POST-HOC ROBUSTIFICATION OF PATHOLOGY FOUNDATION MODELS WITH A VISION-LANGUAGE JUDGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Pathology foundation models are meant to describe tissue, yet their embeddings also record where and how a slide was acquired. Institution, scanner and staining leave batch effects strong enough that tiles can group by domain instead of by biology, which weakens transfer to new sites. Standard mitigation methods require large amounts of labeled training data, while label-free methods often struggle to remove domain-specific bias while preserving genuine biological signal. Multimodal large language models (MLLMs) can make robust judgements on biological similarity in images, but their utility in computational pathology is underexplored. To address the robustness issue in pathology foundation models, we propose RObustify PAthology models with a vision-Language judge (ROPAL), a label-free protocol in which MLLMs supervise their adaptation toward biological similarity. ROPAL asks an MLLM which tissue tiles are biologically alike across institutions and uses these judgements to fine-tune the foundation model, so that its embeddings group tissue by biology rather than by the domain it came from. We demonstrate the effectiveness of ROPAL through extensive empirical analysis across a diverse range of datasets, diseases, tasks, base models and robustification baselines. We release the code, training corpus and fine-tuned checkpoints.

1 INTRODUCTION

Pathology foundation models are the default feature extractor in computational pathology (Chen et al., 2024b; Zimmermann et al., 2024), yet their embeddings also record how a slide was acquired. Networks trained on TCGA recover the submitting institution with high accuracy (Howard et al., 2021; Dehkharghanian et al., 2021), almost every public model encodes the medical center more strongly than the tissue type (de Jong et al., 2025), and hospital and scanner signatures persist after stain normalization and where downstream AUCs look stable (Kömen et al., 2024; Thiringer et al., 2026). Retrieval can then return tiles from the query's own institution ahead of same-tissue tiles from elsewhere (Kömen et al., 2026; Grisi et al., 2026; Zhang et al., 2026c). When site correlates with the label, these batch effects become shortcuts (Geirhos et al., 2020), and PathoROB traces the robustness deficits of all 20 foundation models it evaluates to diagnostic errors (Kömen et al., 2026).

Existing corrections fall into two groups. Methods that protect biological signal train with task or class labels or multi-scanner pairs (Ganin et al., 2015; Zhang, 2026; Falahkheirkhah et al., 2023; Carloni et al., 2025; Vardi et al., 2026), and other work changes distillation, pretraining or fine-tuning objectives (Filiot et al., 2025; Kapse et al., 2026; Filiot et al., 2026). Label-free methods often struggle to remove domain-specific bias while preserving biological signal. Stain normalization acts on colour (Reinhard et al., 2001; Macenko et al., 2009), and site signatures persist beyond staining (Keller et al., 2023). ComBat needs an unseen site's features and batch metadata at test time (Johnson et al., 2007). Adversarial invariance and concept erasure remove site-predictive information, which can take biology correlated with the site along with it (Dincer et al., 2020; Ravfogel et al., 2020; Belrose et al., 2023; Su et al., 2025). What is missing is a label-free signal that says which tiles are biologically alike across institutions.

General-purpose multimodal large language models can provide it. Shown a few tiles, they can name the biological odd one out while ignoring stain, scanner and preparation. We show such a judge

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

triplets of two tiles from one slide and one from another institution, keep those in which it pairs the other-institution tile with a same-slide tile, and train a light adapter on an existing foundation model with the resulting cross-domain preferences. The adapter is then used without the judge.

Contributions.

- We introduce a protocol for improving the acquisition robustness of pathology foundation models using unlabeled histology images, without requiring annotated training data.
- We show empirically that general-purpose multimodal large language models can supply robustness supervision for pathology foundation models, by turning their judgments of biological similarity into training signals.
- We curate and release a training corpus spanning 26 disease cohorts and more than 60,000 patches for adapting pathology foundation models to domain shifts, along with the implementation code and fine-tuned model checkpoints.
- We benchmark different robustification methods on state-of-the-art pathology foundation models across eleven datasets, more than 25 institutions, more than 4,000 slides and more than 2.8 million image patches.

2 Related work

Measuring acquisition bias. Histology models trained on large public slide collections (Weinstein et al., 2013) can predict the submitting site of an image (Howard et al., 2021; Dehkharghanian et al., 2021). Robustness indices, per-sample margins and registered multi-scanner and multi-staining resources show that public pathology foundation models represent the medical center, scanner and staining strongly (de Jong et al., 2025; Kömen et al., 2026; Grisi et al., 2026; Ochi et al., 2024; Ryu et al., 2025; Carloni et al., 2025). We use PathoROB, PLISM, SCORPION and MOSAIC (Zhang et al., 2026c) to evaluate the effectiveness of different approaches for improving the robustness of histology models.

Mitigation grouped by the information it uses. Image-level methods normalize or vary stain appearance (Reinhard et al., 2001; Macenko et al., 2009; Shen et al., 2022; Shaban et al., 2018; Wagner et al., 2021; Kang et al., 2023), and center signal persists in foundation-model features after stain normalization (Kömen et al., 2024). Feature-level methods use batch or nuisance labels: ComBat harmonization gives the largest robustness-index gain among the post-hoc methods of PathoROB (Johnson et al., 2007; Murchan et al., 2024; Kömen et al., 2026), adversarial invariance and concept erasure remove site information from a representation (Edwards & Storkey, 2015; Louizos et al., 2015; Xie et al., 2017; Dincer et al., 2020; Cavinato et al., 2023; Belrose et al., 2023; Ravfogel et al., 2020; Zhao et al., 2019), and domain alignment is combined with task labels (Ganin et al., 2015; Zhang, 2026; Falahkheirkhah et al., 2023; Vardi et al., 2026). Paired scans, self-supervised continued training, robustness fine-tuning and concept-level interventions adapt the encoder or its features (Carloni et al., 2025; Su et al., 2025; Ye et al., 2025; Zhang et al., 2026b; Bøe et al., 2026; Filiot et al., 2026; Kim et al., 2026; Chen et al., 2024c; Huang et al., 2025; Henriksen et al., 2026; Zhu et al., 2026). ROPAL uses unlabeled tiles grouped by slide and institution and obtains the biological similarity signal from a multimodal judge.

Learning from similarity judgments. Human odd-one-out judgments yield interpretable object embeddings and support representation learning (Hebart et al., 2019; 2020; Zheng et al., 2019; Muttenthaler et al., 2022b; Mohammadi et al., 2020), and aligning vision models with human similarity judgments improves few-shot learning, robustness, perceptual metrics and downstream tasks (Muttenthaler et al., 2023; 2025; Fu et al., 2023; Sundaram et al., 2024; Garcia & Isik, 2025). Relational distillation and feature regularization keep the distances or features of a reference model during adaptation (Chen et al., 2017; Park et al., 2019; Tung & Mori, 2019; Mukhoti et al., 2023). We apply similarity-judgment adaptation to acquisition robustness in pathology, with triplets built across slides and institutions and a multimodal LLM in place of human raters.

Multimodal LLMs as teachers. Multimodal LLM judges show human-like discernment in pairwise comparisons (Chen et al., 2024a), and VLM preferences have trained reward models, image-

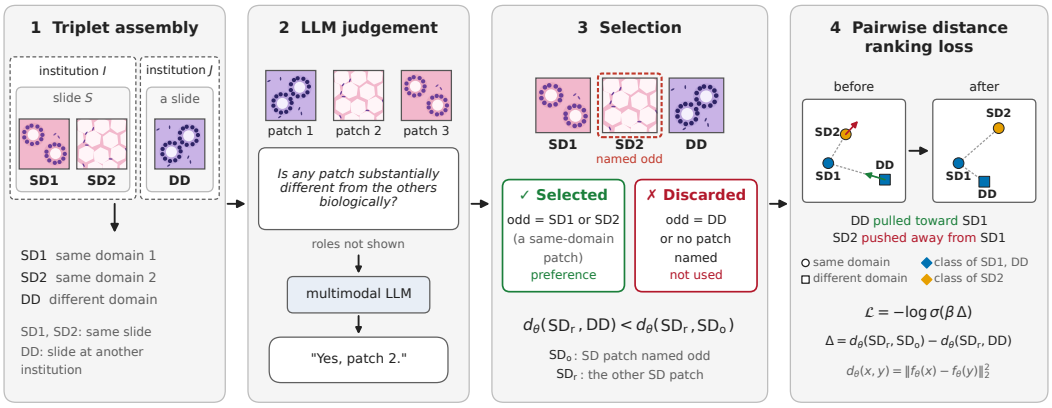


Figure 1: Overview of the protocol. (1) A triplet combines two same-domain patches from one slide, SD1 and SD2, and a different-domain patch DD from another institution. (2) Without knowing the roles, a multimodal LLM is asked whether any patch is substantially different from the others biologically. (3) The triplet is selected when it names SD1 or SD2 and discarded when it names DD or no patch. (4) The pairwise distance ranking loss moves DD closer to the remaining same-domain patch $\mathrm{SD_r}$ than to the one named odd, $\mathrm{SD_o}$ (marker shape: domain, colour: class).

compression models, image encoders and multimodal embeddings (Wang et al., 2024d; Sargent et al., 2025; He et al., 2025; Gu et al., 2025), while preference optimization has been applied to contrastive visual representations (Afzali et al., 2024). Other work makes the MLLM part of the embedding model, for example by encoding MLLM descriptions of histology patches, which reduces institution effects (Zhang et al., 2026d), or by training MLLM-based embedding models and retrieval data (Zhou et al., 2025b; Jiang et al., 2024; Zhou et al., 2025a). Our work uses an MLLM as the teacher of odd-one-out preferences on pathology tiles and distills them into an existing pathology foundation model, which is then used without the MLLM.

Appendix A.9 gives this section in full and Appendix A.10 an extended survey.

3 Method

ROPAL (RObustify PAthology models with a vision-Language judge) takes as input a pathology foundation model f_0 with frozen weights and unlabeled tiles whose only metadata are their source slide and which slides share an institution. No tissue-class, disease or diagnostic labels are used. The output is an adapted encoder f_θ, with $d_\theta(x, y) = \|f_\theta(x) - f_\theta(y)\|_2^2$. Figure 1 summarizes the protocol and Appendix A.1.1 gives the details.

3.1 Triplets and the role-blind judge

A triplet consists of two same-domain tiles, SD1 and SD2, from one slide and a different-domain tile DD from a slide at another institution. A multimodal LLM judge sees the three tiles in random order, without their roles, and answers whether two of them are substantially more similar in biological content to each other than to the third and, if so, which tile is the odd one out. A triplet is selected when the judge names SD1 or SD2 as the odd one. With $\mathrm{SD_o}$ the same-domain tile named odd and $\mathrm{SD_r}$ the other one, it gives the preference $d(\mathrm{SD_r}, \mathrm{DD}) < d(\mathrm{SD_r}, \mathrm{SD_o})$. A triplet is discarded when the judge names DD or no tile. Slide and institution membership only assemble the triplets and are given neither to the judge nor to the loss. The corpus holds the kept triplets of 26 TCGA cohorts (Weinstein et al., 2013).

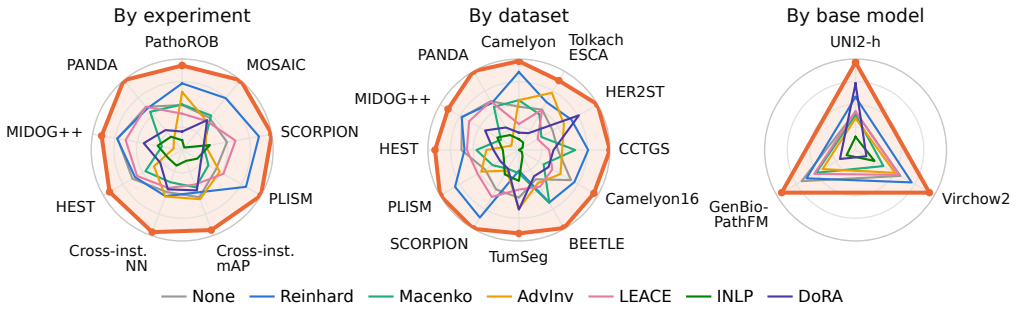


Figure 2: Mean normalized rank of each method by experiment, dataset and base model (outer ring = best). Methods are ranked within each per-dataset, per-base row of the main tables (APD by closeness to 0, ties averaged), and rank r of n scores $(n - r)/(n - 1)$.

3.2 Losses

Each kept triplet enters a Bradley–Terry (BT) likelihood (Bradley & Terry, 1952; Hunter, 2003) on the difference of squared distances,

$$\mathcal{L}_{\mathrm{BT}} = -\mathbb{E}\big[\log \sigma\big(\beta\,(d_\theta(\mathrm{SD_r}, \mathrm{SD_o}) - d_\theta(\mathrm{SD_r}, \mathrm{DD}))\big)\big], \tag{1}$$

where β scales the initial logits to order one. A second term keeps confident within-slide distance orderings of the frozen encoder. For two pairs p and q of tiles from one slide, with frozen margin $m_0 = d_0(q) - d_0(p)$ and adapted margin $m_\theta = d_\theta(q) - d_\theta(p)$,

$$\mathcal{L}_{\mathrm{rank}} = \mathbb{E}_{(p,q):\,|m_0|>\tau}\big[\log\big(1 + e^{-\gamma\,\mathrm{sgn}(m_0)\,m_\theta}\big)\big], \tag{2}$$

where τ is the median of $|m_0|$ over the slide's pairs of pairs and γ normalizes the margins. All entries labelled *ROPAL* use

$$\mathcal{L}_{\mathrm{ROPAL}} = \mathcal{L}_{\mathrm{BT}} + \lambda\,\mathcal{L}_{\mathrm{rank}}, \tag{3}$$

where λ weights the rank term. A low-rank adapter (Hu et al., 2021) on the frozen encoder is trained with this objective, and at test time the adapted encoder f_θ embeds each sample independently, without the judge.

4 Experiments

The experiments move from robustness benchmarks to scanner and stain invariance and then to downstream tasks.

Figure 2 summarizes how ROPAL ranks against the baselines across the main experiments (Sections 4.2–4.4), their datasets and base models, as an overview of the results detailed in the following subsections. ROPAL has the best mean rank on all 9 experiments, on all 12 datasets and on all 3 base models, and it ranks first in 88 of the 117 result rows. Among the other methods, Reinhard has the highest mean score (0.64, against 0.95 for ROPAL).

4.1 Setup

Problem setup. All methods train only on unlabeled TCGA tiles with their slide and institution membership and the pretrained base encoder, use no biological labels or evaluation data, and embed each test sample independently with a fixed mapping (Appendix A.2.1).

Base models and baselines. We adapt UNI2-h (Chen et al., 2024b), Virchow2 (Zimmermann et al., 2024) and GenBio-PathFM (Kapse et al., 2026). We compare with stain normalization (Reinhard (Reinhard et al., 2001), Macenko (Macenko et al., 2009)), removal of site information from the frozen embeddings (AdvInv (Dincer et al., 2020; Edwards & Storkey, 2015), LEACE (Belrose et al., 2023), INLP (Ravfogel et al., 2020)) and self-supervised parameter-efficient tuning (Pathology

Table 1: PathoROB robustness index (RI) and in-domain average performance drop (APD, percentage points), averaged over Camelyon and Tolkach ESCA. Higher is better. Bold: best; underlined: second best. Appendix Table 7 gives the full table.

Metric	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
RI	UNI2-h	0.734	<u>0.890</u>	0.882	0.862	0.720	0.687	0.864	**0.926**
RI	Virchow2	0.881	**0.939**	0.839	0.914	0.890	0.846	0.098	<u>0.935</u>
RI	GenBio-PathFM	0.911	0.930	0.903	<u>0.941</u>	0.874	0.640	0.406	**0.953**
APD	UNI2-h	-2.55	<u>-0.83</u>	-1.01	-1.82	-2.78	-5.10	-1.68	**-0.73**
APD	Virchow2	-1.09	<u>-0.98</u>	-2.18	-2.49	-2.51	-3.58	-10.65	**-0.67**
APD	GenBio-PathFM	-0.95	<u>-0.76</u>	-0.90	-1.29	-2.17	-5.71	-9.97	**-0.65**

DoRA (Bøe et al., 2026)), all under the same constraints. Their configurations are in Appendix A.11. In every table and figure, None is the frozen base encoder and ROPAL the encoder adapted with Eq. 3.

4.2 Robustness benchmarks

4.2.1 PathoROB robustness index on every base

PathoROB (Kömen et al., 2026) measures whether patch embeddings of Camelyon and Tolkach ESCA group by tissue class rather than by acquisition center, through the robustness index (RI) and the average performance drop (APD) of in-domain classifiers under confounded training (Table 1, protocol in Appendix A.2.2).

ROPAL has the highest robustness index on 2 of the 3 bases and the best APD on all 3 bases. The exception is RI on Virchow2 (Reinhard). Its robustness index exceeds 0.5 on every base, so same-class neighbours from other centers outnumber other-class neighbours from the same center. Label-free supervision from MLLM judgements therefore reduces the center bias of every base model, although a baseline is more robust on some of them.

4.2.2 Cross-domain similarity on MOSAIC

PathoROB scores neighbourhoods and classifiers. MOSAIC (Zhang et al., 2026c) tests the same property one comparison at a time: it compares a reference tile with a same-class tile from another slide (HER2ST, CCTGS) or institution (Camelyon16, BEETLE) and a different-class tile from the reference's slide or institution, and an embedding is correct when the same-class tile is closer. Figure 3 reports this cross-domain accuracy (protocol, within-domain and reverse-confound results in Appendix A.2.3).

ROPAL has the highest cross-domain accuracy on all 3 base models, 0.129 to 0.196 above the frozen encoder, and the highest within-domain accuracy on all 3 base models. Its cross-domain gain therefore does not come at the cost of within-domain accuracy. Across all methods and bases the two accuracies rise together (Pearson $r = 0.88$). In the reverse-confound test (Appendix A.5.3), where the same-class tile shares the reference's slide or institution, the same-domain bias (reverse minus forward accuracy) is 0.050 to 0.086 for ROPAL against 0.201 to 0.241 for the frozen encoders, while its reverse accuracy stays within 0.021 of theirs. The preference for same-domain tiles therefore shrinks on every base but does not vanish. Training on MLLM preferences thus teaches the encoder to rank biological similarity ahead of institutional differences while keeping within-domain discrimination, although a smaller same-domain preference remains.

4.3 Scanner and stain invariance

Institution effects mix scanner, staining and preparation. SCORPION (Ryu et al., 2025) and PLISM (Ochi et al., 2024) separate the first two: they provide registered scans of the same tissue under different scanners or stainings, and each tile embedding retrieves the scans of the same tissue under the other conditions by cosine similarity. Figure 4 reports recall at one on SCORPION and top-10 retrieval across stainings on PLISM (mean average precision and the smaller PLISM tile subset in Appendix Figure 6, protocols in Appendix A.2.4).

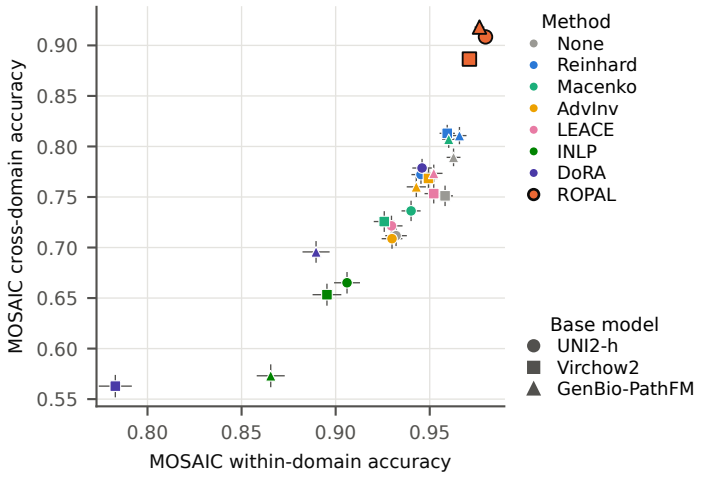


Figure 3: MOSAIC within-domain against cross-domain accuracy per base model (marker) and method (colour), each the mean over HER2ST, CCTGS, Camelyon16 and BEETLE, with bootstrap standard deviations (values in Appendix Tables 8 and 23).

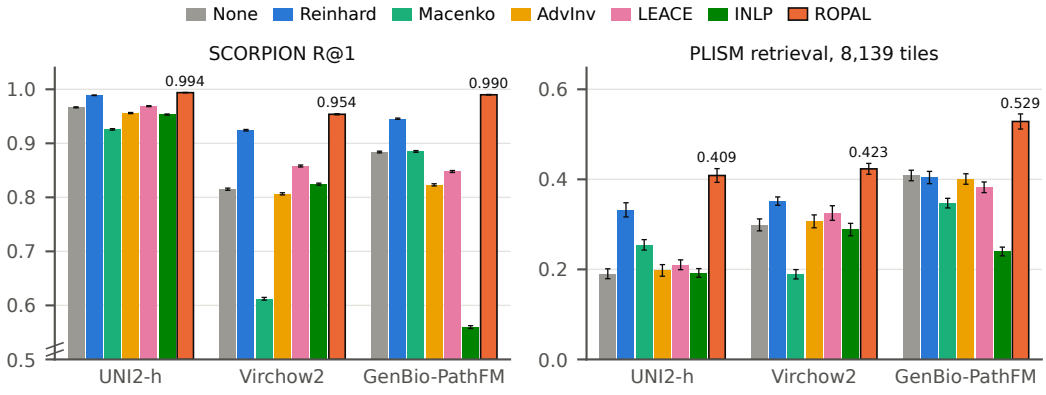


Figure 4: Scanner invariance on SCORPION (left: R@1 of retrieving the other scanners' scans of each tile) and stain invariance on PLISM (right: median top-10 retrieval accuracy across stainings at 8,139 tiles per slide). Higher is better, and error bars are bootstrap standard deviations. The SCORPION y-axis does not start at 0 (break marks). SCORPION mAP and PLISM at 2,713 tiles are in Appendix Figure 6.

ROPAL gives the best value on all 6 metric and base pairs of SCORPION and 6 tile-count and base pairs of PLISM (Figure 4 and Appendix Figure 6). Retrieval of the same tissue across scanners and across stainings is therefore more reliable after adaptation on every base. The robustness learned from cross-institution MLLM preferences thus extends to scanner and staining shifts.

4.4 Downstream tasks

Invariance helps only if the embeddings keep the biology that downstream models read, which four tasks across institutions and studies test.

4.4.1 Cross-institution retrieval and classification

The first task is closest to the training signal. On Camelyon16, BEETLE and TumSeg (Sancéré et al., 2025), tile embeddings retrieve same-class tiles and classify tiles by their nearest candidate across institutions, with or without institution–class confounding. Table 2 reports retrieval mean average precision under confounding and Figure 5 nearest-neighbour accuracy (protocol in Appendix A.2.5).

Table 2: Cross-institution retrieval mean average precision (percent) under institution–class confounding, averaged over UNI2-h, Virchow2 and GenBio-PathFM. Higher is better. Bold: best; underlined: second best. Appendix Table 11 gives the full table.

Dataset	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
Camelyon16	75.1	<u>78.4</u>	72.6	75.2	74.3	60.3	62.8	**86.1**
BEETLE	62.8	64.3	63.7	64.1	<u>64.3</u>	59.7	62.7	**72.7**
TumSeg	50.9	50.2	50.6	52.3	50.4	50.2	<u>52.4</u>	**53.2**

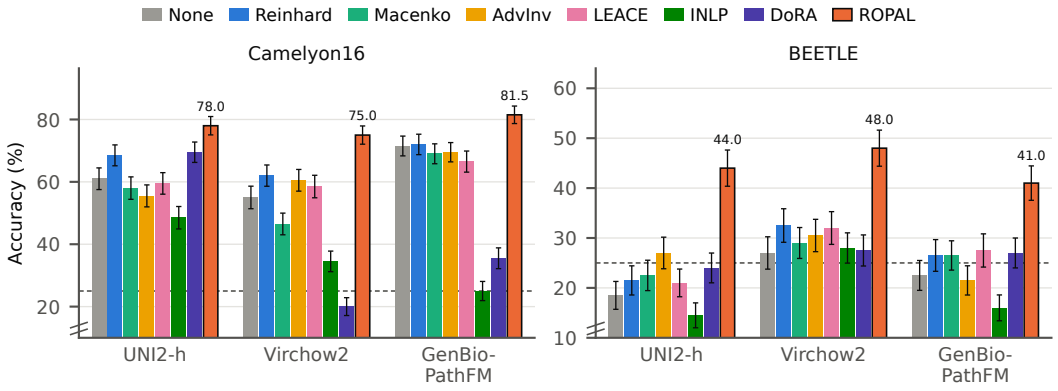


Figure 5: Fixed-candidate nearest-neighbour accuracy (percent) with institution–class confounding on Camelyon16 and BEETLE, per base model and method; each panel has its own accuracy axis. Error bars are bootstrap standard deviations and the dashed line marks chance (25).

Averaged over base models, ROPAL has the highest retrieval mAP on all 3 datasets, 2.3 to 11.0 points above the frozen encoder. Same-class tiles from other institutions are thus ranked ahead of different-class tiles from the query's own institution more often. The robustness gain therefore carries over to retrieving the same tissue class across institutions.

With institution–class confounding, ROPAL has the highest nearest-neighbour accuracy on all 6 dataset and base pairs. Its mean gain over the frozen encoder is 18.7 points. Tiles are thus matched more by tissue class than by institution even where the two are confounded, which is where site bias would otherwise act as a shortcut. Appendix Figure 7 shows the same comparison without confounding.

4.4.2 Cross-study gene-expression prediction on HEST

Gene expression is a finer biological readout than tissue class. On HEST-1k (Jaume et al., 2024), a nearest-neighbour regressor on patch embeddings predicts the gene expression of each spot, with the studies of each data-generating source held out in turn, and is scored by the Pearson correlation between predicted and measured expression (Table 3, protocol in Appendix A.2.6).

Table 3: HEST cross-study gene-expression prediction with kNN regression: Pearson correlation between predicted and measured expression, averaged over ten organs. Higher is better. Bold: best; underlined: second best. Appendix Table 13 gives the full table.

Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
UNI2-h	0.156	0.164	0.156	0.130	0.157	0.135	<u>0.167</u>	**0.206**
Virchow2	<u>0.177</u>	0.175	0.114	0.140	0.165	0.101	0.026	**0.201**
GenBio-PathFM	<u>0.160</u>	0.159	0.159	0.124	0.147	0.069	0.067	**0.198**

Averaged over organs, ROPAL gives the highest correlation on all 3 base models. Adaptation therefore keeps, and here increases, the gene-expression information that a nearest-neighbour regressor

Table 4: MIDOG++ mitotic figure classification: AUROC on the held-out tumor type or scanner (mean and worst fold), averaged over UNI2-h, Virchow2 and GenBio-PathFM. Higher is better. Bold: best; underlined: second best. Appendix Table 14 gives the full table.

Evaluation	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
Tumor type out – Mean AUROC	0.837	0.835	0.769	0.691	0.833	0.760	0.744	**0.844**
Tumor type out – Worst fold	0.785	0.784	0.695	0.631	0.779	0.690	0.692	**0.792**
Scanner out – Mean AUROC	0.838	0.839	0.768	0.698	0.834	0.761	0.733	**0.847**
Scanner out – Worst fold	0.782	0.790	0.697	0.629	0.772	0.710	0.681	**0.800**

Table 5: PANDA cross-site ISUP grading: balanced accuracy averaged over the two transfer directions. Higher is better. Bold: best; underlined: second best.

Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
UNI2-h	0.483	0.515	0.516	0.364	0.494	0.437	0.536	**0.580**
Virchow2	0.493	0.477	0.477	0.433	0.521	0.471	0.173	**0.537**
GenBio-PathFM	0.537	0.531	0.478	0.464	0.526	0.356	0.255	**0.543**

Table 6: Ablations, each cell the mean over UNI2-h and Virchow2 (per base in Appendix Table 21). Lower is better for Same-domain bias. Bold: best; underlined: second best.

Ablation	Setting	MOSAIC cross	MOSAIC within	Same-domain bias	SCORPION R@1	SCORPION mAP	PLISM 8,139	PLISM 2,713
Default	ROPAL	**0.896**	**0.972**	0.075	0.969	0.918	0.378	0.553
Choice of LLM	Qwen3.8-27B judge	0.887	0.968	0.079	**0.975**	**0.929**	0.381	0.554
Loss	BT only (no rank term)	0.888	0.964	**0.074**	0.973	0.923	**0.390**	**0.559**
	Domain adversarial + rank	0.612	0.842	0.263	0.487	0.332	0.012	0.023
Judge selection	Random	0.660	0.923	0.284	0.813	0.633	0.087	0.163
	MNN	0.848	0.962	0.128	0.937	0.850	0.289	0.456
	Odd-one-out score	0.812	0.964	0.170	0.928	0.819	0.268	0.430

can read from the patch embeddings. The robustness gained by ROPAL thus does not come at the cost of biological signal.

4.4.3 Mitotic figure classification across domains on MIDOG++

MIDOG++ (Aubreville et al., 2023) tests a fine morphological distinction under domain shift: a linear probe on its crops separates mitotic figures from hard-negative look-alikes, with one tumor type or one scanner held out at a time (Table 4, protocol in Appendix A.2.7).

Averaged over base models, ROPAL has the highest AUROC on all 4 evaluations, 0.007 to 0.018 above the frozen encoder. Macenko, AdvInv, INLP and DoRA fall below the frozen encoder in every evaluation. The gains therefore carry over to recognizing mitotic figures in unseen domains.

4.4.4 Cross-site prostate grading on PANDA

Finally, PANDA moves from tiles to slides: an attention-based aggregator of tile embeddings assigns the ISUP grade of PANDA prostate biopsies (Bulten et al., 2022), trained at one site and tested at the other (Table 5, protocol in Appendix A.2.8).

ROPAL has the highest balanced accuracy on all 3 base models. The robustness gain therefore also carries over to prostate grading at an unseen site.

5 Ablation studies

Each ablation changes the judge LLM, the loss or the selection of training triplets in a matched setting on UNI2-h and Virchow2 (Table 6, details in Appendix A.3.1). An open-weight judge stays within 0.011 of gpt-5.6-sol on every task, the rank term is not significantly better than BT only (one-sided sign test, $p = 0.87$) and replacing the preference term with a site adversary lowers every score.

The Random, MNN and Odd-one-out score selections fall below ROPAL on every task, so MLLM judgements are the source of the robustness supervision. MOSAIC cross accuracy rises with every increase in corpus size on both bases (Appendix Figure 8), so within the tested range more label-free MLLM judgements give a more robust encoder.

6 LIMITATIONS

On registered PLISM slide pairs that differ only in scanner, the matching tile must be among the ten nearest tiles in the other scan (Appendix Table 20). Here ROPAL gives the best value on 2 of the 6 tile-count and base pairs. Reinhard is best in the other 4. Invariance to scanner changes on PLISM therefore remains a limitation of ROPAL. Scans of one staining image the same section and differ mainly in global lightness, colour and contrast, which the per-channel match of Reinhard normalization removes, unlike the differences between stainings (examples in Appendix Figure 9, Appendix A.5.4).

7 CONCLUSION

We adapted pathology foundation models with role-blind multimodal judgments of biological similarity, collected without labels and learned as cross-domain preferences. The adapted encoders improve on the frozen encoders on robustness benchmarks, scanner and stain invariance and downstream tasks, and the ablations trace the gains to the judge's selection of training triplets. In future work, we plan a large-scale manual histopathological annotation with pathologist audit of the training corpus to better understand the biological basis of these gains.

USE OF GENERATIVE AI

In this work, we used generative AI tools to implement methods, including the data processing, training and evaluation code, to clean and reformat datasets, to provide feedback on the design of experiments and to help interpret results. We have not used generative AI tools to generate synthetic data sets, develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, or propose or refine hypotheses, and translation and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create and edit software code and scientific figures, suggest experimental parameters, draft and edit parts of the paper, summarize and search the literature and format references. Separately from this disclosure, a multimodal LLM is part of the proposed method: its odd-one-out judgments are the training supervision (Section 3). We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Dyah Adila, Changho Shin, Linrong Cai, and Frederic Sala. Zero-shot robustification of zero-shot models. In *International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2309.04344.

Amirabbas Afzali, Borna Khodabandeh, Ali Rasekh, Mahyar JafariNodeh, Sepehr kazemi, and Simon Gottschalk. Aligning Visual Contrastive learning models via Preference Optimization. In *International Conference on Learning Representations*, volume abs/2411.08923, 2024. doi: 10.48550/arXiv.2411.08923.

Sameer Agarwal, Josh Wills, Lawrence Cayton, Gert R. G. Lanckriet, David J. Kriegman, and Serge J. Belongie. Generalized non-metric multidimensional scaling. In *International Conference on Artificial Intelligence and Statistics*, pp. 11–18, 2007.

Hafsa Akebli and Vincenzo Della Mea. Domain Generalization of Histopathology Foundation Models in Multicenter, Multi-Scanner Cohorts: A Comparative Benchmark. *Journal of Imaging*, 12, 2026. doi: 10.3390/jimaging12080381.

Maximilian Alber, Timo Milbich, Alexandra Carpen-Amarie, Stephan Tietz, Jonas Dippel, Lukas Muttenthaler, Beatriz Perez Cancer, A. Benetti, P. Korfiatis, Elias Eulig, J'erome Luscher, Jia-Sen Wu, Sayed Abid Hashimi, G. Dernbach, S. Schallenberg, Neelay Shah, Moritz Krugener, A. Jammoria, J. Matras, P. Duffy, M. Redlon, Philipp Jurmeister, David Horst, Lukas Ruff, Klaus-Robert Muller, Frederick Klauschen, and Andrew Norgan. Atlas 2 – foundation models for clinical deployment. *arXiv.org*, abs/2601.05148, 2026. doi: 10.48550/arXiv.2601.05148.

Saghir Alfasly, Ghazal Alabtah, Sobhan Hemati, Krishna Rani Kalari, and H. R. Tizhoosh. Zero-Shot whole slide image retrieval in histopathology using embeddings of foundation models. *arXiv.org*, abs/2409.04631, 2024. doi: 10.48550/arXiv.2409.04631.

Jonas Ammeling, Jonathan Ganz, Emely Rosbach, Ludwig Lausser, Christof A. Bertram, Katharina Breininger, and Marc Aubreville. Benchmarking foundation models for mitotic figure classification. *Machine Learning for Biomedical Imaging*, abs/2508.04441:38–58, 2025. doi: 10.59275/j.melba.2026-a3eb.

Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Z. Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, Jonas Frisén, Camilla Engblom, and Joakim Lundeberg. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature Communications*, 12(1):6012, 2021. doi: 10.1038/s41467-021-26271-2.

Martín Arjovsky, L. Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. arXiv preprint arXiv:1907.02893, 2019.

Duygu Arslan, Sina Sehlaver, Erce Guder, M. A. Temena, Alper Bahcekapili, Umut Ozdemir, Duriye Ozer Turkay, Gunes Guner, Servet Guresci, C. Sokmensuer, Emre Akbas, and A. Acar. Colorectal cancer tumor grade segmentation: A new dataset and baseline results. *Heliyon*, 11(4), 2025. doi: 10.1016/j.heliyon.2025.e42467.

Maryam Asadi-Aghbolaghi, Amirali Darbandsari, Allen Zhang, Alberto Contreras-Sanz, Jeffrey Boschman, Pouya Ahmadvand, Martin Köbel, David Farnell, David G. Huntsman, Andrew Churg, Peter C. Black, Gang Wang, C. Blake Gilks, Hossein Farahani, and Ali Bashashati. Learning generalizable AI models for multi-center histopathology image classification. *npj Precision Oncology*, 8, 2024. doi: 10.1038/s41698-024-00652-4.

Maria Attarian, Brett D. Roads, and Michael C. Mozer. Transforming neural network visual representations to predict human judgments of similarity. *arXiv.org*, abs/2010.06512, 2020.

