probe's flag rate, the fraction of the scored tissue region that it marks as artifact;

- **M2:** the disagreement between the probe and H2all, as two fractions of the scored tissue region: grid cells flagged by the probe but not by H2all, and the reverse;
- **M3:** how far the slide's UNI2-h token embeddings lie from those of the training slides, as the median Mahalanobis distance (covariance estimated with Ledoit–Wolf shrinkage [22]).

Each set was compared with the development benchmark by the area under the ROC curve (AUROC), over the slides known to be clean (dataset class label or empty artifact mask) and over all slides. A pooled AUROC measured how well each monitor separated the sets where the probe fails from those where it holds.

2.6 Endpoints, calibration and statistics

Endpoints. Per-type sensitivity was measured inside a saturation-based tissue region (SAT), the scored region of [15], and clean-slide false positives (FP) within each dataset's tissue region. Pooled F1, precision and Dice were computed on artifact slides. Because SAT omits pale and blurred tissue, every conclusion was re-scored in an entropy-based tissue region [23] (Table S2). SlideGuard returns per-slide artifact Dice, precision, recall and artifact-free Dice. Masks were submitted once per arm, with arms, rules and analyses fixed before the first submission.

FP-matched comparison. The probe's sensitivity was traced against clean FP along its threshold curves, and generative arms were compared with it at their own FP.

Calibration. Probe calibration was measured at the token level on the benchmark, at the original read and after resampling, with tokens labelled as in probe training (Section 2.2). We report reliability tables (15 bins), expected calibration error (ECE) and Brier score [24]. Calibration was not computed for the generative detectors, whose scores are not probabilities, nor for the per-fold probes, whose models were not stored.

Uncertainty. Intervals come from a paired bootstrap over slides or cases, stratified by clean and artifact status (20,000 replicates, 95% percentile intervals). Because pooled metrics weight slides by their annotated area, we report Kish's effective sample size, $(\Sigma w)^2/\Sigma w^2$, with w the annotated pixels per slide [25]. A difference is called higher or lower only when its interval excludes zero for both seed pairs. Analyses added after the SlideInspect results are labelled post hoc.

2.7 Reproducibility and use of AI tools

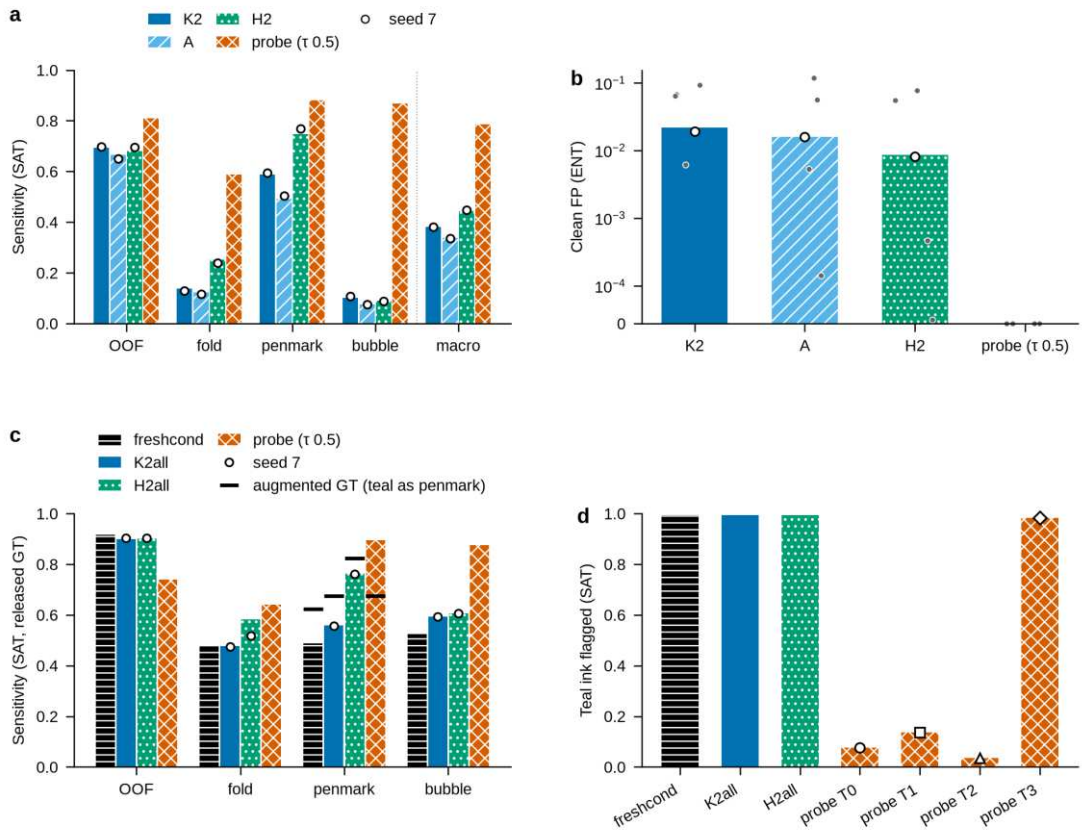
Every evaluation ran once from a version-controlled code state, and each run had to reproduce previously stored results before any new measurement. Code, model adapters, probe parameters and evaluation outputs are deposited on Zenodo (https://doi.org/10.5281/zenodo.23448122). Files derived from Phikon-v2 are shared under the Owkin non-commercial licence, as its ShareAlike term requires; embeddings are not released. A model card and dataset datasheet are given in Online Resource 2, and the completed TRIPOD+AI checklist [26] in Online Resource 5.

Use of AI tools. We used Claude (Anthropic; August–October 2026) through the claude.ai application and through claude.ai/code sessions with terminal access to our computing server, with default settings. The code sessions mainly used Claude Opus 5 and Opus 5.5, occasionally Claude Sonnet 5.5 and Claude Fable 5; the model was not recorded per application session. It drafted and revised the analysis, figure and table scripts, an internal laboratory log and the manuscript text; it summarised papers and drafted reference entries. The authors decided which analyses to adopt, approved every script and revised every section of the text. Every script was approved as a complete diff before it ran. Every reported number was traced to a stored output, and every reference was retrieved and checked by the authors. No histology image was generated or altered by an AI tool. Online Resource 6 gives the full disclosure with representative prompts. The authors take full responsibility for the content.

3 Results

3.1 In distribution the probe is the better detector

In four-fold cross-validation the probe found most annotated artifacts and produced no false positives at all on clean tissue (Fig. 2a). At its native threshold its mean per-type sensitivity was 0.787 (out-of-focus 0.810, fold 0.589, pen marking 0.880, air bubble 0.869), and it flagged none of the 3,446,579 unannotated grid cells in tissue of the four clean slides. The best diffusion detector, H2, reached 0.443 at a clean false-positive (FP) fraction of 0.0086, and K2 reached 0.380 at 0.0218. On the artifact slides the probe's precision and Dice were 0.905 and 0.880, against 0.654 and 0.466 for H2.



