HiLO: Hierarchical Label-Aware Optimal Transport for Cross-Domain Alignment of Pathology Foundation Model Embeddings

Anonymous authors
Paper under double-blind review

Abstract

Pathology foundation models produce strong patch embeddings for H&E whole-slide images (WSIs). However, their embeddings remain highly predictive of non-biological factors such as source centre, scanner, and staining protocol. Consequently, downstream models may fail to rely on biologically meaningful features, which limits their clinical applicability. Optimal transport (OT) offers a principled way to align feature distributions across domains. However, standard OT algorithms exhibit quadratic memory and time complexity in the number of points, making them highly computationally expensive for WSIs. We propose HiLO, a hierarchical label-aware optimal transport framework for cross-domain feature alignment that can scale efficiently to large datasets. Building on hierarchical refinement, we compute globally optimal one-to-one correspondences between foundation-model patch embeddings from different source domains in linear memory. We embed label information into the transport cost by augmenting features with a label embedding. This preserves the low-rank cost factorization required for scalability and suppresses cross-class matches under label shift. The resulting one-to-one correspondences define a Monge map that transports patch embeddings from each source domain to the target domain, on which any downstream model can be trained. We evaluate our method on two publicly available pathology WSI datasets AGGC2022 and BEETLE. All experiments follow a leave-one-domain-out protocol, where each held-out domain is a scanner or centre depending on the dataset, and we report both in-domain (ID) and out-of-domain (OOD) performance. Compared with state-of-the-art OT, stain normalisation, and domain generalisation methods, HiLO achieves higher performance, while also attaining lower total transport cost than the OT baselines. In conclusion, HiLO is a scalable and flexible framework for cross-domain feature alignment that is applicable to large-scale pathology WSIs and can be extended to a wide range of downstream applications.

1 Introduction

Whole-slide images (WSIs) exhibit substantial non-biological variation across medical centres, caused by differences in slide preparation (e.g. different staining protocols) and image acquisition (e.g. different scanners). Pathology foundation models (FMs) are pre-trained on large and diverse WSI datasets to produce generalised patch embeddings, which has the potential to transfer across a range of downstream tasks. However, such embeddings encode all signals present in the training data, capturing non-biological variation alongside the biological signal of interest. In practice, non-biological variation can be more prominent than the biological signal, causing a lightweight classification layer (such as an MLP) to exploit it as a shortcut (Chen et al., 2024; Vorontsov et al., 2024). Indeed, recent benchmarking studies have systematically investigated this phenomenon, showing that patch embeddings from all major FM retain substantial non-biological (confounding) signals (de Jong et al., 2025; Kömen et al., 2026)(Fig. 1).

Existing approaches address this variation at one of two main stages: before feature extraction, by standardising the appearance of the input images (e.g. stain normalisation), or after feature extraction by modifying how the downstream models are trained on the extracted embeddings (e.g. DG

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.

methods). Stain normalisation is applied before encoding and normalises stain colour against one or more reference images. Stain normalisation attempts to standardise the colour appearance of WSIs, either by matching to a target distribution estimated from one or more images (Reinhard et al., 2001; Macenko et al., 2009; Vahadane et al., 2016) or by learning a mapping between domains (Shaban et al., 2019). These methods are designed to remove colour differences between domains, but the transformation is not guaranteed to preserve biological signals and may introduce spurious patterns that do not correspond to natural biological structures. Another approach is to address the generalisability of downstream model using domain generalisation (DG) methods. These methods define domains as different sources causing non-biological variation (such as medical centres and scanner) and train across several source domains so that performance holds on an unseen one, typically by aligning feature distributions, enforcing invariance, or minimizing worst-domain risk. In computational pathology, applying these methods to a downstream model (trained using FM embeddings) yields only small and inconsistent gains (refer Section 4.3 and Table 2). A key limitation of DG objectives is that they are defined over domain distributions as a whole and offer no guarantee on alignment of individual embeddings.