M. Aubreville, N. Stathonikos, C. Bertram, Robert Klopfleisch, Natalie ter Hoeve, F. Ciompi, Frauke Wilm, C. Marzahl, T. Donovan, A. Maier, Jack Breen, N. Ravikumar, Youjin Chung, Jinah Park, Ramin Nateghi, Fattane Pourakpour, R. Fick, S. Ben Hadj, M. Jahanifar, Adam Shephard, Jakob Dexl, T. Wittenberg, Satoshi Kondo, Maxime W. Lafarge, V. Koelzer, Jing Liang, Yubo Wang, Xing-An Long, Jingxin Liu, Salar Razavi, A. Khademi, Sen Yang, Xiyue Wang, Ramona Erber, Andrea Klang, Karoline Lipnik, Pompei Bolfa, Michael J. Dark, Gabriel Wasinger, M. Veta, and Katharina Breininger. Mitosis domain generalization in histopathology images – the MIDOG challenge. *Medical Image Anal.*, 84:102699, 2022. doi: 10.1016/j.media.2022.102699.

M. Aubreville, Frauke Wilm, N. Stathonikos, Katharina Breininger, T. Donovan, S. Jabari, M. Veta, J. Ganz, J. Ammeling, P. V. van Diest, R. Klopfleisch, and C. Bertram. A comprehensive multi-domain dataset for mitotic figure detection. *Scientific Data*, 10(1):484, 2023. doi: 10.1038/s41597-023-02327-4.

R. G. Bahumanya, M. HarshithV., S. Gowda, and R. AnalaM. Explanation stability of Test-Time adaptation in computational pathology: A Large-Scale benchmark. *arXiv preprint arXiv:2608.07062*, 2026.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, John Kernion, Andy Jones, A. Chen, Anna Goldie, Azalia Mirhoseini, C. McKinnon, Carol Chen, Catherine Olsson, Chris Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, E. Perez, Jamie Kerr, J. Mueller, Jeffrey Ladish, J. Landau, Kamal Ndousse, Kamilė Lukošiūtė, Liane Lovitt, M. Sellitto, Nelson Elhage, Nicholas Schiefer, Noem'i Mercado, Nova Dassarma,

R. Lasenby, Robin Larson, Sam Ringer, Scott Johnston, S. Kravec, S. E. Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, T. Henighan, Tristan Hume, Sam Bowman, Zac Hatfield-Dodds, Benjamin Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom B. Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback. *arXiv.org*, abs/2212.08073, 2022. doi: 10.48550/arXiv.2212.08073.

Rohan Bareja, Francisco Carrillo-Perez, Yuanning Zheng, Marija Pizurica, Tarak Nath Nandi, Lu Tian, Jeanne Shen, Ravi Madduri, and Olivier Gevaert. A benchmark study of vision and pathology foundation models for computational pathology. *Nature Communications*, 17, 2026. doi: 10.1038/s41467-026-76004-6.

Babak Ehteshami Bejnordi, M. Veta, Paul Johannes van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. van der Laak, M. Hermsen, Quirine F. Manson, Maschenka C. A. Balkenhol, Oscar G. F. Geessink, Nikolaos Stathonikos, Marcory van Dijk, P. Bult, F. Beca, Andrew H. Beck, Dayong Wang, A. Khosla, Rishab Gargeya, Humayun Irshad, Aoxiao Zhong, Q. Dou, Quanzheng Li, Hao Chen, Huangjing Lin, P. Heng, C. Hass, Elia Bruni, Quincy Wong, U. Halici, Mustafa Ümit Öner, R. Cetin-Atalay, Matt Berseth, V. Khvatkov, A. Vylegzhanin, Oren Z. Kraus, M. Shaban, N. Rajpoot, Ruqayya Awan, K. Sirinukunwattana, Talha Qaiser, Yee-Wah Tsang, David Tellez, Jonas Annuscheit, P. Hufnagl, Mira Valkonen, K. Kartasalo, L. Latonen, P. Ruusuvuori, Kaisa Liimatainen, Shadi Albarqouni, Bharti Mungal, A. George, S. Demirci, N. Navab, Seiryo Watanabe, S. Seno, Y. Takenaka, H. Matsuda, Hady Ahmady Phoulady, V. Kovalev, A. Kalinovsky, V. Liauchuk, G. Bueno, M. Fernández-Carrobles, I. Serrano, O. Deniz, Daniel Racoceanu, and Rui Venâncio. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Journal of the American Medical Association (JAMA)*, 318(22):2199–2210, 2017. doi: 10.1001/jama.2017.14585.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66044–66063, 2023. doi: 10.48550/arXiv.2306.03819.

A. Bentaieb and G. Hamarneh. Adversarial stain transfer for histopathology image analysis. *IEEE Transactions on Medical Imaging*, 37:792–802, 2018. doi: 10.1109/TMI.2017.2781228.

D. Biderman, J. Portes, Jose Javier Gonzalez Ortiz, M. Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA Learns Less and Forgets Less. *Trans. Mach. Learn. Res.*, 2024, 2024. doi: 10.48550/arXiv.2405.09673.

Vilde Schulerud Bøe, Andreas Kleppe, Sebastian Foersch, Daniel-Christoph Wagner, Lill-Tove Rasmussen Busund, and Adín Ramírez Rivera. Low-rank adaptations for increased generalization in foundation model features. In *Proceedings of the MICCAI Workshop on Computational Pathology*, volume 316 of *Proceedings of Machine Learning Research*, pp. 234–247. PMLR, 2026. Presented at the 2025 workshop; proceedings bibliographic year is 2026.

S. Bonn, M. Zimmermann, G. Sauter, E. Bengtsson, T. Huber, J. Baumbach, M. Lennartz, P. Fuhlert, and A. Witte. Cracks in the Foundation: How Data-Hungry and Sensitive to Domain Shift are Vision Foundation Models for Computational Pathology? *medRxiv*, 2026. doi: 10.64898/2026.01.06.25342815.

Christoph Bosch, John K. L. Wong, Martin Paulikat, Myroslav Zapukhlyak, Bharti Arora, Manasi Aichmüller-Ratnaparkhe, Jens Baumann, Shivani Karn, Rutuja Kamble, Swapnil Karnik, Bhushan Khedkar, Serey Vathana Chhut, Witali Aswolinskiy, and Christian Aichmüller. Diversity over scale: Whole-Slide image variety enables h&e foundation model training with fewer patches. *Journal of Pathology Informatics*, 21:100648, 2025. doi: 10.1016/j.jpi.2026.100648.

Sofiène Boutaj, Marin Scalbert, Pierre Marza, Florent Couzinie-Devy, Maria Vakalopoulou, and Stergios Christodoulidis. Controllable Latent Space Augmentation for Digital Pathology. In *IEEE International Conference on Computer Vision*, pp. 22165–22174, 2025. doi: 10.1109/ICCV51701.2025.02058.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324, 1952. doi: 10.2307/2334029.

Jack Breen, Katie Allen, Kieran Zucker, Lucy Godson, Nicolas M. Orsi, and Nishant Ravikumar. A comprehensive evaluation of histopathology foundation models for ovarian cancer subtype classification. *npj Precision Oncology*, 9, 2024. doi: 10.1038/s41698-025-00799-8.

W. Bulten, K. Kartasalo, Po-Hsuan Cameron Chen, P. Ström, H. Pinckaers, K. Nagpal, Yuannan Cai, D. Steiner, H. V. Van Boven, Robert Vink, Christina A. Hulsbergen van de Kaa, J. A. van der Laak, M. Amin, A. Evans, T. H. van der Kwast, Robert Allan, P. Humphrey, H. Grönberg, H. Samaratunga, B. Delahunt, T. Tsuzuki, T. Häkkinen, L. Egevad, Maggie Demkin, Sohier Dane, Fraser Tan, Masi Valkonen, G. Corrado, L. Peng, C. Mermel, P. Ruusuvuori, G. Litjens, M. Eklund, A. Brilhante, A. Çakır, Xavier Farré, K. Geronatsiou, V. Molinie, E. Pereira, Paromita Roy, G. Saile, P. Salles, E. Schaafsma, J. Tschui, Jorge Billoch-Lima, Guilherme Pereira, M. Zhou, Shujun He, Sejun Song, Qing Sun, H. Yoshihara, Taiki Yamaguchi, K. Ono, Tao Shen, Jianyi Ji, Arnaud Roussel, Kairong Zhou, Tianrui Chai, Nina Weng, Dmitry A. Grechka, M. Shugaev, Raphael Kiminya, V. Kovalev, D. Voynov, V. Malyshev, E. Lapo, M. Campos, Noriaki Ota, Shinsuke Yamaoka, Yusuke Fujimoto, Kentaro Yoshioka, Joni Juvonen, M. Tukiainen, A. Karlsson, Rui Guo, Chia-Lun Hsieh, I. Zubarev, Habib S. T. Bukhar, Wen-Yuan Li, Jiayun Li, W. Speier, C. Arnold, Kyungdoc Kim, Byeonguk Bae, Y. Kim, Hong-Seok Lee, and Jeonghyuk Park. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine*, 28(1):154–163, 2022. doi: 10.1038/s41591-021-01620-2.

Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. Learning to rank using gradient descent. In *International Conference on Machine Learning*, 2005. doi: 10.1145/1102351.1102363.

Péter Bándi, Oscar G. F. Geessink, Quirine F. Manson, Marcory van Dijk, Maschenka C. A. Balkenhol, M. Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, K. Paeng, Aoxiao Zhong, Quanzheng Li, F. G. Zanjani, S. Zinger, K. Fukuta, Daisuke Komura, V. Ovtcharov, Shenghua Cheng, Shao-Qun Zeng, J. Thagaard, A. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halici, Hunter Jackson, Richard Chen, Fabian Both, Jörg K.H. Franke, H. Küsters-Vandevelde, W. Vreuls, P. Bult, B. van Ginneken, J. A. van der Laak, and G. Litjens. 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. doi: 10.1109/TMI.2018.2867350.

Gabriele Campanella, Shengjia Chen, Manbir Singh, Ruchika Verma, Silke Muehlstedt, Jennifer Zeng, Aryeh Stock, Matthew Croken, Brandon Veremis, Abdulkadir Elmas, Ivan Shujski, Noora Neittaanmäki, Kuan lin Huang, Ricky Kwan, Jane Houldsworth, A. Schoenfeld, and C. Vanderbilt. A clinical benchmark of public Self-Supervised pathology foundation models. *Nature Communications*, 16, 2024. doi: 10.1038/s41467-025-58796-1.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *International Conference on Machine Learning*, pp. 129–136, 2007. doi: 10.1145/1273496.1273513.

Gianluca Carloni, Biagio Brattoli, Seongho Keum, Jongchan Park, Taebum Lee, Chang Ho Ahn, and Sergio Pereira. Pathology foundation models are scanner sensitive: Benchmark and mitigation with contrastive ScanGen loss. In *MedAGI*, pp. 44–53, 2025. doi: 10.48550/arXiv.2507.22092.

Lara Cavinato, Michela Carlotta Massi, Martina Sollini, Margarita Kirienko, and Francesca Ieva. Dual adversarial deconfounding autoencoder for joint batch-effects removal from multi-center and multi-scanner radiomics data. *Scientific Reports*, 13(1):18857, 2023. doi: 10.1038/s41598-023-45983-7.

Binghao Chai, Jianan Chen, Paul Cool, Fatine Oumlil, Anna Tollitt, David F. Steiner, Tapabrata Chakraborti, and Adrienne M. Flanagan. Impact of tissue staining and scanner variation on the performance of pathology foundation models: a study of sarcomas and their mimics. *The Journal of Pathology: Clinical Research*, 12(2):e70080, 2026. doi: 10.1002/2056-4538.70080.

Jia-Ren Chang, Min-Sheng Wu, Wei-Hsiang Yu, Chi-Chung Chen, Cheng-Kung Yang, Yen-Yu Lin, and Chao-Yuan Yeh. Stain mix-up: Unsupervised domain generalization for histopathology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 117–126, 2021. doi: 10.1007/978-3-030-87199-4_11.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-Judge: Assessing multimodal LLM-as-a-Judge with Vision-Language benchmark. In *International Conference on Machine Learning*, pp. 6562–6595, 2024a.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30:850–862, 2024b. doi: 10.1038/s41591-024-02857-3.

Yihang Chen, Yu Fu, Yanyan Huang, Le-Quan Yu, and Weiqin Zhao. Free lunch in pathology foundation model: Task-specific model adaptation with concept-guided feature enhancement. In *Advances in Neural Information Processing Systems 37*, volume 37, pp. 79963–79995, 2024c. doi: 10.52202/079017-2541.

Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. DarkRank: Accelerating Deep Metric Learning via Cross Sample Similarities Transfer. In *AAAI Conference on Artificial Intelligence*, volume abs/1707.01220, 2017. doi: 10.1609/aaai.v32i1.11783.

Zhiwei Chen, Yang Hu, Yu Xiao, Yakun Ju, Tian-Yang Zhang, Yingxue Xu, Wei Li, Hao Chen, Jens Rittscher, and Kaixiang Yang. Harnessing Adversarial Distillation to Customise Debiased, Disease-Specific Pathology Foundation Models for Breast Cancer. *arXiv preprint arXiv:2608.01356*, 2026a.

Zongyi Chen, Yu Liang, Jie Lin, and Liansheng Wang. PathView-Bench: Can Multimodal Large Language Models Achieve Fine-grained Multiscale Understanding of Pathology Images? *arXiv.org*, abs/2607.28318, 2026b. doi: 10.48550/arXiv.2607.28318.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Da-Cheng Li, Hao Zhang, Banghua Zhu, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *International Conference on Machine Learning*, pp. 8359–8388, 2024. doi: 10.48550/arXiv.2403.04132.

Yesung Cho, Sungmin Lee, Geong-Ran Lee, Minkyung Lee, Jongbae Park, and Dongmyung Shin. G2L:from Giga-Scale to Cancer-Specific Large-Scale pathology foundation models via knowledge distillation. *arXiv.org*, abs/2510.11176, 2025. doi: 10.48550/arXiv.2510.11176.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Neural Information Processing Systems*, volume abs/1706.03741, pp. 4299–4307, 2017.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023. doi: 10.48550/arXiv.2302.00070.

Edwin D. de Jong, Eric Marcus, and Jonas Teuwen. Current pathology foundation models are unrobust to medical center differences. *arXiv.org*, abs/2501.18055, 2025. doi: 10.48550/arXiv.2501.18055.

Taher Dehkharghanian, Azam Asilian Bidgoli, Abtin Riasatian, Pooria Mazaheri, Clinton J. V. Campbell, Liron Pantanowitz, H. R. Tizhoosh, and Shahryar Rahnamayan. Biased data, biased AI: deep networks predict the acquisition site of TCGA images. *Diagnostic Pathology*, 18, 2021. doi: 10.1186/s13000-023-01355-3.

Billy Dickson, Sahaj Singh Maini, Craig Sanders, Robert M. Nosofsky, and Zoran Tiganj. Comparing perceptual judgments in large multimodal models and humans. *Behavior Research Methods*, 57, 2025. doi: 10.3758/s13428-025-02728-w.

Ayse B. Dincer, Joseph D. Janizek, and Su-In Lee. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics*, 36(Supplement_2):i573–i582, 2020. doi: 10.1093/bioinformatics/btaa796.

Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, Ming Y. Lu, Andrew Zhang, Anurag J. Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, Drew F. K. Williamson, Harry Robertson, Bowen Chen, Cristina Almagro-Pérez, Paul Doucet, Sharifa Sahai, Chengkuan Chen, Christina S. Chen, Daisuke Komura, Akihiro Kawabe, Mieko Ochi, Shinya Sato, Tomoyuki Yokose, Yohei Miyagi, Shumpei Ishikawa, Georg Gerber, Tingying Peng, Long Phi Le, and Faisal Mahmood. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 31: 3749–3761, 2025. doi: 10.1038/s41591-025-03982-3.

Leonhard Donle, Michael Phillips, Farieda Gaber, Siddhi Ramesh, Matteo Sacco, Sampsa Hautaniemi, Anni Virtanen, Keno Bressem, Lisa Adams, Kelsey Goon, Elena Nevins, Ryan A. Robinett, Sara Kochanny, Sasha Hassan, James Dolezal, Alexander T. Pearson, and Ernst Lengyel. FEATMAP: Targeted Correction of Acquisition Signatures Harmonizes Medical Foundation Model Embeddings and Enables Robust Task Generalization. *bioRxiv*, 2026. doi: 10.64898/2026.07.02.736184.

David Jacob Drexlin, Jonas Dippel, Julius Hense, Niklas Prenißl, Grégoire Montavon, Frederick Klauschen, and Klaus-Robert Müller. MeDi: Metadata-Guided diffusion models for mitigating biases in tumor classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 379–388, 2025. doi: 10.48550/arXiv.2506.17140.

Changde Du, Kaicheng Fu, Bincheng Wen, Yi Sun, Jie Peng, Wei Wei, Ying Gao, Shengpei Wang, Chuncheng Zhang, Jinpeng Li, Shuang Qiu, Le Chang, and Huiguang He. Human-like object concept representations emerge naturally in multimodal large language models. *Nature Machine Intelligence*, 7:860–875, 2024. doi: 10.1038/s42256-025-01049-z.

Joshua L. Ebbert and Dennis Della Corte. PANDA-PLUS-Bench: A Clinical Benchmark for Evaluating Robustness of AI Foundation Models in Prostate Cancer Diagnosis. *AI in Medicine*, abs/2512.14922, 2025. doi: 10.48550/arXiv.2512.14922.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations*, volume abs/1511.05897, 2015.

Kenzo Enda, Yoshitaka Oda, Zen ichi Tanei, Kenichi Satoh, Hiroaki Motegi, S. Terasaka, Shigeru Yamaguchi, T. Ogawa, Lei Wang, Masumi Tsuda, and Shinya Tanaka. Transfer learning strategies for pathological foundation models: A systematic evaluation in brain tumor classification. *Pathology international (Print)*, 76, 2025. doi: 10.1111/pin.70098.

Pablo Meseguer Esbri, Rocío del Amor, and Valery Naranjo. Benchmarking histopathology foundation models in a multi-center dataset for skin cancer subtyping. In *Annual Conference on Medical Image Understanding and Analysis*, pp. 16–28, 2025. doi: 10.48550/arXiv.2506.18668.

Amir Mohammad Ezzati, Alireza Malekhosseini, Armin Khosravi, and Mohammad Hossein Rohban. All centers are at most a few tokens apart: Knowledge distillation with domain invariant prompt tuning. *arXiv preprint arXiv:2511.22739*, 2025. doi: 10.48550/arXiv.2511.22739.

Kianoush Falahkheirkhah, Alex Xijie Lu, David Alvarez-Melis, and Grace Huynh. Domain adaptation using optimal transport for invariant learning using histopathology datasets. In *International Conference on Medical Imaging with Deep Learning*, volume abs/2303.02241 of *Proceedings of Machine Learning Research*, pp. 1765–1782. PMLR, 2023. doi: 10.48550/arXiv.2303.02241.

Khrystyna Faryna, Jeroen van der Laak, and Geert Litjens. Tailoring automated data augmentation to h&e-stained histopathology. In *International Conference on Medical Imaging with Deep Learning*, volume 143, pp. 168–178, 2021.

Dyke Ferber, Georg Wolflein, I. Wiest, Marta Ligero, Srividhya Sainath, N. G. Laleh, O. E. Nahhas, Gustav Muller-Franzes, Dirk Jāger, D. Truhn, and J. Kather. In-context learning enables multimodal large language models to classify cancer pathology images. *Nature Communications*, 15, 2024. doi: 10.1038/s41467-024-51465-9.

Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv.org*, abs/2409.09173, 2024. doi: 10.48550/arXiv.2409.09173.

Alexandre Filiot, Nicolas Dop, Oussama Tchita, Auriane Riou, Rémy Dubois, Thomas Peeters, Daria Valter, Marin Scalbert, Charlie Saillard, Geneviève Robin, and Antoine Olivier. Distilling foundation models for robust and efficient models in digital pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume abs/2501.16239, pp. 162–172, 2025. doi: 10.48550/arXiv.2501.16239.

Alexandre Filiot, Oskar Thaeter, Benoit Schmauch, and Lionel Guillou. Robustifying pathology foundation models via fine-tuning. *arXiv.org*, abs/2607.22861, 2026. doi: 10.48550/arXiv.2607.22861.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. In *Neural Information Processing Systems*, volume abs/2306.09344, 2023. doi: 10.5555/3666122.3668330.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):189–209, 2015. doi: 10.1007/978-3-319-58347-1_10.

Kathy Garcia and Leyla Isik. Behavioral Geometric Supervision Aligns Video Foundation Models with Human Social Perception, 2025. Version 2, revised May 2026; first posted in 2025.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Yu Gong, Greg Mori, and Frederick Tung. RankSim: Ranking Similarity Regularization for Deep Imbalanced Regression. In *International Conference on Machine Learning*, pp. 7634–7649, 2022. doi: 10.48550/arXiv.2205.15236.

Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Computer Vision and Pattern Recognition*, pp. 19338–19347, 2022. doi: 10.1109/CVPR52729.2023.01853.

Christian Grashei, Christian Brechenmacher, Rao Muhammad Umer, Jing-Song Liu, C. Marr, Ewa Szczurek, and P. Schüffler. Pathryoshka: Compressing pathology foundation models via Multi-Teacher knowledge distillation with nested embeddings. *arXiv.org*, abs/2511.23204, 2025. doi: 10.48550/arXiv.2511.23204.

Clément Grisi, Jeroen van der Laak, and Geert Litjens. A distributional robustness margin for pathology foundation models. *arXiv preprint arXiv:2607.25497*, 2026.

Qiao Gu, Dongsub Shim, and Florian Shkurti. Preserving linear separability in continual learning by backward feature projection. In *Computer Vision and Pattern Recognition*, pp. 24286–24295, 2023. doi: 10.1109/CVPR52729.2023.02326.

Tiancheng Gu, Kaicheng Yang, Kaichen Zhang, Xiang An, Ziyong Feng, Yueyi Zhang, Weidong Cai, Jiankang Deng, and Lidong Bing. UniME-V2: MLLM-as-a-Judge for Universal Multimodal Embedding Learning. In *AAAI Conference on Artificial Intelligence*, volume abs/2510.13515, 2025. doi: 10.48550/arXiv.2510.13515.

Fredrik K. Gustafsson and Mattias Rantalainen. Evaluating computational pathology foundation models for prostate cancer grading under distribution shifts. *arXiv.org*, abs/2410.06723, 2024. doi: 10.48550/arXiv.2410.06723.

Jinwei Han, Zhiwen Lin, Zhongyi Sun, Ying-Guo Gao, Ke Yan, Shouhong Ding, Yuan Gao, and Gui-Song Xia. Anchor-based Robust Finetuning of Vision-Language Models. In *Computer Vision and Pattern Recognition*, pp. 26909–26918, 2024. doi: 10.1109/CVPR52733.2024.02542.

Eric He, Akash Gupta, Adian Liusie, Vatsal Raina, Piotr Molenda, Shirom Chabra, and Vyas Raina. Embedding the teacher: Distilling vLLM preferences for scalable image retrieval, 2025.

Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *bioRxiv*, 14:e0223792, 2019. doi: 10.1371/journal.pone.0223792.

Martin N. Hebart, Charles Y. Zheng, Francisco Pereira, and Chris I. Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgments. *Nature Human Behaviour*, 4:1173–1185, 2020. doi: 10.1038/s41562-020-00951-3.

Narayan Hegde, J. Hipp, Yun Liu, M. Emmert-Buck, Emily Reif, D. Smilkov, Michael Terry, C. Cai, M. Amin, C. Mermel, Phil Q. Nelson, L. Peng, G. Corrado, and Martin C. Stumpe. Similar image search for histopathology: SMILY. *npj Digital Medicine*, 2:56, 2019. doi: 10.1038/s41746-019-0131-z.

Audun L. Henriksen, O. Skrede, L. van der Schee, Enric Domingo, K. Cyll, S. de Raedt, I. Kostolomov, Jennifer Hay, W. Kildal, J. Kalsnes, Robert W Williams, M. Pradhan, J. A. Nesheim, H. Askautrud, M. Isaksen, K. S. de Gordoa, M. Cuatrecasas, Joanne Edwards, T. Group, A. Nesbakken, N. Shepherd, Ian Tomlinson, Daniel-Christoph Wagner, R. Kerr, T. Hveem, K. Liestøl, Yoshiaki Nakamura, Marco Novelli, M. Miyo, Sebastian Forsch, D. Church, M. Lacle, David J. Kerr, and A. Kleppe. Enabling clinical use of foundation models for computational pathology. *arXiv preprint arXiv:2602.22347*, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv.org*, abs/1503.02531, 2015.

Md. Ziaul Hoque, Anja Keskinarkaus, Pia Nyberg, and Tapio Seppänen. Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison. *Information Fusion*, 102:101997, 2023. doi: 10.1016/j.inffus.2023.101997.

Frederick M. Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I. Olopade, Jakob N. Kather, Nicole Cipriani, Robert L. Grossman, and Alexander T. Pearson. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications*, 12, 2021. doi: 10.1038/s41467-021-24698-1.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank adaptation of large language models. In *International Conference on Learning Representations*, volume abs/2106.09685, 2021.

Yanyan Huang, Weiqin Zhao, Zhengyu Zhang, Yihang Chen, Yu Fu, Feng Wu, Yuming Jiang, Li Liang, Shujun Wang, and Lequan Yu. Knowledge-guided adaptation of pathology foundation models effectively improves cross-domain generalization and demographic fairness. *Nature Communications*, 16:11485, 2025. doi: 10.1038/s41467-025-66300-y.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 29:2307–2316, 2023. doi: 10.1038/s41591-023-02504-3.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The Platonic Representation Hypothesis. In *International Conference on Machine Learning*, volume abs/2405.07987, pp. 20617–20642, 2024. doi: 10.48550/arXiv.2405.07987.

David R. Hunter. MM algorithms for generalized Bradley-Terry models. *Annals of Statistics*, 32(1): 384–406, 2003. doi: 10.1214/AOS/1079120141.

Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1M: One Million Image-Text Pairs for Histopathology. In *Neural Information Processing Systems*, volume 36, pp. 37995–38017, 2023. doi: 10.48550/arXiv.2306.11207.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, volume abs/1802.04712, pp. 2132–2141, 2018.

Mostafa Jahanifar, Manahil Raza, Kesi Xu, Trinh Thi Le Vuong, Robert Jewsbury, Adam Shephard, Neda Zamanitajeddin, Jin Tae Kwak, Shan E Ahmed Raza, Fayyaz Minhas, and Nasir Rajpoot. Domain generalization in computational pathology: Survey and guidelines. *ACM Computing Surveys*, 57(11):1–37, 2023. doi: 10.1145/3724391.

Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Pérez, Sophia J. Wagner, Anurag J. Vaidya, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, and Faisal Mahmood. HEST-1k: A dataset for spatial transcriptomics and histology image analysis. In *Neural Information Processing Systems*, volume abs/2406.16192, pp. 53798–53833, 2024. doi: 10.48550/arXiv.2406.16192.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks. In *International Conference on Learning Representations*, volume abs/2410.05160, 2024. doi: 10.48550/arXiv.2410.05160.

W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8 1:118–27, 2007. doi: 10.1093/BIOSTATISTICS/KXJ037.

kaiko.ai, Ioannis Gatopoulos, Nicolas Känzig, Roman Moser, and Sebastian Otálora. eva: Evaluation framework for pathology foundation models. In *Medical Imaging with Deep Learning*, 2024. MIDL 2024 short paper; OpenReview FNBQOPj18N.

Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking Self-Supervised Learning on Diverse Pathology Datasets. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3344–3354, 2023. doi: 10.1109/CVPR52729.2023.00326.

Saarthak Kapse, Mehmet Aygün, Elijah Cole, Emma Lundberg, Le Song, and Eric P. Xing. GenBio-PathFM: A State-of-the-Art foundation model for histopathology. *bioRxiv*, 2026. doi: 10.64898/2026.03.17.712534.

Mikhail Karasikov, J. Doorn, Nicolas Kanzig, Melis Erdal Cesur, Hugo M. Horlings, Robert Berke, Fei Tang, and Sebastian Otálora. Training state-of-the-art pathology foundation models with orders of magnitude less data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 573–583, 2025. doi: 10.48550/arXiv.2504.05186.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A. Valous, Dyke Ferber, Lina Jansen, Constantino Carlos Reyes-Aldasoro, Inka Zörnig, Dirk Jäger, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, and Niels Halama. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16: e1002730, 2019. doi: 10.1371/journal.pmed.1002730.

Piotr Keller, Muhammad Dawood, and Fayyaz ul Amir Minhas. Do tissue source sites leave identifiable signatures in whole slide images beyond staining? In *TML4H*, pp. 1–10. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-39539-0_1.

Chanwoo Kim, Jakub Kaczmarzyk, Deepika Savant, Zhen Zhao, Peter K. Koo, and Su-In Lee. Dissecting and directing pathology foundation models. *bioRxiv*, 2026. doi: 10.64898/2026.06.12.731496. Preprint.

Gangsu Kim and Won-Ki Jeong. LaGuadia: Language-Guided adaptive distillation from pathology foundation models. *arXiv.org*, abs/2607.11257, 2026. doi: 10.48550/arXiv.2607.11257.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations. In *International Conference on Learning Representations*, volume abs/2204.02937, 2022. doi: 10.48550/arXiv.2204.02937.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting

in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114:3521–3526, 2016. doi: 10.1073/pnas.1611835114.

Pang Wei Koh, Shiori Sagawa, H. Marklund, S. Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, R. L. Phillips, Irena Gao, Tony Lee, Etiene David, I. Stavness, Wei Guo, Berton A. Earnshaw, I. Haque, Sara Beery, J. Leskovec, A. Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, volume 139, pp. 5637–5664, 2020.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-Tuning can distort pretrained features and underperform Out-of-Distribution. In *International Conference on Learning Representations*, volume abs/2202.10054, 2022.

Jonah Kömen, Hannah Marienwald, Jonas Dippel, and Julius Hense. Do histopathological foundation models eliminate batch effects? a comparative study. *arXiv.org*, abs/2411.05489, 2024. doi: 10.48550/arXiv.2411.05489.

Jonah Kömen, Edwin D. de Jong, Julius Hense, Hannah Marienwald, Jonas Dippel, Philip Naumann, Eric Marcus, Lukas Ruff, Maximilian Alber, Jonas Teuwen, Frederick Klauschen, and Klaus-Robert Müller. Towards robust foundation models for digital pathology. *Nature Communications*, 17, 2026. doi: 10.1038/s41467-026-73923-2.

Maxime W. Lafarge, Josien P. W. Pluim, Koen A. J. Eppenhof, Pim Moeskops, and Mitko Veta. Domain-Adversarial neural networks to address the appearance variability of histopathology images. In M. Jorge Cardoso, Tal Arbel, Gustavo Carneiro, Tanveer Syeda-Mahmood, João Manuel R. S. Tavares, Mehdi Moradi, Andrew Bradley, Hayit Greenspan, João Paulo Papa, Anant Madabhushi, Jacinto C. Nascimento, Jaime S. Cardoso, Vasileios Belagiannis, and Zhi Lu (eds.), *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, volume 10553 of *Lecture Notes in Computer Science*, pp. 83–91, Cham, 2017. Springer International Publishing. ISBN 978-3-319-67558-9. doi: 10.1007/978-3-319-67558-9_10.

M. Law, N. Thomé, and M. Cord. Learning a distance metric from relative comparisons between quadruplets of images. *International Journal of Computer Vision*, 121:65–94, 2016. doi: 10.1007/s11263-016-0923-4.

Harrison Lee, Samrat Phatale, H. Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. RLAIF: Scaling reinforcement learning from human feedback with AI feedback. *arXiv.org*, abs/2309.00267, 2023. doi: 10.48550/arXiv.2309.00267.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernández Abrego, Wei Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha R. Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. Gecko: Versatile Text Embeddings Distilled from Large Language Models. arXiv preprint arXiv:2403.20327, 2024a.

Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. Prometheus-Vision: Vision-Language Model as a Judge for Fine-Grained Evaluation. In *Annual Meeting of the Association for Computational Linguistics*, pp. 11286–11315, 2024b. doi: 10.48550/arXiv.2401.06591.

Sua Lee, Sanghee Park, and Jinbae Im. MM-JudgeBias: A Benchmark for Evaluating Compositional Biases in MLLM-as-a-Judge. *Annual Meeting of the Association for Computational Linguistics*, pp. 25336–25373, 2026. doi: 10.48550/arXiv.2604.18164.

Yoonho Lee, Annie S. Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical Fine-Tuning improves adaptation to distribution shifts. In *International Conference on Learning Representations*, volume abs/2210.11466, 2022. doi: 10.48550/arXiv.2210.11466.

Carlijn M. Lems, L. Tessier, J. Bokhorst, M. van Rijthoven, W. Aswolinskiy, Matteo Pozzi, Natalie Klubickova, Suzanne Dintzis, Michela Campora, Maschenka C. A. Balkenhol, Peter Bult, J. Spronck, Thomas Detone, M. Barbareschi, E. Munari, G. Bogina, J. Wesseling, E. Lips, F. Ciompi, Frédérique Meeuwsen, and J. A. van der Laak. A multicentric dataset for training and benchmarking breast cancer segmentation in h&e slides. *arXiv.org*, abs/2510.02037, 2025. doi: 10.48550/arXiv.2510.02037.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791, 2024a. doi: 10.48550/arXiv.2411.16594.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5400–5409, 2018a. doi: 10.1109/CVPR.2018.00566. MMD-AAE aligns source-domain feature distributions inside an adversarial autoencoder, the DG-side sibling of the AD-AE baseline the paper labels DANN.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference Distillation for Large Visual Language Models. *arXiv.org*, abs/2312.10665, 2023. doi: 10.48550/arXiv.2312.10665.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. VL-RewardBench: A Challenging Benchmark for Vision-Language Generative Reward Models. In *Computer Vision and Pattern Recognition*, pp. 24657–24668, 2024b. doi: 10.1109/CVPR52734.2025.02296.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning*, volume abs/1802.01483, pp. 2830–2839, 2018b.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2016. doi: 10.1007/978-3-319-46493-0_37.

Siyu Lin, Haowen Zhou, Mark Watson, Ramaswamy Govindan, Richard J. Cote, and Changhuei Yang. Impact of stain variation and color normalization for prognostic predictions in pathology. *Scientific Reports*, 15, 2025. doi: 10.1038/s41598-024-83267-w.

Yu Ling, Weimin Tan, and Bo Yan. Self-Supervised Digital Histopathology Image Disentanglement for Arbitrary Domain Stain Transfer. *IEEE Transactions on Medical Imaging*, 42(12):3625–3638, 2023. doi: 10.1109/TMI.2023.3298361.

G. Litjens, Péter Bándi, Babak Ehteshami Bejnordi, Oscar G. F. Geessink, Maschenka C. A. Balkenhol, P. Bult, Altuna Halilovic, M. Hermsen, Rob van de Loo, Rob Vogels, Quirine F. Manson, N. Stathonikos, A. Baidoshvili, P. V. van Diest, C. Wauters, Marcory van Dijk, and J. A. van der Laak. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience*, 7(6):giy065, 2018. doi: 10.1093/gigascience/giy065.

Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models. In *Conference of the European Chapter of the Association for Computational Linguistics*, pp. 139–151, 2023. doi: 10.18653/v1/2024.eacl-long.8.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, volume abs/1502.02791, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

Shenghan Lou, Genshen Mo, Xiao Zhang, Hao Wang, Hao Li, Keru Ma, Huiying Li, Xinyue Zhang, Meihong Yan, Haonan Xie, Yuze Huang, Chuangqi Li, Siyuan Ma, Hongxue Meng, Lei Cao, and Peng Han. Uncertainty-aware and causal test-time adaptive foundation model for robust colorectal cancer pathology diagnosis. *npj Digital Medicine*, 8, 2025. doi: 10.1038/s41746-025-02149-1.