208

209 **Fig. 2** In-distribution comparison at the true 10× read. Bars are distinguished by colour and fill pattern
210 (legend); white circles mark the seed-7 replicate of each generative arm. **a** Cross-validation (16 slides,
211 four clean): sensitivity per artifact type inside the scored tissue region (SAT; Section 2.6), and the macro
212 mean of the four types, for the label-free baseline (K2), K2 with the anchor term and no hinge (A) and
213 the hinge model (H2), all at seed 42, and for the UNI2-h linear probe at its native threshold τ = 0.5, one
214 probe per fold. **b** Cross-validation: clean-slide false-positive fraction inside entropy-detected tissue
215 excluding teal ink (ENT), outside the annotation. Bars are pooled over the four clean slides, grey dots
216 are individual clean slides, and white circles are the seed-7 pooled values. The y axis is symmetric-
217 logarithmic (linear below 10^{-4}), so that zero values remain visible. The probe flags no clean grid cell (0
218 of 3,446,579), so it has no bar. **c** Development benchmark (24 slides): sensitivity per artifact type inside
219 SAT under the released annotations, for the flagship configuration of [15] (freshcond, hatched with
220 horizontal lines), the label-free baseline and the hinge model trained on all 16 slides (K2all, H2all; seed
221 42), and the UNI2-h probe at τ = 0.5. In the pen-marking group, black dashes give each arm's sensitivity
222 under the ink-augmented reference, in which teal marker ink counts as pen marking. **d** Development
223 benchmark: fraction of unannotated teal ink flagged by each arm's final mask, inside SAT, on the four
224 development slides that contain teal. The probe variants differ only in how training teal was labelled: T0
225 (the probe of panel c), teal grid cells excluded from both classes; T1, annotated teal admitted as
226 positive; T2, exclusion removed; T3, all training teal labelled as pen marking

The development benchmark gave the same picture (Table 1, Fig. 2b). The probe reached a pooled F1 of 0.816 with 2,724 false-positive grid cells on the 11 clean slides, whereas the diffusion detector of [15] reached 0.674 with 528,429. The F1 difference, +0.142, had an interval that included zero (−0.033 to +0.361); its lower end came from a single, mostly out-of-focus slide. Against GrandQC, on the 21 slides it processed, the probe was higher by +0.125 (+0.082 to +0.181). The hinge improved the diffusion detector: H2 − K2 was +0.059 and +0.063 in the two seed pairs, with both intervals above zero.

Table 1 Development benchmark (24 slides: 11 clean, 13 with artifacts), protocol of [15], true 10× read. F1: pooled F1 over grid cells inside SAT, under the released annotations (rel) and under the ink-augmented reference (aug, teal ink counted as pen marking). ΔF1: paired difference against the flagship configuration of [15] (freshcond), with 95% intervals from a slide bootstrap stratified by clean and artifact slides (20,000 replicates). Clean FP: flagged grid cells (8 × 8 µm; Section 2.1) on the 11 clean slides; the ratio is arm / freshcond. The generative arms use leave-one-slide-out selection over the percentile menu of [15], so their clean FP can differ between rel and aug. Probe rule 1 is τ = 0.5; rule 2 is τ selected by leave-one-slide-out cross-validation (LOOCV). Bottom block: paired contrasts between arms, in the two ΔF1 columns. † GrandQC (no class 3) produced output on 21 slides (8 clean); its ΔF1 and clean-FP ratio are against freshcond on the same 21 slides

Arm	F1 (rel)	ΔF1 vs freshcond (rel) [95% CI]	F1 (aug)	ΔF1 vs freshcond (aug) [95% CI]	Clean FP cells (rel)	Clean-FP ratio vs freshcond (rel) [95% CI]	Clean FP cells (aug)
freshcond	0.6741	—	0.7815	—	528,429	—	259,807
K2all	0.6946	+0.0205 [−0.0031, +0.0485]	0.7983	+0.0169 [+0.0005, +0.0364]	351,079	0.664 [0.269, 0.947]	185,547
K2all_s7	0.6895	+0.0154 [−0.0104, +0.0441]	0.7956	+0.0141 [−0.0092, +0.0360]	394,239	0.746 [0.519, 0.842]	159,157
H2all	0.7536	+0.0795 [+0.0221, +0.1599]	0.8371	+0.0556 [+0.0016, +0.1122]	125,176	0.237 [0.137, 0.276]	125,176
H2all_s7	0.7528	+0.0787 [+0.0256, +0.1583]	0.8404	+0.0589 [+0.0096, +0.1141]	156,281	0.296 [0.141, 0.445]	94,601
probe rule 1 (τ 0.5)	0.8157	+0.1416 [−0.0329, +0.3611]	0.7677	−0.0138 [−0.0880, +0.0810]	2,724	0.005 [0.000, 0.013]	2,724
probe rule 2 (LOOCV τ)	0.8027	+0.1287 [−0.0625, +0.3609]	0.7582	−0.0233 [−0.1033, +0.0723]	2,724	0.005 [0.000, 0.013]	2,724
GrandQC (21 slides)†	0.6905	+0.0069 [−0.1752, +0.2213]	0.7664	−0.0203 [−0.1953, +0.1442]	330,174	1.227	330,174
Paired contrasts (ΔF1 [95% CI])							
H2all − K2all		+0.0589 [+0.0195, +0.1154]		+0.0387 [−0.0003, +0.0782]			
H2all_s7 − K2all_s7		+0.0633 [+0.0261, +0.1194]		+0.0447 [+0.0121, +0.0832]			
H2all − probe rule 1		−0.0621 [−0.2069, +0.0630]		+0.0694 [−0.0046, +0.1148]			
H2all_s7 − probe rule 1		−0.0629 [−0.2056, +0.0634]		+0.0727 [−0.0007, +0.1185]			
probe rule 1 − GrandQC (21)†		+0.1251 [+0.0819, +0.1809]		+0.0012 [−0.1063, +0.1326]			

The advantage depended on what had been labelled. When teal marker ink was counted as pen marking, H2's F1 rose to 0.837–0.840 and the probe's fell to 0.768. The diffusion detectors flagged 99.4% of the unannotated teal ink, the probe only 7.6%. The ink never reached the probe's training labels; a probe retrained with it labelled flagged 98.4% of it (Fig. 3).

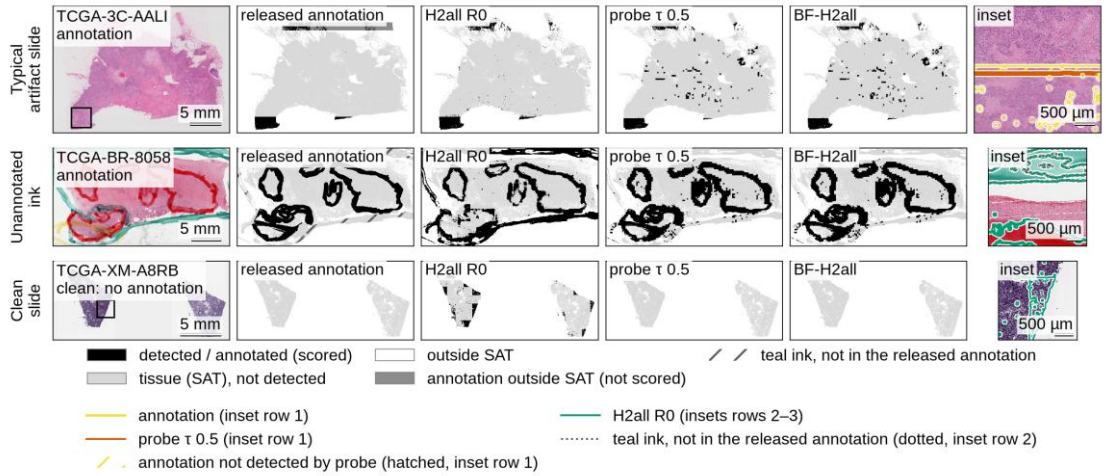


Fig. 3 Whole-slide examples on the development benchmark (true 10× read). Three slides were chosen by rule: top, the artifact slide with the median per-slide probe F1 (TCGA-3C-AALI, 7th of 13); middle, the slide with the most teal ink inside SAT outside the released annotation (TCGA-BR-8058); bottom, the clean slide with the largest H2all clean false positives (TCGA-XM-A8RB). Columns: the slide; the released annotation; H2all (R0 65/75); the probe (τ 0.5); BF-H2all (α 0.5); an inset, whose location is boxed in the first column. Masks are shown inside SAT: black, detected (or annotated); light grey, tissue not detected; white, outside SAT. In the annotation column, mid grey marks annotation outside SAT (not scored); in the middle row, hatching marks teal ink not in the released annotation. Insets show the largest region of annotation missed by the probe (top), of teal ink outside the annotation (middle) and of H2all false positives (bottom). Per-slide values inside SAT: top, F1 0.8985 (H2all), 0.7009 (probe), 0.6812 (BF-H2all), with 58% of the annotated grid cells outside SAT; middle, F1 0.5636, 0.8680 and 0.8699; bottom, clean false positives of 69,074 grid cells (H2all), 0 (probe) and 0 (BF-H2all). Scale bars 5 mm (500 µm in insets). Images shown without colour or contrast adjustment. Slides from The Cancer Genome Atlas (TCGA Research Network, https://www.cancer.gov/tcga; open-access data)