Optimal transport (OT) operates at individual embedding-level, computing the coupling between embeddings of two domains at minimum total transport cost (Peyré & Cuturi, 2019). This coupling is constrained to be one-to-one, where each embedding is assigned to exactly one embedding in the other domain, referred to as a match. Computing the exact OT coupling using linear solvers scales cubically with the number of samples, making it computationally expensive for large datasets (Peyré & Cuturi, 2019). Although entropic regularisation (Cuturi, 2013) reduces this to a sequence of matrix-scaling iterations (with quadratic cost), the resulting coupling is dense and distributes mass across many pairs rather than assigning individual matches. HiRef (Halmos et al., 2025) addresses this limitation by recursively solving low-rank OT subproblems, each of which co-clusters embeddings from two domains into progressively finer clusters until every embedding is paired. This yields a one-to-one coupling in log-linear time and linear memory. However, HiRef is label-agnostic and can produce cross-class matches that pair patch embeddings belonging to different classes together. In WSI analysis, this is a particular concern as distinct pathologies often share similar morphological patterns and therefore lie close in embedding space, making them likely candidates for such cross-class matches. Aligning such matches pulls embeddings of different classes closer and thus reduces class separability.

To bridge the gap between scalability and class supervision, we propose HiLO, a hierarchical label-aware optimal transport framework for the cross-domain alignment of patch embeddings. HiLO augments each patch embedding with its label such that the squared Euclidean distance between augmented embeddings adds a penalty for every cross-class match, thereby discouraging them. Since the augmented transport cost used by HiLO is squared Euclidean distance, the cost matrix can be factorised into low-rank components, enabling the computation of label-aware one-to-one matches at negligible additional cost. Moreover, since HiLO operates in the embedding space, it is agnostic to the choice of backbone encoder and the downstream task, allowing it to be combined with other domain generalisation methods.

We evaluate HiLO on patch embeddings from 10 state-of-the-art (SOTA) FMs under a leave-one-domain-out (LODO) protocol. We evaluate HiLO on two publicly available H&E-stained WSI datasets acquired using multiple scanners, AGGC2022 (AGGC22 Organisers, 2022) (prostate cancer) and BEETLE (Lems et al., 2025) (breast cancer). In AGGC2022, each tissue slide is scanned on 6 scanners, resulting in multiple digitised versions of same slide. On AGGC2022, HiLO achieves the highest out-of-domain (OOD) macro-F1 among stain normalisation and OT methods on [K1] of the 10 FMs (Table 1) and outperforms the strongest DG method on 6 of them. Furthermore, combining HiLO with existing DG methods leads to additional improvements of OOD macro-F1. On BEETLE, HiLO also achieves high OOD macro-F1 among stain normalisation and OT methods (Table 2). Our comprehensive experiments and ablation studies across 10 FMs and 2 datasets show that HiLO consistently outperforms stain normalisation and existing OT methods, matches or exceeds DG methods, and further improves their OOD performance when combined.

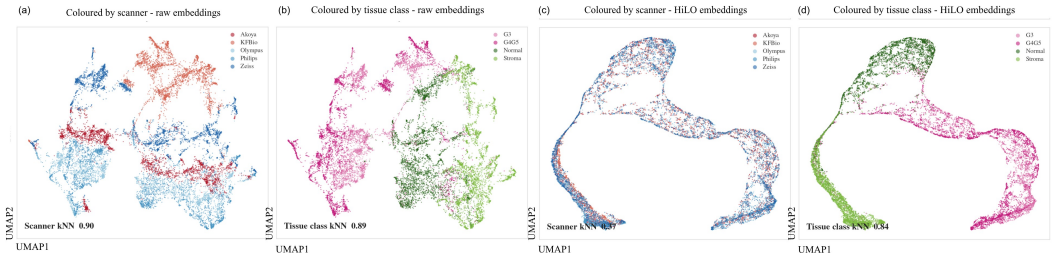


Figure 1: UMAP of Virchow patch embeddings on AGGC2022, where the same specimens are digitized on six scanners, coloured by scanner and tissue class. Raw patch embeddings show better clustering by scanner (a), whereas patch embeddings aligned by HiLO mix across scanners (c). Tissue-class structure is largely preserved by the alignment (b, d), and in the aligned embeddings the tissue classes form more contiguous regions (d). The accuracy of a k-nearest-neighbour classifier quantifies this effect. Scanner accuracy drops from 0.90 to 0.37, whereas tissue-class accuracy decreases only from 0.89 to 0.84. KFBio scanner is held out and 14,000 patches are plotted above.