Christos Louizos, Kevin Swersky, Yujia Li, M. Welling, and R. Zemel. The variational fair autoencoder. In *International Conference on Learning Representations*, 2015.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush U Patel, Amr Soliman, Chengkuan Chen, Tong Ding, Judy J. Wang, Georg K. Gerber, Ivy Liang, L. Le, Anil V. Parwani, Luca L. Weishaupt, and Faisal Mahmood. A multimodal generative ai copilot for human pathology. *Nature*, 634:466–473, 2024. doi: 10.1038/s41586-024-07618-3.

Martin Q. Ma, Yao-Hung Hubert Tsai, Paul Pu Liang, Han Zhao, Kun Zhang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Conditional contrastive learning for improving fairness in self-supervised learning, 2021. Conditions InfoNCE sampling on a sensitive attribute so the learned representation does not encode it, directly analogous to conditioning on institution to remove site signatures.

Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. A method for normalizing histology slides for quantitative analysis. In *IEEE International Symposium on Biomedical Imaging*, pp. 1107–1110, 2009. doi: 10.1109/ISBI.2009.5193250.

Niccolò Marini, Manfredo Atzori, Sebastian Otálora, Stéphane Marchand-Maillet, and Henning Müller. H&E-adversarial network: A convolutional neural network to learn stain-invariant features through hematoxylin & eosin regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 601–610, 2021. doi: 10.1109/ICCVW54120.2021.00073. Adversarial regression of the per-image H&E stain matrix, computed from the image itself, to remove stain information without domain labels; a label-free stain-invariance objective missing from the related work.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L. Griffiths. Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14:21445, 2023. doi: 10.1038/s41598-024-72071-1.

Volodymyr Mavrych, E. Yousef, Ahmed Yaqinuddin, A. Shaikh, and O. Bolgova. Visual recognition limitations in multimodal large language models: A comparative analysis of histological image interpretation. *PLOS Digital Health*, 5, 2026. doi: 10.1371/journal.pdig.0001306.

Vaibhav Mishra and William Lotter. Comparing computational pathology foundation models using representational similarity analysis. In *Proceedings of machine learning research*, volume 297 of *Proceedings of Machine Learning Research*, pp. 1007–1022. PMLR, 2025. doi: 10.48550/arXiv.2509.15482.

Salman Mohammadi, Anders Kirk Uhrenholt, and Bjørn Sand Jensen. Odd-One-Out Representation Learning. *arXiv.org*, abs/2012.07966, 2020.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, A. Galstyan, and G. V. Steeg. Invariant representations without adversarial training. In *Neural Information Processing Systems*, pp. 9102–9111, 2018.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 28 of *PMLR*, pp. 10–18, 2013. DICA learns a fixed feature transformation that minimizes distributional variance across source domains using only domain labels (an unsupervised variant needs no task labels), a plug-and-play projection baseline on frozen features.

Jishnu Mukhoti, Yarin Gal, Philip H. S. Torr, and Puneet K. Dokania. Fine-tuning can cripple your foundation model; preserving features may be the solution. *Trans. Mach. Learn. Res.*, abs/2308.13320, 2023. doi: 10.48550/arXiv.2308.13320.

Pierre Murchan, Pilib Ó Broin, Anne-Marie Baird, Orla Sheils, and Stephen P. Finn. Deep feature batch correction using ComBat for machine learning applications in computational pathology. *Journal of Pathology Informatics*, 15:100396, 2024. doi: 10.1016/j.jpi.2024.100396.

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. In *International Conference on Learning Representations*, volume abs/2211.01201, 2022a. doi: 10.48550/arXiv.2211.01201.

Lukas Muttenthaler, Charles Y. Zheng, Patrick McClure, Robert A. Vandermeulen, Martin N. Hebart, and Francisco Pereira. VICE: Variational Interpretable Concept Embeddings. In *Neural Information Processing Systems*, volume abs/2205.00756, 2022b. doi: 10.48550/arXiv.2205.00756.

Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A. Vandermeulen, Katherine Hermann, Andrew K. Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. In *Neural Information Processing Systems*, volume abs/2306.04507, pp. 50978–51007, 2023. doi: 10.48550/arXiv.2306.04507.

Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, M. Mozer, Klaus-Robert Muller, Thomas Unterthiner, and Andrew Kyle Lampinen. Aligning machine and human visual representations across abstraction levels. *Nature*, 647:349–355, 2025. doi: 10.1038/s41586-025-09631-6.

Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. Hibou: A family of foundational vision transformers for pathology. *arXiv.org*, abs/2406.05074, 2024. doi: 10.48550/arXiv.2406.05074.

Peter Neidlinger, Omar S. M. El Nahhas, Hannah Sophie Muti, Tim Lenz, Michael Hoffmeister, Hermann Brenner, Marko van Treeck, Rupert Langer, Bastian Dislich, Hans Michael Behrens, Christoph Röcken, Sebastian Foersch, Daniel Truhn, Antonio Marra, Oliver Lester Saldanha, and Jakob Nikolas Kather. Benchmarking foundation models as feature extractors for weakly supervised computational pathology. *Nature Biomedical Engineering*, 10:1113–1123, 2025. doi: 10.1038/s41551-025-01516-3.

Hai Cao Truong Nguyen and David Joon Ho. fmMAP: A framework reducing Site-Bias batch effect from foundation models in pathology. In *MICCAI Workshop on Computational Pathology with Multimodal Data (COMPAYL)*, 2025. OpenReview o247Qbp9qJ.

Tan H. Nguyen, Dinkar Juyal, Jin Li, Aaditya Prakash, Shima Nofallah, Chintan Shah, Sai Chowdary Gullapally, Limin Yu, Michael Griffin, Anand Sampat, John Abel, Justin Lee, and Amaro Taylor-Weiner. ContriMix: Scalable stain color augmentation for domain generalization without domain labels in digital pathology. In *COMPAY@MICCAI*, pp. 121–130, 2023.

Justin D. Norman, Michael U. Rivera, and D. A. Hughes. Reliability without Validity: A Systematic, Large-Scale Evaluation of LLM-as-a-Judge Models Across Agreement, Consistency, and Bias. arXiv preprint arXiv:2606.19544, 2026.

Mieko Ochi, Daisuke Komura, Takumi Onoyama, Koki Shinbo, Haruya Endo, Hiroto Odaka, Miwako Kakiuchi, Hiroto Katoh, Tetsuo Ushiku, and Shumpei Ishikawa. Registered multi-device/staining histology image dataset for domain-agnostic machine learning models. *Scientific Data*, 11, 2024. doi: 10.1038/s41597-024-03122-5.

Sejoon Oh, Berk Ustun, Julian McAuley, and Srijan Kumar. FINEST: Stabilizing Recommendations by Rank-Preserving Fine-Tuning. *ACM Transactions on Knowledge Discovery from Data*, 18:1–22, 2024. doi: 10.1145/3695256.

Daisuke Ono, Dennis W. Dickson, and Shunsuke Koga. Evaluating the efficacy of few-shot learning for GPT-4Vision in neurodegenerative disease histopathology: A comparative analysis with convolutional neural network model. *Neuropathology and Applied Neurobiology*, 50:e12997, 2024. doi: 10.1111/nan.12997.

M. Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra,

Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, H. Jégou, J. Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, abs/2304.07193, 2023. doi: 10.48550/arXiv.2304.07193.

Sebastian Otálora, Manfredo Atzori, Vincent Andrearczyk, Amjad Khan, and Henning Müller. Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology. *Frontiers in Bioengineering and Biotechnology*, 7, 2019. doi: 10.3389/fbioe.2019.00198.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Neural Information Processing Systems*, volume abs/2203.02155, pp. 27730–27744, 2022. doi: 10.52202/068431-2011.

Shi Pan, Jianan Chen, and Maria Secrier. Teaching pathology foundation models to accurately predict gene expression with parameter efficient knowledge transfer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 605–613, 2025. doi: 10.48550/arXiv.2504.07061.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational Knowledge Distillation. In *Computer Vision and Pattern Recognition*, pp. 3962–3971, 2019. doi: 10.1109/CVPR.2019.00409.

Nikolaos Passalis and A. Tefas. Learning deep representations with probabilistic knowledge transfer. In *European Conference on Computer Vision*, pp. 283–299, 2018. doi: 10.1007/978-3-030-01252-6_17.

Joshua C. Peterson, Joshua T. Abbott, and Thomas L. Griffiths. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Sciences*, 42 8(8):2648–2669, 2017. doi: 10.1111/cogs.12670.

Zhen Qin, R. Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. Large language models are effective text rankers with pairwise ranking prompting. In *NAACL-HLT*, pp. 1504–1518, 2023. doi: 10.48550/arXiv.2306.17563.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume abs/2103.00020, pp. 8748–8763, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Neural Information Processing Systems*, volume abs/2305.18290, pp. 53728–53741, 2023. doi: 10.52202/075280-2338.

Lavish Ramchandani, Gunjan Deotale, and Dev Kumar Das. Parameter-efficient fine-tuning (PEFT) of Vision Foundation Models for Atypical Mitotic Figure Classification. In *MIDOG@MICCAI*, pp. 198–203, 2025. doi: 10.48550/arXiv.2509.16935.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, 2020. doi: 10.18653/v1/2020.acl-main.647. Iterative nullspace projection removes linearly decodable attribute information (here, site/scanner) from frozen embeddings with only attribute labels; a simple projection baseline missing from the comparison.

E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21:34–41, 2001. doi: 10.1109/38.946629.

Daniel Reisenbüchler, Lucas Luttner, N. Schaadt, F. Feuerhake, and D. Merhof. Unsupervised latent stain adaptation for computational pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 755–765, 2024. doi: 10.1007/978-3-031-72120-5_70.

Jian Ren, I. Hacihaliloglu, E. Singer, D. Foran, and X. Qi. Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 11071, pp. 201–209, 2018. doi: 10.1007/978-3-030-00934-2_23.

Brett D. Roads and Bradley C. Love. Enriching ImageNet with human similarity judgments and psychological embeddings. In *Computer Vision and Pattern Recognition*, pp. 3546–3556, 2020. doi: 10.1109/CVPR46437.2021.00355.

Jeongun Ryu, Heon Song, Seungeun Lee, Soo Ick Cho, Jiwon Shin, Kyunghyun Paeng, and Sérgio Pereira. SCORPION: Addressing Scanner-Induced variability in histopathology. In *UNSURE@MICCAI*, pp. 102–111, 2025. doi: 10.48550/arXiv.2507.20907.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *arXiv.org*, volume abs/1911.08731, 2019.

Lucas Sancéré, C. Lorenz, Doris Helbig, O. Persa, Sonja Dengler, A. Kreuter, M. Laimer, Roland Lang, Anne Fröhlich, Jennifer Landsberg, J. Brägelmann, and Katarzyna Bozek. Histo-Miner: Deep learning based tissue features extraction pipeline from H&E whole slide images of cutaneous squamous cell carcinoma. *PLoS Comput. Biol.*, 22(1):e1013907–e1013907, 2025. doi: 10.1371/journal.pcbi.1013907.

Craig Sanders, Billy Dickson, Sahaj Singh Maini, Robert M. Nosofsky, and Zoran Tiganj. Vision-language models learn the geometry of human perceptual space. arXiv preprint arXiv:2510.20859, 2025.

Kyle Sargent, Ruiqi Gao, Philipp Henzler, Charles Herrmann, Aleksander Holynski, Fei-Fei Li, Jiajun Wu, and J. Zhang. VLIC: Vision-Language Models As Perceptual Judges for Human-Aligned Image Compression. arXiv preprint arXiv:2512.15701, 2025.

L. Schoenpflug, N. J. van den Berg, S. Andani, N. Horeweg, J. Barkey Wolf, T. Bosse, Viktor Koelzera, and M. Lafarge. A protocol for evaluating robustness to H&E staining variation in computational pathology models. *Journal of Pathology Informatics*, 22, 2026. doi: 10.1016/j.jpi.2026.100702.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Computer Vision and Pattern Recognition*, pp. 815–823, 2015. doi: 10.1109/CVPR.2015.7298682.

Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. In *Neural Information Processing Systems*, pp. 41–48, 2003.

Birgid Schömig-Markiefka, Alexey E. Pryalukhin, W. Hulla, A. Bychkov, J. Fukuoka, A. Madabhushi, V. Achter, Lech Nieroda, R. Buettner, A. Quaas, and Y. Tolkach. Quality control stress test for deep learning-based diagnostic model in digital pathology. *Modern Pathology*, 34:2098–2108, 2021. doi: 10.1038/s41379-021-00859-x.

M. Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. Staingan: Stain style transfer for digital histological images. In *IEEE International Symposium on Biomedical Imaging*, pp. 953–956, 2018. doi: 10.1109/ISBI.2019.8759152.

Abubakr Shafique, Amanda Dy, Xiaoli Qin, N. Alshamlan, Dimitri Androutsos, Susan J. Done, and April Khademi. Lightweight self supervised learning framework for domain generalization in histopathology. *Scientific Reports*, 15(1):36631, 2025. doi: 10.1038/s41598-025-20411-0.

Abubakr Shafique, Xiaoli Qin, Amanda Dy, N. Alshamlan, Dimitri Androutsos, Susan J. Done, and April Khademi. Reliability of foundation models for image retrieval in histopathology. *npj Imaging*, 4(1):41, 2026. doi: 10.1038/s44303-026-00174-7.

Yiqing Shen, Yulin Luo, Dinggang Shen, and Jing Ke. RandStainNA: Learning Stain-Agnostic features from histology slides by bridging stain augmentation and normalization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume abs/2206.12694, pp. 212–221, 2022. doi: 10.1007/978-3-031-16434-7_21.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Neural Information Processing Systems*, pp. 1849–1857, 2016.

Karin Stacke, G. Eilertsen, Jonas Unger, and C. Lundström. Measuring domain shift for deep learning in histopathology. *IEEE Journal of Biomedical and Health Informatics*, 25:325–336, 2020. doi: 10.1109/JBHI.2020.3032060.

Stella Su, Marc Harary, Scott J. Rodig, and William Lotter. AdvDINO: Domain-adversarial self-supervised representation learning for spatial proteomics. In Yuankai Huo, Mingchen Gao, Chang-Fu Kuo, Yueming Jin, and Ruining Deng (eds.), *arXiv.org*, volume abs/2508.04955 of *Proceedings of Machine Learning Research*, pp. 3704–3722. PMLR, 2025. doi: 10.48550/arXiv.2508.04955.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J. Cueva, Erin Grant, I. Groen, Jascha Achterberg, J. B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, M. Hebart, Nathan Cloos, N. Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew Kyle Lampinen, Klaus-Robert Muller, M. Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *Trans. Mach. Learn. Res.*, abs/2310.13018, 2023. doi: 10.48550/arXiv.2310.13018.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV Workshops*, pp. 443–450, 2016. doi: 10.1007/978-3-319-49409-8_35.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, P. Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent. In *Conference on Empirical Methods in Natural Language Processing*, volume abs/2304.09542, pp. 14918–14937, 2023. doi: 10.48550/arXiv.2304.09542.

Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Zhongyi Shui, Kai Zhang, Jingxiong Li, Xingheng Lyu, Tao Lin, and Lin Yang. PathGen-1.6M: 1.6 Million Pathology Image-text Pairs Generation through Multi-agent Collaboration. *arXiv.org*, abs/2407.00203, 2024. doi: 10.48550/arXiv.2407.00203.

Yuxuan Sun, Hao Wu, Chenglu Zhu, Yixuan Si, Qizi Chen, Yunlong Zhang, Kai Zhang, Jingxiong Li, Jiatong Cai, Yuhan Wang, Lin Sun, Tao Lin, and Lin Yang. PathBench: Advancing the Benchmark of Large Multimodal Models for Pathology Image Understanding at Patch and Whole Slide Level. *IEEE Transactions on Medical Imaging*, 44:4087–4097, 2025. doi: 10.1109/TMI.2025.3584857.

Shobhita Sundaram, Stephanie Fu, Lukas Muttenthaler, Netanel Y. Tamir, Lucy Chai, Simon Kornblith, Trevor Darrell, and Phillip Isola. When does perceptual alignment benefit vision representations? In *Neural Information Processing Systems*, volume abs/2410.10817, pp. 55314–55341, 2024. doi: 10.48550/arXiv.2410.10817.

David Tellez, G. Litjens, Péter Bándi, W. Bulten, J. Bokhorst, F. Ciompi, and J. A. van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Anal.*, 58:101544, 2019. doi: 10.1016/j.media.2019.101544.

Erik Thiringer, Fredrik K. Gustafsson, Kajsa Ledesma Eriksson, and Mattias Rantalainen. Scanner-Induced domain shifts undermine the robustness of pathology foundation models. *arXiv.org*, abs/2601.04163, 2026. doi: 10.48550/arXiv.2601.04163.

Junjiao Tian, Xiaoliang Dai, Chih-Yao Ma, Zecheng He, Yen-Cheng Liu, and Z. Kira. Trainable projected gradient method for robust fine-tuning. In *Computer Vision and Pattern Recognition*, pp. 7836–7845, 2023. doi: 10.1109/CVPR52729.2023.00757.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and Mitigating Position Bias of Multi-image Vision-Language Models. *Computer Vision and Pattern Recognition*, pp. 10599–10609, 2025. doi: 10.1109/CVPR52734.2025.00991.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, volume abs/1910.10699, 2019.

Alexandre Tiard, A. Wong, D. J. Ho, Yang Wu, Eliram Nof, Alvin C. Goh, Stefano Soatto, and S. Nadeem. Stain-invariant self supervised learning for histopathology image analysis. *arXiv.org*, abs/2211.07590, 2022. doi: 10.48550/arXiv.2211.07590.

Y. Tolkach, Lisa Marie Wolgast, A. Damanakis, Alexey E. Pryalukhin, S. Schallenberg, W. Hulla, Marie-Lisa Eich, W. Schröder, A. Mukhopadhyay, Moritz Fuchs, S. Klein, C. Bruns, Reinhard Büttner, F. Gebauer, Birgid Schömig-Markiefka, and A. Quaas. Artificial intelligence for tumour tissue detection and histological regression grading in oesophageal adenocarcinomas: a retrospective algorithm development and validation study. *The Lancet Digital Health*, 5 5(5):e265–e275, 2023. doi: 10.1016/s2589-7500(23)00027-4.

Cheng-Chang Tsai, Yuan-Chih Chen, and Chun-Shien Lu. Test-time stain adaptation with diffusion models for histopathology image classification. In *Computer Vision – ECCV 2024*, pp. 257–275, 2024. doi: 10.1007/978-3-031-72761-0_15.

Yao-Hung Hubert Tsai, Tianqin Li, Martin Q. Ma, Han Zhao, Kun Zhang, Louis-Philippe Morency, and Ruslan Salakhutdinov. Conditional contrastive learning with kernel. In *International Conference on Learning Representations (ICLR)*, 2022. Contrastive learning conditioned on a nuisance variable (sampling positives/negatives within the same condition) removes that variable's information without task labels, a label-free alternative to the judge-filtered triplet objective.

Frederick Tung and Greg Mori. Similarity-Preserving Knowledge Distillation. In *IEEE International Conference on Computer Vision*, pp. 1365–1374, 2019. doi: 10.1109/ICCV.2019.00145.

Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. Structure-Preserving color normalization and sparse stain separation for histological images. *IEEE Transactions on Medical Imaging*, 35:1962–1971, 2016. doi: 10.1109/TMI.2016.2529665.

Laurens van der Maaten and Kilian Weinberger. Stochastic triplet embedding. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2012. doi: 10.1109/MLSP.2012.6349720.

M. van Rijthoven, W. Awolinsky, L. Tessier, R. Salgado, J. A. van der Laak, F. Ciompi, and TIGER-LC Consortium. Tumor-infiltrating lymphocytes in breast cancer through artificial intelligence: biomarker analysis from the results of the TIGER challenge. *medRxiv*, 2025. doi: 10.1101/2025.02.28.25323078.

Ben Vardi, Dana Schonberger, Yuval Friedmann, Zohar Yakhini, Iris Barshack, Alexander Loebel, and Ariel Shamir. Discrepancy minimization improves Cross-Hospital robustness in digital pathology. *arXiv.org*, abs/2605.25175, 2026. doi: 10.48550/arXiv.2605.25175.

E. Vorontsov, A. Bozkurt, Adam Casson, George Shaikovski, Michał Zelechowski, Kristen A. Severson, Eric Zimmermann, J. Hall, Neil A. Tenenholtz, Nicolò Fusi, Ellen Yang, Philippe Mathieu, A. van Eck, Donghun Lee, Julian Viret, Eric Robert, Y. Wang, J. Kunz, Matthew C H Lee, Jan H Bernhard, R. Godrich, Gerard Oakley, Ewan Millar, Matthew G Hanna, Hannah Y Wen, Juan A Retamero, William A. Moye, Razik Yousfi, C. Kanan, D.S. Klimstra, B. Rothrock, Siqi Liu, and Thomas J Fuchs. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30:2924–2935, 2024. doi: 10.1038/s41591-024-03141-0.

Trinh Thi Le Vuong, Quoc Dang Vu, Mostafa Jahanifar, Simon Graham, Jin Tae Kwak, and Nasir Rajpoot. IMPaSh: A novel domain-shift resistant representation for colorectal cancer tissue classification. In *Computer Vision – ECCV 2022 Workshops*, Lecture Notes in Computer Science, pp. 543–555, 2023. doi: 10.1007/978-3-031-25066-8_31. Self-supervised contrastive learning with patch shuffling and stain perturbation to obtain domain-shift-resistant tissue representations without labels; a label-free stain-augmented SSL baseline not discussed.

Sophia J. Wagner, Nadieh Khalili, Raghav Sharma, Melanie Boxberg, Carsten Marr, Walter de Back, and Tingying Peng. Structure-Preserving Multi-domain Stain Color Augmentation Using Style-Transfer with Disentangled Representations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume abs/2107.12357 of *Lecture Notes in Computer Science*, pp. 257–266. Springer, 2021. doi: 10.1007/978-3-030-87237-3_25.

Chong Wang, Shuxin Li, Jing Ke, Chen Zhang, and Yiqing Shen. RandStainNA++: Enhance random stain augmentation and normalization through foreground and background differentiation. *IEEE Journal of Biomedical and Health Informatics*, 28:3660–3671, 2024a. doi: 10.1109/JBHI.2024.3379280.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully Test-Time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tao Yan, Wenjie Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. MuirBench: A Comprehensive Benchmark for Robust Multi-image Understanding. In *International Conference on Learning Representations*, volume abs/2406.09411, 2024b. doi: 10.48550/arXiv.2406.09411.

Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1386–1393, 2014. doi: 10.1109/CVPR.2014.180.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving Text Embeddings with Large Language Models. In *Annual Meeting of the Association for Computational Linguistics*, pp. 11897–11916, 2023a. doi: 10.48550/arXiv.2401.00368.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large Language Models are not Fair Evaluators. In *Annual Meeting of the Association for Computational Linguistics*, volume abs/2305.17926, 2023b. doi: 10.48550/arXiv.2305.17926.

Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, Fang Wang, Yulong Peng, Junyou Zhu, Jing Zhang, Christopher R. Jackson, Jun Zhang, Deborah Dillon, Nancy U. Lin, Lynette Sholl, Thomas Denize, David Meredith, Keith L. Ligon, Sabina Signoretti, Shuji Ogino, Jeffrey A. Golden, MacLean P. Nasrallah, Xiao Han, Sen Yang, and Kun-Hsing Yu. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634:970–978, 2024c. doi: 10.1038/s41586-024-07894-z.

Yufei Wang, Zhanyi Sun, Jesse Zhang, Zhou Xian, Erdem Biyik, David Held, and Zackory Erickson. RL-VLM-F: Reinforcement learning from vision language foundation model feedback. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51484–51501. PMLR, 2024d. doi: 10.48550/arXiv.2402.03681.

Zhi Wang, Geelon So, and Ramya Korlakai Vinayak. Metric Learning from Limited Pairwise Preference Comparisons. In *Conference on Uncertainty in Artificial Intelligence*, pp. 3571–3602, 2024e. doi: 10.48550/arXiv.2403.19629.

Kian R. Weihrauch, Thomas A. Buckley, W. Lotter, and A. K. Manrai. How seemingly inconsequential design choices dictate performance of llms in pathology. arXiv preprint arXiv:2606.12407, 2026.

Kilian Q. Weinberger, John Blitzer, and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Neural Information Processing Systems*, volume 10, pp. 1473–1480, 2005. doi: 10.5555/1577069.1577078.

J. Weinstein, E. Collisson, G. Mills, K. Shaw, B. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and Joshua M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45:1113–1120, 2013. doi: 10.1038/ng.2764.

Tengjin Weng, Wenhao Jiang, Jingyi Wang, Ming Li, Lin Ma, and Ming Zhong. OddGridBench: Exposing the Lack of Fine-Grained Visual Discrepancy Sensitivity in Multimodal Large Language Models. *arXiv.org*, abs/2603.09326, 2026. doi: 10.48550/arXiv.2603.09326.

Anja Witte, Maximilian Lennartz, Jan Baumbach, Guido Sauter, Stefan Bonn, Patrick Fuhlert, and Marina Zimmermann. EXPOSE: Explainable and Domain-Robust Embeddings from Pathology Vision Foundation Models using Sparse Autoencoders. *arXiv preprint arXiv:2608.28191*, 2026.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, R. Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Computer Vision and Pattern Recognition*, pp. 7949–7961, 2021. doi: 10.1109/CVPR52688.2022.00780.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, volume abs/2203.05482, pp. 23965–23998, 2022. doi: 10.48550/arXiv.2203.05482.

Yonghuang Wu, Wenwen Zeng, Xuan Xie, Chengqian Zhao, Guoqing Wu, and Jinhua Yu. SAMPO-Path: Segmentation Intent-Aligned preference optimization for pathology foundation model segmentation. *arXiv preprint arXiv:2508.02464*, 2025.

G. Wölflein, Dyke Ferber, A. Meneghetti, Omar S. M. El Nahhas, D. Truhn, Zunamys I. Carrero, David J. Harrison, Ognjen Arandjelovíc, and J. Kather. A good feature extractor is all you need for weakly supervised pathology slide classification. In *ECCV Workshops*, pp. 68–87, 2024. doi: 10.1007/978-3-031-91721-9_5.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: theory and algorithm. In *International Conference on Machine Learning*, pp. 1192–1199, 2008. doi: 10.1145/1390156.1390306.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *Neural Information Processing Systems*, 2017.

Qizhe Xie, E. Hovy, Minh-Thang Luong, and Quoc V. Le. Self-training with noisy student improves ImageNet classification. In *Computer Vision and Pattern Recognition*, pp. 10684–10695, 2019. doi: 10.1109/cvpr42600.2020.01070.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Hao-Qi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. LLLaVA-Critic: Learning to Evaluate Multimodal Models. In *Computer Vision and Pattern Recognition*, pp. 13618–13628, 2024. doi: 10.1109/CVPR52734.2025.01271.

Hanwen Xu, N. Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, C. Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu-Hsin Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chun yue Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, R. Weerasinghe, Bill J. Wright, Ari Robicsek, B. Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630:181–188, 2024. doi: 10.1038/s41586-024-07441-w.

Dhyey Yajnik, Amina Asif, and Fayyaz Minhas. The Good, the Bad, and the Brittle: Benchmarking Robustness and Generalisation of Histopathology Foundation Models. *arXiv.org*, abs/2607.04401, 2026. doi: 10.48550/arXiv.2607.04401.

Rikiya Yamashita, Jin Long, Snikitha Banda, Jeanne Shen, and Daniel L. Rubin. Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. *IEEE Transactions on Medical Imaging*, 40:3945–3954, 2021. doi: 10.1109/TMI.2021.3101985.

Pengshuai Yang, Xiaoxu Yin, Haiming Lu, Zhongliang Hu, Xuegong Zhang, Rui Jiang, and Hairong Lv. CS-CO: A Hybrid Self-Supervised Visual Representation Learning Method for H&E-stained Histopathological Images. *Medical Image Anal.*, 81:102539, 2022. doi: 10.1016/j.media.2022.102539.

Haili Ye, Yuan-Yuan Yang, Shunzhi Zhu, Dahan Wang, Xu-Yao Zhang, Xin Yang, and He Huang. Stain-adaptive self-supervised learning for histopathology image analysis. *Pattern Recognition*, 161:111242, 2025. doi: 10.1016/j.patcog.2024.111242.

Lu Yu, V. O. Yazici, Xialei Liu, Joost van de Weijer, Yongmei Cheng, and Arnau Ramisa. Learning metrics from teachers: Compact networks for image embedding. In *Computer Vision and Pattern Recognition*, pp. 2902–2911, 2019. doi: 10.1109/CVPR.2019.00302.

Tianyu Yu, Hao-Ye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. RLAIF-V: Open-Source AI Feedback Leads to Super GPT-4V Trustworthiness. In *Computer Vision and Pattern Recognition*, pp. 19985–19995, 2024. doi: 10.1109/CVPR52734.2025.01861.

Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating data processing and benchmarking of AI models for pathology. *arXiv.org*, abs/2502.06750, 2025. doi: 10.48550/arXiv.2502.06750.

Chengyang Zhang, Wenchuan Zhang, Bo Li, Xinyu Liu, Jiaming Yang, Mengran Li, Chenxun Deng, Jie Chen, Yang Zhang, Wei Ju, Yuhao Yi, Hong Bu, and Jiancheng Lv. Do Pathology Vision-Language Models Truly See Pathology? *arXiv.org*, abs/2607.21065, 2026a. doi: 10.48550/arXiv.2607.21065.

Mengliang Zhang. Quantifying and mitigating hospital domain bias in pathology foundation models using adversarial feature disentanglement. In Yuankai Huo, Mingchen Gao, Chang-Fu Kuo, Yueming Jin, and Ruining Deng (eds.), *Proceedings of The 9th International Conference on Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, pp. 3866–3884. PMLR, 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Xiao-Long Zhang, Jian-Wei Zhang, Selim Sevim, Emek Demir, Ece Eksi, and Xubo Song. Histopathology image normalization via latent manifold compaction. arXiv preprint arXiv:2602.24251, 2026b.

Yishu Zhang, Yun Li, and Daiwei Zhang. Multimodal LLMs outperform pathology foundation models in cross-domain histological similarity. In *Advances in Neural Information Processing Systems*, 2026c.

Yishu Zhang, Shushan Wu, Zhen-Ze Zhang, Didong Li, Huaxiu Yao, Yun Li, Iain Carmichael, K. Hoadley, Hongtu Zhu, Di Wu, and Daiwei Zhang. Mitigating batch effects in histopathology via Language-Mediated robust embedding generation. CVPR 2026, Findings poster, 2026d.

Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J. Gordon. On Learning Invariant Representations for Domain Adaptation. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7523–7532. PMLR, 2019.

Charles Y. Zheng, Francisco Pereira, Chris I. Baker, and Martin N. Hebart. Revealing interpretable object representations from human behavior. In *International Conference on Learning Representations*, volume abs/1901.02915, 2019.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Da-Cheng Li, Eric P. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Neural Information Processing Systems*, volume abs/2306.05685, pp. 46595–46623, 2023a. doi: 10.52202/075280-2020.

Zangwei Zheng, Mingyu Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *IEEE International Conference on Computer Vision*, pp. 19068–19079, 2023b. doi: 10.1109/ICCV51070.2023.01752.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. MegaPairs: Massive data synthesis for universal multimodal retrieval. In *Annual Meeting of the Association for Computational Linguistics*, pp. 19076–19095. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.acl-long.935.

Qifeng Zhou, Wenliang Zhong, Thao M. Dang, Hehuan Ma, Saiyang Na, Yuzhi Guo, and Junzhou Huang. Histopathology Multi-modal Embedding for Pathology Composed Retrieval, 2025b.

Hanwei Zhu, Xiangjie Sui, Baoliang Chen, Xuelin Liu, Peilin Chen, Yuming Fang, and Shiqi Wang. 2AFC Prompting of Large Multimodal Models for Image Quality Assessment. *IEEE transactions on circuits and systems for video technology (Print)*, 34:12873–12878, 2024a. doi: 10.1109/TCSVT.2024.3434999.

Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai, Weisi Lin, and Shiqi Wang. Adaptive Image Quality Assessment via Teaching Large Multimodal Model to Compare. In *Neural Information Processing Systems*, volume abs/2405.19298, 2024b. doi: 10.48550/arXiv.2405.19298.

Xinyu Zhu, Zhiguo Jiang, Kun Wu, Jun Shi, and Yushan Zheng. Adapting pathology foundation models for continual cross-center WSI retrieval. *Medical Image Anal.*, 113:104213, 2026. doi: 10.1016/j.media.2026.104213.

Eric Zimmermann, E. Vorontsov, Julian Viret, Adam Casson, Michał Zelechowski, George Shaikovski, Neil A. Tenenholtz, J. Hall, David S. Klimstra, Razik Yousfi, Thomas J Fuchs, Nicolò Fusi, Siqi Liu, and Kristen A. Severson. Virchow2: Scaling Self-Supervised mixed magnification models in pathology. *arXiv.org*, abs/2408.00738, 2024. doi: 10.48550/arXiv.2408.00738.

A APPENDIX

A.1 METHOD DETAILS

A.1.1 METHOD

ROPAL (RObustify PAthology models with a vision-Language judge) takes as input a pathology foundation model f_0 with frozen weights and unlabeled tiles from several slides and institutions. The output is an adapted encoder f_θ trained to improve acquisition robustness while preserving useful within-domain structure. The only metadata required for corpus construction are each tile's source slide and which slides belong to the same institution; institution names need not be known. No tissue-class, disease or diagnostic labels are required. The LLM judge supplies the biological-similarity judgments used for training. Institution membership is used only to assemble comparisons, is not supplied to the judge or adaptation loss, and is not required at inference. No scanner labels or registered image pairs are required. Write $d_\theta(x, y) = \|f_\theta(x) - f_\theta(y)\|_2^2$. Figure 1 summarizes the protocol.

Triplet construction A triplet consists of two same-domain tiles, SD1 (same-domain 1) and SD2 (same-domain 2), from one slide S at institution I, and a different-domain tile DD from a slide at a different institution. Slide and institution membership define these roles without assigning biological classes or declaring any two tiles similar. For each candidate triplet, the LLM judge determines whether two patches are substantially more similar in biological content to each other than to the third patch. Its verdict establishes whether a usable relative-similarity preference is available.

The role-blind odd-one-out judge The three tiles are shuffled in a fixed random order derived from the triplet index and shown to the judge as tiles 1, 2 and 3 with the prompt, verbatim: *"You are shown three histology tiles, in order: tile 1, tile 2, tile 3. Condition: two of these three tiles are substantially similar to each other compared to the third, judging biological content and ignoring non-biological artifacts (stain, scanner, preparation). Is the condition true? If true, which tile is the odd one out? Reply with ONLY a JSON object: {"condition": true/false, "odd": 1/2/3 or null}."* Nothing tells the judge which two tiles share a slide.

The judge is thus asked the general question of whether one patch is biologically substantially different from the other two; it is not told which patches share a slide. After the answer, the reported tile is mapped back to its hidden role. A triplet is selected, and enters the corpus, when the condition holds and the odd patch is a same-domain tile, SD1 or SD2. Write $\mathrm{SD_o}$ for the same-domain tile named odd and $\mathrm{SD_r}$ for the other same-domain tile. Such a verdict says that $\mathrm{SD_r}$ and DD are similar and $\mathrm{SD_o}$ differs from both, which establishes two distance rankings,

$$d(\mathrm{SD_r}, \mathrm{DD}) < d(\mathrm{SD_r}, \mathrm{SD_o}), \qquad d(\mathrm{SD_r}, \mathrm{DD}) < d(\mathrm{DD}, \mathrm{SD_o}), \tag{4}$$

whereas the relation between $d(\mathrm{SD_r}, \mathrm{SD_o})$ and $d(\mathrm{DD}, \mathrm{SD_o})$ cannot be determined from the verdict and is treated as a missing value. A triplet is discarded when the judge names DD as the odd one or finds no patch that differs from the other two. This leaves all three pairwise distance relations missing, and the triplet contributes nothing to training. When fine-tuning a model with ROPAL, only non-missing pairwise distance relations enter the Bradley–Terry (BT) loss; the loss uses the ranking relative to $\mathrm{SD_r}$, $d(\mathrm{SD_r}, \mathrm{DD}) < d(\mathrm{SD_r}, \mathrm{SD_o})$, one comparison per selected triplet. In other words, among all comparisons between two of the $\binom{n}{2}$ pairwise distances of n patches, the LLM judge labels only a subset, and those labels enter the training objective; every other comparison of distances is treated as having a missing label and is skipped during training.