3.2 On a blind benchmark at matching magnification the probe kept its precision

SlideGuard slides were read at 0.91–1.01 µm per pixel without upsampling, so the probe saw tissue at its training scale. On the 28 non-TCGA slides its mean precision was 0.291, against 0.102 for H2 (difference +0.189, interval +0.092 to +0.301), while artifact Dice did not differ (+0.054, −0.016 to +0.130). In every cohort the probe had higher precision (Fig. 4b) and lower recall (0.11–0.33 against 0.37–0.76) than both diffusion detectors.

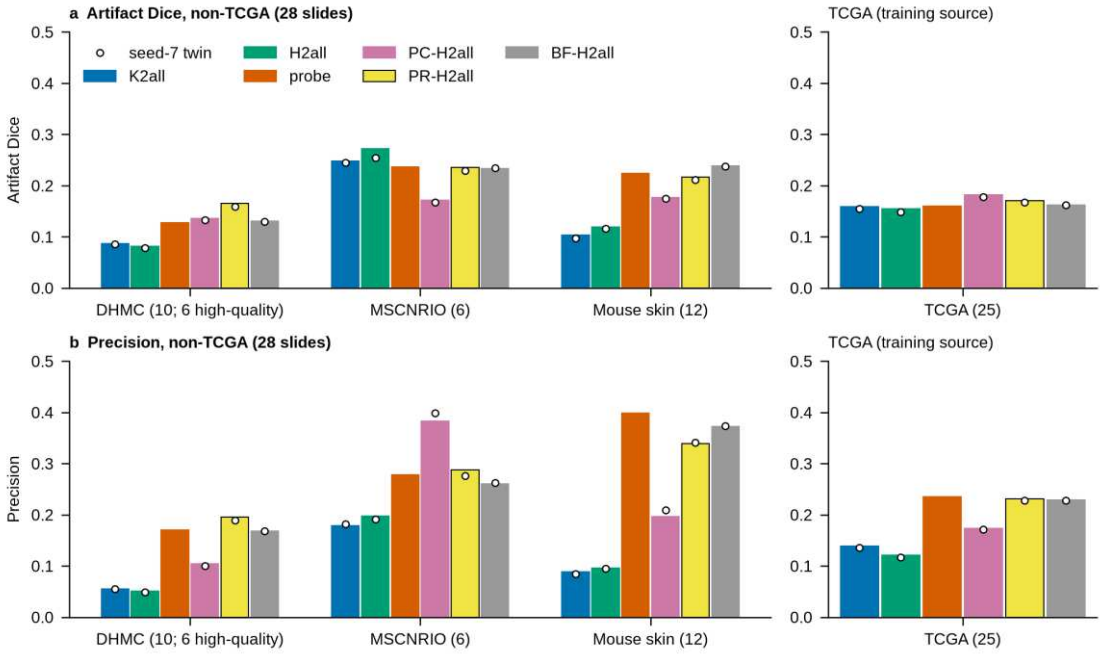


Fig. 4 SlideGuard: per-cohort artifact Dice (**a**) and precision (**b**). Left: the 28 non-TCGA slides by cohort (DHMC 10, of which 6 are high quality; MSCNRIO 6; mouse skin 12). Right: TCGA (25), drawn separately because our training slides and the diffusion backbone's pretraining data come from TCGA. Bars are per-slide means for K2all, H2all, the probe (τ 0.5) and the PC, PR and BF rules built on H2all (seed 42); white circles mark the seed-7 replicates. The DHMC means include the six high-quality slides, which carry little or no annotated artifact, so artifact Dice and precision on them are 0 or near 0 for every arm; Table 2 scores those six slides separately. Every slide was read at 0.91–1.01 µm per pixel without up-sampling

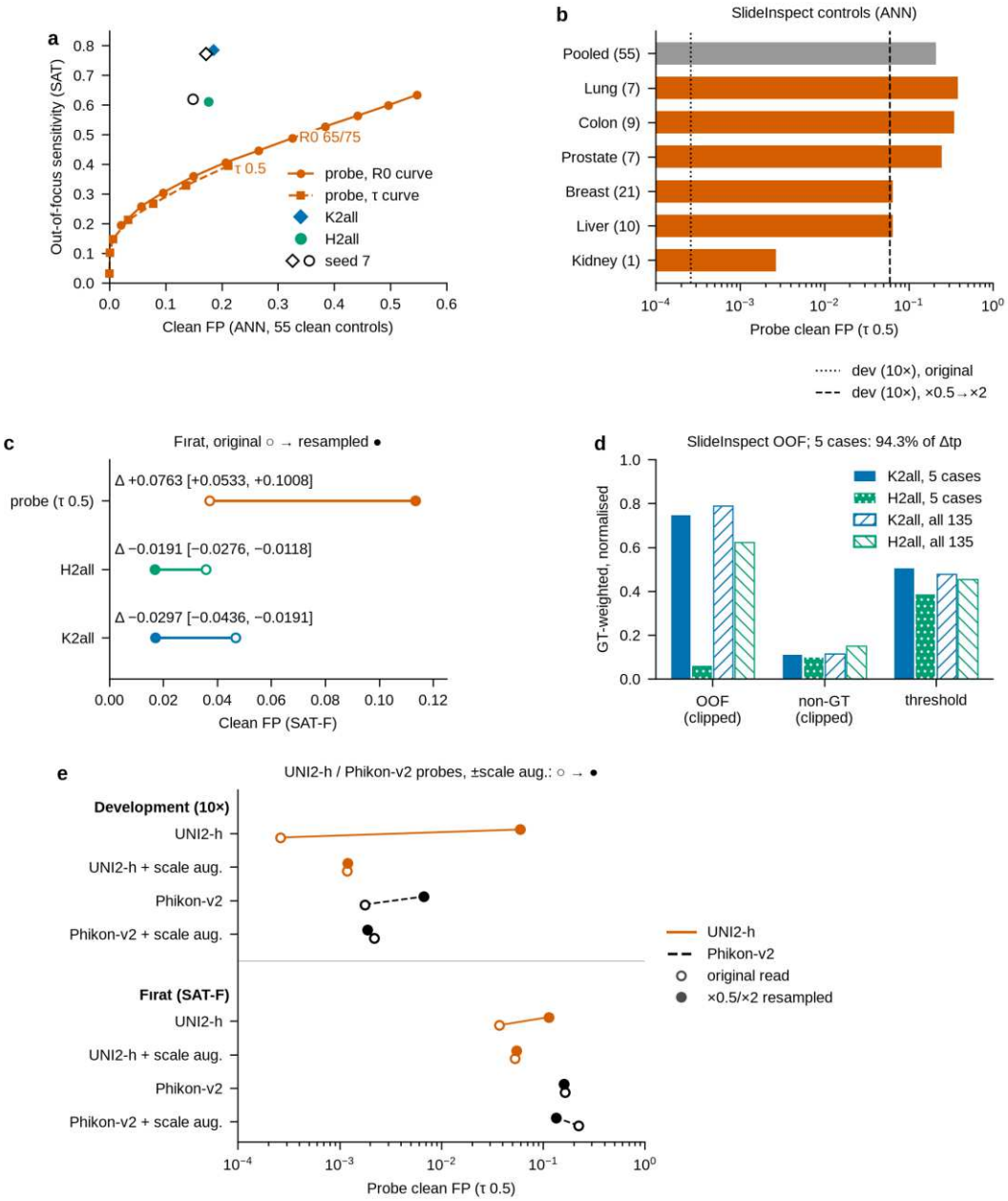
Table 2 places our detectors beside the seven methods published with the benchmark [5], on the same 47 slides. On the human cohorts every detector fell within the published range. On mouse skin the diffusion detectors fell below it (0.097–0.119 against a lowest published value of 0.154), while the probe fell within it (0.224). On the six high-quality slides the probe's artifact-free Dice, 0.9996, was above every published value (highest 0.9982). These placements are descriptive, because the published methods report no per-slide values.