2 Related Work

Limitations of pathology FMs. de Jong et al. (2025) investigated the embedding space of 10 SOTA FMs and identified that the embeddings are organised more strongly by medical centre than by biological class. To quantify this effect, they introduced a new metric, robustness index (RI), which analyses the k nearest neighbours of each patch embedding and computes the ratio between the proportion of neighbouring patches belonging to the same biological class and the proportion that share the same medical centre. Notably, only Virchow2 Zimmermann et al. (2024) was able to attain an RI marginally above one, indicating that it was the only FM whose embeddings are organised by biological class than medical centre. This work was later extended by the same authors in the PathoROB bechnmark study Kömen et al. (2026) which analysed 20 FMs using a large scale WSI dataset spanning 28 biological classes and 34 medical centres. They concluded that all evaluated FMs encode centre-specific signatures that can propagate into downstream models, and if biological labels are correlated with medical centres (during training), serious misclassifications can occur such as tumour tissue being predicted as normal. Akebli & Della Mea (2026) benchmarked 10 FMs on 2 multi-centre, multi-scanner cohorts and reported similar observations with consistent performance degradation on unseen scanners and centres. Collectively, these findings establish that sensitivity to non-biological variation is a critical and unresolved limitation of current pathology FMs, with direct consequences for their reliability in clinical deployment.

Stain normalisation and FM fine-tuning. Classical stain normalisation methods differ in the colour representation space it operates, namely in perceptual colour space or stain vector space. Reinhard normalisation converts images to the $l\alpha\beta$ colour space and matches the mean and standard deviation of each channel to those of the reference(Reinhard et al., 2001). Modified Reinhard normalisation improves luminance mapping by preserving the mean luminance of the source image and preventing loss of contrast through revised statistical matching (Roy et al., 2021). Macenko normalisation estimates stain vectors from optical density using singular value decomposition (Macenko et al., 2009) and Vahadane normalisation estimates them through sparse non-negative matrix factorization with a sparsity constraint (Vahadane et al., 2016). ScanGen trains a lightweight projection head on the embeddings of a frozen encoder with a contrastive objective that aligns scans of same specimen (Carloni et al., 2025). GLMP (Zhang et al., 2026) prompts a multimodal language model to describe clusters of image patches in biological terms and replaces the patch embeddings with combination of text embeddings of these descriptions.

Domain generalisation. Empirical risk minimization (ERM) (serves as baseline for our DG experiments), pools all source domains and trains a model that minimizes the average training loss over all samples without distinguishing between domains. All DG methods extend ERM using domain labels to discourage the networks' reliance on domain-specific signals. *Feature alignment-based methods* attempt to align the feature distributions of the source domains. CORAL matches the second-order statistics of source domains (Sun & Saenko, 2016), whereas DANN trains a feature extractor adversarially against a classifier that predicts the source domain (Ganin et al., 2016). *Invariance-based*

methods constraint a predictor to perform consistently across source domains. IRM learns an classifier that is optimal for every source domain (Arjovsky et al., 2019), and VREx penalizes the variance of the training risks across source domains (Krueger et al., 2021). GroupDRO minimizes the risk of the worst-performing source domain (Sagawa et al., 2020) and Fish encourages the model generalise across all source domains by maximizing the inner product between their gradients (Shi et al., 2022). Finally, Mixup acts on the data by training on convex combinations of samples and labels drawn from different source domains (with sample pairs formed at random) (Zhang et al., 2018).