The judge is OpenAI's `gpt-5.6-sol` at high or xhigh reasoning effort. Each triplet is judged independently, with the three tiles and the same prompt. Judgments are cached with the tile order and raw response. The released judgments retain the model, reasoning effort, image order and parsed JSON response of every record.

The corpus We curate the adaptation corpus from unlabeled slides using the slide and institution relationships described above. Biological matching and odd-one-out decisions are left to the LLM judge, without human-provided similarity labels or tissue annotations; a judged triplet is kept only when the judge names one of its two same-domain tiles as the odd one. A strength of the approach is that it is label-free. It does not curate a dataset with biologically validated labels or seek a teacher

whose labels are biologically validated. It transfers the similarity measurement of the LLM into the embedding model, and the robustness and downstream experiments (Section 4) measure the effect of this transfer on empirical performance against the frozen encoders and the baselines. Slide and institution membership serve only to assemble the comparisons. The slides come from The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013). Triplets were judged in 26 TCGA cohorts, about 184,000 judgments in all, giving 60,985 kept pairs (Table 35, Appendix A.8). The earlier 12-cohort configuration, trained on a subset of 42,820 pairs and labelled ROPAL (12-cohort) in the archive, appears only in archived comparisons.

Losses A kept triplet supplies the non-missing relation $d(\mathrm{SD_r}, \mathrm{DD}) < d(\mathrm{SD_r}, \mathrm{SD_o})$: DD should be closer to $\mathrm{SD_r}$ than $\mathrm{SD_o}$ is. The student is trained under a Bradley-Terry likelihood (Bradley & Terry, 1952; Hunter, 2003) on the difference of squared distances, using the pairwise logistic ranking form also employed by RankNet (Burges et al., 2005),

$$\mathcal{L}_{\mathrm{BT}} = -\mathbb{E}\big[\log \sigma\big(\beta\,(d_\theta(\mathrm{SD_r}, \mathrm{SD_o}) - d_\theta(\mathrm{SD_r}, \mathrm{DD}))\big)\big], \tag{5}$$

with $\beta = 2/\bar{m}_0$, where $\bar{m}_0$ is the mean absolute distance difference of the frozen encoder over the first three training batches, so that the initial logits are of order one whatever the feature scale.

A second term retains confident within-slide distance orderings of the frozen encoder. Each step draws $m = 8$ tiles from one slide of the training pools, giving $P = 28$ within-slide pairs with frozen squared distances d_0 and adapted squared distances d_θ. We compare all $\binom{28}{2} = 378$ unordered pairs of these distances. For a pair of pairs (p, q) the frozen margin is $m_0 = d_0(q) - d_0(p)$ and the adapted margin $m_\theta = d_\theta(q) - d_\theta(p)$, and

$$\mathcal{L}_{\mathrm{rank}} = \mathbb{E}_{(p,q):\,|m_0|>\tau}\big[\log\big(1 + e^{-\gamma\,\mathrm{sgn}(m_0)\,m_\theta}\big)\big], \qquad \gamma = 2/\mathbb{E}_{|m_0|>\tau}|m_0|, \tag{6}$$

where τ is the median of $|m_0|$ over the slide's pairs of pairs. The frozen distances are read from cached embeddings.

All table entries labelled *ROPAL* use the same loss configuration across the reported base encoders:

$$\mathcal{L}_{\mathrm{ROPAL}} = \mathcal{L}_{\mathrm{BT}} + \lambda\,\mathcal{L}_{\mathrm{rank}}, \tag{7}$$

where λ is the weight of the rank term (value in Appendix A.11). The triplet term is implemented as binary cross-entropy with logits on the scaled difference of squared embedding distances. These checkpoints use 60,985 kept triplets from 26 cohorts and one training epoch.

Only a low-rank adapter is trained (Hu et al., 2021): LoRA modules of rank 16, scaling 32 and dropout 0.05 in the attention and feed-forward projections of the last four transformer blocks. Training uses AdamW (Loshchilov & Hutter, 2017) at 10^{-4}, mixed precision and one epoch over batches of 12 pairs.

Reporting convention We report cross-domain robustness and within-domain performance separately. Each table specifies its evaluation subset and aggregation.

A.2 Experimental protocol details

A.2.1 Setup

Problem setup. All methods are evaluated under the same constraints. For training, a method may use unlabeled H&E tiles from TCGA with their slide and institution membership, together with the pretrained base encoder. ROPAL additionally queries a general-purpose multimodal LLM on these tiles. No method uses biological labels or other manual annotations, and no method is fitted on images or statistics of the evaluation cohorts. At inference, each test sample is embedded independently by a fixed mapping, without access to the test population, a calibration cohort or batch labels. Statistics of the sample's own pixels, as used by stain normalization, are allowed. Labels used by the downstream evaluation protocols, such as probe training splits and retrieval galleries, belong to the evaluation. The evaluation cohorts are external to the training data, with one qualification: 7 of the 35 BEETLE slides are TCGA breast cancer slides from 7 patients. None of these slides or patients is in our training corpus, the TCGA pool used by AdvInv, LEACE, INLP and Pathology DoRA, or the stain-normalization references. Four of their six tissue-source sites contribute other patients to training.

Students and pools. The student encoders are UNI2-h (Chen et al., 2024b), Virchow2 (Zimmermann et al., 2024) and GenBio-PathFM (Kapse et al., 2026). Training pools are unlabeled TCGA tiles, about 3,000 per cohort with 6 to 34 submitting institutions each.

MOSAIC. Separately from the unlabeled adaptation corpus, we evaluate on MOSAIC (Zhang et al., 2026c), which covers four public non-TCGA histopathology cohorts: HER2ST (Andersson et al., 2021), CCTGS (Arslan et al., 2025), Camelyon16 (Bejnordi et al., 2017; Litjens et al., 2018) and BEETLE (Lems et al., 2025). Each comparison pairs a reference H&E patch with a same-class patch from another slide or institution and a different-class patch from the reference's slide or institution. The within-domain sets of MOSAIC keep all three patches in the same slide or institution. The tissue annotations score the MOSAIC comparisons. The adaptation corpus and the adapters do not use them.

Evaluation. MOSAIC accuracy is the measure of Section 4.2.2 and is also measured in the ablations. Cross-domain accuracy is measured on the cross-slide (HER2ST, CCTGS) and cross-institution (Camelyon16, BEETLE) sets of MOSAIC, and within-domain accuracy on its sets that hold the slide or institution fixed. PathoROB (Kömen et al., 2026) is the external robustness benchmark with its robustness index, average performance drop under confounded training, excluding TCGA-LUSC from the evaluation while retaining the published leaderboard feature, against the robustification baselines it evaluates (Reinhard and Macenko). Preservation is checked on HEST-1k (Jaume et al., 2024); retrieval and classification protocols on annotated patches assess downstream performance. Appendix A.12 gives protocols.

Baselines. Every comparison applies the same frozen encoder to the same evaluation samples. None is the released encoder on the original patches. Reinhard (Reinhard et al., 2001) shifts and scales each LAB channel of a patch to the mean and standard deviation of a fixed reference before extraction, and Macenko (Macenko et al., 2009) estimates a patch's H&E stain vectors from the singular value decomposition of its optical densities and re-expresses its stain concentrations with a fixed reference stain matrix before extraction. LEACE (Belrose et al., 2023) applies a fixed affine map to the frozen embeddings, fitted in closed form on the same TCGA pool as AdvInv so that the tissue-source site is no longer linearly decodable while the embeddings move as little as possible in mean square. INLP (Ravfogel et al., 2020) removes the site information iteratively with the authors' code: a linear site classifier is trained on the same TCGA pool, the embeddings are projected onto its nullspace, and the step is repeated. The authors run a fixed number of iterations. We stop when the held-out site accuracy of the next classifier falls to the majority-site rate, when its projection would remove every remaining dimension, or after 10 rounds. The composed projection is applied as a fixed linear map. Pathology DoRA (Bøe et al., 2026) tunes rank-16 DoRA adapters in every block of the encoder with DINO self-distillation under strong colour and blur augmentation on the same TCGA pool, using the authors' released training code with a budget of 1,100 steps. AdvInv adapts the frozen embeddings themselves, as described next.

Appendix Table 36 lists these baselines, ROPAL and other published methods against the constraints of the problem setup.

Baseline configuration. The baselines are evaluated as they would be deployed without retuning: each uses its published default settings or a direct integration of the authors' code, and every departure from the original method is listed in Appendix Table 37. The AdvInv adversarial weight is the only baseline setting chosen from a screen, and it was chosen to limit a collapse of the latent space (Appendix A.11). No other baseline setting was selected with evaluation data. The comparisons therefore describe these methods at the tested settings, not their performance after per-task tuning.

AdvInv is the adversarial-invariance autoencoder of Dincer et al. (2020) and Edwards & Storkey (2015), fitted at native width to frozen patch embeddings using TCGA tissue-source-site labels and feature reconstruction. Its 62,445-image pool is sampled uniformly without replacement from the original randomly extracted TCGA patch inventory.

A.2.2 PATHOROB ROBUSTNESS INDEX ON EVERY BASE

Task. PathoROB evaluates tissue preservation across acquisition centers using patch embeddings from Camelyon and Tolkach ESCA, both external to the adaptation corpus. ROPAL uses the complete 60,985-comparison, 26-cohort training corpus with the objective of Section 3.2. UNI2-h, Virchow2 and GenBio reuse the identical full-corpus checkpoints used in the other main comparisons.

The CAMELYON collection (Litjens et al., 2018; Bándi et al., 2019) retains its two in-domain centers for RI and APD. Tolkach ESCA (Tolkach et al., 2023) retains 62 slides and 13,800 patches across WNS, CHA and UKK. We use the leaderboard feature view: CLS plus mean patch tokens where applicable, with native pooled GenBio embeddings.

RI measures the fraction of nearest neighbors matching tissue class across centers among neighbors matching either that relation or center across tissue classes. APD measures the change in in-domain classifier accuracy across the benchmark's confounding splits relative to its balanced split.

AdvInv uses the same general 62,445-image pool as the other comparisons (47,926 fitting and 14,519 participant-separated validation images), fitted on the PathoROB leaderboard feature view of each base. Selection does not consult LLM verdicts, the proposed method's corpus selection or feature availability. Evaluation uses Camelyon and Tolkach ESCA, with no fitting on their images or population statistics.

A.2.3 CROSS-DOMAIN SIMILARITY ON MOSAIC

Rationale. This test is a minimal assessment of the geometry of each foundation model's embedding space. A reference tile is compared with a tile of the same tissue class from another slide or institution and a tile of a different class from the reference's own slide or institution. If a model's embeddings cannot rank the biological distance ahead of the institutional difference, this indicates sensitivity to institution-specific batch effects. Together with benchmarks such as PathoROB, the test explicitly assesses a robustness property that standard benchmarks often find hard to measure. MOSAIC (Zhang et al., 2026c) uses the non-TCGA HER2ST, CCTGS, Camelyon16 and BEETLE cohorts (Section 4.1), held out from adaptation, with tissue classes annotated by pathologists (Appendix A.5.1). The format deliberately mirrors the judgements that built the TCGA adaptation corpus, so the test also shows whether the adapted encoders carry these judgements to other data sources, while the other experiments use tasks of other formats. Slide and institution membership only define the MOSAIC task. They are not shown to the judge or provided to the encoder as inputs, labels or targets, and at test time each embedding is computed from the image alone. The baselines (stain normalization, adversarial invariance, linear concept erasure and self-supervised parameter-efficient tuning) are evaluated on MOSAIC in the same way.

Task. The encoder is correct when it places the same-class tile closer to the reference than the different-class tile. The same-class tile comes from another slide (HER2ST, CCTGS) or institution (Camelyon16, BEETLE). The within-domain sets of MOSAIC, which hold the slide or institution fixed, are reported in Appendix A.5.2. A reverse-confound test, in which the same-class tile shares the reference's slide or institution and the different-class tile comes from another one, is reported in Appendix A.5.3. ROPAL keeps the reverse accuracy within 0.03 of the frozen encoders, while the same-domain bias (reverse minus forward accuracy) falls from 0.20–0.24 to 0.05–0.09, so a small preference for same-domain tiles remains.

A.2.4 SCANNER AND STAIN INVARIANCE

Task. Registered scans of the same tissue under different acquisition conditions test whether an embedding depends on the tissue rather than on the scanner or the staining. Both benchmarks retrieve matched tissue across conditions with cosine similarity.

SCORPION varies only the scanner: the same slides are scanned on five scanners. We divide each of SCORPION's 480 registered regions into sixteen tiles, yielding 7,680 physical tile locations and 38,400 scanned images. Each image is a query against the other 38,399 images pooled across all scanners, including its own scanner. The four other scans of the same location are relevant. Recall at one is the fraction of queries with a relevant image at rank one. Average precision is the mean precision at the four relevant ranks; mAP averages this quantity over all queries. Both metrics weight queries equally, without a separate scanner-pair average (Table 9).

PLISM provides 91 registered slides of the same tissue under 13 stainings and 7 scanners. We measure top-10 retrieval across stainings at 8,139 and 2,713 tiles per slide (Table 10). Scanner-only retrieval is reported separately in Section 6 (Table 20).

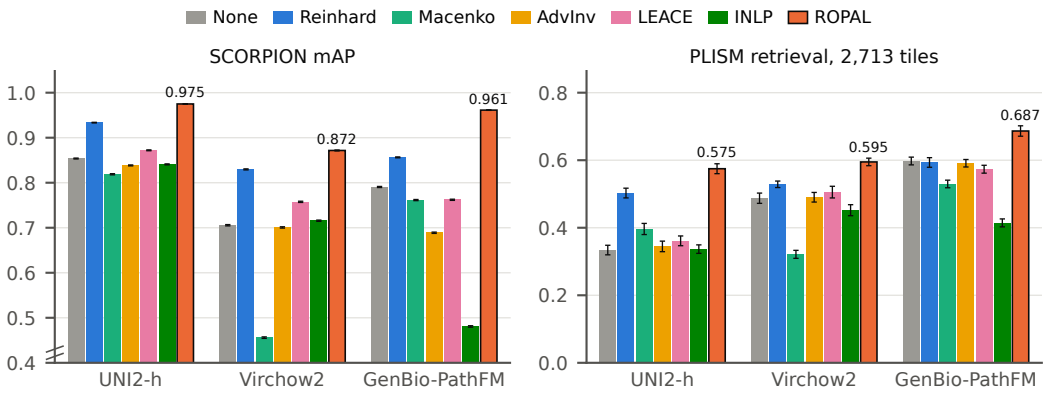


Figure 6: Scanner invariance on SCORPION (left: mAP of retrieving the other scanners' scans of each tile) and stain invariance on PLISM (right: median top-10 retrieval accuracy across stainings at 2,713 tiles per slide), the panels not shown in Figure 4. Higher is better, and error bars are bootstrap standard deviations. The SCORPION y-axis does not start at 0 (break marks).

A.2.5 Cross-institution retrieval and classification

Task. Retrieval and nearest-neighbour classification across institutions on Camelyon16 (tumor against normal), BEETLE (invasive epithelium, non-invasive epithelium and necrosis, four institutions) and TumSeg (tumor against non-tumor, two centers), with and without institution–class confounding.

With confounding (Present, tile protocol), each dataset's tiles are those of its cross-institution triplets, each pairing a reference tile with a same-class tile from another institution and a different-class tile from the same institution. The gallery of each query holds same-class tiles from other institutions as targets and different-class tiles from the query's institution as distractors, and each of 200 nearest-neighbour queries has one correct candidate from another institution and three wrong candidates from its own institution.

Without confounding (Absent, patch protocol), each transfer direction pairs a training institution with a held-out institution. Each held-out-institution patch retrieves the training institution's patches, and each of 200 held-out-institution queries per direction has four candidates from the training institution, one of its class and three of other classes. Values are averaged over the transfer directions. Camelyon16 uses 40 slides, ten per class and institution, with up to 2,000 patches per slide. BEETLE uses 3,830 labelled tiles from its four institutions, with every ordered pair of institutions as a transfer direction. TumSeg uses 48 slides, 24 per center, with 256 μm patches at 1.25$\times$.

Retrieval counts same-class items as relevant and is scored by mean average precision. Nearest-neighbour classification selects the candidate with the highest cosine similarity (chance 25%). Tables 2 and 12 compare all methods on UNI2-h, Virchow2 and GenBio-PathFM, with the same tiles, patches and candidate sets for every method.

Without confounding (Figure 7), ROPAL has the highest nearest-neighbour accuracy on 4 of the 6 dataset and base pairs. Its mean gain over the frozen encoder is 6.7 points, against 18.7 with confounding, so the gain is concentrated where class and institution are confounded.

A.2.6 Cross-study gene-expression prediction on HEST

Task. Spatial transcriptomics pairs every spot on a slide with an H&E patch and a gene-expression profile. The task is to predict each gene's expression at a spot from the patch alone, with a nearest-neighbour regressor fitted on the encoder's features, and to score how well the predicted spatial pattern of each gene correlates with the measured one.

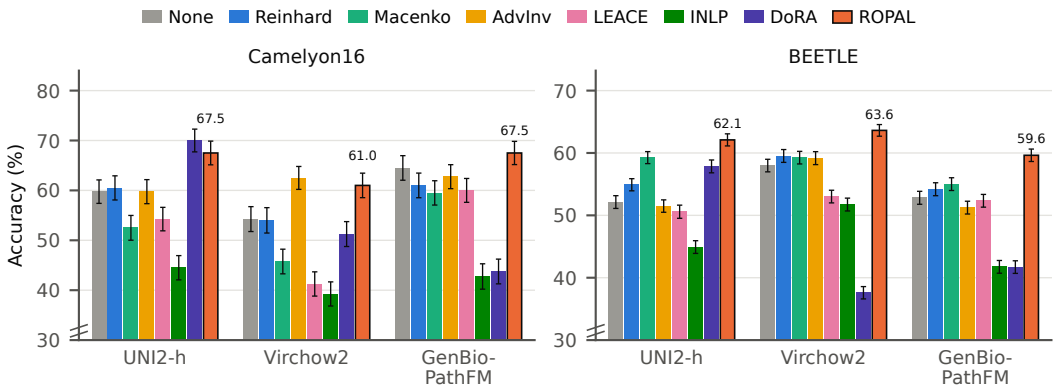


Figure 7: Fixed-candidate nearest-neighbour accuracy (percent) without institution–class confounding, per dataset, base model and method (the confounding-present comparison is Figure 5). Error bars are bootstrap standard deviations and the dashed line marks chance (25).

The cross-study evaluation on HEST-1k (Jaume et al., 2024) groups studies by their data-generating source and holds out one source group per fold. Per organ, the model is fitted on all other groups' slides and scored on every slide of the held-out group. The score is computed within each slide, with truth and prediction scaled to $[0, 1]$ per gene and the Pearson correlation averaged over the slide's 1,000 most variable genes, then averaged over the held-out group's slides and across source groups. Table 3 compares UNI2-h, Virchow2 and GenBio-PathFM under kNN regression with the stain-normalization baselines, per organ and averaged over the ten organs. The adapted model is the checkpoint trained on the 26-cohort corpus.

A.2.7 Mitotic figure classification across domains on MIDOG++

Task. Decide whether a 224 px crop centred on a candidate cell shows a mitotic figure or a hard-negative look-alike.

We use the MIDOG++ collection (Aubreville et al., 2023) for this cell-level task: crops centred on 11,937 mitotic figures and 14,349 hard negatives from seven tumor types and four scanners, evaluated leave-one-tumor-type-out and leave-one-scanner-out with a linear probe on every base (Table 4).

A.2.8 Cross-site prostate grading on PANDA

Task. Assign the ISUP grade (0 to 5, an ordinal scale of prostate cancer severity) to a prostate needle biopsy from its tiles.

PANDA (Bulten et al., 2022) is a two-site biopsy collection on which frozen foundation models are documented to degrade under Radboud to Karolinska transfer. We subsample 1,389 biopsies per site stratified by ISUP grade, train an aggregator on one site and test on the other, and report balanced accuracy averaged over the two transfer directions (Table 5). All encoders use this fixed biopsy subset, identical tiles and splits.

A.3 Ablation details

A.3.1 Ablation studies

All conditions use the same fixed evaluation manifests. The ablations report MOSAIC accuracy (Section 4.2.2, Table 8). MOSAIC cross (Average) and MOSAIC within (Average) are the unweighted mean accuracies over the four cross-domain and four within-domain sets of MOSAIC in Tables 8 and 23. The ablations also report the Same-domain bias, SCORPION and PLISM rows of the judge-selection ablation (Section A.3.1).

The ablations use UNI2-h unless stated otherwise and train fresh adapters in a matched setting: an identical 20,000-comparison manifest sampled proportionally across 26 cohorts and processed in one pass of 19,992 comparisons in 1,666 complete minibatches, with initialization, temperature, triplet order, rank-input draws, forward batches and optimization settings matched between conditions.

Loss components This comparison removes the rank term $\mathcal{L}_{\text{rank}}$ (Eq. 2) from the ROPAL objective (Eq. 3). BT only trains with $\mathcal{L}_{\text{BT}}$ (Eq. 1) alone, and ROPAL (BT + rank) is the ROPAL condition of the matched setting. Within each base, both arms share every other setting of the matched setting.

Judge selection This comparison isolates the judge's selection of training triplets from the preference it labels. Each arm trains a fresh adapter with the ROPAL objective (Eq. 3) and on the same number of triplets per cohort. ROPAL uses the triplets selected by the judge, each with the preference $d(\text{SD}_\text{r}, \text{DD}) < d(\text{SD}_\text{r}, \text{SD}_\text{o})$ of Section 3.1. It is the ROPAL condition of the matched setting, whose checkpoints are reused. Random draws every triplet from all TCGA tiles of the cohort using only slide and institution membership: two same-domain tiles SD1 and SD2 from one slide and a different-domain tile DD from a different institution, independent of the judge, and imposes $d(\text{SD1}, \text{DD}) < d(\text{SD1}, \text{SD2})$, the ROPAL preference with SD2 as SD_o. MNN and Odd-one-out score select triplets with the frozen encoder instead of the judge and impose the same preference. MNN pairs tiles of the same cohort from different institutions that are mutual neighbours: each is among the other's 20 most similar tiles from other institutions of the cohort (cosine similarity of frozen UNI2-h features). One tile of the pair, chosen at random, is SD1, the other is DD, and SD2 is drawn from the 5% of the other tiles of SD1's slide least similar to SD1 (at least one tile). Odd-one-out score ranks the triplets presented to the judge by $s = \min\{\text{sim}(\text{SD1}, \text{DD}) - \text{sim}(\text{SD1}, \text{SD2}),\ \text{sim}(\text{SD1}, \text{DD}) - \text{sim}(\text{DD}, \text{SD2})\}$ under frozen UNI2-h features and keeps the highest-scoring triplets of each cohort. Both use the same triplets for every base. Within each base, initialization, temperature, minibatch size, rank-input draws, optimization settings and evaluation manifests are identical. Only the selection of the triplets differs. MOSAIC cross and MOSAIC within are the MOSAIC accuracies of Section 4.2.2 and Section A.5.2, Same-domain bias is the reverse minus forward accuracy of the reverse-confound test of Section A.5.3 (lower is better, and 0 means no same-domain preference), and the SCORPION rows follow the scanner-invariance protocol of Section 4.3. The PLISM rows give the staining retrieval of Table 10 (Section 4.3) computed only on the 13 stainings of a single scanner (78 of the 546 inter-staining slide pairs), so they are not comparable with the values of Table 10.

Choice of judge LLM This comparison measures how the identity of the judge LLM affects the adapted encoder. The production judge is gpt-5.6-sol at high reasoning effort, and the alternative is the open-weight Qwen3.8-27B. Both judges assess the same candidate triplets with the same role-blind prompt and the same randomized tile order for every triplet, and the same selection rule applies (Section 3.1): a triplet enters the corpus when the judge finds two tiles substantially more similar to each other than to the third and names a same-domain tile, SD1 or SD2, as the odd one. As a rough estimate of judging cost, Qwen3.8-27B uses about 11,000 tokens (prompt and completion, including reasoning) per triplet and gpt-5.6-sol about 3,000, from each judge's mean tokens per verdict and its keep rate. Each judge's corpora have the two smallest corpus sizes of the corpus-size ablation (Table 19), and the larger corpus extends the smaller one. Each arm trains a fresh adapter with the ROPAL objective (Eq. 3) and imposes the same training preference, $d(\text{SD}_\text{r}, \text{DD}) < d(\text{SD}_\text{r}, \text{SD}_\text{o})$. Within a base, initialization, temperature, minibatch size, rank-input draws, optimization settings and evaluation manifests are identical; the gpt-5.6-sol arms are the checkpoints of the two smallest corpora in Table 19. The rows are those of Table 16. Both judges improve on Frozen in every row. The open-weight Qwen3.8-27B gives results close to the proprietary gpt-5.6-sol: over the 28 paired values, the mean absolute difference is 0.005 (largest 0.015), and gpt-5.6-sol is better in 20 of them.

Contribution of preference supervision We replace the BT term with cross-entropy for a source-site classifier trained through gradient reversal (Ganin et al., 2015), retaining the $\lambda\,\mathcal{L}_{\text{rank}}$ term of the ROPAL objective (Eq. 3). Let $h_\theta(x)$ denote the adapted embedding of tile x, $s(x) \in \{1, \ldots, K\}$ its TCGA tissue-source-site code ($K = 542$), and $g_\phi : \mathbb{R}^D \to \mathbb{R}^K$ a classifier with parameters $\phi = (W_1, b_1, W_2, b_2)$,

$$g_\phi(h) = W_2\,\text{GELU}\big(W_1\,\text{LN}(h) + b_1\big) + b_2, \qquad W_1 \in \mathbb{R}^{256\times D},\ W_2 \in \mathbb{R}^{K\times 256}, \tag{8}$$

where LN is layer normalization without affine parameters and D is the embedding width of the base encoder. Writing $g_\phi(h)_k$ for the k-th logit, the site cross-entropy over the three tiles of every judged comparison $(\mathrm{SD1}, \mathrm{SD2}, \mathrm{DD})$ is

$$\mathcal{L}_{\mathrm{CE}}(\theta, \phi) = -\mathbb{E}_{(\mathrm{SD1},\mathrm{SD2},\mathrm{DD})}\left[\frac{1}{3}\sum_{x\in\{\mathrm{SD1},\mathrm{SD2},\mathrm{DD}\}} \log \frac{\exp\left(g_\phi(h_\theta(x))_{s(x)}\right)}{\sum_{k=1}^{K}\exp\left(g_\phi(h_\theta(x))_k\right)}\right], \tag{9}$$

where the expectation runs over the judged comparisons of a minibatch, so a comparison contributes its three tiles with equal weight and no judge label enters the term. Gradient reversal inserts R_η between h_θ and g_ϕ, where $R_\eta(h) = h$ in the forward pass and $\partial R_\eta/\partial h = -\eta I$ in the backward pass, with $\eta = 1$. Writing $\mathcal{L}_{\mathrm{adv}}(\theta, \phi) = \mathcal{L}_{\mathrm{CE}}$ evaluated on $g_\phi(R_\eta(h_\theta(x)))$, one gradient step on $\mathcal{L}_{\mathrm{adv}}$ moves ϕ to decrease the cross-entropy and θ to increase it, so the domain-adversarial control solves

$$\min_\theta \ \big[-\eta\, \mathcal{L}_{\mathrm{CE}}(\theta, \phi) + \lambda\, \mathcal{L}_{\mathrm{rank}}(\theta)\big] \quad \text{with} \quad \min_\phi \ \mathcal{L}_{\mathrm{CE}}(\theta, \phi), \tag{10}$$

implemented as the single loss $\mathcal{L}_{\mathrm{control}} = \mathcal{L}_{\mathrm{adv}} + \lambda\, \mathcal{L}_{\mathrm{rank}}$ with one AdamW step over (θ, ϕ) per minibatch. Neither the judge preferences nor $\mathcal{L}_{\mathrm{BT}}$ (Eq. 1) enter this objective; $\mathcal{L}_{\mathrm{rank}}$ is Eq. 2. For UNI2-h and Virchow2, this comparison uses exactly the training dataset and evaluation manifests of the matched setting: the same 19,992 image triplets across 26 cohorts and the same MOSAIC sets. Its 1,666 updates, adapter initialization, image order and rank-input draws match that setting's ROPAL condition for each base, whose checkpoint and scores are reused here. The classifier receives the three adapted tile embeddings separately and predicts the TCGA tissue-source-site code used for source grouping in the corpus. The classifier has input layer normalization, a 256-unit GELU hidden layer and a source-site output layer. Gradient reversal has coefficient 1; the classifier and adapter learning rates are 10^{-3} and 10^{-4}. The classifier is discarded after adaptation.

This domain-adversarial control uses no tissue-class labels or judge preferences in its loss and omits original DANN's supervised task loss. Both arms use the same training corpus.

A further arm keeps the BT preference term and adds the same source-site classifier and cross-entropy to the full objective, so that all adapted conditions can be written with the same terms:

$$\text{ROPAL (BT + rank)}: \ \min_\theta \ \mathcal{L}_{\mathrm{BT}}(\theta) + \lambda\, \mathcal{L}_{\mathrm{rank}}(\theta), \tag{11}$$

$$\text{Domain adversarial + rank}: \ \min_\theta \ -\eta\, \mathcal{L}_{\mathrm{CE}}(\theta, \phi) + \lambda\, \mathcal{L}_{\mathrm{rank}}(\theta), \quad \min_\phi \ \mathcal{L}_{\mathrm{CE}}(\theta, \phi), \tag{12}$$

$$\text{BT + domain adversarial + rank}: \ \min_\theta \ \mathcal{L}_{\mathrm{BT}}(\theta) - \eta\, \mathcal{L}_{\mathrm{CE}}(\theta, \phi) + \lambda\, \mathcal{L}_{\mathrm{rank}}(\theta), \quad \min_\phi \ \mathcal{L}_{\mathrm{CE}}(\theta, \phi) \tag{13}$$

with $\eta = 1$ throughout. The adversarial arms are implemented with the gradient-reversal layer, so the encoder receives $-\eta\, \partial\mathcal{L}_{\mathrm{CE}}/\partial\theta$ while the classifier receives $+\partial\mathcal{L}_{\mathrm{CE}}/\partial\phi$. All arms share each base's initialization, minibatch streams, classifier initialization, learning rates and 1,666 updates; the classifier is discarded after adaptation.

Corpus size With the ROPAL objective (Eq. 3), nested, cohort-stratified prefixes contain 1,248, 2,496, 5,004, 9,996 and 19,992 distinct judged comparisons across all 26 cohorts. Their cohort proportions are approximately matched and the rank-input pool is fixed. Conditions use the same initialization and minibatch sequence, with 104, 208, 417, 833 and 1,666 optimizer updates, respectively. This measures one-pass data-and-compute scaling. Counts denote distinct judged comparisons; an image can participate in multiple comparisons.

Every condition is trained on UNI2-h and on Virchow2 with the corpus of the production judge (gpt-5.6-sol).

A.4 Full result tables

The main text shows summaries of some result tables. This subsection gives each of them in full, with its complete caption.

Full caption of Figure 1. Overview of the protocol. (1) A candidate triplet combines two same-domain patches, SD1 and SD2, from one slide and a different-domain patch DD from a slide at another institution, using only slide and institution membership. The cartoon patches show the tissue class by their content and the domain by their stain tint. (2) A multimodal LLM sees the three patches in random order, without their roles, and is asked whether any patch is substantially different from the others biologically. (3) The triplet is selected when the LLM names SD1 or SD2 as the odd one, and discarded when it names DD or no patch. A selected triplet gives the preference $d_\theta(\mathrm{SD_r}, \mathrm{DD}) < d_\theta(\mathrm{SD_r}, \mathrm{SD_o})$, where $\mathrm{SD_o}$ is the same-domain patch named odd and $\mathrm{SD_r}$ the other one. (4) The pairwise distance ranking loss (Eq. 1) moves DD closer to $\mathrm{SD_r}$ than $\mathrm{SD_o}$ in the embedding space of the encoder f_θ, where marker shape shows the domain and colour the class.

Full caption of Table 5. PANDA cross-site ISUP grading: 1,389 biopsies per site, a grade-stratified subsample after the eva noisy-label exclusion; attention-based multiple-instance learning with a six-class head. Reinhard and Macenko use frozen encoders on stain-normalized versions of the same tiles, with fixed reference targets and unchanged splits and aggregator settings. Values are cross-site average balanced accuracy. Each value is the unweighted arithmetic mean of Radboud to Karolinska and Karolinska to Radboud transfer, computed before rounding. None: frozen base; ROPAL: 26-cohort checkpoint. Bold: best; underlined: second best.

Table 7: Full version of Table 1. Modified PathoROB on Camelyon and Tolkach ESCA: robustness index (RI; higher is better) and in-domain average performance drop (APD), in percentage points (at most zero; higher is better). Each metric is shown for Camelyon, for Tolkach ESCA and as Average, the unweighted mean of the two dataset scores before rounding. APD is the mean over a dataset's confounded splits of the relative in-domain accuracy change clipped at zero, $\min(0, \mathrm{acc}_s/\mathrm{acc}_0 - 1)$, with each split's accuracy averaged over repetitions, so gains under confounding do not offset drops. ROPAL uses the full 60,985-comparison adaptation corpus. Bold: best; underlined: second best.

Metric	Dataset	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
RI	Camelyon	UNI2-h	0.544	0.839	0.819	0.770	0.512	0.506	0.780	**0.888**
RI	Camelyon	Virchow2	0.808	**0.923**	0.725	0.870	0.826	0.760	0.003	0.904
RI	Camelyon	GenBio-PathFM	0.862	0.898	0.848	0.919	0.790	0.426	0.125	**0.935**
RI	Tolkach ESCA	UNI2-h	0.923	0.940	0.945	0.955	0.928	0.869	0.949	**0.963**
RI	Tolkach ESCA	Virchow2	0.954	0.954	0.953	0.958	0.954	0.932	0.193	**0.967**
RI	Tolkach ESCA	GenBio-PathFM	0.959	0.961	0.958	0.963	0.958	0.853	0.687	**0.970**
RI	Average	UNI2-h	0.734	0.890	0.882	0.862	0.720	0.687	0.864	**0.926**
RI	Average	Virchow2	0.881	**0.939**	0.839	0.914	0.890	0.846	0.098	0.935
RI	Average	GenBio-PathFM	0.911	0.930	0.903	0.941	0.874	0.640	0.406	**0.953**
APD	Camelyon	UNI2-h	-4.63	-1.62	-1.79	-3.62	-5.56	-9.15	-3.33	**-1.45**
APD	Camelyon	Virchow2	-2.03	-1.74	-3.94	-4.48	-4.66	-5.84	-12.91	**-1.24**
APD	Camelyon	GenBio-PathFM	-1.90	-1.53	-1.81	-2.58	-4.28	-8.80	-11.42	**-1.06**
APD	Tolkach ESCA	UNI2-h	-0.48	-0.05	-0.23	-0.02	-0.01	-1.06	-0.03	**+0.00**
APD	Tolkach ESCA	Virchow2	-0.15	-0.22	-0.42	-0.50	-0.36	-1.31	-8.39	**-0.10**
APD	Tolkach ESCA	GenBio-PathFM	**+0.00**	**+0.00**	**+0.00**	**+0.00**	-0.07	-2.62	-8.53	-0.24
APD	Average	UNI2-h	-2.55	-0.83	-1.01	-1.82	-2.78	-5.10	-1.68	**-0.73**
APD	Average	Virchow2	-1.09	-0.98	-2.18	-2.49	-2.51	-3.58	-10.65	**-0.67**
APD	Average	GenBio-PathFM	-0.95	-0.76	-0.90	-1.29	-2.17	-5.71	-9.97	**-0.65**

A.5 Additional experiment results

A.5.1 Validation of the LLM judgements

The adaptation corpus is built from the judgements of a multimodal LLM. As a validation check of this supervision signal, we compare the judgements with the tissue-class annotations of six datasets: HER2ST (Andersson et al., 2021), CCTGS (Arslan et al., 2025), TIGER (van Rijthoven et al., 2025), Camelyon16 (Bejnordi et al., 2017; Litjens et al., 2018), BEETLE (Lems et al., 2025) and TCGA-LUSC (Weinstein et al., 2013). TCGA-LUSC is a TCGA cohort. The other five are not.