Table 2 SlideGuard: our arms next to the seven published QC methods of [5]. Columns: artifact Dice by cohort on the 47 slides without the six high-quality DHMC slides (TCGA 25, MSCNRIO 6, DHMC 4, all human 35, mouse skin 12); last column, artifact-free Dice on the six high-quality DHMC slides (identified by the organizers, personal communication). Our arms: mean over slides [95% slide-bootstrap interval, 20,000 replicates]; probe-guided rules are built on H2all (seed-7 and K2-based arms in Table S7). Published values are cited from [5] (default settings there), not re-run; † marks patch-level methods. No difference to a published method is tested, because no per-slide values exist for them; our values are read against the published range only

Method	TCGA (25)	MSCNRIO (6)	DHMC non-HQ (4)	Human (35)	Mouse skin (12)	HQ DHMC (6), artifact-free Dice
Published methods (values cited, not re-run)						
HistoQC	0.1275	0.1992	0.1007	0.1367	0.1766	0.9283
Sem. Seg.	0.4206	0.3844	0.0370	0.3706	0.2037	0.7716
GrandQC (5×)	0.3357	0.5271	0.0703	0.3382	0.2566	0.9965
GrandQC (10×)	0.3536	0.3261	0.2326	0.3351	0.2445	0.9974
MoE DCNN	0.1393	0.3102	0.1272	0.1672	0.2267	0.9813
WSIsegQC	0.2004	0.2086	0.5440	0.2411	0.2230	0.9982
PathProfiler†	0.0928	0.2030	0.0405	0.1057	0.1538	0.5805
TCNN†	0.2259	0.3395	0.4772	0.2741	0.2015	0.9960
This study: mean over slides [95% CI]						
K2all	0.1584 [0.0778, 0.2557]	0.2479 [0.1321, 0.3643]	0.2162 [0.0099, 0.4225]	0.1804 [0.1124, 0.2566]	0.1033 [0.0451, 0.1717]	0.9258 [0.8695, 0.9648]
H2all	0.1545 [0.0798, 0.2444]	0.2718 [0.1503, 0.3901]	0.2022 [0.0113, 0.3930]	0.1800 [0.1161, 0.2514]	0.1189 [0.0511, 0.1979]	0.9286 [0.8796, 0.9653]
probe	0.1600 [0.0696, 0.2628]	0.2365 [0.1403, 0.3409]	0.3194 [0.0016, 0.6371]	0.1913 [0.1149, 0.2779]	0.2235 [0.1072, 0.3514]	0.9996 [0.9990, 1.0000]
PC-H2all	0.1823 [0.1003, 0.2768]	0.1715 [0.1235, 0.2250]	0.3398 [0.0225, 0.6570]	0.1984 [0.1277, 0.2784]	0.1770 [0.0894, 0.2778]	0.9858 [0.9822, 0.9900]
PR-H2all	0.1702 [0.0785, 0.2741]	0.2358 [0.1816, 0.2995]	0.4123 [0.1473, 0.6773]	0.2091 [0.1324, 0.2926]	0.2171 [0.0945, 0.3630]	0.9994 [0.9984, 0.9999]
BF-H2all	0.1616 [0.0703, 0.2660]	0.2338 [0.1446, 0.3232]	0.3264 [0.0035, 0.6494]	0.1928 [0.1160, 0.2795]	0.2382 [0.1116, 0.3803]	0.9996 [0.9990, 1.0000]

3.3 Under a magnification shift the ranking reversed

SlideInspect. On its 55 clean controls, distributed at 5× and upsampled, the probe flagged 21% of tissue, and at matched false positives H2's out-of-focus sensitivity exceeded the probe's by 0.23–0.28 (Fig. 5a). This shift, however, mixes magnification with stain, scanner and organ.



295

296 **Fig. 5** Behaviour under acquisition shift. **a** Out-of-focus detection on SlideInspect (289 cases, 55 clean
297 controls; native 5× material up-sampled ×2 to the working resolution). The y axis is out-of-focus
298 sensitivity inside SAT, pooled over the 289 cases (135 with out-of-focus annotation); the x axis is the
299 false-positive fraction inside the annotator's tissue mask (ANN), over the 55 clean controls. For the
300 probe, the solid line uses the label-free R0 rule (percentile pairs 45/55 to 95/99) and the dashed line
301 fixed thresholds τ = 0.5 to 0.99. The generative arms, K2all (diamond) and H2all (circle), are shown at R0
302 65/75, seed 42, with open markers for their seed-7 replicates. **b** Probe clean false positives (τ = 0.5) on
303 the SlideInspect clean controls by organ (number of controls in brackets), inside ANN; the pooled value
304 is in grey. The vertical lines are the probe's clean false positives on the 11 clean development slides,
305 inside SAT: dotted at the original read, dashed after ×0.5 → ×2 resampling of every patch (Δ-B).
306 Logarithmic axis. **c** Fırat clean false positives inside SAT-F (Table 3) before (open circles) and after (filled

circles) the same resampling (Δ-F), for K2all, H2all and the probe. Each change is given with its 95% interval (paired slide bootstrap stratified by clean and fold slides, 20,000 replicates). **d** Out-of-focus scores of K2all and H2all on SlideInspect. The bars are means weighted by each case's out-of-focus annotation and normalised to each arm's own R0 levels in SAT (p65 → 0, p75 → 1), for three quantities: the clipped score (the map on which Otsu operates) inside the out-of-focus annotation, the clipped score in SAT tissue outside all artifact annotations, and the per-case Otsu threshold. Filled bars are the five cases with the largest true-positive loss of H2all relative to K2all, which carry 94.3% of that loss; open hatched bars are all 135 out-of-focus cases. **e** Probe clean false positives before (open) and after (filled) the same resampling, for the UNI2-h probe (vermillion, solid) and the Phikon-v2 probe (black, dashed), each trained without and with scale augmentation: top, the 11 clean development slides at the true 10× read, inside SAT; bottom, the Fırat clean images, inside SAT-F. Logarithmic axis. Both evaluations use the resampling operator on which augmentation was trained

Controlled resampling. To isolate magnification we resampled every patch to half resolution and back, keeping the tissue unchanged. On the development benchmark the probe's clean false positives rose from 2,724 to 619,578 grid cells, a factor of 227, and its F1 fell from 0.82 to 0.49 (Fig. 5b). On Fırat at the original magnification, the probe and H2 produced the same clean FP fraction (0.037 and 0.036), and the probe led on fold sensitivity at matched false positives. After resampling, the probe's clean FP fraction rose to 0.114, in every organ and on 119 of 150 clean slides, with most of the rise from kidney and small intestine (Table S3); H2's fell to 0.017. The difference-in-differences was +0.095 (+0.071 to +0.122), and at matched false positives the order reversed: H2's fold sensitivity exceeded the probe's by 0.15–0.18 (Table 3, Fig. 6).