Optimal transport. OT computes a coupling between two domains that minimizes the total transport cost, where a coupling is a joint distribution over pairs of patch embeddings from the two domains (Peyré & Cuturi, 2019). Courty et al. (2017) applied OT for unsupervised domain adaptation and incorporated the class labels of the source domain through a group-lasso regularizer. OT has since been applied to domain adaptation in histopathology (Falahkheirkhah et al., 2023), but only at modest scale, because exact OT solvers scale at least cubically with the number of samples (Kuhn, 1955; Peyré & Cuturi, 2019), making it computationally expensive to apply to entire WSI datasets. Entropy regularizers are added to the transport objective to make it strictly convex and efficiently solvable using Sinkhorn iterations (Cuturi, 2013), but it produces a dense coupling whose memory grows quadratically with the number of samples. Low-rank OT (Scetbon et al., 2021) reduces the memory of the coupling to linear in the number of samples by constraining the rank, provided that the cost matrix admits a low-rank factorization. FRLC (Halmos et al., 2024) reparameterizes the low-rank coupling through a latent coupling between its two factors, which decomposes the optimization into three OT subproblems while retaining linear memory. However, neither low-rank OT nor FRLC yields a one-to-one correspondence, since a low-rank coupling groups samples into a small number of shared clusters rather than pairing them individually. HiRef (Halmos et al., 2025) computes one-to-one coupling in linear memory by recursively refining the co-clusters of low-rank OT solutions until each contains a single matched pair. HiLO formulates cross-domain matching as a label-aware transport problem by encoding class labels directly in the transport cost so that the resulting couplings favour same-class matches, and solves it at scale using hierarchical refinement of low-rank couplings. The augmented cost defined in HiLO remains a squared Euclidean distance, allowing it to compute label-aware one-to-one couplings over entire clusters of patch embeddings at no additional cost.

3 METHOD

Figure 2 summarizes HiLO, which consists of three stages. For each pair of source domains, HiLO first draws two pools of patch embeddings with equal class counts from a m embedding space (Figure 2a). Then it computes a label-aware one-to-one coupling between the pools using hierarchical refinement (Figure 2b). Finally, the resulting matches are used to train an alignment network jointly with the downstream model (Figure 2c). Section 3.1 defines the problem and training objective. Sections 3.3 describe the construction and computation of the couplings and Section **??** describes training and inference.

3.1 PROBLEM SETTING AND OBJECTIVE

A frozen FM ϕ produces a patch embedding $\mathbf{z} = \phi(x) \in \mathbb{R}^d$ for each image patch x, and $\bar{\mathbf{z}} = \mathbf{z}/\|\mathbf{z}\|_2$ denotes the patch embedding normalised to unit length. Each patch embedding belongs to a class $y \in \{1, \dots, C\}$. Training uses labelled patch embeddings from a set of source domains $\mathcal{S}$, and evaluation uses patch embeddings from a held-out domain whose labels are unavailable during training.

HiLO learns an alignment $f_\theta : \mathbb{R}^d \to \mathbb{R}^{d'}$, which takes a normalised patch embedding $\bar{\mathbf{z}}$ and returns the aligned patch embedding $f_\theta(\bar{\mathbf{z}})$, together with a downstream model g_ψ that predicts the class from the aligned patch embedding. The alignment is supervised by a set $\mathcal{M}$ of matches, each of which pairs two patch embeddings from different source domains. HiLO minimizes

$$\mathcal{L}(\theta, \psi) \;=\; \mathcal{L}_{\text{task}}(\theta, \psi) \;+\; \lambda_c \, \mathcal{L}_{\text{align}}(\theta), \qquad \mathcal{L}_{\text{align}}(\theta) \;=\; \frac{1}{|\mathcal{B}|} \sum_{(\bar{\mathbf{z}}, \bar{\mathbf{z}}') \in \mathcal{B}} \left\| f_\theta(\bar{\mathbf{z}}) - f_\theta(\bar{\mathbf{z}}') \right\|_2^2, \quad (1)$$

where $\mathcal{L}_{\text{task}}$ is the cross-entropy on labelled patch embeddings from the source domains, $\mathcal{B}$ is a mini-batch of matches drawn from $\mathcal{M}$, and $\lambda_c \geq 0$ is the alignment weight. Minimizing $\mathcal{L}_{\text{align}}$ removes