Sets. The HER2ST, CCTGS, Camelyon16 and BEETLE sets are the cross-domain sets of MOSAIC (Zhang et al., 2026c) (Section 4.1), and TIGER and TCGA-LUSC add two further sets of the same form. Each holds a reference tile, a tile of the same tissue class taken from a different slide (the cross-slide HER2ST, CCTGS and TIGER sets) or a different institution (the cross-institution Camelyon16,

Table 8: Figure 3 in the main text shows this table. Cross-domain accuracy of the embeddings on MOSAIC, per non-TCGA cohort. The same-class candidate comes from another slide or institution and the different-class candidate from the reference's slide or institution. None is the frozen base. ROPAL is the checkpoint trained on the 26-cohort corpus with the objective of Section 3.2. MOSAIC cross (Average) is the unweighted mean of HER2ST, CCTGS, Camelyon16 and BEETLE, computed before rounding. MOSAIC forward–reverse (Average) is the mean of forward and reverse accuracy on the reverse-confound sets (Appendix A.5.3), averaged over the same four datasets. Within-domain results are in Table 23. Higher is better. Bold: best; underlined: second best.

Metric	Dataset	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
MOSAIC cross	HER2ST	UNI2-h	0.676	0.725	0.674	0.658	0.662	0.649	0.729	**0.855**
MOSAIC cross	HER2ST	Virchow2	0.693	0.750	0.604	0.764	0.683	0.658	0.694	**0.870**
MOSAIC cross	HER2ST	GenBio-PathFM	0.695	0.742	0.706	0.777	0.684	0.557	0.821	**0.884**
MOSAIC cross	CCTGS	UNI2-h	0.694	0.785	0.777	0.736	0.729	0.621	0.815	**0.954**
MOSAIC cross	CCTGS	Virchow2	0.858	0.875	0.841	0.859	0.838	0.644	0.651	**0.963**
MOSAIC cross	CCTGS	GenBio-PathFM	0.878	0.883	0.887	0.831	0.841	0.583	0.710	**0.967**
MOSAIC cross	Camelyon16	UNI2-h	0.789	0.848	0.771	0.776	0.780	0.735	0.848	**0.964**
MOSAIC cross	Camelyon16	Virchow2	0.776	0.892	0.709	0.767	0.803	0.650	0.377	**0.955**
MOSAIC cross	Camelyon16	GenBio-PathFM	0.901	0.901	0.888	0.767	0.901	0.610	0.682	**0.973**
MOSAIC cross	BEETLE	UNI2-h	0.688	0.731	0.723	0.665	0.715	0.655	0.723	**0.860**
MOSAIC cross	BEETLE	Virchow2	0.678	0.735	0.749	0.684	0.690	0.661	0.530	**0.758**
MOSAIC cross	BEETLE	GenBio-PathFM	0.682	0.717	0.747	0.665	0.667	0.542	0.571	**0.848**
MOSAIC cross	Average	UNI2-h	0.712	0.772	0.736	0.709	0.721	0.665	0.779	**0.908**
MOSAIC cross	Average	Virchow2	0.751	0.813	0.726	0.768	0.753	0.653	0.563	**0.886**
MOSAIC cross	Average	GenBio-PathFM	0.789	0.811	0.807	0.760	0.773	0.573	0.696	**0.918**
MOSAIC forward–reverse	Average	UNI2-h	0.823	0.858	0.839	0.824	0.825	0.780	0.866	**0.934**
MOSAIC forward–reverse	Average	Virchow2	0.861	0.891	0.828	0.864	0.851	0.775	0.699	**0.923**
MOSAIC forward–reverse	Average	GenBio-PathFM	0.876	0.888	0.884	0.850	0.863	0.726	0.799	**0.939**

Table 9: Figure 4 in the main text shows this table. Scanner-only invariance on SCORPION: 7,680 physical tile locations scanned on five scanners (38,400 images). Each query retrieves from all other images, with four relevant scanner copies of the same location. R@1 and mAP are unweighted query means. Per base: frozen (None), frozen with Reinhard or Macenko stain normalization of every tile, and the checkpoint trained on the 26-cohort corpus with the objective of Section 3.2 (ROPAL). Rows pair each metric with a base model; columns compare methods. Higher is better. Bold: best; underlined: second best.

Metric	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
R@1	GenBio-PathFM	0.884	0.946	0.885	0.823	0.848	0.560	0.394	**0.990**
R@1	UNI2-h	0.967	0.989	0.926	0.956	0.969	0.953	0.981	**0.994**
R@1	Virchow2	0.815	0.924	0.612	0.807	0.858	0.824	0.000	**0.954**
mAP	GenBio-PathFM	0.791	0.856	0.761	0.689	0.762	0.481	0.165	**0.961**
mAP	UNI2-h	0.854	0.933	0.819	0.838	0.872	0.841	0.919	**0.975**
mAP	Virchow2	0.706	0.830	0.456	0.701	0.758	0.716	0.002	**0.872**

Table 10: Figure 4 in the main text shows this table. Staining invariance on PLISM (plismbench, 91 slides of the same tissue under 13 stainings and 7 scanners) for the indicated base encoders, at 8,139 and 2,713 tiles per slide. Retrieval: median top-10 retrieval accuracy over inter-staining pairs. None: frozen; Reinhard and Macenko: frozen encoder on stain-normalized tiles; Higher is better. ROPAL: the checkpoint trained on the 26-cohort corpus with the objective of Section 3.2. Bold: best; underlined: second best.

Tiles per slide	Base model	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
8,139	GenBio-PathFM	0.409	0.404	0.347	0.401	0.382	0.240	0.166	**0.529**
8,139	UNI2-h	0.190	0.332	0.254	0.198	0.210	0.192	0.291	**0.409**
8,139	Virchow2	0.299	0.352	0.189	0.307	0.325	0.288	0.005	**0.423**
2,713	GenBio-PathFM	0.598	0.593	0.530	0.591	0.573	0.414	0.268	**0.687**
2,713	UNI2-h	0.334	0.503	0.396	0.345	0.361	0.337	0.459	**0.575**
2,713	Virchow2	0.487	0.529	0.321	0.490	0.506	0.452	0.011	**0.595**

Table 11: Full version of Table 2. Cross-institution retrieval mean average precision. Institution–class confounding Present: the gallery of each query holds same-class targets from other institutions and different-class distractors from the query's institution (tile protocol). Absent: held-out-institution patches retrieve the training institution's patches (patch protocol, mean over transfer directions). Camelyon16 (tumor against normal), BEETLE (three tissue classes, four institutions) and TumSeg (Sancéré et al., 2025) (tumor against non-tumor), in percent. Reinhard and Macenko use stain-normalized inputs to the frozen encoder. ROPAL: the checkpoint trained on the 26-cohort corpus with the objective of Section 3.2. Bold: best; underlined: second best.

Institution–class confounding	Dataset	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
Present	Camelyon16	UNI2-h	73.8	78.0	72.7	71.2	72.2	69.6	80.6	**86.0**
Present	Camelyon16	Virchow2	69.6	76.5	65.5	73.5	69.8	57.6	49.6	**84.1**
Present	Camelyon16	GenBio-PathFM	81.7	80.8	79.6	80.9	80.8	53.8	58.1	**88.0**
Present	BEETLE	UNI2-h	60.3	61.7	61.7	63.0	60.9	57.6	64.1	**73.2**
Present	BEETLE	Virchow2	65.3	66.7	65.1	67.2	67.6	65.2	63.8	**73.0**
Present	BEETLE	GenBio-PathFM	62.8	64.5	64.2	62.1	64.4	56.3	60.3	**72.1**
Present	TumSeg	UNI2-h	53.3	52.1	50.0	54.6	52.8	52.4	54.4	**56.8**
Present	TumSeg	Virchow2	46.1	46.4	49.2	46.4	46.2	46.7	**50.5**	48.1
Present	TumSeg	GenBio-PathFM	53.3	52.0	52.4	**55.9**	52.1	51.5	52.1	54.7
Absent	Camelyon16	UNI2-h	89.9	89.5	88.8	89.4	88.3	84.8	**92.8**	92.3
Absent	Camelyon16	Virchow2	90.5	90.0	86.5	**90.9**	86.9	82.8	82.5	90.7
Absent	Camelyon16	GenBio-PathFM	**92.4**	91.5	91.3	90.4	90.9	83.8	84.0	92.2
Absent	BEETLE	UNI2-h	57.0	58.6	60.1	56.9	56.7	52.1	60.7	**61.7**
Absent	BEETLE	Virchow2	61.9	61.7	**62.7**	61.5	58.9	57.1	44.3	62.4
Absent	BEETLE	GenBio-PathFM	57.0	57.5	58.2	56.7	57.2	49.1	46.9	**59.5**
Absent	TumSeg	UNI2-h	62.7	61.2	57.9	62.2	62.6	62.1	63.1	**64.3**
Absent	TumSeg	Virchow2	58.3	58.2	58.3	58.5	58.1	58.0	58.4	**59.1**
Absent	TumSeg	GenBio-PathFM	61.4	60.8	59.9	**61.7**	61.3	61.1	**61.7**	**61.7**

Table 12: Figure 5 in the main text shows this table. Fixed-candidate nearest-neighbour accuracy (chance 25%). Institution–class confounding Present: the correct candidate is at another institution and the three wrong candidates are at the query's institution (tile protocol). Absent: all four candidates come from the training institution (patch protocol, mean over transfer directions). Camelyon16 (tumor against normal), BEETLE (three tissue classes, four institutions) and TumSeg (Sancéré et al., 2025) (tumor against non-tumor), in percent. Reinhard and Macenko use stain-normalized inputs to the frozen encoder. ROPAL: the checkpoint trained on the 26-cohort corpus with the objective of Section 3.2. Bold: best; underlined: second best.

Institution–class confounding	Dataset	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
Present	Camelyon16	UNI2-h	61.0	68.5	58.0	55.5	59.5	48.5	69.5	**78.0**
Present	Camelyon16	Virchow2	55.0	62.0	46.5	60.5	58.5	34.5	20.0	**75.0**
Present	Camelyon16	GenBio-PathFM	71.5	72.0	69.0	69.5	66.5	25.0	35.5	**81.5**
Present	BEETLE	UNI2-h	18.5	21.5	22.5	27.0	21.0	14.5	24.0	**44.0**
Present	BEETLE	Virchow2	27.0	32.5	29.0	30.5	32.0	28.0	27.5	**48.0**
Present	BEETLE	GenBio-PathFM	22.5	26.5	26.5	21.5	27.5	16.0	27.0	**41.0**
Present	TumSeg	UNI2-h	33.0	29.5	19.0	**38.0**	32.5	31.5	31.5	**38.0**
Present	TumSeg	Virchow2	20.0	22.5	21.0	20.0	20.5	22.0	**28.0**	22.5
Present	TumSeg	GenBio-PathFM	34.5	28.5	28.0	32.5	34.0	31.0	29.0	**35.5**
Absent	Camelyon16	UNI2-h	59.7	60.5	52.5	59.7	54.2	44.5	**70.0**	67.5
Absent	Camelyon16	Virchow2	54.2	54.0	45.8	**62.5**	41.2	39.2	51.2	61.0
Absent	Camelyon16	GenBio-PathFM	64.5	61.0	59.5	62.7	60.0	42.8	43.8	**67.5**
Absent	BEETLE	UNI2-h	52.1	54.9	59.3	51.5	50.6	44.9	57.8	**62.1**
Absent	BEETLE	Virchow2	58.0	59.5	59.3	59.2	53.0	51.7	37.6	**63.6**
Absent	BEETLE	GenBio-PathFM	52.8	54.2	55.0	51.2	52.3	41.7	41.7	**59.6**
Absent	TumSeg	UNI2-h	37.8	33.5	29.2	39.5	37.8	36.8	37.8	**41.8**
Absent	TumSeg	Virchow2	25.8	25.2	27.2	24.5	25.5	25.5	27.3	**28.0**
Absent	TumSeg	GenBio-PathFM	32.5	28.7	27.5	**33.8**	32.8	32.8	31.5	33.2

BEETLE and TCGA-LUSC sets), and a tile of a different class taken from the reference's own slide or institution. The HER2ST classes are cancer, breast glands, connective tissue, adipose tissue and immune infiltrate. The CCTGS classes are tumor, normal mucosa and four other colorectal compartments. The TIGER classes are tumor, stroma, healthy glands and necrosis. Camelyon16 has tumor and normal lymph-node tissue. BEETLE has invasive epithelium, non-invasive epithelium and necrosis. TCGA-LUSC has tumor and normal lung tissue.

Table 13: Full version of Table 3. HEST cross-study gene-expression prediction with kNN ($k = 20$), per organ, for UNI2-h, Virchow2 and GenBio-PathFM: frozen (None), frozen on Reinhard- or Macenko-normalized tiles, and ROPAL, the checkpoint trained on the 26-cohort corpus (Table 8). Human Visium and legacy ST slides, whole-transcriptome, one slide per subject. Studies are grouped by their data-generating source; per organ, each source group is held out in turn from supervised training. Features predict log1p counts of the candidate genes (union over slides of each slide's 1,000 most variable genes). Metric: per slide, truth and prediction scaled to $[0, 1]$ per gene, Pearson per gene within the slide, mean over the slide's 1,000 most variable genes; then the mean over the held-out group's slides and the unweighted mean over source groups (macro r). Organs with at least ten slides are shown; the last block, Average, is the unweighted mean over these ten organs. Higher is better. Bold: best; underlined: second best.

Organ	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
Breast	UNI2-h	0.157	0.195	0.130	0.151	0.114	0.100	**0.204**	0.192
–	Virchow2	0.203	0.176	0.083	0.100	0.132	0.060	0.049	**0.206**
–	GenBio-PathFM	0.130	0.144	**0.161**	0.091	0.099	0.035	0.067	**0.161**
Bowel	UNI2-h	0.240	0.287	0.291	0.251	0.242	0.233	0.280	**0.326**
–	Virchow2	0.250	0.279	0.174	0.170	0.254	0.160	-0.013	**0.313**
–	GenBio-PathFM	0.281	0.265	0.262	0.260	0.269	0.041	0.051	**0.330**
Kidney	UNI2-h	0.083	0.076	0.084	0.077	**0.095**	0.086	0.075	0.087
–	Virchow2	0.091	0.074	0.046	0.078	0.076	0.057	0.046	**0.105**
–	GenBio-PathFM	**0.090**	**0.090**	0.069	0.052	0.069	0.043	0.072	0.083
Brain	UNI2-h	0.039	0.051	0.016	0.043	0.054	0.028	0.061	**0.073**
–	Virchow2	0.061	0.054	0.019	0.055	0.059	0.042	0.018	**0.074**
–	GenBio-PathFM	0.035	0.014	0.029	**0.058**	0.022	0.021	0.007	0.049
Heart	UNI2-h	0.040	0.057	0.043	0.029	0.040	0.043	0.055	**0.061**
–	Virchow2	0.036	0.050	0.038	0.035	0.030	0.026	0.033	**0.052**
–	GenBio-PathFM	0.050	**0.054**	0.047	0.022	0.038	0.015	0.019	0.050
Prostate	UNI2-h	0.167	0.144	0.147	0.151	0.167	0.103	0.183	**0.244**
–	Virchow2	0.162	0.130	0.075	0.164	0.173	0.108	-0.095	**0.222**
–	GenBio-PathFM	0.168	0.163	0.170	0.128	0.205	0.049	-0.004	**0.242**
Lung	UNI2-h	0.097	0.117	0.126	0.078	0.131	0.121	0.102	**0.251**
–	Virchow2	0.114	0.133	0.088	0.153	0.116	-0.009	0.024	**0.173**
–	GenBio-PathFM	0.070	0.116	0.100	0.057	0.070	0.061	0.047	**0.209**
Liver	UNI2-h	0.151	0.134	0.144	0.093	0.163	0.150	**0.166**	**0.166**
–	Virchow2	0.157	**0.168**	0.099	0.057	0.161	0.088	0.050	0.122
–	GenBio-PathFM	0.174	0.151	0.158	0.109	**0.176**	0.099	0.032	0.156
Skin	UNI2-h	**0.480**	0.465	0.446	0.299	0.469	0.415	0.477	0.470
–	Virchow2	0.486	0.460	0.350	0.414	0.475	0.346	0.156	**0.508**
–	GenBio-PathFM	0.463	0.460	0.460	0.385	0.421	0.256	0.243	**0.485**
Uterus	UNI2-h	0.109	0.114	0.134	0.125	0.093	0.072	0.072	**0.192**
–	Virchow2	0.214	0.222	0.173	0.169	0.179	0.129	-0.013	**0.234**
–	GenBio-PathFM	0.141	0.131	0.131	0.081	0.102	0.070	0.140	**0.213**
Average	UNI2-h	0.156	0.164	0.156	0.130	0.157	0.135	0.167	**0.206**
–	Virchow2	0.177	0.175	0.114	0.140	0.165	0.101	0.026	**0.201**
–	GenBio-PathFM	0.160	0.159	0.159	0.124	0.147	0.069	0.067	**0.198**

Judges. We apply three multimodal LLM judges to these sets: gpt-5.6-sol at high reasoning effort, the judge model used to construct the adaptation corpus (Section 3.1), and GLM 5.3 Flash and DeepSeek V4.1 Flash. Every judge receives the same prompt and the same per-triplet randomization of tile order as in corpus construction. A judgement is correct when the judge names the different-class tile as the odd one out. Any other verdict, including the report that no two tiles are substantially more similar, counts as incorrect.

A.5.2 WITHIN-DOMAIN SIMILARITY ON MOSAIC

Table 23 complements the cross-domain MOSAIC results of Section 4.2.2 with the within-domain sets of MOSAIC on the same four datasets. These hold the domain fixed: the reference and both candidates come from the same slide (HER2ST, CCTGS) or institution (Camelyon16, BEETLE), so the test measures tissue discrimination without an institutional difference. The encoder is correct

Table 14: Full version of Table 4. MIDOG++ mitotic figure classification (11,937 mitotic and 14,349 hard-negative 224 px crops, seven tumor types, four scanners): linear probe at native resolution (0.23 to 0.25 μm/px), held-out domain AUROC, mean and worst fold. Reinhard and Macenko use frozen encoders with the fixed stain references. ROPAL: the checkpoint trained on the 26-cohort corpus with the objective of Section 3.2. Bold: best; underlined: second best.

Evaluation	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
Tumor type out – Mean AUROC	UNI2-h	0.843	0.845	0.798	0.705	0.838	0.801	**0.851**	0.849
Tumor type out – Mean AUROC	Virchow2	**0.834**	0.833	0.697	0.698	0.832	0.749	0.569	**0.834**
Tumor type out – Mean AUROC	GenBio-PathFM	0.833	0.827	0.811	0.672	0.830	0.729	0.813	**0.850**
Tumor type out – Worst fold	UNI2-h	0.775	0.779	0.733	0.656	0.762	0.726	**0.789**	0.784
Tumor type out – Worst fold	Virchow2	0.792	**0.797**	0.607	0.638	0.795	0.669	0.537	0.788
Tumor type out – Worst fold	GenBio-PathFM	0.786	0.776	0.745	0.598	0.781	0.676	0.751	**0.803**
Scanner out – Mean AUROC	UNI2-h	0.845	0.851	0.797	0.719	0.835	0.803	0.854	**0.857**
Scanner out – Mean AUROC	Virchow2	**0.837**	0.836	0.686	0.698	0.834	0.756	0.547	0.832
Scanner out – Mean AUROC	GenBio-PathFM	0.833	0.829	0.821	0.678	0.834	0.724	0.799	**0.851**
Scanner out – Worst fold	UNI2-h	0.796	0.807	0.744	0.623	0.766	0.746	0.808	**0.813**
Scanner out – Worst fold	Virchow2	0.788	0.797	0.607	0.641	0.780	0.726	0.506	**0.800**
Scanner out – Worst fold	GenBio-PathFM	0.761	0.765	0.741	0.624	0.771	0.658	0.728	**0.788**

Table 15: Loss components on UNI2-h and Virchow2. BT only omits the rank term of the ROPAL objective. MOSAIC rows give absolute accuracy. Higher is better except for Same-domain bias. MOSAIC cross (Average) and MOSAIC within (Average) are the unweighted means over the four cross-domain and four within-domain sets of MOSAIC in Tables 8 and 23, as in Table 16. Same-domain bias (lower is better), R@1 and mAP on SCORPION and PLISM retrieval are defined as in Table 16. Bold: best; underlined: second best.

Metric	Dataset	Base	Frozen	BT only	ROPAL (BT + rank)
MOSAIC cross	Average	UNI2-h	0.712	0.884	**0.892**
MOSAIC cross	Average	Virchow2	0.751	0.893	**0.900**
MOSAIC within	Average	UNI2-h	0.932	0.966	**0.974**
MOSAIC within	Average	Virchow2	0.958	0.963	**0.970**
Same-domain bias	Average	UNI2-h	0.241	**0.077**	0.078
Same-domain bias	Average	Virchow2	0.223	**0.071**	0.072
R@1	SCORPION	UNI2-h	0.967	**0.994**	0.993
R@1	SCORPION	Virchow2	0.815	**0.952**	0.944
mAP	SCORPION	UNI2-h	0.854	**0.970**	**0.970**
mAP	SCORPION	Virchow2	0.706	**0.876**	0.866
Retrieval	PLISM, 8,139 tiles	UNI2-h	0.197	**0.380**	0.362
Retrieval	PLISM, 8,139 tiles	Virchow2	0.283	**0.400**	0.394
Retrieval	PLISM, 2,713 tiles	UNI2-h	0.334	**0.554**	0.540
Retrieval	PLISM, 2,713 tiles	Virchow2	0.468	0.564	**0.566**

when it places the same-class tile closer to the reference than the different-class tile. On these, ROPAL has the highest average on each of the three bases.

A.5.3 Reverse-confound test on the MOSAIC cohorts

The cross-domain sets of MOSAIC (Table 8) place the same-class tile in another slide or institution and the different-class tile in the reference's own slide or institution. This mirrors the structure of the adaptation corpus, so a model that simply preferred tiles from other domains would also score well. The reverse test swaps the two: the same-class tile comes from the reference's own slide (HER2ST, CCTGS) or institution (Camelyon16, BEETLE), and the different-class tile from another one. A model that relies on domain similarity scores higher on the reverse comparisons than on the forward ones. The mean of the two accuracies balances the confound, and the same-domain bias (reverse minus forward accuracy) gives the direction and size of the remaining domain preference, with 0 meaning no same-domain preference.

The forward comparisons come from the MOSAIC cross-domain sets of Table 8 and the reverse comparisons from the same four cohorts. Both directions have 550, 550, 223 and 487 comparisons for HER2ST, CCTGS, Camelyon16 and BEETLE.

Table 16: Selection of training triplets by the judge versus other selections on UNI2-h and Virchow2. All arms use the ROPAL objective. MOSAIC cross and MOSAIC within give absolute accuracy on the cross-domain and within-domain sets of MOSAIC in Tables 8 and 23, each the unweighted mean of the four sets before rounding. Same-domain bias is the reverse minus forward accuracy of the reverse-confound test (Section A.5.3), the unweighted mean of the four datasets before rounding, where 0 means no same-domain preference. R@1 and mAP on SCORPION follow Section 4.3, and PLISM retrieval is the staining retrieval of Table 10 at 8,139 and 2,713 tiles per slide, computed only on the stainings of a single scanner and therefore not comparable with Table 10. Lower is better for Same-domain bias and higher is better for every other row. Bold: best; underlined: second best.

Metric	Dataset	Base	Frozen	Random	MNN	Odd-one-out score	ROPAL
MOSAIC cross	Average	UNI2-h	0.712	0.620	<u>0.823</u>	0.773	**0.892**
MOSAIC cross	Average	Virchow2	0.751	0.701	<u>0.872</u>	0.851	**0.900**
MOSAIC within	Average	UNI2-h	0.932	0.897	0.957	<u>0.960</u>	**0.974**
MOSAIC within	Average	Virchow2	0.958	0.949	0.966	<u>0.968</u>	**0.970**
Same-domain bias	Average	UNI2-h	0.241	0.290	<u>0.151</u>	0.208	**0.078**
Same-domain bias	Average	Virchow2	0.223	0.278	<u>0.105</u>	0.133	**0.072**
R@1	SCORPION	UNI2-h	0.967	0.917	<u>0.978</u>	0.969	**0.993**
R@1	SCORPION	Virchow2	0.815	0.709	<u>0.896</u>	0.886	**0.944**
mAP	SCORPION	UNI2-h	0.854	0.673	<u>0.901</u>	0.862	**0.970**
mAP	SCORPION	Virchow2	0.706	0.593	<u>0.799</u>	0.777	**0.866**
Retrieval	PLISM, 8,139 tiles	UNI2-h	0.197	0.120	<u>0.242</u>	0.217	**0.362**
Retrieval	PLISM, 8,139 tiles	Virchow2	0.283	0.054	<u>0.336</u>	0.320	**0.394**
Retrieval	PLISM, 2,713 tiles	UNI2-h	0.334	0.220	<u>0.401</u>	0.365	**0.540**
Retrieval	PLISM, 2,713 tiles	Virchow2	0.468	0.106	<u>0.511</u>	0.496	**0.566**

A.5.4 Scanner differences on PLISM

Reinhard normalization does well on scanner-only retrieval (Section 6) because the scans of one staining image the same tissue section: the scanners differ mainly in lightness, colour and contrast, and the per-channel mean and standard-deviation match of Reinhard normalization removes most of this difference. The stainings are separate sections whose colour channels relate differently, which a per-channel match does not correct, and across stainings the comparison of Figure 4 applies. The statistics below support this reading.

Figure 9 shows four tiles of the 8,139-tile PLISM subset at staining GMH under the seven scanners, raw and after the Reinhard normalization of the baseline. The tiles follow fixed rules on the per-tile top-10 outcomes of UNI2-h over the 21 scanner pairs in both directions (42 queries per tile). On every 40th tile of the subset (204 tiles), the median CIELAB lightness of the scanners ranges from 70.4 (S360) to 85.7 (GT450), the median a* from 9.0 (GT450) to 24.6 (S360) and the lightness standard deviation from 8.5 (P) to 16.8 (AT2). After integer alignment to the S60 scan (median offsets of 1 pixel for S210, S360 and SQ and 7 to 14 pixels for GT450, P and AT2), the median pixel correlations with S60, which the per-channel affine map of Reinhard normalization leaves unchanged, are 0.92 to 0.95 for L*, 0.90 to 0.94 for a* and 0.83 to 0.88 for b*, and 0.79 to 0.84 for high-pass-filtered lightness. The scanners thus differ mainly in global lightness, colour and contrast, with smaller local differences in registration and sharpness. The 13 stainings at scanner S60 are separate sections (median pixel correlations of 0.01 to 0.39 with the GMH section), their median lightness ranges from 24.8 (HR) to 78.4 (KRH), and the within-tile correlations between the CIELAB channels, which Reinhard normalization does not change, differ from those of the GMH scan by a median absolute 0.078, against 0.039 across scanners. The examples also show the limits of the normalization: tile C, about half unstained background, is retrieved in 18 of 42 queries after Reinhard normalization and in all 42 with ROPAL, and tile D is retrieved in at most 10 queries by any method.

A.6 Correlation of MOSAIC accuracy with other tasks

MOSAIC (Section 4.2.2, Table 8) is a minimal test of whether a model's embedding geometry places biological similarity ahead of institutional differences. This analysis measures whether a model's

Table 17: Choice of judge LLM on UNI2-h and Virchow2. Each arm trains on the corpus kept by one judge under the same ROPAL objective and training preference; the smaller and larger corpora are the two smallest corpus sizes of the corpus-size ablation. MOSAIC rows give absolute accuracy. Higher is better except for Same-domain bias. MOSAIC cross (Average) and MOSAIC within (Average) are the unweighted means over the four cross-domain and four within-domain sets of MOSAIC in Tables 8 and 23, as in Table 16. Same-domain bias (lower is better), R@1 and mAP on SCORPION and PLISM retrieval are defined as in Table 16. Bold: best; underlined: second best.

Metric	Dataset	Base	Frozen	Qwen3.8-27B	gpt-5.6-sol (ROPAL)
MOSAIC cross	Average	UNI2-h (smaller corpus)	0.712	**0.734**	0.731
MOSAIC cross	Average	UNI2-h (larger corpus)	0.712	0.760	**0.775**
MOSAIC cross	Average	Virchow2 (smaller corpus)	0.751	0.779	**0.791**
MOSAIC cross	Average	Virchow2 (larger corpus)	0.751	**0.838**	**0.838**
MOSAIC within	Average	UNI2-h (smaller corpus)	0.932	**0.959**	0.957
MOSAIC within	Average	UNI2-h (larger corpus)	0.932	0.957	**0.964**
MOSAIC within	Average	Virchow2 (smaller corpus)	0.958	0.959	**0.960**
MOSAIC within	Average	Virchow2 (larger corpus)	0.958	0.960	**0.966**
Same-domain bias	Average	UNI2-h (smaller corpus)	0.241	**0.237**	0.240
Same-domain bias	Average	UNI2-h (larger corpus)	0.241	0.201	**0.187**
Same-domain bias	Average	Virchow2 (smaller corpus)	0.223	0.200	**0.185**
Same-domain bias	Average	Virchow2 (larger corpus)	0.223	0.140	**0.137**
R@1	SCORPION	UNI2-h (smaller corpus)	0.967	**0.972**	**0.972**
R@1	SCORPION	UNI2-h (larger corpus)	0.967	0.978	**0.979**
R@1	SCORPION	Virchow2 (smaller corpus)	0.815	0.869	**0.871**
R@1	SCORPION	Virchow2 (larger corpus)	0.815	**0.909**	0.901
mAP	SCORPION	UNI2-h (smaller corpus)	0.854	0.864	**0.868**
mAP	SCORPION	UNI2-h (larger corpus)	0.854	0.889	**0.894**
mAP	SCORPION	Virchow2 (smaller corpus)	0.706	**0.766**	**0.766**
mAP	SCORPION	Virchow2 (larger corpus)	0.706	**0.813**	0.801
Retrieval	PLISM, 8,139 tiles	UNI2-h (smaller corpus)	0.197	0.203	**0.206**
Retrieval	PLISM, 8,139 tiles	UNI2-h (larger corpus)	0.197	0.217	**0.223**
Retrieval	PLISM, 8,139 tiles	Virchow2 (smaller corpus)	0.283	0.315	**0.319**
Retrieval	PLISM, 8,139 tiles	Virchow2 (larger corpus)	0.283	0.336	**0.338**
Retrieval	PLISM, 2,713 tiles	UNI2-h (smaller corpus)	0.334	0.346	**0.350**
Retrieval	PLISM, 2,713 tiles	UNI2-h (larger corpus)	0.334	0.372	**0.380**
Retrieval	PLISM, 2,713 tiles	Virchow2 (smaller corpus)	0.468	0.494	**0.497**
Retrieval	PLISM, 2,713 tiles	Virchow2 (larger corpus)	0.468	0.506	**0.512**

Table 18: Preference supervision versus domain-adversarial adaptation on UNI2-h and Virchow2. All adapted conditions include the same $\lambda \mathcal{L}_{\text{rank}}$ term. BT + domain adversarial + rank adds the reversed-gradient source-site cross-entropy to the BT objective. MOSAIC rows give absolute accuracy. Higher is better except for Same-domain bias. MOSAIC cross (Average) and MOSAIC within (Average) are the unweighted means over the four cross-domain and four within-domain sets of MOSAIC in Tables 8 and 23, as in Table 16. Same-domain bias (lower is better), R@1 and mAP on SCORPION are defined as in Table 16. Bold: best; underlined: second best.

Metric	Dataset	Base	Frozen	Domain adversarial + rank	BT + domain adversarial + rank	ROPAL (BT + rank)
MOSAIC cross	Average	UNI2-h	0.712	0.580	0.805	**0.892**
MOSAIC cross	Average	Virchow2	0.751	0.643	0.678	**0.900**
MOSAIC within	Average	UNI2-h	0.932	0.898	0.870	**0.974**
MOSAIC within	Average	Virchow2	0.958	0.787	0.780	**0.970**
Same-domain bias	Average	UNI2-h	0.241	0.352	**0.051**	0.078
Same-domain bias	Average	Virchow2	0.223	0.175	0.120	**0.072**
R@1	SCORPION	UNI2-h	0.967	0.763	0.426	**0.993**
R@1	SCORPION	Virchow2	0.815	0.211	0.464	**0.944**
mAP	SCORPION	UNI2-h	0.854	0.538	0.144	**0.970**
mAP	SCORPION	Virchow2	0.706	0.126	0.326	**0.866**

Table 19: Corpus-size ablation with 26 cohorts on UNI2-h and Virchow2. Columns specify the number of distinct judged comparisons processed by ROPAL. MOSAIC rows give absolute accuracy. Higher is better except for Same-domain bias. MOSAIC cross (Average) and MOSAIC within (Average) are the unweighted means over the four cross-domain and four within-domain sets of MOSAIC in Tables 8 and 23, as in Table 16. Same-domain bias (lower is better), R@1 and mAP on SCORPION and PLISM retrieval are defined as in Table 16. Bold: best; underlined: second best.

Metric	Dataset	Base	Frozen	ROPAL (1,248)	ROPAL (2,496)	ROPAL (5,004)	ROPAL (9,996)	ROPAL (19,992)
MOSAIC cross	Average	UNI2-h	0.712	0.731	0.775	0.840	0.871	**0.892**
MOSAIC cross	Average	Virchow2	0.751	0.791	0.838	0.859	0.878	**0.900**
MOSAIC within	Average	UNI2-h	0.932	0.957	0.964	0.967	0.969	**0.974**
MOSAIC within	Average	Virchow2	0.958	0.960	0.966	**0.972**	0.969	0.970
Same-domain bias	Average	UNI2-h	0.241	0.240	0.187	0.135	0.104	**0.078**
Same-domain bias	Average	Virchow2	0.223	0.185	0.137	0.128	0.098	**0.072**
R@1	SCORPION	UNI2-h	0.967	0.972	0.979	0.988	0.991	**0.993**
R@1	SCORPION	Virchow2	0.815	0.871	0.901	0.907	0.933	**0.944**
mAP	SCORPION	UNI2-h	0.854	0.868	0.894	0.936	0.957	**0.970**
mAP	SCORPION	Virchow2	0.706	0.766	0.801	0.804	0.845	**0.866**
Retrieval	PLISM, 8,139 tiles	UNI2-h	0.197	0.206	0.223	0.277	0.323	**0.362**
Retrieval	PLISM, 8,139 tiles	Virchow2	0.283	0.319	0.338	0.350	0.377	**0.394**
Retrieval	PLISM, 2,713 tiles	UNI2-h	0.334	0.350	0.380	0.441	0.498	**0.540**
Retrieval	PLISM, 2,713 tiles	Virchow2	0.468	0.497	0.512	0.525	0.554	**0.566**

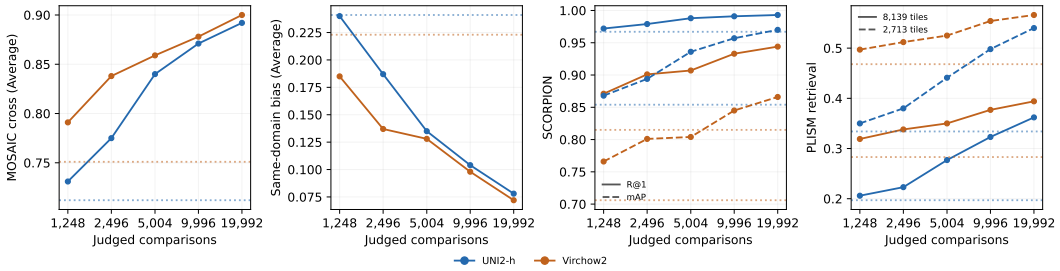


Figure 8: Corpus size on UNI2-h (blue) and Virchow2 (orange): MOSAIC cross accuracy, same-domain bias (lower is better), SCORPION R@1 (solid) and mAP (dashed) and PLISM retrieval at 8,139 (solid) and 2,713 (dashed) tiles per slide against the number of judged comparisons (Appendix Table 19). Dotted lines are the frozen references. PLISM uses the stainings of a single scanner, so its values are not comparable with Figure 4.