Table 3 External fold detection on the Fırat set (2,127 images from 302 slides: 150 clean, 152 with folds; 2,125 images scored after excluding two fold images with empty masks), read at 0.98 µm/px with the scale-bar rectangle excluded. Primary region SAT-F, the saturation-based tissue region (SAT) minus the scale-bar rectangle; secondary region COV-F, the whole image field minus the rectangle. Fold sensitivity and clean-slide false-positive fraction are pooled over grid cells; 95% intervals from a slide bootstrap stratified by clean and fold slides (20,000 replicates). **a** Original magnification: levels, contrasts between arms (H2 − K2 on fold sensitivity and clean FP; probe vs H2all clean FP), and fold sensitivity of each arm minus the probe's at the arm's own clean FP, read on the probe's label-free R0 curve. **b** Controlled resampling (Δ-F; each patch ×0.5 → ×2, bilinear; tissue gating, coverage and regions unchanged): generative R0 levels recomputed label-free, probe at τ = 0.5; the difference-in-differences of the change in clean FP (probe − H2all), the probe's resampled / original clean-FP ratio, and the FP-matched contrast on the probe's resampled R0 curve. τ-curve counterparts are in Table S4. ‡ 3 of 20,000 bootstrap replicates, whose target fell below the curve, are excluded. A dash marks a value not reported for that region

a Original magnification: levels

Arm	Fold sensitivity, SAT-F	Clean FP, SAT-F	Fold sensitivity, COV-F	Clean FP, COV-F

Arm	Fold sensitivity, SAT-F	Clean FP, SAT-F	Fold sensitivity, COV-F	Clean FP, COV-F
K2all	0.387 [0.336, 0.439]	0.0467 [0.0311, 0.0667]	0.430 [0.379, 0.483]	0.0891 [0.0699, 0.1103]
K2all_s7	0.373 [0.324, 0.424]	0.0472 [0.0322, 0.0662]	0.426 [0.376, 0.478]	0.0987 [0.0783, 0.1207]
H2all	0.446 [0.393, 0.498]	0.0359 [0.0236, 0.0510]	0.486 [0.435, 0.538]	0.0602 [0.0469, 0.0749]
H2all_s7	0.422 [0.374, 0.472]	0.0353 [0.0238, 0.0493]	0.477 [0.429, 0.528]	0.0625 [0.0486, 0.0783]
probe (τ 0.5)	0.533 [0.478, 0.583]	0.0372 [0.0245, 0.0523]	0.698 [0.643, 0.747]	0.0675 [0.0488, 0.0899]

a Original magnification: contrasts

Contrast	SAT-F [95% CI]	COV-F [95% CI]
H2all − K2all, fold sens	+0.0582 [+0.0436, +0.0726]	+0.0564 [+0.0387, +0.0741]
H2all_s7 − K2all_s7, fold sens	+0.0487 [+0.0353, +0.0617]	+0.0508 [+0.0333, +0.0677]
H2all − K2all, clean FP	−0.0109 [−0.0186, −0.0049]	−0.0289 [−0.0382, −0.0207]
H2all_s7 − K2all_s7, clean FP	−0.0119 [−0.0200, −0.0056]	−0.0362 [−0.0466, −0.0266]
probe − H2all, clean FP	+0.0013 [−0.0105, +0.0128]	+0.0074 [−0.0117, +0.0280]
probe − H2all_s7, clean FP	+0.0018 [−0.0107, +0.0143]	—
H2all − probe, fold sens at matched FP (R0)	−0.082 [−0.179, +0.007]	−0.173 [−0.262, −0.078]
H2all_s7 − probe (R0)	−0.103 [−0.203, −0.012]	−0.190 [−0.283, −0.090]
K2all − probe (R0)	−0.202 [−0.286, −0.109]	—
K2all_s7 − probe (R0)	−0.219 [−0.304, −0.124]	—

b Resampling (Δ-F): levels and changes

Arm	Clean FP, original → resampled	ΔClean FP [95% CI]	Fold sensitivity, original → resampled	ΔFold sensitivity [95% CI]
SAT-F				
K2all	0.0467 → 0.0170	−0.0297 [−0.0436, −0.0191]	0.3874 → 0.3707	−0.0166 [−0.0604, +0.0319]
H2all	0.0359 → 0.0168	−0.0191 [−0.0276, −0.0118]	0.4455 → 0.4705	+0.0250 [−0.0094, +0.0619]
probe (τ 0.5)	0.0372 → 0.1135	+0.0763 [+0.0533, +0.1008]	0.5331 → 0.6422	+0.1091 [+0.0972, +0.1220]
COV-F				
K2all	0.0891 → 0.0293	−0.0598 [−0.0748, −0.0465]	0.4300 → 0.3678	−0.0622 [−0.1103, −0.0078]
H2all	0.0602 → 0.0218	−0.0384 [−0.0471, −0.0301]	0.4864 → 0.5200	+0.0336 [+0.0001, +0.0702]
probe (τ 0.5)	0.0675 → 0.1946	+0.1271 [+0.0927, +0.1663]	0.6981 → 0.8295	+0.1315 [+0.1112, +0.1539]

b Resampling (Δ-F): contrasts

Contrast	SAT-F [95% CI]	COV-F [95% CI]
Difference-in-differences: change in clean FP, probe − H2all	+0.0954 [+0.0707, +0.1219]	+0.1654 [+0.1285, +0.2070]
Probe clean FP, resampled / original (ratio)	3.05 [2.51, 3.79]	2.88 [2.41, 3.52]
H2all − probe, fold sens at matched resampled FP (R0)‡	+0.1659 [+0.0571, +0.2585]	+0.1603 [+0.0480, +0.2599]
K2all − probe (R0)	+0.0638 [−0.0466, +0.1572]	−0.0428 [−0.1724, +0.0765]