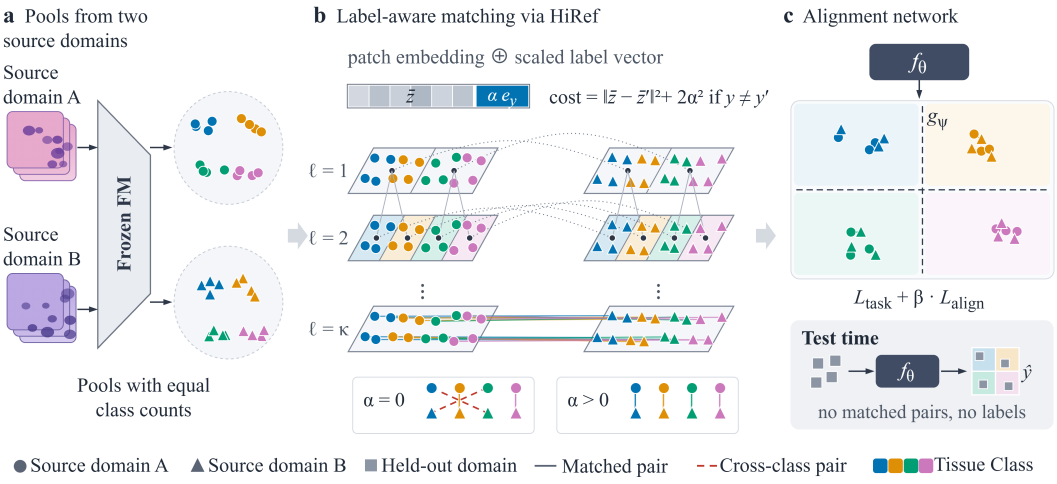


Figure 2: Overview of HiLO. (a) For each pair of source domains,HiLO draws two pools of embeddings with equal class counts. (b) HiLO appends a one-hot label vector, multiplied by the label weight α, to each patch embedding, and computes a matching between the two pools by recursively refining co-clusters. With $\alpha = 0$, the matching contains cross-class pairs (red), whereas with $\alpha > 0$ the matched pairs connect patch embeddings of the same class. (c) The matched pairs supervise the alignment network f_θ, which is trained jointly with the downstream model g_ψ by minimizing $\mathcal{L}_{\text{task}} + \beta\,\mathcal{L}_{\text{align}}$. At test time, patch embeddings from the held-out domain pass through f_θ and g_ψ without matched pairs or labels.

the differences between the two patch embeddings of each match. These differences should reflect the domain rather than the class, so each match should pair patch embeddings of the same class that are otherwise as similar as possible. Section 3.2 constructs such matches with label-aware OT, and Section 3.3 computes them at scale with HiRef.

3.2 Label-aware one-to-one correspondences

For every pair of source domains $\{A, B\} \subset \mathcal{S}$, HiLO draws a pool of N patch embeddings from each of the two source domains, uniformly at random and such that both pools contain the same number of patch embeddings of each class. A one-to-one correspondence between the two pools is a permutation σ of $\{1, \ldots, N\}$ that matches $\mathbf{z}_i^A$ with $\mathbf{z}_{\sigma(i)}^B$. To make the one-to-one correspondence label-aware, HiLO augments each patch embedding with a scaled label embedding,

$$\tilde{\mathbf{z}} \;=\; \big[\,\bar{\mathbf{z}}\;;\;\lambda_y \mathbf{e}_y\,\big] \in \mathbb{R}^{d+C}, \tag{2}$$

where $\mathbf{e}_y$ is the one-hot vector of class y and $\lambda_y \geq 0$ is the scale of the label embedding. The transport cost between two augmented patch embeddings is their squared Euclidean distance,

$$c(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}') \;=\; \|\bar{\mathbf{z}} - \bar{\mathbf{z}}'\|_2^2 \;+\; 2\lambda_y^2\, \mathbb{I}[\,y \neq y'\,]. \tag{3}$$