Table 20: Scanner invariance on PLISM for the indicated base encoders, at 8,139 and 2,713 tiles per slide. The same registered tissue is scanned under 13 stainings and 7 scanners (91 slides). Scanner: median top-10 matched-tile retrieval accuracy over inter-scanner slide pairs. None: frozen; Reinhard and Macenko: frozen encoder on stain-normalized tiles; Higher is better. ROPAL: the checkpoint trained on the 26-cohort corpus with the objective of Section 3.2. Bold: best; underlined: second best.

Tiles per slide	Base model	None	Reinhard	Macenko	LEACE	INLP	ROPAL
8,139	GenBio-PathFM	0.709	0.800	0.687	0.654	0.365	**0.919**
8,139	UNI2-h	0.501	**0.881**	0.655	0.540	0.523	0.789
8,139	Virchow2	0.610	**0.821**	0.424	0.615	0.573	0.798
2,713	GenBio-PathFM	0.848	0.900	0.818	0.813	0.556	**0.958**
2,713	UNI2-h	0.629	**0.925**	0.763	0.657	0.649	0.868
2,713	Virchow2	0.744	**0.902**	0.556	0.740	0.702	0.881

MOSAIC accuracy tracks its performance in the other main experiments. A model is one base encoder with one method. For each (experiment, metric) pair, the models with results in both Table 8 (MOSAIC cross (Average) over HER2ST, CCTGS, Camelyon16 and BEETLE) and the experiment table enter, and the Pearson correlation r between the two displayed metric values summarizes the pair. Every metric is higher-is-better, including the signed PathoROB APD, for which a less negative drop is better. The points of different pairs share bases and methods, so the p-values are descriptive.

Table 21: Per-base version of Table 6, with the frozen encoder as reference. Lower is better for Same-domain bias and higher for every other column. The frozen MOSAIC and SCORPION values are those of Tables 8, 23, 24 and 9. PLISM columns use the stainings of a single scanner and are not comparable with Table 10. Bold: best; underlined: second best, over both bases.

Base	Setting	MOSAIC cross	MOSAIC within	Same-domain bias	SCORPION R@1	SCORPION mAP	PLISM 8,139	PLISM 2,713
UNI2-h	ROPAL	0.892	**0.974**	0.078	0.993	**0.970**	0.362	0.540
	Qwen3.8-27B judge (choice of LLM)	0.887	0.967	0.078	0.993	0.968	0.359	0.533
	BT only (no rank term) (loss)	0.884	0.966	0.077	**0.994**	**0.970**	0.380	0.554
	Domain adversarial + rank (loss)	0.580	0.898	0.352	0.763	0.538	0.023	0.044
	Random (judge selection)	0.620	0.897	0.290	0.917	0.673	0.120	0.220
	MNN (judge selection)	0.823	0.957	0.151	0.978	0.901	0.242	0.401
	Odd-one-out score (judge selection)	0.773	0.960	0.208	0.969	0.862	0.217	0.365
	Frozen encoder	0.712	0.932	0.241	0.967	0.854	0.197	0.334
Virchow2	ROPAL	**0.900**	0.970	0.072	0.944	0.866	0.394	0.566
	Qwen3.8-27B judge (choice of LLM)	0.888	0.969	0.081	0.957	0.890	**0.403**	**0.575**
	BT only (no rank term) (loss)	0.893	0.963	**0.071**	0.952	0.876	0.400	0.564
	Domain adversarial + rank (loss)	0.643	0.787	0.175	0.211	0.126	0.001	0.002
	Random (judge selection)	0.701	0.949	0.278	0.709	0.593	0.054	0.106
	MNN (judge selection)	0.872	0.966	0.105	0.896	0.799	0.336	0.511
	Odd-one-out score (judge selection)	0.851	0.968	0.133	0.886	0.777	0.320	0.496
	Frozen encoder	0.751	0.958	0.223	0.815	0.706	0.283	0.468

Table 22: Accuracy of LLM judgements against the annotated tissue classes. A judgement is correct when the judge names the different-class tile as the odd one out. Average is the unweighted mean of the six datasets, computed before rounding. gpt-5.6-sol is the judge model used for corpus construction. Higher is better.

Dataset	gpt-5.6-sol	GLM 5.3 Flash	DeepSeek V4.1 Flash
HER2ST	0.790	0.746	0.738
CCTGS	0.849	0.841	0.793
TIGER	0.691	0.606	0.514
Camelyon16	0.933	0.870	0.843
BEETLE	0.739	0.684	0.593
TCGA-LUSC	0.754	0.776	0.702
Average	0.793	0.754	0.697

Table 23: Within-domain accuracy of the embeddings on MOSAIC, for the datasets of Table 8. All three tiles come from the same slide (HER2ST, CCTGS) or institution (Camelyon16, BEETLE). Methods and bases are those of Table 8. MOSAIC within (Average) is the unweighted mean of the four datasets, computed before rounding. Higher is better. Bold: best; underlined: second best.

Metric	Dataset	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
MOSAIC within	HER2ST	UNI2-h	0.894	0.912	0.889	0.870	0.898	0.888	0.910	**0.965**
MOSAIC within	HER2ST	Virchow2	0.917	0.919	0.818	0.910	0.912	0.880	0.737	**0.951**
MOSAIC within	HER2ST	GenBio-PathFM	0.922	0.934	0.906	0.906	0.912	0.856	0.915	**0.951**
MOSAIC within	CCTGS	UNI2-h	0.969	0.974	0.971	0.967	0.965	0.940	0.976	**0.988**
MOSAIC within	CCTGS	Virchow2	**0.990**	**0.990**	0.983	0.989	0.986	0.936	0.835	**0.990**
MOSAIC within	CCTGS	GenBio-PathFM	0.987	**0.988**	0.985	0.976	0.980	0.896	0.902	0.985
MOSAIC within	Camelyon16	UNI2-h	0.912	0.938	0.951	0.934	0.903	0.858	0.934	**0.991**
MOSAIC within	Camelyon16	Virchow2	0.965	0.969	0.938	0.947	0.942	0.823	0.765	**0.973**
MOSAIC within	Camelyon16	GenBio-PathFM	0.978	0.973	0.991	0.938	0.956	0.814	0.819	**0.996**
MOSAIC within	BEETLE	UNI2-h	0.954	0.958	0.949	0.950	0.953	0.938	0.964	**0.974**
MOSAIC within	BEETLE	Virchow2	0.961	0.960	0.964	0.951	0.968	0.943	0.795	**0.970**
MOSAIC within	BEETLE	GenBio-PathFM	0.964	0.968	0.958	0.952	0.961	0.896	0.923	**0.975**
MOSAIC within	Average	UNI2-h	0.932	0.945	0.940	0.930	0.930	0.906	0.946	**0.980**
MOSAIC within	Average	Virchow2	0.958	0.959	0.926	0.949	0.952	0.895	0.783	**0.971**
MOSAIC within	Average	GenBio-PathFM	0.963	0.966	0.960	0.943	0.952	0.865	0.890	**0.976**

The pairs cover the main experiment tables (Tables 1, 9, 10, 2, 12, 3, 4 and 5). The cross-institution tables contribute each metric for each dataset and institution–class confounding condition. Figures 10 and 11 show the values, and Table 25 lists the correlations.

Table 24: Reverse-confound test on the MOSAIC cohorts for the methods and bases of Table 8. Forward: the same-class candidate comes from another slide or institution and the different-class candidate from the reference's. Reverse: the same-class candidate comes from the reference's slide or institution and the different-class candidate from another one. Average is the mean of Forward and Reverse, and Same-domain bias is Reverse minus Forward (0 means no same-domain preference). All values are unweighted means of HER2ST, CCTGS, Camelyon16 and BEETLE, computed before rounding. Forward is measured on the reverse-confound sets, which hold fewer comparisons than the cross-domain sets of Table 8, so it differs from MOSAIC cross there. Binomial standard deviations are at most 0.012 (Forward), 0.009 (Reverse), 0.008 (Average) and 0.015 (Same-domain bias). Higher is better for Forward, Reverse and Average, and lower is better for Same-domain bias. Bold: best; underlined: second best.

Evaluation	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
Forward	UNI2-h	0.702	0.764	0.726	0.705	0.716	0.656	<u>0.772</u>	**0.902**
Forward	Virchow2	0.749	<u>0.809</u>	0.725	0.762	0.747	0.643	0.555	**0.880**
Forward	GenBio-PathFM	0.776	<u>0.802</u>	0.797	0.752	0.761	0.564	0.692	**0.914**
Reverse	UNI2-h	0.944	0.953	0.951	0.942	0.934	0.904	<u>0.960</u>	**0.965**
Reverse	Virchow2	<u>0.972</u>	**0.974**	0.931	0.966	0.955	0.906	0.843	0.966
Reverse	GenBio-PathFM	**0.977**	<u>0.975</u>	0.970	0.948	0.965	0.888	0.907	0.964
Average	UNI2-h	0.823	0.858	0.839	0.824	0.825	0.780	<u>0.866</u>	**0.934**
Average	Virchow2	0.861	<u>0.891</u>	0.828	0.864	0.851	0.775	0.699	**0.923**
Average	GenBio-PathFM	0.876	<u>0.888</u>	0.884	0.850	0.863	0.726	0.799	**0.939**
Same-domain bias	UNI2-h	0.241	0.189	0.226	0.237	0.218	0.249	<u>0.188</u>	**0.062**
Same-domain bias	Virchow2	0.223	<u>0.164</u>	0.207	0.204	0.208	0.263	0.288	**0.086**
Same-domain bias	GenBio-PathFM	0.201	<u>0.173</u>	<u>0.173</u>	0.196	0.204	0.324	0.214	**0.050**

Table 25: Pearson correlation r and p-value between MOSAIC accuracy (Table 8, MOSAIC cross (Average)) and each experiment metric, over the n models with results in both tables, computed on the displayed values and sorted by r. Table: source table of the metric.

Experiment	Metric	Table	n	Pearson r	p
Cross-institution retrieval and classification	Fixed-candidate NN accuracy, Camelyon16, confounding present	12	24	0.92	2.4×10^{-10}
Cross-institution retrieval and classification	Retrieval mAP, Camelyon16, confounding present	2	24	0.91	7.3×10^{-10}
HEST	Gene expression (Average)	3	24	0.87	2.4×10^{-8}
PLISM	Staining retrieval, 8,139 tiles	10	24	0.84	3.1×10^{-7}
Cross-institution retrieval and classification	Retrieval mAP, Camelyon16, confounding absent	2	24	0.83	4.7×10^{-7}
Cross-institution retrieval and classification	Fixed-candidate NN accuracy, BEETLE, confounding absent	12	24	0.82	8.6×10^{-7}
PLISM	Staining retrieval, 2,713 tiles	10	24	0.80	3.4×10^{-6}
Cross-institution retrieval and classification	Retrieval mAP, BEETLE, confounding present	2	24	0.76	1.9×10^{-5}
PANDA	Cross-site balanced accuracy	5	24	0.74	3.2×10^{-5}
PathoROB	APD (Average, signed)	1	24	0.74	3.7×10^{-5}
Cross-institution retrieval and classification	Retrieval mAP, BEETLE, confounding absent	2	24	0.73	5.7×10^{-5}
PathoROB	RI (Average)	1	24	0.72	7.2×10^{-5}
Cross-institution retrieval and classification	Fixed-candidate NN accuracy, Camelyon16, confounding absent	12	24	0.69	2.1×10^{-4}
Cross-institution retrieval and classification	Fixed-candidate NN accuracy, BEETLE, confounding present	12	24	0.69	2.1×10^{-4}
SCORPION	mAP	9	24	0.67	3.8×10^{-4}
SCORPION	R@1	9	24	0.65	6.0×10^{-4}
MIDOG++	Scanner out, mean AUROC	4	24	0.63	0.001
MIDOG++	Scanner out, worst-fold AUROC	4	24	0.62	0.001
MIDOG++	Tumor type out, worst-fold AUROC	4	24	0.62	0.001
MIDOG++	Tumor type out, mean AUROC	4	24	0.61	0.002
Cross-institution retrieval and classification	Retrieval mAP, TumSeg, confounding absent	2	24	0.19	0.362
Cross-institution retrieval and classification	Retrieval mAP, TumSeg, confounding present	2	24	0.18	0.394
Cross-institution retrieval and classification	Fixed-candidate NN accuracy, TumSeg, confounding present	12	24	0.12	0.589
Cross-institution retrieval and classification	Fixed-candidate NN accuracy, TumSeg, confounding absent	12	24	0.09	0.692

A.7 Standard deviations of table results

Tables 26–34 repeat the result tables of the Experiments section with the standard deviation (SD) of every estimate over 1,000 bootstrap resamples of the evaluation units named in each caption, computed from the saved per-sample outputs without rerunning any model. Estimates are the values of the corresponding main table. An entry without an SD has no saved per-sample outputs; a dash marks a pending result whose expected finish time is shown in the main table. An asterisk marks a best method whose margin over the second best exceeds 1.96 times the combined SD of the two estimates, treating them as independent.

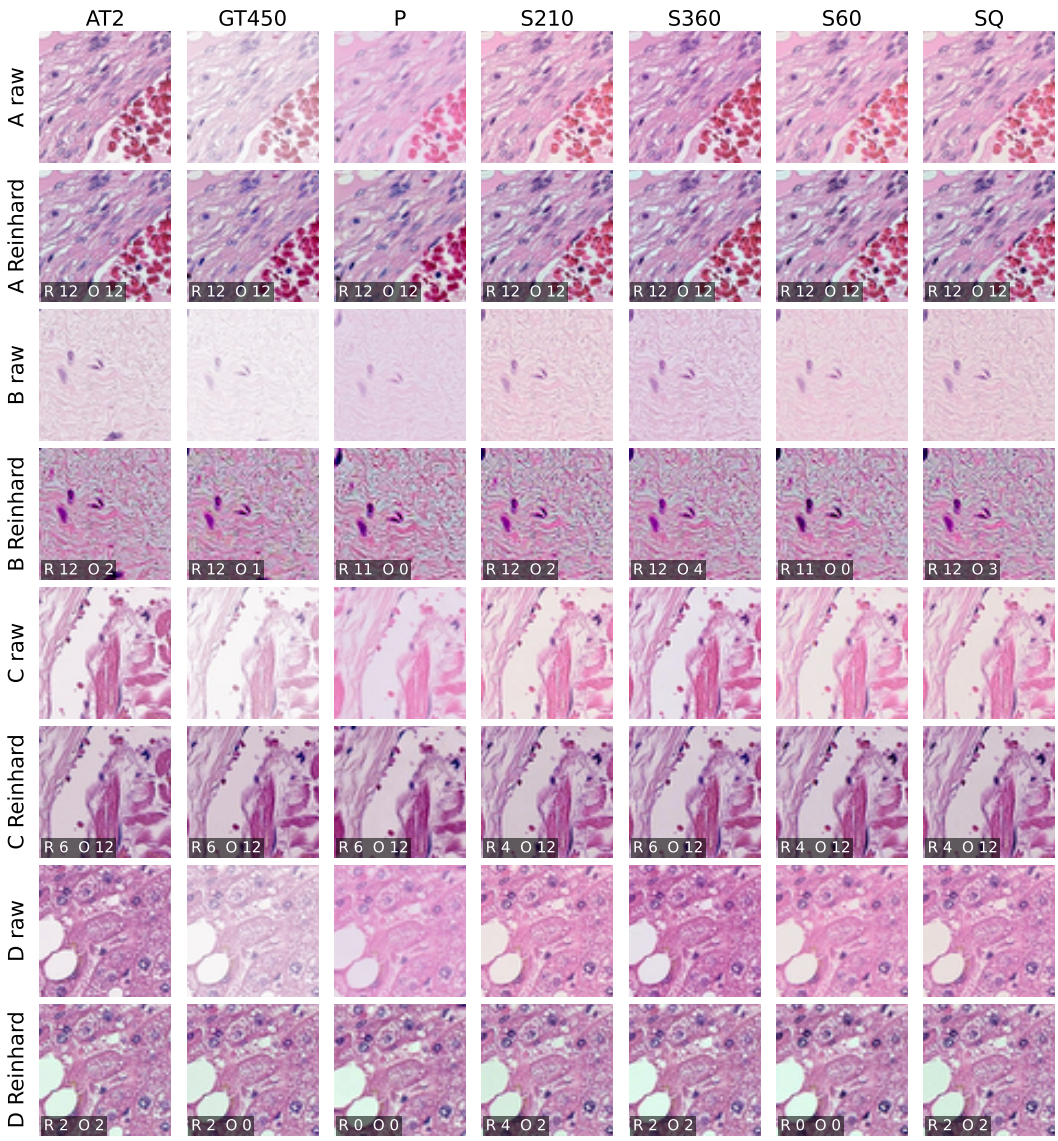


Figure 9: PLISM tiles at staining GMH under the seven scanners (columns), raw and after Reinhard normalization (rows). Labels give, for UNI2-h, the number of the 12 queries involving that scanner in which Reinhard (R) and ROPAL (O) rank the matching tile among the top 10. A: retrieved in all 42 queries by Reinhard and ROPAL, in 10 by None; B: largest advantage of Reinhard over ROPAL (41 against 6); C: largest advantage of ROPAL (42 against 18); D: fewest Reinhard hits (6).

A.8 THE ODD-ONE-OUT CORPUS

The adaptation pools contain unlabeled tiles. Curation requires only their source slides and the grouping of slides by institution; the LLM judge provides the biological comparisons. Pools hold about 3,000 tiles per cohort from 6 (KICH, UVM) to 34 (LUSC, SARC) submitting institutions. Kept counts for the fourteen cohorts outside the 12-cohort subset are rounded from their rates. TCGA-ESCA shares an organ with PathoROB's Tolkach esophageal set but is a different source with different patients; TCGA-LUSC training tiles come from patients absent from the TCGA-LUSC evaluation set.

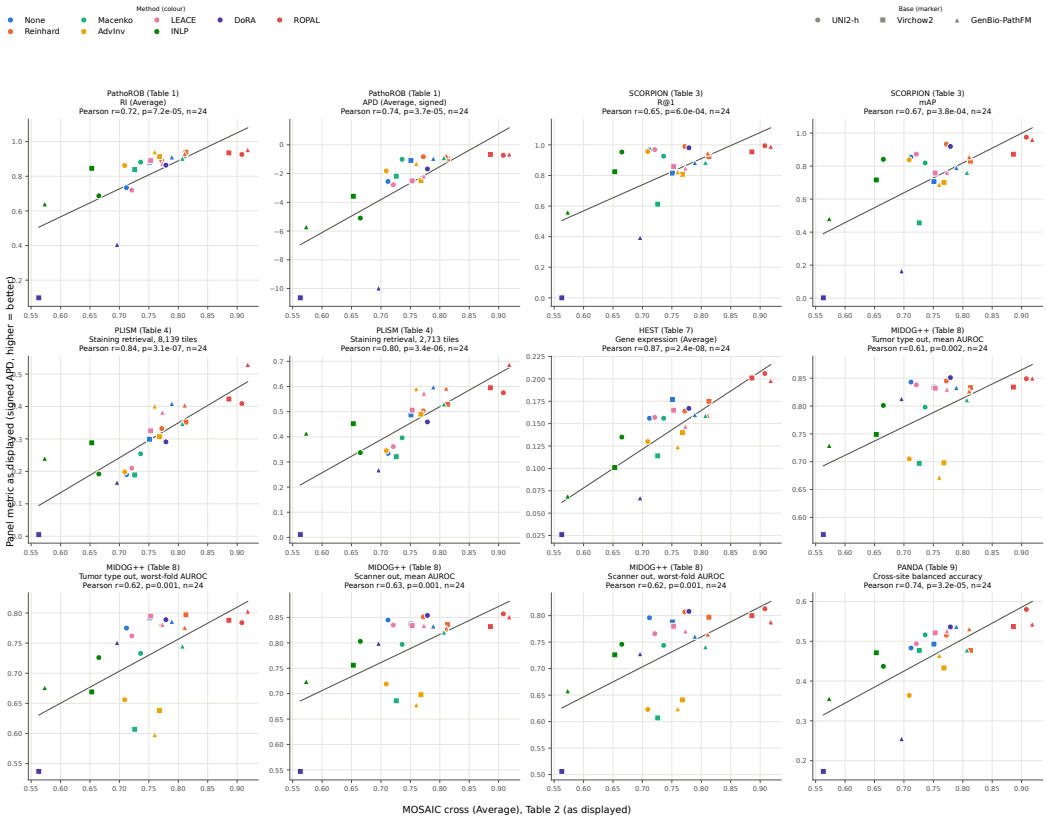


Figure 10: MOSAIC accuracy against the robustness, invariance and downstream-task metrics of PathoROB, SCORPION, PLISM, HEST, MIDOG++ and PANDA. Each point is one model, coloured by method and shaped by base encoder. Both axes show the displayed values, and the line is the least-squares fit. Panel titles give the Pearson r, its p-value and n (Table 25).

Table 26: Table 1 with bootstrap standard deviations. SD over 1,000 bootstrap resamples: RI resamples reference patches and recomputes SO/(SO+OS) at the benchmark k; APD resamples classifier repetitions and recomputes the clipped mean, in percentage points. Average combines the two dataset SDs as half the root sum of squares. Bold: best; underlined: second best. An asterisk marks a best method whose margin over the second best exceeds 1.96 times the combined bootstrap SD, treating the two estimates as independent.

Metric	Dataset	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
RI	Camelyon	UNI2-h	0.544 ± 0.013	0.839 ± 0.006	0.819 ± 0.006	0.770 ± 0.006	0.512 ± 0.014	0.506 ± 0.011	0.780	**0.888 ± 0.004***
RI	Camelyon	Virchow2	0.808 ± 0.007	**0.923 ± 0.003***	0.725 ± 0.005	0.870 ± 0.004	0.826 ± 0.006	0.760 ± 0.008	0.003	0.904 ± 0.004
RI	Camelyon	GenBio-PathFM	0.862 ± 0.005	0.898 ± 0.004	0.848 ± 0.006	0.919 ± 0.003	0.790 ± 0.008	0.426 ± 0.011	0.125	**0.935 ± 0.002***
RI	Tolkach ESCA	UNI2-h	0.923 ± 0.003	0.940 ± 0.003	0.945 ± 0.003	0.955 ± 0.002	0.928 ± 0.003	0.869 ± 0.003	0.949	**0.963 ± 0.002***
RI	Tolkach ESCA	Virchow2	0.954 ± 0.002	0.954 ± 0.002	0.953 ± 0.002	0.958 ± 0.002	0.954 ± 0.002	0.932 ± 0.002	0.193	**0.967 ± 0.002***
RI	Tolkach ESCA	GenBio-PathFM	0.959 ± 0.002	0.961 ± 0.002	0.958 ± 0.002	0.963 ± 0.002	0.958 ± 0.002	0.853 ± 0.004	0.687	**0.970 ± 0.001***
RI	Average	UNI2-h	0.734 ± 0.007	0.890 ± 0.003	0.882 ± 0.003	0.862 ± 0.003	0.720 ± 0.007	0.687 ± 0.006	0.864	**0.926 ± 0.002***
RI	Average	Virchow2	0.881 ± 0.004	**0.939 ± 0.002**	0.839 ± 0.003	0.914 ± 0.002	0.890 ± 0.003	0.846 ± 0.004	0.098	0.935 ± 0.002
RI	Average	GenBio-PathFM	0.911 ± 0.003	0.930 ± 0.002	0.903 ± 0.003	0.941 ± 0.002	0.874 ± 0.004	0.640 ± 0.006	0.406	**0.953 ± 0.001***
APD	Camelyon	UNI2-h	-4.63 ± 0.63	-1.62 ± 0.47	-1.79 ± 0.32	-3.62 ± 1.72	-5.56 ± 1.52	-9.15 ± 2.22	-3.33	**-1.45 ± 0.24**
APD	Camelyon	Virchow2	-2.03 ± 0.60	-1.74 ± 0.42	-3.94 ± 0.51	-4.48 ± 1.15	-4.66 ± 0.90	-5.84 ± 0.49	-12.91	**-1.24 ± 0.23**
APD	Camelyon	GenBio-PathFM	-1.90 ± 0.45	-1.53 ± 0.28	-1.81 ± 0.36	-2.58 ± 1.04	-4.28 ± 0.99	-8.80 ± 1.53	-11.42	**-1.06 ± 0.19**
APD	Tolkach ESCA	UNI2-h	-0.48 ± 0.28	-0.05 ± 0.17	-0.23 ± 0.17	-0.02 ± 0.24	-0.01 ± 0.25	-1.06 ± 0.49	-0.03	**+0.00 ± 0.10**
APD	Tolkach ESCA	Virchow2	-0.15 ± 0.10	-0.22 ± 0.11	-0.42 ± 0.14	-0.50 ± 0.29	-0.36 ± 0.34	-1.31 ± 0.25	-8.39	**-0.10 ± 0.09**
APD	Tolkach ESCA	GenBio-PathFM	**+0.00 ± 0.12**	**+0.00 ± 0.07**	**+0.00 ± 0.11**	**+0.00 ± 0.20**	-0.07 ± 0.55	-2.62 ± 0.66	-8.53	-0.24 ± 0.09
APD	Average	UNI2-h	-2.55 ± 0.34	-0.83 ± 0.25	-1.01 ± 0.18	-1.82 ± 0.87	-2.78 ± 0.77	-5.10 ± 1.14	-1.68	**-0.73 ± 0.13**
APD	Average	Virchow2	-1.09 ± 0.30	-0.98 ± 0.22	-2.18 ± 0.26	-2.49 ± 0.59	-2.51 ± 0.48	-3.58 ± 0.28	-10.65	**-0.67 ± 0.12**
APD	Average	GenBio-PathFM	-0.95 ± 0.23	-0.76 ± 0.14	-0.90 ± 0.19	-1.29 ± 0.53	-2.17 ± 0.57	-5.71 ± 0.83	-9.97	**-0.65 ± 0.10**

A.9 Related work (full main-text version)

We organize prior work by idea. Appendix A.10 gives an extended survey.

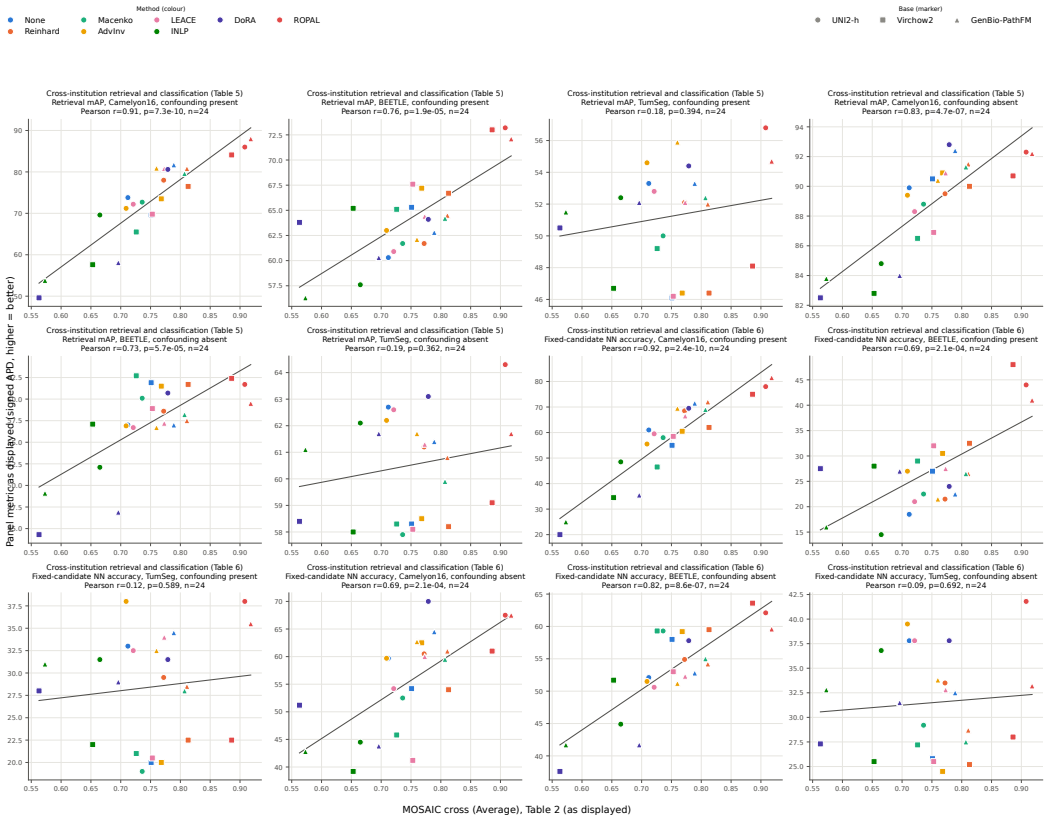


Figure 11: MOSAIC accuracy against the cross-institution retrieval and classification metrics. Each point is one model, coloured by method and shaped by base encoder. Both axes show the displayed values, and the line is the least-squares fit. Panel titles give the Pearson r, its p-value and n (Table 25).

Table 27: Table 8 with bootstrap standard deviations. SD over 1,000 binomial bootstrap resamples of each MOSAIC set. MOSAIC cross (Average) combines the four cross-domain set SDs as the root sum of squares divided by four. MOSAIC forward–reverse (Average) takes half the root sum of squares of the forward and reverse SDs of each dataset and combines the four datasets in the same way. Accuracies are means of binary outcomes, so their bootstrap resamples the outcomes with replacement (binomial resampling of the stored proportion and count); averages over independent sets combine the set SDs as the root sum of squares divided by the number of sets. Bold: best; underlined: second best. An asterisk marks a best method whose margin over the second best exceeds 1.96 times the combined bootstrap SD, treating the two estimates as independent.

Metric	Dataset	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
MOSAIC cross	HER2ST	UNI2-h	0.676 ± 0.017	0.725 ± 0.016	0.674 ± 0.017	0.658 ± 0.017	0.662 ± 0.017	0.649 ± 0.017	0.729 ± 0.016	**0.855 ± 0.013***
MOSAIC cross	HER2ST	Virchow2	0.693 ± 0.017	0.750 ± 0.015	0.604 ± 0.017	0.764 ± 0.015	0.683 ± 0.017	0.658 ± 0.017	0.694 ± 0.017	**0.870 ± 0.012***
MOSAIC cross	HER2ST	GenBio-PathFM	0.695 ± 0.017	0.742 ± 0.016	0.706 ± 0.017	0.777 ± 0.015	0.684 ± 0.017	0.557 ± 0.017	0.821 ± 0.014	**0.884 ± 0.011***
MOSAIC cross	CCTGS	UNI2-h	0.694 ± 0.012	0.785 ± 0.010	0.777 ± 0.011	0.736 ± 0.011	0.729 ± 0.012	0.621 ± 0.013	0.815 ± 0.010	**0.954 ± 0.005***
MOSAIC cross	CCTGS	Virchow2	0.858 ± 0.009	0.875 ± 0.009	0.841 ± 0.010	0.859 ± 0.009	0.838 ± 0.010	0.644 ± 0.013	0.651 ± 0.012	**0.963 ± 0.005***
MOSAIC cross	CCTGS	GenBio-PathFM	0.878 ± 0.009	0.883 ± 0.008	0.887 ± 0.008	0.831 ± 0.010	0.841 ± 0.009	0.583 ± 0.013	0.710 ± 0.012	**0.967 ± 0.005***
MOSAIC cross	Camelyon16	UNI2-h	0.789 ± 0.027	0.848 ± 0.024	0.771 ± 0.029	0.776 ± 0.027	0.780 ± 0.028	0.735 ± 0.030	0.848 ± 0.024	**0.964 ± 0.013***
MOSAIC cross	Camelyon16	Virchow2	0.776 ± 0.029	0.892 ± 0.021	0.709 ± 0.030	0.767 ± 0.028	0.803 ± 0.027	0.650 ± 0.031	0.377 ± 0.033	**0.955 ± 0.013***
MOSAIC cross	Camelyon16	GenBio-PathFM	0.901 ± 0.021	0.901 ± 0.020	0.888 ± 0.022	0.767 ± 0.029	0.901 ± 0.020	0.610 ± 0.033	0.682 ± 0.032	**0.973 ± 0.011***
MOSAIC cross	BEETLE	UNI2-h	0.688 ± 0.022	0.731 ± 0.020	0.723 ± 0.020	0.665 ± 0.021	0.715 ± 0.020	0.655 ± 0.021	0.723 ± 0.020	**0.860 ± 0.016***
MOSAIC cross	BEETLE	Virchow2	0.678 ± 0.021	0.735 ± 0.020	0.749 ± 0.020	0.684 ± 0.022	0.690 ± 0.021	0.661 ± 0.021	0.530 ± 0.022	**0.758 ± 0.020**
MOSAIC cross	BEETLE	GenBio-PathFM	0.682 ± 0.021	0.717 ± 0.021	0.747 ± 0.019	0.665 ± 0.022	0.667 ± 0.020	0.542 ± 0.022	0.571 ± 0.022	**0.848 ± 0.017***
MOSAIC cross	Average	UNI2-h	0.712 ± 0.010	0.772 ± 0.009	0.736 ± 0.010	0.709 ± 0.010	0.721 ± 0.010	0.665 ± 0.011	0.779 ± 0.009	**0.908 ± 0.006***
MOSAIC cross	Average	Virchow2	0.751 ± 0.010	0.813 ± 0.008	0.726 ± 0.010	0.768 ± 0.010	0.753 ± 0.010	0.653 ± 0.011	0.563 ± 0.011	**0.886 ± 0.007***
MOSAIC cross	Average	GenBio-PathFM	0.789 ± 0.009	0.811 ± 0.009	0.807 ± 0.009	0.760 ± 0.010	0.773 ± 0.009	0.573 ± 0.011	0.696 ± 0.011	**0.918 ± 0.006***
MOSAIC forward–reverse	Average	UNI2-h	0.823 ± 0.006	0.858 ± 0.006	0.839 ± 0.006	0.824 ± 0.006	0.825 ± 0.006	0.780 ± 0.007	0.866 ± 0.006	**0.934 ± 0.004***
MOSAIC forward–reverse	Average	Virchow2	0.861 ± 0.006	0.891 ± 0.005	0.828 ± 0.006	0.864 ± 0.006	0.851 ± 0.006	0.775 ± 0.007	0.699 ± 0.008	**0.923 ± 0.004***
MOSAIC forward–reverse	Average	GenBio-PathFM	0.876 ± 0.005	0.888 ± 0.005	0.884 ± 0.005	0.850 ± 0.006	0.863 ± 0.005	0.726 ± 0.007	0.799 ± 0.007	**0.939 ± 0.004***

Measuring acquisition bias. Histology models trained on large public slide collections (Weinstein et al., 2013) can predict the submitting site of an image (Howard et al., 2021; Dehkharghanian et al., 2021). The robustness index compares same-biology neighbours from other medical cen-

Table 28: Table 9 with bootstrap standard deviations. SD over 1,000 bootstrap resamples of the 38,400 queries, recomputing R@1 and mAP per query from the saved features; both metrics of one base and method share the same resamples. Bold: best; underlined: second best. An asterisk marks a best method whose margin over the second best exceeds 1.96 times the combined bootstrap SD, treating the two estimates as independent.