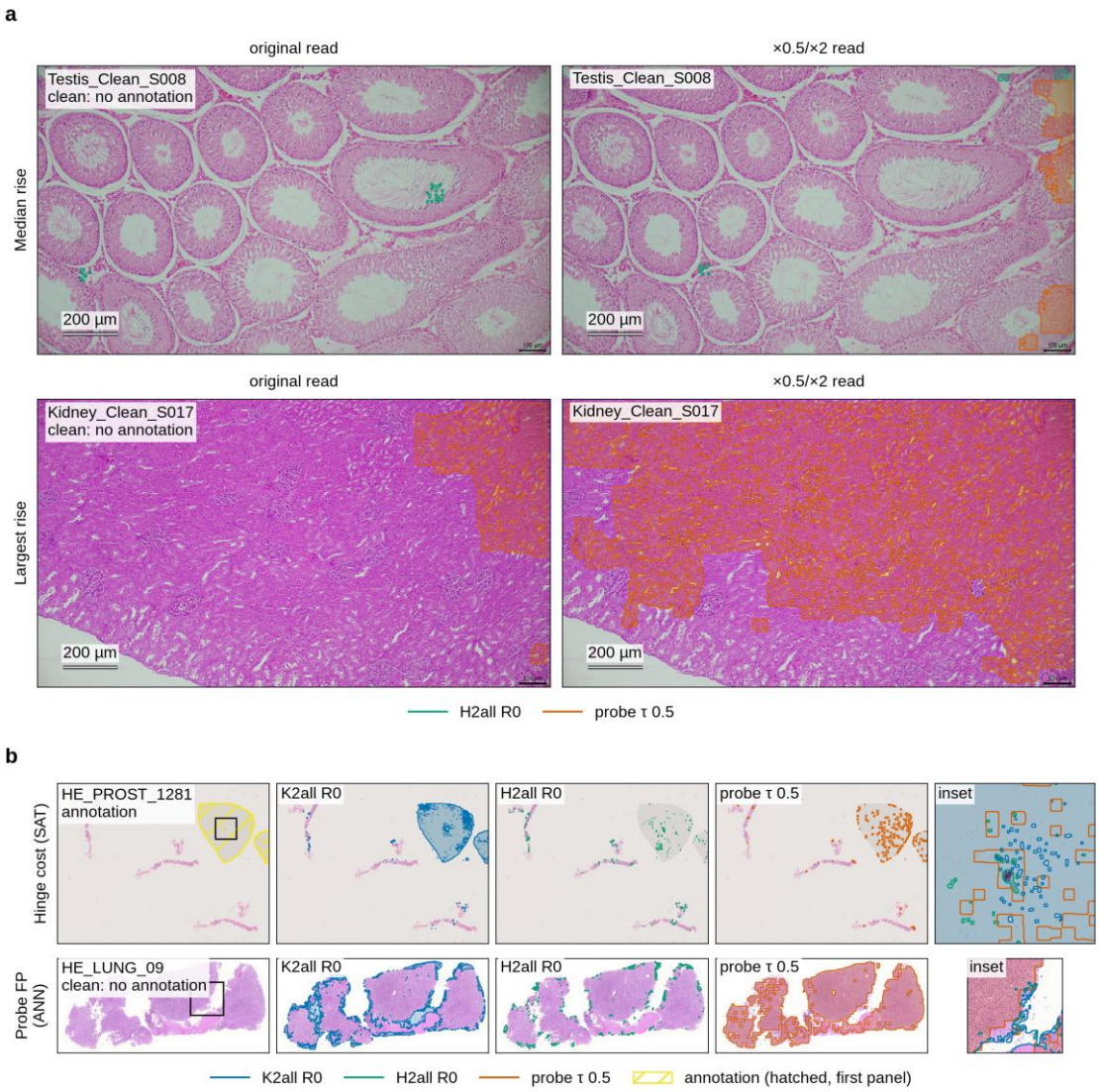


Fig. 6 Examples under acquisition shift. **a** Fırat clean images, original read (left) and after ×0.5 → ×2 resampling (right), with probe τ 0.5 (vermillion) and H2all R0 (green) overlays. Two slides were chosen by rule among the 119 clean slides whose probe-flagged area rose: the lower-median increase (Testis_Clean_S008, +3,372 grid cells) and the largest increase (Kidney_Clean_S017, +187,476 grid cells); within each slide, the image with the largest increase is shown. Flagged fraction of the image's SAT-F region, original → resampled: testis, probe 0 → 0.057, H2all 0.002 → 0.003; kidney, probe 0.110 → 0.746, H2all 0 → 0. The 200 µm scale bars are ours; the small 100 µm bar is part of the original images. Images from the Fırat dataset, Version 2 [19], CC BY 4.0; overlays added. **b** SlideInspect (native 5×, up-sampled ×2). Top: HE_PROST_1281, the case with the largest loss of out-of-focus true positives from K2all to H2all; out-of-focus sensitivity inside SAT: K2all 0.960, H2all 0.007, probe 0.149 (annotation hatched in the first panel). Bottom: HE_LUNG_09, the clean control with the largest probe false-positive fraction; false-positive fraction inside ANN: K2all 0.141, H2all 0.024, probe 0.770. Row label: Probe FP, probe false positives. Insets show the largest region of annotation detected by K2all and missed by H2all (top) and of probe false positives (bottom). No scale bar (pixel size not distributed with the

dataset). Images from SlideInspect [16], reproduced with permission of the dataset authors; these images are not covered by the article's Creative Commons licence

The hinge's cost appeared here. H2 lost 0.15–0.17 out-of-focus sensitivity against K2 on SlideInspect. Five cases, where H2 reconstructed the blur as normal tissue, accounted for 94–97% of the loss.

3.4 The effect depended on the encoder, and training reduced it

The same resampling raised the Phikon-v2 probe's clean false positives on the development benchmark 3.8-fold, against 227-fold for UNI2-h (Table S6). On Fırat its clean FP did not rise, but it was already high at our base magnification: 0.163, against 0.037 for UNI2-h.

Training the UNI2-h probe on resampled copies as well removed most of the effect. On the resampled benchmark its clean false positives fell from 619,578 to 12,596 grid cells and F1 returned to 0.80; on Fırat the difference-in-differences fell from +0.095 to +0.021. The price was paid at the original magnification: F1 −0.037 (−0.091 to −0.003) and 4.6 times more clean false positives (12,459 grid cells).

3.5 Resampling also broke the probe's calibration

At the original read the probe's probabilities were close to calibrated, with an expected calibration error (ECE) of 0.051. After resampling the ECE rose to 0.296 (difference +0.246, +0.151 to +0.324) and the Brier score rose from 0.039 to 0.196 (Table 4; Online Resource 4, Fig. S1). On clean slides the mean predicted artifact probability rose from 0.032 to 0.200, and the whole distribution moved up: 8.7% of clean tokens exceeded 0.5, against 0.04% before. The scale-augmented probe kept its calibration (ECE 0.031 and 0.033), with a slightly worse Brier score at the original read (0.046 against 0.039).

Table 4 Calibration of the UNI2-h probe at the token level on the development benchmark (24 slides, true 10× read), at the original read and after ×0.5/×2 resampling (Δ-B). probe_all is the probe used throughout and probe_aug the scale-augmented probe. ECE is the expected calibration error over 15 equal-width bins, and Brier the Brier score. On the 11 clean slides the observed artifact rate is zero, so the mean predicted probability equals the clean-slide ECE, and the interval shown is that ECE's interval. The 95% intervals come from a paired slide bootstrap stratified by clean and artifact slides (20,000 replicates). Labels follow the probe's training rule. The second block gives the paired differences, Δ-B minus original, over all slides

Levels

Probe	Read	Tokens	ECE [95% CI]	Brier [95% CI]	Mean p, 11 clean slides [95% CI]
probe_all	original	61,781	0.0507 [0.0292, 0.0806]	0.0391 [0.0290, 0.0534]	0.0322 [0.0189, 0.0475]
probe_all	resampled (Δ-B)	61,781	0.2963 [0.1834, 0.3993]	0.1960 [0.1263, 0.2688]	0.1999 [0.1276, 0.2817]
probe_aug	original	61,781	0.0312 [0.0148, 0.0687]	0.0460 [0.0312, 0.0660]	0.0316 [0.0119, 0.0558]
probe_aug	resampled (Δ-B)	61,781	0.0326 [0.0165, 0.0641]	0.0410 [0.0273, 0.0600]	0.0332 [0.0146, 0.0561]

Paired differences

Probe	Comparison	ΔECE [95% CI]	ΔBrier [95% CI]
probe_all	resampled − original, all slides	+0.2456 [+0.1510, +0.3237]	+0.1569 [+0.0833, +0.2310]

Probe	Comparison	ΔECE [95% CI]	ΔBrier [95% CI]
probe_aug	resampled − original, all slides	+0.0014 [−0.0113, +0.0092]	−0.0050 [−0.0093, +0.0001]

3.6 Rules combining the probe and a reconstruction detector did not beat the probe alone

Of six rules that use the probe to guide the diffusion error, only Bayesian fusion (BF) met the selection criterion on the development benchmark. BF reached F1 0.808 and 0.809 in the two seed pairs, against the probe's 0.816, and neither difference excluded zero. On SlideInspect every probe-guided rule fell to the probe's out-of-focus sensitivity (0.33–0.41, against 0.61–0.78 for R0). On SlideGuard, BF and PC raised H2's precision by 0.17–0.18 and 0.10, but no rule's artifact Dice separated from the probe's (Table 5).