The first term is the distance between the normalised patch embeddings and lies in $[0, 4]$, and the second term adds the constant $2\lambda_y^2$ to every cross-class match. For each pair of source domains, HiLO computes the one-to-one correspondence that minimizes the total transport cost,

$$\sigma_{AB} \;=\; \arg\min_{\sigma} \sum_{i=1}^{N} c\big(\tilde{\mathbf{z}}_i^A, \tilde{\mathbf{z}}_{\sigma(i)}^B\big). \tag{4}$$

Setting $\lambda_y = 0$ recovers label-agnostic OT, and increasing λ_y penalizes cross-class matches. Because both pools contain the same number of patch embeddings of each class, a sufficiently large λ_y eliminates cross-class matches.

Proposition 1. *If $\lambda_y^2 > 2$, every minimizer of Equation (4) contains no cross-class match.*

The proof is given in Appendix A. The matches of all pairs of source domains form the set

$$\mathcal{M} \;=\; \bigcup_{\{A,B\} \subset \mathcal{S}} \big\{\big(\bar{\mathbf{z}}_i^A, \bar{\mathbf{z}}_{\sigma_{AB}(i)}^B\big)\big\}_{i=1}^{N}. \tag{5}$$

3.3 SCALABLE COMPUTATION WITH HIREF

Solving Equation (4) with an exact assignment solver requires time cubic in N. HiLO instead uses HiRef (Halmos et al., 2025), which requires the cost matrix to admit a low-rank factorization. Every augmented patch embedding has squared norm $1 + \lambda_y^2$, so the cost matrix between two pools factorizes as

$$\mathbf{C} \;=\; 2(1+\lambda_y^2)\,\mathbf{1}\mathbf{1}^\top \;-\; 2\,\tilde{\mathbf{Z}}^A(\tilde{\mathbf{Z}}^B)^\top, \tag{6}$$

where the rows of $\tilde{\mathbf{Z}}^A, \tilde{\mathbf{Z}}^B \in \mathbb{R}^{N\times(d+C)}$ are the augmented patch embeddings of the two pools, and the rank of $\mathbf{C}$ is at most $d + C + 1$. HiRef solves a low-rank OT problem on this factorized cost matrix to split the two pools into co-clusters, and it repeats this step within each co-cluster until every co-cluster contains a single match. Under the conditions stated by Halmos et al. (2025), this recursion recovers the minimizer of Equation (4) in $\mathcal{O}(N \log N)$ time and $\mathcal{O}(N)$ memory. The label embedding therefore makes the one-to-one correspondence label-aware without losing the factorization on which HiRef depends. HiLO computes $\mathcal{M}$ once before training.

3.4 TRAINING AND INFERENCE

HiLO trains f_θ and g_ψ jointly by minimizing Equation (1) over mini-batches of labelled patch embeddings and matches. The task loss prevents f_θ from collapsing all patch embeddings to a constant. When HiLO is combined with a DG method, $\mathcal{L}_{\text{task}}$ becomes the objective of that DG method, which retains the cross-entropy and either adds a term to it or reweights the cross-entropies of the source domains, and $\mathcal{L}_{\text{align}}$ is added unchanged. At test time, the prediction for a patch embedding $\mathbf{z}$ from the held-out domain is $g_\psi(f_\theta(\bar{\mathbf{z}}))$. The alignment requires neither matches nor labels from the held-out domain, and the FM is never updated.

4 EXPERIMENTS

We first compare HiLO with methods that operate on individual patches or patch embeddings, namely stain normalisation and other OT methods (Section 4.2). Then we compare HiLO with SOTA DG methods which are trained with objectives defined over entire source domains (Section 4.3). Finally, we combine HiLO with DG methods to evaluate whether the two approaches are complimentary (Section 4.4). In summary, our experiments are designed to comprehensively evaluate HiLO against all major categories of existing SOTA methods for mitigating non-biological variation in FM patch embeddings.