Metric	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
R@1	GenBio-PathFM	0.884 ± 0.002	0.946 ± 0.001	0.885 ± 0.002	0.823 ± 0.002	0.848 ± 0.002	0.560 ± 0.003	0.394	**0.990 ± 0.001***
R@1	UNI2-h	0.967 ± 0.001	0.989 ± 0.001	0.926 ± 0.001	0.956 ± 0.001	0.969 ± 0.001	0.953 ± 0.001	0.981	**0.994 ± 0.000***
R@1	Virchow2	0.815 ± 0.002	0.924 ± 0.001	0.612 ± 0.003	0.807 ± 0.002	0.858 ± 0.002	0.824 ± 0.002	0.000	**0.954 ± 0.001***
mAP	GenBio-PathFM	0.791 ± 0.001	0.856 ± 0.001	0.761 ± 0.001	0.689 ± 0.001	0.762 ± 0.001	0.481 ± 0.002	0.165	**0.961 ± 0.001***
mAP	UNI2-h	0.854 ± 0.001	0.933 ± 0.001	0.819 ± 0.001	0.838 ± 0.001	0.872 ± 0.001	0.841 ± 0.001	0.919	**0.975 ± 0.000***
mAP	Virchow2	0.706 ± 0.001	0.830 ± 0.001	0.456 ± 0.002	0.701 ± 0.001	0.758 ± 0.001	0.716 ± 0.001	0.002	**0.872 ± 0.001***

Table 29: Table 10 with bootstrap standard deviations. SD over 1,000 bootstrap resamples of the 546 inter-staining slide pairs (same scanner, different staining), recomputing the median top-10 retrieval accuracy on each resample. Bold: best; underlined: second best. An asterisk marks a best method whose margin over the second best exceeds 1.96 times the combined bootstrap SD, treating the two estimates as independent.

Tiles per slide	Base model	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
8,139	GenBio-PathFM	0.409 ± 0.012	0.404 ± 0.014	0.347 ± 0.011	0.401 ± 0.012	0.382 ± 0.012	0.240 ± 0.010	0.166	**0.529 ± 0.017***
8,139	UNI2-h	0.190 ± 0.011	0.332 ± 0.016	0.254 ± 0.012	0.198 ± 0.013	0.210 ± 0.011	0.192 ± 0.010	0.291	**0.409 ± 0.015***
8,139	Virchow2	0.299 ± 0.013	0.352 ± 0.009	0.189 ± 0.010	0.307 ± 0.014	0.325 ± 0.016	0.288 ± 0.014	0.005	**0.423 ± 0.012***
2,713	GenBio-PathFM	0.598 ± 0.011	0.593 ± 0.014	0.530 ± 0.011	0.591 ± 0.011	0.573 ± 0.012	0.414 ± 0.012	0.268	**0.687 ± 0.015***
2,713	UNI2-h	0.334 ± 0.014	0.503 ± 0.014	0.396 ± 0.016	0.345 ± 0.016	0.361 ± 0.014	0.337 ± 0.013	0.459	**0.575 ± 0.015***
2,713	Virchow2	0.487 ± 0.015	0.529 ± 0.010	0.321 ± 0.012	0.490 ± 0.014	0.506 ± 0.017	0.452 ± 0.016	0.011	**0.595 ± 0.011***

Table 30: Table 2 with bootstrap standard deviations. SD in percentage points over 1,000 bootstrap resamples. Confounding present: retrieval resamples the queries with their fixed galleries and fixed-candidate NN accuracy the 200 query-candidate sets. Confounding absent: within each transfer direction the retrieval queries or the query-candidate sets are resampled, and the directions are averaged in every draw. Bold: best; underlined: second best. An asterisk marks a best method whose margin over the second best exceeds 1.96 times the combined bootstrap SD, treating the two estimates as independent.

Institution–class confounding	Dataset	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
Present	Camelyon16	UNI2-h	73.8 ± 0.5	78.0 ± 0.4	72.7 ± 0.4	71.2 ± 0.6	72.2 ± 0.5	69.6 ± 0.5	80.6 ± 0.4	**86.0 ± 0.4***
Present	Camelyon16	Virchow2	69.6 ± 0.5	76.5 ± 0.5	65.5 ± 0.6	73.5 ± 0.5	69.8 ± 0.5	57.6 ± 0.6	49.6 ± 0.6	**84.1 ± 0.5***
Present	Camelyon16	GenBio-PathFM	81.7 ± 0.4	80.8 ± 0.4	79.6 ± 0.4	80.9 ± 0.4	80.8 ± 0.4	53.8 ± 0.6	58.1 ± 0.4	**88.0 ± 0.4***
Present	BEETLE	UNI2-h	60.3 ± 0.5	61.7 ± 0.5	61.7 ± 0.5	63.0 ± 0.5	60.9 ± 0.5	57.6 ± 0.4	64.1 ± 0.5	**73.2 ± 0.5***
Present	BEETLE	Virchow2	65.3 ± 0.5	66.7 ± 0.5	65.1 ± 0.5	67.2 ± 0.6	67.6 ± 0.5	65.2 ± 0.5	63.8 ± 0.4	**73.0 ± 0.6***
Present	BEETLE	GenBio-PathFM	62.8 ± 0.5	64.5 ± 0.5	64.2 ± 0.5	62.1 ± 0.4	64.4 ± 0.5	56.3 ± 0.4	60.3 ± 0.4	**72.1 ± 0.5***
Present	TumSeg	UNI2-h	53.3 ± 0.3	52.1 ± 0.3	50.0 ± 0.3	54.6 ± 0.3	52.8 ± 0.3	52.4 ± 0.3	54.4 ± 0.3	**56.8 ± 0.4***
Present	TumSeg	Virchow2	46.1 ± 0.3	46.4 ± 0.3	49.2 ± 0.3	46.4 ± 0.3	46.2 ± 0.3	46.7 ± 0.3	**50.5 ± 0.4***	48.1 ± 0.3
Present	TumSeg	GenBio-PathFM	53.3 ± 0.3	52.0 ± 0.3	52.4 ± 0.3	**55.9 ± 0.3***	52.1 ± 0.3	51.5 ± 0.3	52.1 ± 0.4	54.7 ± 0.3
Absent	Camelyon16	UNI2-h	89.9 ± 0.1	89.5 ± 0.1	88.8 ± 0.0	89.4 ± 0.1	88.3 ± 0.1	84.8 ± 0.1	**92.8 ± 0.0**	92.3 ± 0.0
Absent	Camelyon16	Virchow2	90.5 ± 0.0	90.0 ± 0.0	86.5 ± 0.1	**90.9 ± 0.1***	86.9 ± 0.0	82.8 ± 0.1	82.5 ± 0.1	90.7 ± 0.0
Absent	Camelyon16	GenBio-PathFM	**92.4 ± 0.0**	91.5 ± 0.0	91.3 ± 0.0	90.4 ± 0.1	90.9 ± 0.0	83.8 ± 0.1	84.0 ± 0.1	92.2 ± 0.0
Absent	BEETLE	UNI2-h	57.0 ± 0.4	58.6 ± 0.4	60.1 ± 0.4	56.9 ± 0.4	56.7 ± 0.4	52.1 ± 0.4	60.7 ± 0.4	**61.7 ± 0.4**
Absent	BEETLE	Virchow2	61.9 ± 0.4	61.7 ± 0.4	**62.7 ± 0.3**	61.5 ± 0.4	58.9 ± 0.4	57.1 ± 0.4	44.3 ± 0.3	62.4 ± 0.4
Absent	BEETLE	GenBio-PathFM	57.0 ± 0.4	57.5 ± 0.4	58.2 ± 0.4	56.7 ± 0.4	57.2 ± 0.4	49.1 ± 0.4	46.9 ± 0.3	**59.5 ± 0.4***
Absent	TumSeg	UNI2-h	62.7 ± 0.1	61.2 ± 0.1	57.9 ± 0.1	62.2 ± 0.1	62.6 ± 0.1	62.1 ± 0.1	63.1 ± 0.1	**64.3 ± 0.1***
Absent	TumSeg	Virchow2	58.3 ± 0.1	58.2 ± 0.1	58.3 ± 0.1	58.5 ± 0.1	58.1 ± 0.1	58.0 ± 0.1	58.4 ± 0.1	**59.1 ± 0.1***
Absent	TumSeg	GenBio-PathFM	61.4 ± 0.1	60.8 ± 0.1	59.9 ± 0.1	**61.7 ± 0.1**	61.3 ± 0.1	61.1 ± 0.1	**61.7 ± 0.1**	**61.7 ± 0.1**

ters with other-biology neighbours from the same center, and de Jong et al. (2025) find that nearly all public pathology foundation models represent the medical center strongly. PathoROB extends this measurement to four multi-center datasets and 20 foundation models, adds a performance drop under confounded training and a clustering score, and links robustness deficits to diagnostic errors (Kömen et al., 2026). CRoMa replaces the index with a signed per-sample margin (Grisi et al., 2026). Registered multi-scanner and multi-staining resources isolate single acquisition factors (Ochi et al., 2024; Ryu et al., 2025), and a multi-scanner benchmark shows that foundation models remain sensitive to the scanner (Carloni et al., 2025). We use PathoROB (Kömen et al., 2026), PLISM (Ochi et al., 2024), SCORPION (Ryu et al., 2025) and MOSAIC (Zhang et al., 2026c) to evaluate the effectiveness of different approaches for improving the robustness of histology models.

Table 31: Table 12 with bootstrap standard deviations. SD in percentage points over 1,000 bootstrap resamples. Confounding present: retrieval resamples the queries with their fixed galleries and fixed-candidate NN accuracy the 200 query-candidate sets. Confounding absent: within each transfer direction the retrieval queries or the query-candidate sets are resampled, and the directions are averaged in every draw. Bold: best; underlined: second best. An asterisk marks a best method whose margin over the second best exceeds 1.96 times the combined bootstrap SD, treating the two estimates as independent.

Institution–class confounding	Dataset	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
Present	Camelyon16	UNI2-h	61.0 ± 3.5	68.5 ± 3.3	58.0 ± 3.6	55.5 ± 3.5	59.5 ± 3.5	48.5 ± 3.6	69.5 ± 3.3	**78.0 ± 2.9**
Present	Camelyon16	Virchow2	55.0 ± 3.6	62.0 ± 3.4	46.5 ± 3.5	60.5 ± 3.5	58.5 ± 3.6	34.5 ± 3.3	20.0 ± 2.9	**75.0 ± 2.9***
Present	Camelyon16	GenBio-PathFM	71.5 ± 3.2	72.0 ± 3.3	69.0 ± 3.2	69.5 ± 3.1	66.5 ± 3.4	25.0 ± 3.1	35.5 ± 3.3	**81.5 ± 2.8***
Present	BEETLE	UNI2-h	18.5 ± 2.8	21.5 ± 2.9	22.5 ± 3.0	27.0 ± 3.2	21.0 ± 2.8	14.5 ± 2.5	24.0 ± 3.0	**44.0 ± 3.6***
Present	BEETLE	Virchow2	27.0 ± 3.2	32.5 ± 3.4	29.0 ± 3.1	30.5 ± 3.2	32.0 ± 3.3	28.0 ± 3.0	27.5 ± 3.1	**48.0 ± 3.6***
Present	BEETLE	GenBio-PathFM	22.5 ± 3.0	26.5 ± 3.2	26.5 ± 2.9	21.5 ± 2.9	27.5 ± 3.3	16.0 ± 2.6	27.0 ± 3.0	**41.0 ± 3.4***
Present	TumSeg	UNI2-h	33.0 ± 3.4	29.5 ± 3.2	19.0 ± 2.8	**38.0 ± 3.5**	32.5 ± 3.2	31.5 ± 3.3	31.5 ± 3.4	**38.0 ± 3.3**
Present	TumSeg	Virchow2	20.0 ± 2.7	22.5 ± 2.9	21.0 ± 2.8	20.0 ± 2.7	20.5 ± 2.9	22.0 ± 2.9	**28.0 ± 3.1**	22.5 ± 3.0
Present	TumSeg	GenBio-PathFM	34.5 ± 3.3	28.5 ± 3.1	28.0 ± 3.3	32.5 ± 3.4	34.0 ± 3.4	31.0 ± 3.3	29.0 ± 3.2	**35.5 ± 3.5**
Absent	Camelyon16	UNI2-h	59.7 ± 2.4	60.5 ± 2.4	52.5 ± 2.5	59.7 ± 2.4	54.2 ± 2.4	44.5 ± 2.4	**70.0 ± 2.3**	67.5 ± 2.4
Absent	Camelyon16	Virchow2	54.2 ± 2.5	54.0 ± 2.5	45.8 ± 2.5	**62.5 ± 2.3**	41.2 ± 2.4	39.2 ± 2.4	51.2 ± 2.5	61.0 ± 2.5
Absent	Camelyon16	GenBio-PathFM	64.5 ± 2.5	61.0 ± 2.5	59.5 ± 2.4	62.7 ± 2.4	60.0 ± 2.4	42.8 ± 2.5	43.8 ± 2.5	**67.5 ± 2.3**
Absent	BEETLE	UNI2-h	52.1 ± 1.0	54.9 ± 1.0	59.3 ± 1.0	51.5 ± 1.0	50.6 ± 1.1	44.9 ± 1.0	57.8 ± 1.0	**62.1 ± 1.0***
Absent	BEETLE	Virchow2	58.0 ± 1.0	59.5 ± 1.0	59.3 ± 1.0	59.2 ± 1.0	53.0 ± 1.0	51.7 ± 1.0	37.6 ± 1.0	**63.6 ± 0.9***
Absent	BEETLE	GenBio-PathFM	52.8 ± 1.0	54.2 ± 1.0	55.0 ± 1.0	51.2 ± 1.0	52.3 ± 1.0	41.7 ± 1.0	41.7 ± 1.0	**59.6 ± 1.0***
Absent	TumSeg	UNI2-h	37.8 ± 2.4	33.5 ± 2.4	29.2 ± 2.3	39.5 ± 2.4	37.8 ± 2.4	36.8 ± 2.4	37.8 ± 2.3	**41.8 ± 2.5**
Absent	TumSeg	Virchow2	25.8 ± 2.2	25.2 ± 2.2	27.2 ± 2.2	24.5 ± 2.1	25.5 ± 2.2	25.5 ± 2.2	27.3 ± 2.2	**28.0 ± 2.4**
Absent	TumSeg	GenBio-PathFM	32.5 ± 2.3	28.7 ± 2.2	27.5 ± 2.2	**33.8 ± 2.4**	32.8 ± 2.4	32.8 ± 2.4	31.5 ± 2.3	33.2 ± 2.4

Table 32: Table 3 with bootstrap standard deviations. SD over 1,000 bootstrap resamples of the per-slide Pearson scores: within each held-out source group the slides are resampled with replacement and averaged, then the group means are averaged; Average reuses the same draws, so its SD is that of the mean of the ten organ draws (joint resampling over independent organs). Bold: best; underlined: second best. An asterisk marks a best method whose margin over the second best exceeds 1.96 times the combined bootstrap SD, treating the two estimates as independent.

Organ	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
Breast	UNI2-h	0.157 ± 0.023	0.195 ± 0.020	0.130 ± 0.025	0.151 ± 0.020	0.114 ± 0.026	0.100 ± 0.029	**0.204**	0.192 ± 0.021
–	Virchow2	0.203 ± 0.033	0.176 ± 0.023	0.083 ± 0.025	0.100 ± 0.037	0.132 ± 0.017	0.060 ± 0.020	0.049	**0.206 ± 0.024**
–	GenBio-PathFM	0.130 ± 0.023	0.144 ± 0.031	**0.161 ± 0.028**	0.091 ± 0.010	0.099 ± 0.019	0.035 ± 0.026	0.067	**0.161 ± 0.022**
Bowel	UNI2-h	0.240 ± 0.019	0.287 ± 0.022	0.291 ± 0.021	0.251 ± 0.017	0.242 ± 0.020	0.233 ± 0.017	0.280	**0.326 ± 0.021**
–	Virchow2	0.250 ± 0.015	0.279 ± 0.023	0.174 ± 0.017	0.170 ± 0.018	0.254 ± 0.015	0.160 ± 0.011	-0.013	**0.313 ± 0.017**
–	GenBio-PathFM	0.281 ± 0.022	0.265 ± 0.020	0.262 ± 0.022	0.260 ± 0.017	0.269 ± 0.019	0.041 ± 0.016	0.051	**0.330 ± 0.021**
Kidney	UNI2-h	0.083 ± 0.013	0.076 ± 0.012	0.084 ± 0.013	0.077 ± 0.011	**0.095 ± 0.012**	0.086 ± 0.013	0.075	0.087 ± 0.012
–	Virchow2	0.091 ± 0.013	0.074 ± 0.013	0.046 ± 0.011	0.078 ± 0.013	0.076 ± 0.013	0.057 ± 0.012	0.046	**0.105 ± 0.013**
–	GenBio-PathFM	**0.090 ± 0.015**	**0.090 ± 0.013**	0.069 ± 0.013	0.052 ± 0.011	0.069 ± 0.012	0.043 ± 0.012	0.072	0.083 ± 0.013
Brain	UNI2-h	0.039 ± 0.019	0.051 ± 0.024	0.016 ± 0.020	0.043 ± 0.019	0.054 ± 0.015	0.028 ± 0.023	0.061	**0.073 ± 0.011**
–	Virchow2	0.061 ± 0.020	0.054 ± 0.016	0.019 ± 0.009	0.055 ± 0.007	0.059 ± 0.020	0.042 ± 0.019	0.018	**0.074 ± 0.012**
–	GenBio-PathFM	0.035 ± 0.012	0.014 ± 0.014	0.029 ± 0.011	**0.058 ± 0.011**	0.022 ± 0.010	0.021 ± 0.007	0.007	0.049 ± 0.017
Heart	UNI2-h	0.040 ± 0.003	0.057 ± 0.003	0.043 ± 0.004	0.029 ± 0.003	0.040 ± 0.004	0.043 ± 0.004	0.055	**0.061 ± 0.004**
–	Virchow2	0.036 ± 0.003	0.050 ± 0.003	0.038 ± 0.003	0.035 ± 0.003	0.030 ± 0.003	0.026 ± 0.003	0.033	**0.052 ± 0.004**
–	GenBio-PathFM	0.050 ± 0.003	**0.054 ± 0.004**	0.047 ± 0.004	0.022 ± 0.003	0.038 ± 0.003	0.015 ± 0.003	0.019	0.050 ± 0.004
Prostate	UNI2-h	0.167 ± 0.033	0.144 ± 0.030	0.147 ± 0.032	0.151 ± 0.034	0.167 ± 0.030	0.103 ± 0.038	0.183	**0.244 ± 0.036**
–	Virchow2	0.162 ± 0.042	0.130 ± 0.033	0.075 ± 0.021	0.164 ± 0.027	0.173 ± 0.041	0.108 ± 0.031	-0.095	**0.222 ± 0.034**
–	GenBio-PathFM	0.168 ± 0.036	0.163 ± 0.034	0.170 ± 0.032	0.128 ± 0.020	0.205 ± 0.036	0.049 ± 0.013	-0.004	**0.242 ± 0.037**
Lung	UNI2-h	0.097 ± 0.016	0.117 ± 0.034	0.126 ± 0.029	0.078 ± 0.032	0.131 ± 0.025	0.121 ± 0.022	0.102	**0.251 ± 0.018***
–	Virchow2	0.114 ± 0.030	0.133 ± 0.020	0.088 ± 0.019	0.153 ± 0.016	0.116 ± 0.026	-0.009 ± 0.029	0.024	**0.173 ± 0.027**
–	GenBio-PathFM	0.070 ± 0.027	0.116 ± 0.020	0.100 ± 0.024	0.057 ± 0.026	0.070 ± 0.023	0.061 ± 0.021	0.047	**0.209 ± 0.020***
Liver	UNI2-h	0.151 ± 0.011	0.134 ± 0.019	0.144 ± 0.010	0.093 ± 0.009	0.163 ± 0.015	0.150 ± 0.011	**0.166**	**0.166 ± 0.032**
–	Virchow2	0.157 ± 0.012	**0.168 ± 0.009**	0.099 ± 0.015	0.057 ± 0.014	0.161 ± 0.010	0.088 ± 0.017	0.050	0.122 ± 0.019
–	GenBio-PathFM	0.174 ± 0.019	0.151 ± 0.017	0.158 ± 0.020	0.109 ± 0.024	**0.176 ± 0.027**	0.099 ± 0.042	0.032	0.156 ± 0.015
Skin	UNI2-h	**0.480 ± 0.031**	0.465 ± 0.030	0.446 ± 0.031	0.299 ± 0.028	0.469 ± 0.030	0.415 ± 0.030	0.477	0.470 ± 0.032
–	Virchow2	0.486 ± 0.034	0.460 ± 0.033	0.350 ± 0.029	0.414 ± 0.030	0.475 ± 0.033	0.346 ± 0.035	0.156	**0.508 ± 0.034**
–	GenBio-PathFM	0.463 ± 0.029	0.460 ± 0.026	0.460 ± 0.026	0.385 ± 0.031	0.421 ± 0.025	0.256 ± 0.027	0.243	**0.485 ± 0.031**
Uterus	UNI2-h	0.109 ± 0.012	0.114 ± 0.011	0.134 ± 0.009	0.125 ± 0.010	0.093 ± 0.009	0.072 ± 0.009	0.072	**0.192 ± 0.012***
–	Virchow2	0.214 ± 0.016	0.222 ± 0.017	0.173 ± 0.029	0.169 ± 0.014	0.179 ± 0.012	0.129 ± 0.009	-0.013	**0.234 ± 0.012**
–	GenBio-PathFM	0.141 ± 0.021	0.131 ± 0.010	0.131 ± 0.029	0.081 ± 0.009	0.102 ± 0.015	0.070 ± 0.008	0.140	**0.213 ± 0.017***
Average	UNI2-h	0.156 ± 0.006	0.164 ± 0.007	0.156 ± 0.007	0.130 ± 0.007	0.157 ± 0.007	0.135 ± 0.007	0.167	**0.206 ± 0.007**
–	Virchow2	0.177 ± 0.008	0.175 ± 0.006	0.114 ± 0.006	0.140 ± 0.007	0.165 ± 0.007	0.101 ± 0.007	0.026	**0.201 ± 0.007***
–	GenBio-PathFM	0.160 ± 0.007	0.159 ± 0.007	0.159 ± 0.007	0.124 ± 0.006	0.147 ± 0.007	0.069 ± 0.007	0.067	**0.198 ± 0.007***

Mitigation grouped by the information it uses. Image-level methods use the input image and a fixed reference or augmentation distribution. Reinhard and Macenko normalization match colour statistics or stain vectors to a reference (Reinhard et al., 2001; Macenko et al., 2009), and stain augmentation and learned stain transfer vary appearance during training (Shen et al., 2022; Shaban et al., 2018; Wagner et al., 2021; Kang et al., 2023). Center signal persists in foundation-model fea-

Table 33: Table 4 with bootstrap standard deviations. SD over 1,000 bootstrap resamples of the crops within each held-out fold, recomputing the fold AUROC and then the mean (Mean AUROC) or minimum (Worst fold) over the folds of that axis. Bold: best; underlined: second best. An asterisk marks a best method whose margin over the second best exceeds 1.96 times the combined bootstrap SD, treating the two estimates as independent.

Evaluation	Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
Tumor type out – Mean AUROC	UNI2-h	0.843 ± 0.003	0.845 ± 0.003	0.798 ± 0.003	0.705 ± 0.004	0.838 ± 0.003	0.801 ± 0.003	**0.851**	0.849 ± 0.003
Tumor type out – Mean AUROC	Virchow2	**0.834 ± 0.003**	0.833 ± 0.003	0.697 ± 0.004	0.698 ± 0.004	0.832 ± 0.003	0.749 ± 0.004	0.569	**0.834 ± 0.003**
Tumor type out – Mean AUROC	GenBio-PathFM	0.833 ± 0.003	0.827 ± 0.003	0.811 ± 0.003	0.672 ± 0.004	0.830 ± 0.003	0.729 ± 0.004	0.813	**0.850 ± 0.003***
Tumor type out – Worst fold	UNI2-h	0.775 ± 0.005	0.779 ± 0.005	0.733 ± 0.008	0.656 ± 0.006	0.762 ± 0.005	0.726 ± 0.006	**0.789**	0.784 ± 0.008
Tumor type out – Worst fold	Virchow2	0.792 ± 0.006	**0.797 ± 0.006**	0.607 ± 0.013	0.638 ± 0.006	0.795 ± 0.007	0.669 ± 0.013	0.537	0.788 ± 0.005
Tumor type out – Worst fold	GenBio-PathFM	0.786 ± 0.005	0.776 ± 0.006	0.745 ± 0.005	0.598 ± 0.007	0.781 ± 0.005	0.676 ± 0.007	0.751	**0.803 ± 0.005***
Scanner out – Mean AUROC	UNI2-h	0.845 ± 0.003	0.851 ± 0.003	0.797 ± 0.003	0.719 ± 0.004	0.835 ± 0.003	0.803 ± 0.003	0.854	**0.857 ± 0.003**
Scanner out – Mean AUROC	Virchow2	**0.837 ± 0.003**	0.836 ± 0.003	0.686 ± 0.004	0.698 ± 0.004	0.834 ± 0.003	0.756 ± 0.004	0.547	0.832 ± 0.003
Scanner out – Mean AUROC	GenBio-PathFM	0.833 ± 0.003	0.829 ± 0.003	0.821 ± 0.003	0.678 ± 0.004	0.834 ± 0.003	0.724 ± 0.004	0.799	**0.851 ± 0.003***
Scanner out – Worst fold	UNI2-h	0.796 ± 0.004	0.807 ± 0.004	0.744 ± 0.006	0.623 ± 0.005	0.766 ± 0.006	0.746 ± 0.007	0.808	**0.813 ± 0.005**
Scanner out – Worst fold	Virchow2	0.788 ± 0.004	0.797 ± 0.004	0.607 ± 0.008	0.641 ± 0.005	0.780 ± 0.004	0.726 ± 0.005	0.506	**0.800 ± 0.004**
Scanner out – Worst fold	GenBio-PathFM	0.761 ± 0.004	0.765 ± 0.004	0.741 ± 0.004	0.624 ± 0.005	0.771 ± 0.004	0.658 ± 0.008	0.728	**0.788 ± 0.004***

Table 34: Table 5 with bootstrap standard deviations. SD over 1,000 bootstrap resamples of the biopsies within each transfer direction, recomputing balanced accuracy and averaging the two directions. Bold: best; underlined: second best. An asterisk marks a best method whose margin over the second best exceeds 1.96 times the combined bootstrap SD, treating the two estimates as independent.

Base	None	Reinhard	Macenko	AdvInv	LEACE	INLP	DoRA	ROPAL
UNI2-h	0.483 ± 0.008	0.515 ± 0.009	0.516 ± 0.010	0.364 ± 0.007	0.494 ± 0.009	0.437 ± 0.010	0.536	**0.580 ± 0.011**
Virchow2	0.493 ± 0.009	0.477 ± 0.009	0.477 ± 0.009	0.433 ± 0.009	0.521 ± 0.009	0.471 ± 0.009	0.173	**0.537 ± 0.009**
GenBio-PathFM	0.537 ± 0.011	0.531 ± 0.010	0.478 ± 0.009	0.464 ± 0.010	0.526 ± 0.010	0.356 ± 0.009	0.255	**0.543 ± 0.010**

Table 35: The judged corpus by cohort. Judged: role-blind odd-one-out judgments; kept: condition true and a same-domain tile called odd; rate: kept over judged. The twelve cohorts above the rule form the 42,820-pair training set; all twenty-six form the 60,985-pair set. Cohorts are TCGA project codes.

Cohort	Judged	Kept	Rate	Cohort	Judged	Kept	Rate
BRCA	11,000	7,462	67.8%	KICH	15,233	2,569	16.9%
COAD	11,168	4,428	39.6%	KIRC	8,366	2,430	29.0%
LUSC	12,213	3,885	31.8%	KIRP	8,367	2,053	24.5%
LUAD	11,064	3,843	34.7%	*twelve cohorts: 130,894 judged, 42,820 kept,*			*32.7%*
SKCM	10,456	3,454	33.0%	OV	2,750	1,133	41.2%
UCEC	10,821	3,283	30.3%	BLCA	5,500	2,167	39.4%
UCS	10,865	3,262	30.0%	PRAD	5,500	2,107	38.3%
READ	9,834	3,076	31.3%	PAAD	2,750	1,026	37.3%
UVM	11,507	3,075	26.7%	CESC	2,750	1,007	36.6%
ESCA	2,750	998	36.3%	STAD	2,750	908	33.0%
THCA	5,501	1,964	35.7%	GBM	5,501	1,639	29.8%
HNSC	5,499	1,925	35.0%	SARC	2,750	674	24.5%
LGG	5,499	1,336	24.3%	LIHC	2,750	668	24.3%
TGCT	2,750	608	22.1%	*twenty-six cohorts: ≈184k judged, 60,985 kept*			

tures after stain normalization (Kömen et al., 2024). Feature harmonization uses groups of samples with batch labels: ComBat estimates location and scale corrections for each batch from its features (Johnson et al., 2007), pathology patch embeddings have been harmonized with the training cohort as the reference batch (Murchan et al., 2024), and ComBat gives the largest robustness-index gain among the post-hoc methods evaluated in PathoROB (Kömen et al., 2026). Nuisance labels such as the source site support methods that remove domain information from a representation. Adversarial invariance trains an encoder against a nuisance predictor while a reconstruction or variational term keeps the remaining input information (Edwards & Storkey, 2015; Louizos et al., 2015; Xie et al., 2017; Dincer et al., 2020; Cavinato et al., 2023), and concept erasure removes linearly decodable attributes with a closed-form or iterative projection (Belrose et al., 2023; Ravfogel et al., 2020). Combined with task labels, domain-adversarial training (Ganin et al., 2015) has been applied to frozen pathology features with a hospital adversary (Zhang, 2026), an optimal-transport loss

aligns features across institutions (Falahkheirkhah et al., 2023), and class-conditional discrepancy fine-tuning improves cross-hospital transfer (Vardi et al., 2026). Domain invariance alone need not retain task information when label distributions differ between domains (Zhao et al., 2019). Scans of the same specimen on several scanners support contrastive feature projections (Carloni et al., 2025). Self-supervised continued training uses images alone, with a slide adversary in AdvDINO (Su et al., 2025), a stain adversary in SASSL (Ye et al., 2025), stain-varied views in Latent Manifold Compaction (Zhang et al., 2026b) and DoRA adapters trained by DINO self-distillation in Pathology DoRA (Bøe et al., 2026). Filiot et al. (2026) fine-tune ten pathology foundation models for scanner and staining robustness, report an average PathoROB robustness index increase from 0.72 to 0.87 together with downstream gains, and release the fine-tuned Phikon-v2 (Phaet) and Midnight-12k (Mascaret). Other approaches intervene on sparse-autoencoder concepts (Kim et al., 2026), calibrate features with vision–language concepts (Chen et al., 2024c), use knowledge-guided information bottlenecks (Huang et al., 2025), regularize downstream predictors (Henriksen et al., 2026) or adapt retrieval continually across centers (Zhu et al., 2026). ROPAL uses unlabeled tiles grouped by slide and institution and obtains the biological similarity signal from a multimodal judge.

Learning from similarity judgments. The THINGS odd-one-out datasets measure human object similarity at scale (Hebart et al., 2019; 2020), and SPoSE and VICE learn interpretable object embeddings from these triplets (Zheng et al., 2019; Muttenthaler et al., 2022b). Odd-one-out observations also support representation learning (Mohammadi et al., 2020). gLocal aligns the global structure of network representations with human similarity judgments while preserving local structure, which improves few-shot learning and anomaly detection (Muttenthaler et al., 2023). AligNet transfers human-aligned structure across levels of abstraction by fine-tuning vision foundation models, which improves generalization and out-of-distribution robustness (Muttenthaler et al., 2025). DreamSim collects human comparisons of synthetic image pairs and tunes a perceptual metric, including through LoRA, that agrees better with human perception (Fu et al., 2023). Fine-tuning on human perceptual triplets improves many downstream vision tasks (Sundaram et al., 2024), and behavioral geometric supervision adapts video backbones with LoRA to human odd-one-out judgments (Garcia & Isik, 2025). Relational distillation keeps the rankings, distances or pairwise activations of a reference model (Chen et al., 2017; Park et al., 2019; Tung & Mori, 2019), and LDIFS regularizes features during fine-tuning (Mukhoti et al., 2023). We apply similarity-judgment adaptation to acquisition robustness in pathology, with triplets built across slides and institutions and a multimodal LLM in place of human raters.

Multimodal LLMs as teachers. Multimodal LLM judges show human-like discernment in pairwise comparisons and diverge more from people in scoring and batch ranking (Chen et al., 2024a). VLM preferences have trained reward models with a Bradley–Terry objective (Wang et al., 2024d) and post-trained image-compression models (Sargent et al., 2025). *Embedding the Teacher* distills VLM preference rankings into an image encoder for text–image retrieval, so that retrieval no longer calls the VLM (He et al., 2025), and UniME-V2 uses MLLM semantic scores as soft supervision for multimodal embeddings (Gu et al., 2025). Preference optimization has also been applied to contrastive visual representations (Afzali et al., 2024). Other work makes the MLLM part of the embedding model. GLMP describes histology patches with a general-purpose MLLM and encodes the descriptions with a text encoder, which yields embeddings with reduced institution effects (Zhang et al., 2026d). MLLM4PUE and VLM2Vec train MLLM-based embedding models contrastively (Zhou et al., 2025b; Jiang et al., 2024), and MegaPairs generates image–instruction–image triples for multimodal retrieval (Zhou et al., 2025a). Our work uses an MLLM as the teacher of odd-one-out preferences on pathology tiles and distills them into an existing pathology foundation model, which is then used without the MLLM.

A.10 Extended related work

Table 36: Methods and the constraints of the problem setup (Section 4.1). No manual labels: no biological labels or other manual annotations during training. Independent test embedding: each test sample is embedded by a fixed mapping without access to the test population, a calibration cohort or batch labels. No paired scans: no registered or same-specimen images across acquisitions. Adapts a given encoder: produces patch embeddings from a given pretrained pathology encoder by preprocessing, a feature map or fine-tuning. Author code: code for the published method is available. Our bases: applicable to UNI2-h, Virchow2 and GenBio-PathFM. ✓: satisfied. ✗: not satisfied, with the relevant requirement. Not assessed: not checked for this method.