Table 5 Probe-guided decision rules across four settings. Rows: R0 65/75, the probe-clean threshold (PC; q 0.95, p_c 0.05), probe-restricted Otsu (PR; m 16), Bayesian fusion (BF; α 0.5) and the probe at τ = 0.5. Rules are built on H2all, with the H2all_s7 value in parentheses; their configurations were selected on the development benchmark and frozen before external use. Development benchmark: pooled F1 (released reference) and false-positive grid cells on the 11 clean slides (true 10× read). Fırat: fold sensitivity and clean false-positive fraction (SAT-F; Table 3). SlideInspect: out-of-focus sensitivity and clean-control false-positive fraction. SlideGuard: means over the 28 non-TCGA slides, which include the six high-quality DHMC slides; these carry little or no annotated artifact, so artifact Dice and precision on them are 0 or near 0 for every arm (Table 2 scores them separately)

Rule	Development F1	Development clean FP (cells)	Fırat fold sensitivity	Fırat clean FP	SlideInspect out-of-focus sensitivity	SlideInspect clean FP	SlideGuard Dice	SlideGuard precision
R0 65/75	0.7547 (0.7587)	229,036 (189,584)	0.446 (0.422)	0.036 (0.035)	0.611 (0.620)	0.176 (0.149)	0.1382 (0.1322)	0.1017 (0.0990)
PC (q 0.95, p_c 0.05)	0.4184 (0.4114)	449,387 (433,754)	0.429 (0.408)	0.040 (0.039)	0.332 (0.333)	0.122 (0.136)	0.1613 (0.1580)	0.2037 (0.2107)
PR (m 16)	0.6363 (0.6324)	6,568 (6,744)	0.361 (0.350)	0.036 (0.036)	0.361 (0.359)	0.217 (0.225)	0.2025 (0.1965)	0.2767 (0.2729)
BF (α 0.5)	0.8079 (0.8085)	4,371 (4,056)	0.549 (0.547)	0.041 (0.040)	0.413 (0.414)	0.224 (0.224)	0.1988 (0.1982)	0.2756 (0.2766)
Probe (τ 0.5)	0.8157	2,724	0.533	0.037	0.394	0.211	0.1921	0.2910

3.7 On clean slides, the probe's behavior flagged the sets where it failed

On slides known to be clean, the fraction of tissue the probe flagged (M1) separated the two sets on which it failed, SlideInspect (distributed at 5×) and resampled Fırat, from the clean development slides: AUROC 0.936 (55 against 11 slides) and 0.894 (150 against 11). So did the fraction flagged by the probe but not by H2all (M2). The signal was graded: original Fırat, where the probe performed well, scored 0.79. Over all slides, including those with artifacts, M1 separated the sets less well (AUROC 0.687 and 0.656). Embedding distance to the training slides (M3) rose in every external set but did not separate the sets where the probe failed from those where it held (pooled AUROC 0.455, against 0.69–0.71 for M1 and M2).

3.8 Sensitivity to the scoring region and the read resolution

Of 88 comparison–metric pairs, 71 kept their direction and classification when scored in the entropy region instead of SAT; the probe's in-distribution advantage and the hinge's gains were among them (Table S2). Repeating the in-distribution analyses at the earlier, wrong read changed no conclusion (Table S5).

4 Discussion

At the magnification it was trained on, the foundation-model probe was the better QC detector. It made fewer false positives, ranked artifacts better, and kept its precision on a blind benchmark from four cohorts.

That advantage did not survive a change of scale, that is, of the physical size of tissue that each pixel covers. Resampling the same tissue to half resolution and back multiplied the probe's false positives on clean tissue and shifted its predicted probabilities upward, while the reconstruction detectors were unaffected or improved. The two encoders differed nearly 60-fold in this sensitivity. Phikon-v2 already flagged more clean tissue at our base magnification, consistent with its training at 20× [13]. Encoder features thus depend on scale, and a linear boundary learned at one scale can place clean tissue seen at another scale on the artifact side. UNI's developers reported that its features remained discriminative across input resolutions in region classification [10]; in our setting, by contrast, the probe was trained and applied at different scales.

For practice, this suggests four checks. First, verify the physical resolution of the pyramid level actually read, not its nominal magnification. Second, monitor the fraction of tissue that the probe flags on slides known to be clean. This monitor tracked failure better than M3, the distance of a slide's embeddings from the training slides, which also rose where the probe performed well. It needs known-clean slides: over all slides it separated the sets less well. Third, do not read the probe's probabilities at face value after a change of scanner or protocol, since its calibration degraded with its accuracy. Fourth, keep a reconstruction detector as a safeguard. It is less precise in distribution, but it did not fail under the shifts we tested, and it detects artifacts that were never labelled, such as marker ink outside the annotation scope.

Training on resampled copies (scale augmentation) removed most of the effect but lowered accuracy at the original magnification. It was tested against the same resampling it was trained on, so its benefit under real scanner differences remains to be shown. Combining the detectors moved the reconstruction detector toward the probe's operating point and, with it, toward the probe's failure.

Limitations. All models were trained on 16 TCGA slides, and the generative backbone and the second encoder were pretrained on TCGA images. The benchmark's 13 artifact slides carry very unequal amounts of annotation, so its pooled metrics rest on the equivalent of about six slides (Kish effective

size 6.2). Two of the training slides entered training at the earlier, wrong read. SlideInspect was examined repeatedly, and the resampling, encoder, augmentation, calibration and monitor analyses were added after its results. Only two encoders and one resampling operator were tested. The blind benchmark returns no per-type or clean-slide false-positive scores. Teal ink was excluded from the probe's labels, and calibration was measured on one benchmark only. Finally, the scored tissue region omits some pale and blurred annotated tissue; when re-scored in a second region, 17 of 88 comparisons depended on the region.

5 Conclusions

A linear probe on foundation-model features is an accurate artifact detector when slides are read at the magnification it was trained on. When the scale changes, it fails on clean tissue, its probabilities stop meaning what they say, and rules that lean on it fail with it. Reconstruction detectors were stable under the same shifts. Before deploying such a probe for quality control, laboratories should verify the read resolution, monitor its behavior on clean tissue, and keep a reconstruction detector, which needs no artifact labels, alongside it.

Statements and Declarations

Funding. This work was co-financed by the European Union and national funds through the National Strategic Reference Framework (NSRF) 2021–2027, under the Operational Programme "Competitiveness", Action "Bilateral Scientific and Technological Cooperation Greece–China" (Call Code 12KE, State Aid Code 11434), within the project "Research of Digital Pathology Image-based Intelligence Diagnosis Techniques of Prostate Cancer and Development of DT Software System - PRODIGY" (Grant Code: ΔΣΕΚΡ01-0021403, ERDF / European Regional Development Fund). The funder had no role in the study design, data analysis, decision to publish or preparation of the manuscript.

Competing interests. The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval. This study used only publicly available, de-identified image datasets and annotations: TCGA with the AIRAQc annotations, SlideInspect, the SlideGuard benchmark, the Dartmouth Lung Cancer Histology Dataset and the Fırat tissue-fold dataset. Each was used under its terms of use; no new human data or samples were collected, and no ethics approval was required.

Consent to participate / to publish. Not applicable. No identifiable participant information is included; pathology slide images are exempt from consent under the journal's policy.

Data availability.