4.1 EXPERIMENTAL SETUP

We conduct experiments on three pathology WSI datasets. AGGC2022 (AGGC22 Organisers, 2022) serves as the main dataset and is used in all comparisons. It contains prostatectomy and biopsy specimens from a single hospital, each digitized on six scanners (Akoya, KFBio, Leica, Olympus, Philips, and Zeiss). Patches are labelled as Gleason pattern 3 (G3), Gleason pattern 4 or 5 (G4∪G5), Normal, or Stroma, and the specimens are split into 20 training, 6 validation, and 12 testing. BEETLE (Lems et al., 2025) contains breast biopsies and resections annotated with four tissue classes. We use its development set and treat each scanner as a domain. Of the seven scanners in the development set, we retain the five that contain all four classes. These five scanners come from four sources, and their numbers of patches range from 3,520 to 56,548 (Appendix B). Camelyon17 (Bandi et al., 2019) contains lymph node sections from five centres, each considered as an independant domain. Since Camelyon17 contains only two annotation classes (tumour vs normal), all methods exceed an OOD macro-F1 of 0.992 leaving little room for improvement. Therefore, we only use it to visualize the effect of alignment by comparing the UMAP projections of raw and HiLO-aligned patch embeddings (see Appendix). Patch sizes are FM embedding sizes are listed in Appendix Table 7.

Patch embeddings are extracted using 10 publicly available SOTA FMs, namely UNI (Chen et al., 2024), UNI2 (Chen et al., 2024), PLIP (Huang et al., 2023), H-optimus-1 (Bioptimus, 2025), Virchow (Vorontsov et al., 2024), Virchow2 (Zimmermann et al., 2024), CONCH v1.5 (Lu et al., 2024), Prov-GigaPath (Xu et al., 2024), Midnight (Karasikov et al., 2025), and MUSK (Xiang et al., 2025). The alignment network f_θ is a two-layer perceptron with batch normalisation, a hidden width of 512, and an output dimension of $d' = 128$, and the downstream model g_ψ is a linear classifier. Within

each experimental setting, all methods share the same architecture, optimizer (Adam), number of training epochs (30), mini-batch size (1024 patch embeddings per source domain), and differ only in the training objective or input preprocessing (for stain normalisation).

HiLO uses $\lambda_y = 0.5$, $m_{\max} = 8192$, and mini-batches of $|\mathcal{B}| = 4096$ matches. Within each pair of source domains, the number of patch embeddings of class c drawn from each source domain is $2^{\lfloor \log_2 \min(n_c^A, n_c^B, m_{\max}) \rfloor}$, where n_c^A and n_c^B are the numbers of available patch embeddings of class c, and rounding to a power of two ensures that N admits the rank schedule of HiRef. For each FM and each pair of source domains, the pools are drawn once with a fixed random seed. On AGGC2022, each fold yields between 99k and 197k matches depending on the FM, and HiRef computes the one-to-one correspondence of a pair with $N = 32{,}768$ in about 68 seconds on a single A100.

All experiments follow a leave-one-domain-out (LODO) protocol. Each fold holds out one domain for OOD evaluation, reserves a second domain for validation, and uses the remaining domains as source domains. On AGGC2022, the validation set contains only the validation domain and the validation specimens, which are excluded from training before pools are drawn, so that validation is disjoint from training in both domain and specimen. OOD performance is measured on the held-out domain and ID performance (See Appendix Tables 11 and 10) on the source domains, both restricted to the test specimens. Macro-F1 serves as both the selection and the reporting metric. For fair comparison Gulrajani & Lopez-Paz (2021), each method tunes one hyperparameter over five values and ERM tunes weight decay over the same number of values. Within each combination of fold and seed, the value with the highest validation macro-F1 is selected, and results are averaged over all such combinations.