Method	No manual labels	Independent test embedding	No paired scans	Adapts a given encoder	Author code	Our bases	Evaluated
ROPAL	✓	✓	✓	✓	✓	✓	✓
Reinhard (Reinhard et al., 2001)	✓	✓	✓	✓	✓	✓	✓
Macenko (Macenko et al., 2009)	✓	✓	✓	✓	✓	✓	✓
AdvInv (Edwards & Storkey, 2015; Dincer et al., 2020)	✓	✓	✓	✓	✓	✓	✓
LEACE (Belrose et al., 2023)	✓	✓	✓	✓	✓	✓	✓
INLP (Ravfogel et al., 2020)	✓	✓	✓	✓	✓	✓	✓
Pathology DoRA (Bøe et al., 2026)	✓	✓	✓	✓	✓	✓	✓
ComBat (Johnson et al., 2007; Murchan et al., 2024)	✓	✗ Target-batch features and batch labels	✓	✓	✓	✓	✗
Filiot et al. (Phaet, Mascaret) (Filiot et al., 2026)	Not described	✓	Not described	Not described	✗ Fine-tuning method not described	✗ Released for Phikon-v2 and Midnight-12k	✗
StainGAN (published protocol) (Shaban et al., 2018)	✓	✗ Destination-domain images for fitting	✓	✓	✓	✓	✗
HistAuGAN (Wagner et al., 2021)	✓	✓	✓	✗ Augmentation for a separate training objective	✓	Not assessed	✗
Stain-augmented SSL (Kang et al., 2023)	✓	✓	✓	✗ Pretraining recipe	✗ GMM extension not released	Not assessed	✗
AdvDINO (Su et al., 2025)	✓	✓	✓	✓	✓	✗ Released trainer not a drop-in fit	✗
DANN (published objective) (Ganin et al., 2015)	✗ Source task labels	✗ Unlabeled target images	✓	✓	Not assessed	Not assessed	✗

Histopathology DANN (Lafarge et al., 2017)	✗ Mitosis annotations	Not assessed	✓	✗ Trains a task network	Not assessed	Not assessed	✗
Adversarial adaptor (Zhang, 2026)	✗ Disease labels	✓	✓	✓	Not assessed	Not assessed	✗
OT regularization (Falahkheirkhah et al., 2023)	✗ Source task labels	✓	✓	Not assessed	Not assessed	Not assessed	✗
PFM-LMMD (Vardi et al., 2026)	✗ Class labels	Target images in the adaptation setting	✓	✓	Not assessed	Not assessed	✗
MMD-AAE (Li et al., 2018a)	✗ Task labels	✓	✓	Not assessed	Not assessed	Not assessed	✗
H&E-adversarial network (Marini et al., 2021)	✗ Task labels	✓	✓	Not assessed	Not assessed	Not assessed	✗
DIPT (Ezzati et al., 2025)	✗ Class labels	✓	✓	Not assessed	Not assessed	Not assessed	✗
fmMAP (Nguyen & Ho, 2025)	✗ Class labels in supervised UMAP	Not assessed	✓	✓	Not assessed	Not assessed	✗
PICASSO artifact removal (Kim et al., 2026)	✗ Manual artifact labels	✓	✓	✓	Not assessed	Not assessed	✗
TT-SaD (published setting) (Tsai et al., 2024)	✗ Tumor labels select the reference stain	✓	✓	✓	Not assessed	Not assessed	✗
ScanGen (Carloni et al., 2025)	✗ Task labels	✓	✗ Same-specimen scans	✓	Not assessed	Not assessed	✗
FEATMAP (Donle et al., 2026)	✓	Not assessed	✗ Matched acquisitions	✓	Not assessed	Not assessed	✗
EXPOSE (published scanner protocol) (Witte et al., 2026)	✓	✓	✗ Registered scans	✓	Not assessed	Not assessed	✗
VFAE (Louizos et al., 2015)	✓	✗ Nuisance value at inference	✓	Not assessed	Not assessed	Not assessed	✗
DICA (Muandet et al., 2013)	✓	✓	✓	✓	✗	Not assessed	✗
GLMP (Zhang et al., 2026d)	✓	Uses within-slide context	✓	✗ Separate MLLM and text-encoder representation	Not assessed	Not assessed	✗
Embedding the Teacher (He et al., 2025)	✓	✓	✓	✗ Text-query retrieval task	Not assessed	Not assessed	✗
Biased-prompt debiasing (Chuang et al., 2023)	✓	✓	✓	✗ Acts on text embeddings	Not assessed	✗ Needs an aligned text encoder	✗
RoboShot (Adila et al., 2024)	✓	✓	✓	✓	Not assessed	✗ Needs a joint text–patch space	✗

CCL-K (Tsai et al., 2022)	✓	✓	✓	✗ Objective for training new encoders	Not assessed	Not assessed	✗	
Conditional contrastive learning (Ma et al., 2021)	✓	✓	✓	✗ Objective for training new encoders	Not assessed	Not assessed	✗	
IMPaSh (Vuong et al., 2023)	✓	✓	✓	✗ Pretraining recipe	Not assessed	Not assessed	✗	

This appendix extends the survey of Section 2, following the same ideas: measuring acquisition bias, mitigation grouped by the information it uses, multimodal LLMs as judges and teachers, and learning from similarity judgments and rankings, followed by parameter-efficient fine-tuning and benchmarks.

A.10.1 Measuring acquisition bias

The empirical case that histology models learn acquisition signatures was made by Howard et al. (2021) for TCGA and confirmed by Dehkharghanian et al. (2021), who showed that networks predict the acquisition site of TCGA images with high accuracy. Tellez et al. (2019) quantified how stain augmentation and color normalization affect cross-center generalization for CNNs, and Stacke et al. (2020) proposed representation-level measures of domain shift in histopathology. With the arrival of foundation models the question was asked again and answered the same way. Kömen et al. (2024) found that hospital signatures survive in FM embeddings, including after stain normalization. de Jong et al. (2025) introduced the robustness index, the ratio of same-biology other-center to other-biology same-center neighbors, and found that almost all public PFMs encode the medical center more strongly than the tissue or cancer type.

Kömen et al. (2026) scaled that measurement into PathoROB, four multi-center datasets from three public sources covering 28 biological classes and 34 medical centers, with three metrics: the robustness index, an average performance drop under confounded training, and a clustering score. Across 20 foundation models it documents robustness deficits everywhere, links them to diagnostic errors, and evaluates post-hoc robustification with Reinhard normalization, ComBat, domain-adversarial training and vision-language alignment. We use a participant-disjoint subset of PathoROB as an external robustness benchmark. Grisi et al. (2026) argue that the robustness index has structural limitations that make cross-model comparison unreliable and propose CRoMa, a signed per-sample margin between same-biology other-confounder and same-confounder other-biology neighbors, evaluated across 20 tile-level encoders. MOSAIC accuracy is a per-sample quantity of the same family.

A parallel line isolates single acquisition factors with registered data. Ochi et al. (2024) released PLISM, registered tiles of the same slides across scanners and stainings. Filiot et al. (2025) used it for a robustness analysis and distilled a small model markedly more robust to staining and scanning variation than much larger encoders. Ryu et al. (2025) released SCORPION, a publicly available multi-scanner resource, and Carloni et al. (2025) demonstrated scanner sensitivity in PFMs and proposed the ScanGen loss against it. Thiringer et al. (2026) isolated scanner effects on 384 breast whole-slide images digitized on five devices and found that most PFMs encode scanner variability even where AUC appears stable, which they identify as a calibration failure rather than an accuracy failure. MIDOG remains the standard multi-scanner, multi-domain challenge for mitosis detection (Aubreville et al., 2022), and Lin et al. (2025) show that stain variation and color normalization materially change prognostic predictions. Chai et al. (2026) report the same sensitivity for sarcomas and their mimics, Gustafsson & Rantalainen (2024) evaluate PFMs for prostate grading under distribution shift, and Schoenpflug et al. (2026) propose a protocol for evaluating robustness to H&E staining variation. Jahanifar et al. (2023) survey domain generalization in computational pathology and give guidelines for it.

Recent evaluations extend these measurements to multicenter skin subtyping (Esbri et al., 2025), multicenter and multiscanner pathology cohorts (Akebli & Della Mea, 2026), slide-specific effects in prostate grading (Ebbert & Della Corte, 2025), and controlled acquisition perturbations (Bonn et al., 2026; Yajnik et al., 2026). WILDS places hospital shift within a broader benchmark of

real deployment shifts (Koh et al., 2020). Controlled artifact stress tests also reveal failures of pathology predictors (Schömig-Markiefka et al., 2021). Co-registered scanner comparisons expose scanner-dependent retrieval behavior (Shafique et al., 2026), while representational similarity analysis reveals substantial slide dependence across pathology foundation models (Mishra & Lotter, 2025). These evaluations measure acquisition invariance separately from in-domain predictive performance.

A.10.2 Mitigation grouped by the information it uses

Image-level normalization and augmentation. Reinhard et al. (2001), Macenko et al. (2009) and Vahadane et al. (2016) are the canonical color-transfer, stain-vector and structure-preserving normalizations, and Hoque et al. (2023) review and experimentally compare them. Center signal nevertheless survives stain normalization in foundation model features (Kömen et al., 2024). On the augmentation side, RandStainNA and its successor sample stain styles during training (Shen et al., 2022; Wang et al., 2024a), ContriMix mixes content and attribute codes without domain labels (Nguyen et al., 2023), STRAP applies style transfer from medically irrelevant images (Yamashita et al., 2021), and stain-invariant features have been built by augmentation and by self-supervision (Otálora et al., 2019; Tiard et al., 2022). HistAug learns controllable transformations of cached pathology embeddings, making feature-level augmentation an alternative to repeated image extraction (Boutaj et al., 2025).

Learned image transformations include StainGAN's unpaired stain transfer (Shaban et al., 2018), HistAuGAN's morphology/stain disentanglement for augmentation (Wagner et al., 2021), and self-supervised disentanglement for transfer to unseen stain domains (Ling et al., 2023). The earlier adversarial stain-transfer framework of Bentaieb & Hamarneh (2018) instead jointly trains its generator with a task-specific prediction network. Stain Mix-up constructs label-independent color augmentations (Chang et al., 2021), and Faryna et al. (2021) adapt automated augmentation policies to H&E images. The value of normalization depends on the feature extractor and slide-classification setting (Wölflein et al., 2024).

Feature harmonization and domain adaptation. ComBat (Johnson et al., 2007) is the empirical-Bayes batch correction that Murchan et al. (2024) applied to deep pathology features and that PathoROB evaluates as representation robustification. It estimates location and scale parameters for each batch from the features and batch assignments of that batch. For external inference, Murchan et al. (2024) harmonize target features against the corrected training cohort as a reference batch. DANN trains a task classifier on labeled source examples together with a domain adversary that can use unlabeled target examples (Ganin et al., 2015), and Lafarge et al. (2017) apply it to mitosis detection in histopathology. Zhang (2026) keeps the pathology encoder frozen and trains a projection head with disease classification and a hospital adversary through gradient reversal. The method of Falahkheirkhah et al. (2023) replaces adversarial alignment with an optimal-transport loss between institutions, alongside source classification, and improves generalization to unseen institutions. Earlier prostate-slide adversarial adaptation transfers a source classifier to unlabeled target images (Ren et al., 2018), and ULSA combines source annotations, stain translation and consistency on unlabeled target stains (Reisenbüchler et al., 2024). DAN and Deep CORAL align distribution means and covariances, respectively, alongside source classification (Long et al., 2015; Sun & Saenko, 2016). IRM seeks predictors that remain optimal across training environments, and group DRO minimizes the worst-group risk (Arjovsky et al., 2019; Sagawa et al., 2019). Domain invariance alone need not retain predictive information under label shift (Zhao et al., 2019). Reconstruction-based adversarial methods remove a nuisance variable while reconstruction keeps the remaining input information (Edwards & Storkey, 2015; Dincer et al., 2020). VFAE combines reconstruction with distribution matching in unsupervised and semi-supervised forms (Louizos et al., 2015), and information-theoretic nuisance suppression uses conditional reconstruction (Moyer et al., 2018). Asadi-Aghbolaghi et al. (2024) compare generalization strategies for multi-center histopathology classification. Nguyen & Ho (2025) reduce the site-bias batch effect in FM features post hoc, and Drexlin et al. (2025) generate metadata-guided synthetic data to rebalance under-represented acquisition settings. Huang et al. (2025) adapt PFMs with knowledge guidance and report gains in cross-domain generalization and demographic fairness. Test-time adaptation (Wang et al., 2021) adapts at inference on unlabeled target data, and Bahumanya et al. (2026) benchmark it in computational pathology, reporting that continual methods cause the largest drift in model explanations.

UAD-FM combines uncertainty estimation, causal test-time adaptation and calibration for colorectal pathology (Lou et al., 2025).

Robustness through pretraining, distillation and fine-tuning. CS-CO combines cross-stain prediction with contrastive learning (Yang et al., 2022), and SASSL adds a stain discriminator to self-supervised representation learning (Ye et al., 2025). HistoLite studies lightweight autoencoder representations under scanner shift (Shafique et al., 2025). Latent Manifold Compaction trains a small encoder with stain-varying views and a redundancy-reduction objective, using source images without target data (Zhang et al., 2026b). AdvDINO continues self-supervised training from UNI weights on 5.46 million multiplex immunofluorescence tiles from 435 slides, with new DINO heads and a slide adversary, and learns representations less dominated by slide-specific effects (Su et al., 2025). Filiot et al. (2025) study distillation for acquisition robustness on PLISM. Kapse et al. (2026) changed the pretraining objective to a two-stage JEPA plus DINO strategy and reach a position among the most robust public models on PathoROB, and Alber et al. (2026) report robustness from the scale and diversity of the pretraining corpus, while Bosch et al. (2025) and Karasikov et al. (2025) show that slide diversity rather than raw patch count drives quality. Closer to fine-tuning, ScanGen trains a feature projection with a contrastive loss over same-specimen scanner pairs and supervised task training (Carloni et al., 2025). Vardi et al. (2026) fine-tune PFMs with class-conditional local maximum mean discrepancy and supervised cross-entropy, with consistent patch- and slide-level gains in domain adaptation and domain generalization.

Filiot et al. (2026) fine-tune ten PFMs for acquisition robustness and report gains in both the PathoROB robustness index and downstream performance. They release the fine-tuned Phikon-v2 (Phaet) and Midnight-12k (Mascaret). Kim & Jeong (2026) distill several PFMs into a compact encoder with clinical keywords from pathology reports supplying guidance through a vision-language meta-teacher. Efficiency- oriented distillation of PFMs is also pursued by Grashei et al. (2025) and Cho et al. (2025). SmartStu combines multi-teacher transfer with adversarial distillation to reduce nuisance information in breast-specific encoders (Chen et al., 2026a).

A.10.3 Multimodal LLMs as judges and teachers

Zheng et al. (2023a) studied LLM-as-a-judge, comparing pairwise comparison with single-answer grading and documenting position, verbosity and self-enhancement biases, and Chiang et al. (2024) scaled pairwise human preference evaluation. Li et al. (2024a) survey the area. Chen et al. (2024a) extended judging to multimodal models and found that MLLMs show human-like discernment in pairwise comparison while diverging from human preferences in scoring and batch ranking, which is the reason our queries are pairwise and ranking comparisons rather than absolute scores. LLMs as listwise and pairwise rankers (Sun et al., 2023; Qin et al., 2023) supply the related comparison formats for learning relevance. Liusie et al. (2023) compare relative assessment with absolute scoring, and Wang et al. (2023b) demonstrate that candidate order can change a judge's ranking. These findings motivate controlling presentation order.

Prometheus-Vision trains a rubric-conditioned multimodal response evaluator, and LLaVA-Critic learns to evaluate responses and provide preference feedback (Lee et al., 2024b; Xiong et al., 2024). These are trained evaluators, distinct from using a general-purpose model to judge tissue similarity. VL-RewardBench examines visual reward judgments (Li et al., 2024b); MuirBench tests joint multi-image understanding, including minimally changed unanswerable questions (Wang et al., 2024b). The study by Norman et al. (2026) further separates repeated-judgment consistency from bias and external agreement. Accordingly, we measure pairwise consistency between judges.

Image-quality work provides further comparison-based precedents: 2AFC prompting elicits pairwise quality judgments, while Compare2Score converts comparisons into quality scores (Zhu et al., 2024a;b). Silkie and RLAIF-V use AI feedback to improve multimodal response generation (Li et al., 2023; Yu et al., 2024). These studies concern image quality or answer quality, rather than acquisition-invariant tissue similarity. Triplet judgments from language and multimodal models can also recover interpretable object-concept structure (Du et al., 2024). Language-model comparisons recover aspects of human psychophysical similarity (Marjieh et al., 2023). With visual inputs, comparisons of rock images reveal stronger agreement on some elementary properties than on domain-specific ones (Dickson et al., 2025), and VLM pairwise ratings can be converted into

perceptual spaces through multidimensional scaling (Sanders et al., 2025). These findings motivate comparison-based supervision.

Multi-image position bias and MM-JudgeBias show that image order and irrelevant changes can alter model decisions (Tian et al., 2025; Lee et al., 2026). OddGridBench tests fine visual discrepancies, while PathView-Bench and PathBind examine fine-grained pathology perception and grounding (Weng et al., 2026; Chen et al., 2026b; Zhang et al., 2026a). Their findings motivate separating reproducibility of judge verdicts from the training utility of the resulting preferences.

In pathology specifically, Ferber et al. (2024) showed that GPT-4V with in-context examples approaches specialized models on cancer image classification, and PLIP (Huang et al., 2023) trained a pathology CLIP from social-media image-text pairs. Quilt-1M curates image–text pairs from educational pathology videos and other sources, whereas PathGen-1.6M uses multimodal models to generate pathology captions (Ikezogwo et al., 2023; Sun et al., 2024). PathChat instead trains a pathology-specific generative assistant with visual-language instructions (Lu et al., 2024). Direct histology evaluations reveal variation across models and tissues (Mavrych et al., 2026); a neurohistology study also shows sensitivity to textual context and improvements from labeled in-context examples for selected tau lesions (Ono et al., 2024). PathBench evaluates patch and slide understanding with controls for text-only shortcuts (Sun et al., 2025), while Weihrauch et al. (2026) show that patch size, magnification and joint processing affect WSI performance. Together, these studies characterize pathology understanding across tasks and input protocols and complement the evaluation of relative-similarity preferences. On the distillation side, our setting inherits from teacher-student transfer (Hinton et al., 2015; Tian et al., 2019), self-training with noisy pseudo-labels (Xie et al., 2019), and the substitution of AI feedback for human feedback in Constitutional AI and RLAIF (Bai et al., 2022; Lee et al., 2023). Our teacher provides relative-similarity preferences rather than labels or captions, and the student is a vision encoder rather than a language policy. LLM-derived supervision also trains text embeddings: Gecko uses generated and relabeled retrieval pairs (Lee et al., 2024a), and Wang et al. (2023a) train contrastive embedding models on synthetic text tasks. These studies provide precedents for transferring LLM knowledge into embedding models; we investigate its application to acquisition robustness in pathology. Learning encoders from language supervision (Radford et al., 2021) and the observation that large models converge toward shared representations across modalities (Huh et al., 2024) give one reading of why a general-purpose MLLM can judge histological similarity at all.

A.10.4 Learning from similarity judgments and rankings

ROPAL is closest to alignment with human similarity judgments. Representational similarity analysis (Kriegeskorte et al., 2008) and the THINGS odd-one-out datasets (Hebart et al., 2019; 2020) established the measurement, Roads & Love (2020) enriched ImageNet with human similarity judgments, and Peterson et al. (2017) and Attarian et al. (2020) transform deep features to predict human similarity. LPIPS studies perceptual similarity in deep features (Zhang et al., 2018); DreamSim adapts representations using human comparisons of synthetic images (Fu et al., 2023). gLocal fits a linear transform on frozen features to triplet judgments and improves downstream transfer (Muttenthaler et al., 2022a; 2023), Sundaram et al. (2024) fine-tune backbones on human perceptual triplets, Muttenthaler et al. (2025) show that alignment transfers across abstraction levels, and Sucholutsky et al. (2023) survey the field. We use MLLM judgments in place of human judgments, build the triplets across slides and institutions to target acquisition robustness, and fine-tune the backbone with low-rank adapters.

The paired-comparison model and its estimation are due to Bradley & Terry (1952) and Hunter (2003), and listwise extensions were given by Cao et al. (2007) and Xia et al. (2008). RankNet learns pairwise rankings with a logistic likelihood (Burges et al., 2005). Schultz & Joachims (2003) learn a metric from qualitative relative comparisons; neural image ranking extends this principle to fine-grained visual similarity (Wang et al., 2014), and Law et al. (2016) formulate relative distance constraints over up to four images. Related metric-learning objectives range from LMNN (Weinberger et al., 2005) and the triplet loss of FaceNet (Schroff et al., 2015) to the N-pair loss (Sohn, 2016), generalized non-metric multidimensional scaling (Agarwal et al., 2007) and stochastic triplet embedding (van der Maaten & Weinberger, 2012). Recent metric learning theory also characterizes what can be identified from limited pairwise preferences (Wang et al., 2024e). Our loss is a Bradley-Terry likelihood on distance differences, which behaves like a soft triplet loss with a temperature.

Preference-based training of large models (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023) optimizes the same objective on preference pairs. Wu et al. (2025) adapt DPO to a pathology foundation model for segmentation. We apply preference learning to the geometry of a vision encoder, with an LLM as the preference source and no downstream task in the loop.

Ranking also provides a preservation signal. DarkRank transfers cross-sample rankings, RKD transfers distance and angle relations, and similarity-preserving distillation retains pairwise activation structure (Chen et al., 2017; Park et al., 2019; Tung & Mori, 2019). Probabilistic knowledge transfer matches neighborhood distributions in feature space (Passalis & Tefas, 2018); metric distillation explicitly contrasts absolute embedding matching with matching pairwise distances (Yu et al., 2019). RankSim aligns neighbourhood rankings with label-space order in regression (Gong et al., 2022); FINEST regularizes recommendation rankings against perturbations (Oh et al., 2024).

A.10.5 Parameter-efficient fine-tuning that preserves the source model

LoRA injects low-rank updates into frozen weights (Hu et al., 2021). Kumar et al. (2022) showed that naive fine-tuning distorts pretrained features and underperforms out of distribution, which motivates linear probing before fine-tuning, and Lee et al. (2022) fine-tune selected layers under distribution shift. Weight-space interpolation between fine-tuned and zero-shot models preserves robustness (Wortsman et al., 2021; 2022), and Goyal et al. (2022) fine-tune with the pretraining objective. Explicit regularization toward the starting point is the other standard remedy, whether in weight space (Li et al., 2018b), in function space (Li & Hoiem, 2016) or through parameter importance (Kirkpatrick et al., 2016). LDIFS directly penalizes changes in feature space (Mukhoti et al., 2023), while anchor-based robust vision–language fine-tuning adds auxiliary semantic supervision (Han et al., 2024). LoRA has shown reduced forgetting in language models (Biderman et al., 2024). Last-layer retraining can reduce spurious correlations without adapting the backbone (Kirichenko et al., 2022), further motivating task-specific evaluation of the adaptation–preservation tradeoff. Backward Feature Projection permits coordinate changes by learning a projection before matching old and new features (Gu et al., 2023). TPGM learns layerwise bounds on parameter deviation (Tian et al., 2023), while ZSCL combines distillation on a diverse reference dataset with weight averaging to preserve CLIP transfer (Zheng et al., 2023b). In pathology, Enda et al. (2025) found that linear probing is sufficient for PFM transfer while fine-tuning often degrades performance, and Pan et al. (2025) use parameter-efficient fine-tuning of PFMs for gene-expression prediction. Ramchandani et al. (2025) evaluate LoRA on UNI and Virchow-family models for atypical mitosis classification, including group-based splits. Ammeling et al. (2025) also compare linear probing, supervised LoRA and end-to-end training for mitotic-figure classification across tumor domains.

A.10.6 Benchmarks and base models

Bareja et al. (2026) benchmark 32 vision and pathology foundation models with linear probing on TCGA, CPTAC and out-of-domain and external patch-level datasets. Campanella et al. (2024), Neidlinger et al. (2025) and Breen et al. (2024) benchmark PFMs as feature extractors on clinical tasks. eva (kaiko.ai et al., 2024), HEST-1k (Jaume et al., 2024) and Patho-Bench (Zhang et al., 2025) are the standard evaluation suites. Alfasly et al. (2024) evaluate foundation model embeddings for zero-shot whole-slide image retrieval, the use case behind our cross-institution content-based retrieval task. SMILY earlier demonstrated histology image search with pathologist assessment (Hegde et al., 2019), and Kather et al. (2019) is the multi-center study behind NCT-CRC-HE. Related pathology foundation models include UNI (Chen et al., 2024b), Virchow and Virchow2 (Vorontsov et al., 2024; Zimmermann et al., 2024), GenBio-PathFM (Kapse et al., 2026), Prov-GigaPath (Xu et al., 2024), CHIEF (Wang et al., 2024c), TITAN (Ding et al., 2025), Phikon-v2 (Filiot et al., 2024), Midnight (Karasikov et al., 2025), Hibou (Nechaev et al., 2024) and Atlas 2 (Alber et al., 2026), with DINOv2 (Oquab et al., 2023) as the general-purpose reference and ABMIL (Ilse et al., 2018) for slide-level aggregation.

A.11 Baseline configurations and execution checks

Table 37 gives the configuration of every baseline and of ROPAL, taken from the code and the saved design files of each run. AdvInv, LEACE, INLP and Pathology DoRA are fitted on the same unlabeled TCGA pool of 62,445 randomly extracted patches: 47,926 fitting images and a participant-

separated validation split of 14,519 images. Its 574 tissue-source-site codes include 533 in the fitting images. We used $\lambda = 0.3$ in Eq. 3 for all experiments.

Table 37: Configurations of the baselines and of ROPAL. Site: TCGA tissue-source site. Fitting and validation splits: the 47,926 and 14,519 images of the TCGA pool described above.

Method	Training data	Objective	Key hyperparameters	Schedule and budget	Model selection	Deviations from the original method	Execution check
Reinhard	None. Fixed LAB reference: means and standard deviations of 500 TCGA-LUSC patches, shared with the PathoROB baseline	Match each patch's LAB channel means and standard deviations to the reference	torchstain implementation	None	None	None	One reference file for every table and base
Macenko	None. One fixed TCGA-LUSC reference tile	Stain vectors from the SVD of optical densities, re-expressed with the reference stain matrix	torchstain defaults ($I_o = 240$, $\alpha = 1$, $\beta = 0.15$)	None	None	The patch is left unchanged when its stain estimate fails	Reference tile hash recorded with every extraction
AdvInv	Frozen features of the fitting split, site codes as targets	Feature reconstruction (MSE) minus w times the adversary's site cross-entropy. The adversary minimizes its cross-entropy	Encoder and decoder MLPs with 500 hidden units and dropout 0.1, latent width equal to the input width, adversary MLP with two layers of 100 units, $w = 0.01$, Adam at 10^{-3} for both the autoencoder and the adversary (separate optimizers), batch 128, standardization with fitting-split statistics	10 autoencoder and 8 adversary pretraining epochs (patience 5), then 50 alternating rounds of one encoder minibatch and one full adversary epoch	Pretraining stages by validation reconstruction and adversary cross-entropy. The final alternating round is kept	Adversarial weight 0.01 instead of 1.0. Each stage runs one quarter of the 40/30/200 epochs and rounds of the unchanged protocol, to limit compute time. Applied to frozen encoder features	Adversary site accuracy and weight comparison in Table 40
LEACE	Frozen features of the fitting split, one-hot site concept (533 sites)	Closed-form least-squares concept erasure	concept-erasure 0.2.4: affine eraser, covariance shrinkage, SVD tolerance 0.01, trace constraint	Closed form	None	None	Input and eraser hashes recorded with every result
INLP	Fitting split for the classifiers, validation split as the authors' development set	Iterative nullspace projection of linear site classifiers (authors' code, commit e1edcc1)	SGDClassifier with hinge loss, L2 penalty, $\alpha = 10^{-4}$, intercept, at most 1,000 epochs, tolerance 10^{-3}, one-vs-rest over 533 sites, so a round removes up to 533 directions	Rounds until the validation site accuracy of the next classifier reaches the majority-site rate (0.044), until its projection would remove every remaining dimension, or 10 rounds	The stopping rule	The authors run a fixed number of iterations. SGD with hinge loss (one of the authors' two classifiers) instead of LinearSVC, for run time. An eigendecomposition fallback of the authors' projection formula when its SVD does not converge	Rounds, retained rank and site accuracy in Table 38. A refit of round 1 reproduces the recorded validation accuracy exactly for UNI2-h and Virchow2

Pathology DoRA	Images of the fitting split	DINO v1 self-distillation with 65,536 prototypes, student temperature 0.1, teacher temperature 0.04, centre momentum 0.9	DoRA rank 16, $\alpha = 16$, dropout 0.1 on the qkv projection of every block. AdamW, peak learning rate 0.002 (0.0005 times the batch size, as in the authors' code), minimum 10^{-6}, weight decay 0.04 to 0.4, gradient clipping 3.0. Two global crops (scale 0.32 to 1) and eight local crops (scale 0.05 to 0.32) with flips, colour jitter (0.7, 0.7, 0.7, hue 0.5, $p = 0.9$), greyscale ($p = 0.3$) and Gaussian blur ($p = 0.6$)	1,100 steps at batch 4 (4,400 images, 0.09 of one pass), 110 warm-up steps, cosine decay, teacher momentum 0.996 to 1, last head layer frozen for 11 steps, fp32	Teacher initialized as a copy of the student and updated as an exponential moving average. Features from the teacher backbone at the final step, DINO head discarded	Step budget instead of 100 epochs with 10 warm-up epochs, with all schedules rescaled. Tiles resized to the encoder input before augmentation, and views rescaled to $[0, 1]$ before normalization with the encoder's statistics. Virchow2 DINO-head input is the CLS token (1,280-d) instead of CLS concatenated with the mean patch token (2,560-d) in the authors' configuration. The authors give no configuration for GenBio-PathFM, which uses their generic path	Loss trajectory and feature spread in Table 41
ROPAL	60,985 LLM-judged comparisons from 26 TCGA cohorts, within-slide tiles for the rank term	$\mathcal{L}_{\mathrm{BT}} + \lambda\,\mathcal{L}_{\mathrm{rank}}$ (Section 3.2)	$\lambda = 0.3$, LoRA rank 16, scaling 32, dropout 0.05 in the attention and feed-forward projections of the last four blocks, AdamW at 10^{-4}, batches of 12 comparisons, mixed precision	One epoch	Final checkpoint	Not applicable	Same settings for every base

INLP stopping rule. The authors of INLP run a fixed number of iterations for each experiment. Our encoders have 768 to 4,608 dimensions and each round removes up to 533 directions, so we stop on the validation split instead: when the site accuracy of a newly trained classifier falls to the majority-site rate, when its projection would remove every remaining dimension, or after 10 rounds. Table 38 gives the result for every base. GenBio-PathFM and the PathoROB view of UNI2-h stop at the majority-site rate. The other fits stop because one more round would remove every remaining dimension, with site accuracy still above the majority-site rate. For UNI2-h and Virchow2 we also evaluated the projection after the first round (Table 39). It changes the MOSAIC cross-domain Average of the frozen features by less than 0.02, while the final projection lowers it by 0.04 (UNI2-h) and 0.10 (Virchow2).

Table 38: INLP fits. Rounds: accepted projections. Retained rank: rank of the composed projection. Site accuracy: accuracy on the validation split of a site classifier trained on the final projected features (majority-site rate 0.044). General view: the tile embedding used in every table except PathoROB. PathoROB view: the CLS-plus-mean-patch embedding of the PathoROB leaderboard.

View	Base	Input dimension	Rounds	Retained rank	Site accuracy	Stop
General	UNI2-h	1,536	2	470	0.212	Rank exhausted
General	Virchow2	1,280	2	214	0.117	Rank exhausted
General	GenBio-PathFM	4,608	8	439	0.027	Majority rate
PathoROB	UNI2-h	3,072	5	407	0.040	Majority rate
PathoROB	Virchow2	2,560	4	428	0.053	Rank exhausted
PathoROB	GenBio-PathFM	4,608	8	439	0.027	Majority rate

AdvInv adversarial weight. The adversary learns site information from the frozen features. After pretraining, its cross-entropy on the validation split is 3.1 nats for UNI2-h and Virchow2 and 3.2 to 4.2 nats for the other bases, against 5.41 nats for the entropy of the fitting-split site distribution and $\ln 574 = 6.35$ nats for a uniform guess. After the alternating rounds, the final adversary still predicts the site of 49% (UNI2-h) and 46% (Virchow2) of validation images, against a majority-site rate of

Table 39: MOSAIC cross-domain Average (as in Table 8) of INLP after 0, 1 and all accepted rounds. Zero rounds is the frozen feature set on which INLP is fitted. For UNI2-h it differs from the None entry of Table 8 (0.712), which uses a separate extraction of the frozen encoder. The final projection has two rounds for both bases and is the INLP entry of Table 8.

Base	0 rounds	1 round	Final
UNI2-h	0.709	0.703	0.665
Virchow2	0.751	0.763	0.653

4.4%, so the adapted features retain site information. At the weight 1.0 of the unchanged protocol, the latent space collapsed: the participation-ratio effective rank of the HER2ST embeddings was 1.5 to 2.7 (Table 40). We did not find an established weight for frozen pathology-encoder features, so we screened UNI2-h under the full schedule at the weights 0.1, 0.05, 0.03 and 0.01. The effective rank of the latent embeddings of 4,000 validation images fell as the weight grew: 104.5, 53.0, 14.2 and 3.7 for 0.01, 0.03, 0.05 and 0.1. We use 0.01, the tested weight with the highest rank. The screen also recorded the Camelyon robustness index of PathoROB, which was lower at 0.01 (0.833) than at 0.1 (0.920). The weight 1.0 fits give a higher PathoROB robustness index but lower MOSAIC accuracy and nearly one-dimensional embeddings, so the chosen weight does not maximize the robustness index. Table 40 compares the four existing AdvInv campaigns on the tasks they share. No new fits were run for it.

Table 40: AdvInv at two adversarial weights and two schedule lengths (full: 40 autoencoder epochs, 30 adversary epochs, 200 alternating rounds. Quarter: 10, 8 and 50). RI: PathoROB robustness index, mean of Camelyon and Tolkach ESCA. MOSAIC: cross-domain Average as in Table 8. Rank: participation-ratio effective rank of the HER2ST embeddings. Site accuracy: final adversary on the validation split (majority-site rate 0.044). The weight 0.01, quarter schedule rows are the AdvInv entries of the main tables. †PathoROB adapter fitted on a separate 45,505-image participant-held-out pool. –: not fitted for that feature view.

Weight	Schedule	Base	RI	MOSAIC	Rank	Site accuracy
1.0	Full	UNI2-h	0.940†	0.655	2.3	0.153
1.0	Full	Virchow2	0.833†	0.688	1.5	0.148
1.0	Quarter	UNI2-h	0.945	0.644	2.7	0.256
1.0	Quarter	Virchow2	0.903	0.700	2.7	0.225
0.01	Full	UNI2-h	0.893	–	–	–
0.01	Full	Virchow2	0.928	–	–	–
0.01	Quarter	UNI2-h	0.862	0.709	35.2	0.488
0.01	Quarter	Virchow2	0.914	0.768	19.6	0.461

Pathology DoRA diagnostics. With the authors' configuration, DINO training collapsed for Virchow2 and GenBio-PathFM. For both bases the loss reached 44.36 by about step 300 and stayed there until step 1,100. This is the value of the uniform solution over the 65,536 prototypes in the authors' loss ($4 \ln 65{,}536$). The teacher features show the same collapse: Virchow2 features are nearly one-dimensional around their mean, and GenBio-PathFM features point in nearly the same direction (Table 41). UNI2-h did not collapse. We report the collapsed runs as results, since they come from the authors' configuration transferred to these encoders without retuning, and read them as a compatibility limitation of that configuration for Virchow2 and GenBio-PathFM. Each base was trained once, so we do not report run-to-run variability, and we did not test whether other learning rates or head inputs avoid the collapse.

A.12 Further material

A.12.1 Evaluation protocols

The downstream protocols use the patch labels and acquisition metadata with cosine similarity on L2-normalized features. Cross-institution retrieval is scored by mean average precision (Alfasly

Table 41: Pathology DoRA training loss and teacher feature spread on the first 4,000 HER2ST tiles of MOSAIC. Loss: DINO loss at steps 0, 300 and 1,099 (uniform solution: 44.36). Rank: participation-ratio effective rank. PC1: fraction of variance on the first principal component. Cosine: mean pairwise cosine similarity. Frozen: the base encoder without adaptation.

Base	Loss (0)	Loss (300)	Loss (1,099)	Rank (frozen)	Rank (DoRA)	PC1 (DoRA)	Cosine (frozen)	Cosine (DoRA)
UNI2-h	43.53	18.59	13.90	33.1	30.4	0.122	0.391	0.354
Virchow2	42.98	44.38	44.36	18.7	1.0	0.988	0.504	0.299
GenBio-PathFM	43.14	44.39	44.36	20.6	4.8	0.436	0.499	0.994

et al., 2024), and fixed-candidate nearest-neighbour classification uses four candidates per query. Section 4.4.1 specifies the class–institution-confounded conditions used in the table.

A.12.2 Map of the released material

The released material includes the tile corpus, cached judgments, verbatim judge prompt, provenance records, training configurations, checkpoints and evaluation code. Judgment records identify the displayed tile order and raw response. Evaluation records specify the cohort, split, model and metrics needed to reproduce each reported result.