- **Public datasets:**
  - TCGA diagnostic slides: Genomic Data Commons portal (https://portal.gdc.cancer.gov);

  - the AIRAQc annotations [14], distributed by their authors through Google Drive (https://drive.google.com/drive/folders/1Z3oNTqG2JZdbj--b0te8TMPzFlzhN0oN);
  - SlideInspect [16], public test set: Kaggle (https://www.kaggle.com/datasets/artemis90/slideinspect);
  - the Fırat histology tissue fold dataset [18, 19], Version 2: Zenodo, https://doi.org/10.5281/zenodo.22107895 (CC BY 4.0).
- **Controlled datasets:**
  - SlideGuard [5]: annotations are held by the organizers and scores are obtained through its evaluation server (Codabench competition 16457);
  - the Dartmouth Lung Cancer Histology Dataset [17]: available from its owners under a research-use agreement.
- **Our outputs** are deposited on Zenodo (https://doi.org/10.5281/zenodo.23448122): the analysis code; the LoRA adapters of the reconstruction detectors; the parameters of the UNI2-h and Phikon-v2 probes; the decision-rule configurations; per-slide evaluation outputs; and the source data of the figures and tables.
  - Score maps and embeddings are not included.
  - Our SlideGuard masks are included except for the Dartmouth-Hitchcock slides, for which only scores are reported, in line with the dataset's research-use agreement.
  - License: CC BY-NC 4.0, with notices for files derived from UNI2-h (UNI2-h model licence) and GrandQC (CC BY-NC-SA 4.0). Files derived from Phikon-v2 (probe parameters and evaluation outputs) are shared under the Owkin non-commercial licence, whose ShareAlike term requires the same terms.

Code availability. Analysis code is included in the Zenodo deposit (https://doi.org/10.5281/zenodo.23448122). The model weights used are publicly available from their providers: PixCell-1024, UNI2-h (revision d517a8dd) and Phikon-v2 (revision 2ae989a9). The LoRA adapters trained on PixCell-1024 are in the deposit, with the permission of the PixCell authors (personal communication, 18 Oct 2026); PixCell-1024 itself must be obtained under its own licence.

Author contributions (CRediT):

- **Konstantinos Moutselos:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft.
- **Ilias Maglogiannis:** Conceptualization, Supervision, Funding acquisition, Writing – review and editing.

Use of AI tools. Reported in Methods, Section 2.7.

Acknowledgments. Computing resources were provided by the University of Piraeus, NVIDIA DGX GPU Cluster. We thank Massimo Salvi (Politecnico di Torino) for permission to reproduce images from the SlideInspect dataset; Saeed Hassanpour (Dartmouth) for permission to submit predictions on, report scores for and show slides of the Dartmouth Lung Cancer Histology Dataset; Gabriela Kaczmarek and the SlideGuard organizers for evaluating our submissions and clarifying the benchmark's slide set; and

Srikar Yellapragada and the PixCell authors for permission to share the LoRA adapters trained on PixCell-1024.

Clinical trial number: not applicable.

References

1. Schömig-Markiefka B, Pryalukhin A, Hulla W, et al (2021) Quality control stress test for deep learning-based diagnostic model in digital pathology. Mod Pathol 34:2098–2108. https://doi.org/10.1038/s41379-021-00859-x
2. Wright AI, Dunn CM, Hale M, et al (2021) The effect of quality control on accuracy of digital pathology image analysis. IEEE J Biomed Health Inform 25:307–314. https://doi.org/10.1109/JBHI.2020.3046094
3. Janowczyk A, Zuo R, Gilmore H, et al (2019) HistoQC: an open-source quality control tool for digital pathology slides. JCO Clin Cancer Inform 3:1–7. https://doi.org/10.1200/CCI.18.00157
4. Weng Z, Seper A, Pryalukhin A, et al (2024) GrandQC: a comprehensive solution to quality control problem in digital pathology. Nat Commun 15:10685. https://doi.org/10.1038/s41467-024-54769-y
5. Kaczmarek G, Krawczyk-Borysiak Z, Miller M, et al (2026) SlideGuard: WSI artifact detection benchmark. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2026, Lecture Notes in Computer Science, vol 16895. Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-38470-6_28
6. Wang Z, Zhou Z, Wen Z, et al (2026) DiffusionQC: artifact detection and quality control in histopathology images via diffusion model. In: 2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI). IEEE, London, United Kingdom, pp 1–5. https://doi.org/10.1109/ISBI61048.2026.11515418
7. Linmans J, Raya G, van der Laak J, Litjens G (2024) Diffusion models for out-of-distribution detection in digital pathology. Med Image Anal 93:103088. https://doi.org/10.1016/j.media.2024.103088
8. Yellapragada S, Graikos A, Li Z, et al (2025) PixCell: a generative foundation model for digital histopathology images. arXiv. https://doi.org/10.48550/arXiv.2506.05127
9. Moutselos K, Maglogiannis I (2026) The effect of tissue detection on false positives of diffusion-based artifact detection in histopathology. arXiv. https://doi.org/10.48550/arXiv.2609.40083
10. Chen RJ, Ding T, Lu MY, et al (2024) Towards a general-purpose foundation model for computational pathology. Nat Med 30:850–862. https://doi.org/10.1038/s41591-024-02857-3
11. Guo L, Boucher D, Ouimet MI, Sun C (2026) Data-efficient pathology whole-slide image segmentation with slide-level conditioning. npj Imaging. https://doi.org/10.1038/s44303-026-00201-7
12. Mahmood Lab (2025) UNI2-h model weights, revision d517a8dd. Hugging Face. https://huggingface.co/MahmoodLab/UNI2-h
13. Filiot A, Jacob P, Mac Kain A, Saillard C (2024) Phikon-v2, a large and public feature extractor for biomarker prediction. arXiv. https://doi.org/10.48550/arXiv.2409.09173

14. Gautam K, Raipuria G, Singhal N (2025) AIRAQc: pre-analytical tool for accurate identification and quantification of artefacts in histopathology. In: Medical Imaging with Deep Learning – Short Papers. https://openreview.net/forum?id=XNNsQqs1UP

15. Moutselos K, Maglogiannis I (2026) Reliable benchmarking of artifact detection in computational pathology: a reproducibility and uncertainty analysis. arXiv. https://doi.org/10.48550/arXiv.2608.30835

16. Scotto M, Patti R, L'Imperio V, et al (2026) SlideInspect: from pixel-level artifact detection to actionable quality metrics in digital pathology. Int J Imaging Syst Technol 36:e70292. https://doi.org/10.1002/ima.70292

17. Wei JW, Tafe LJ, Linnik YA, et al (2019) Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. Sci Rep 9:3358. https://doi.org/10.1038/s41598-019-40041-7

18. Koparir OF, Tarakci Gencer B, Sengur A (2026) Deep learning-assisted quality control of histology teaching slides: detection and localization of tissue fold artifacts in H&E-stained images. Bioengineering 13:937. https://doi.org/10.3390/bioengineering13080937

19. Koparir OF, Tarakçı Gençer B, Sengur A (2026) Histology tissue fold dataset with pixel-level annotations for artificial intelligence-based detection and segmentation of tissue fold artifacts in H&E-stained teaching slides. Version 2, Zenodo. https://doi.org/10.5281/zenodo.22107895

20. Hu EJ, Shen Y, Wallis P, et al (2022) LoRA: low-rank adaptation of large language models. In: International Conference on Learning Representations

21. Moutselos K, Maglogiannis I (2026) Conditioning noise is a free regularizer for LoRA fine-tuning: no pathology encoder required for diffusion-based artifact detection in histopathology. arXiv. https://doi.org/10.48550/arXiv.2609.16032

22. Ledoit O, Wolf M (2004) A well-conditioned estimator for large-dimensional covariance matrices. J Multivar Anal 88:365–411. https://doi.org/10.1016/S0047-259X(03)00096-4

23. Zheng Y, Abila E, Chrenková E, et al (2026) LazySlide: accessible and interoperable whole-slide image analysis. Nat Methods 23:728–731. https://doi.org/10.1038/s41592-026-03044-7

24. Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. In: Precup D, Teh YW (eds) Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 70. PMLR, pp 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

25. Kish L (1965) Survey sampling. John Wiley & Sons, New York

26. Collins GS, Moons KGM, Dhiman P, et al (2024) TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 385:e078378. https://doi.org/10.1136/bmj-2023-078378

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf
- ESM2.pdf
- ESM3.pdf
- ESM4.pdf
- ESM5.pdf
- ESM6.pdf