4.2 Comparison with stain normalisation and OT methods - AGGC2022 dataset

Table 1 shows that HiLO improves OOD macro-F1 over ERM on all 10 FMs, by 0.036 on average, improving by 0.041 on UNI, 0.037 on Virchow, 0.042 on CONCH v1.5, and 0.048 on Prov-GigaPath. HiLO achieves the highest OOD macro-F1 among stain normalisation and OT methods on 6 of the 10 FMs (UNI, UNI2, Virchow, Virchow2, CONCH v1.5, and Prov-GigaPath). On CONCH v1.5, HiLO reaches 0.817, which is the best performance ahead of low-rank OT at 0.812. On the remaining 4 FMs, HiLO is outperformed by low-rank OT on H-optimus-1 (0.794 vs. 0.789), mini-batch Sinkhorn on Midnight (0.764 vs. 0.755), FRLC on MUSK (0.658 vs. 0.649), and Vahadane on PLIP (0.620 vs. 0.577).

All OT methods in Table 1 train the same alignment network with the same alignment loss and differ only in how the matches are computed. Mini-batch Sinkhorn falls below ERM on 5 of 10 FMs. Low-rank OT and FRLC exceed ERM on 9 of 10 and 8 of 10 FMs, respectively.

Stain normalisation also improves OOD macro-F1 over ERM on most FMs, with the largest gains on PLIP, from 0.542 to 0.620, and on CONCH v1.5, from 0.775 to 0.807. HiLO outperforms the best stain normalisation method on 9 of the 10 FMs, the exception being PLIP, where Vahadane reaches 0.620 compared with 0.577 for HiLO.

The label embedding contributes consistently to the performance of HiLO. On 8 of 10 FMs, HiLO outperforms its label-agnostic counterpart (with $\lambda_y = 0$) by between 0.003 and 0.018, with the largest improvement seen on Virchow, where OOD macro-F1 rises from 0.722 to 0.740. The label embedding achieves this by suppressing cross-class matches, which account for fewer than 1% of all matches at $\lambda_y = 0.5$.

Figure 1 illustrates how HiLO alignment changes the structure of patch embeddings. Before alignment, Virchow patch embeddings form more prominent clusters for each scanner, whereas after HiLO alignment these clusters are no longer discernible. We quantify this effect with k-nearest-neighbour classifiers that predict either the scanner or the class of each patch embedding, and we refer to their accuracies as scanner accuracy and class accuracy (Appendix C). Averaged over the 10 FMs, the alignment reduces scanner accuracy from 0.889 to 0.374, while class accuracy remains at 0.814. HiLO thus removes 75% of the scanner accuracy above chance and retains 87% of the class accuracy above chance.

Table 1: OOD macro-F1 on AGGC2022 for stain normalisation and OT methods across 10 FMs. For each method hyperparameters are selected on validation set and the results are averaged over all combinations of fold and seed.

Method	UNI	UNI2	PLIP	H-optimus-1	Virchow	Virchow2	CONCH v1.5	Prov-GigaPath	Midnight	MUSK
ERM	0.731	0.727	0.542	0.749	0.703	0.747	0.775	0.711	0.730	0.617
Reinhard	0.731	0.748	0.613	0.743	0.706	0.752	0.774	0.731	0.750	0.629
Modified Reinhard	0.744	0.751	0.614	0.746	0.703	0.758	0.772	0.726	0.748	0.614
Macenko	0.736	0.749	0.614	0.748	0.683	0.737	0.792	0.728	0.721	0.633
Vahadane	0.729	0.753	0.620	0.718	0.717	0.729	0.807	0.708	0.706	0.620
Mini-batch Sinkhorn	0.726	0.703	0.574	0.717	0.693	0.706	0.808	0.715	0.764	0.647
Low-rank OT	0.758	0.700	0.575	0.794	0.739	0.765	0.812	0.732	0.744	0.651
FRLC	0.727	0.706	0.572	0.769	0.710	0.769	0.807	0.737	0.748	0.658
HiLO ($\lambda_y = 0$)	0.769	0.758	0.584	0.779	0.722	0.761	0.810	0.750	0.757	0.641
HiLO	0.772	0.762	0.577	0.789	0.740	0.772	0.817	0.759	0.755	0.649

Table 2: OOD macro-F1 on AGGC2022 for domain generalisation methods across 10 FMs. For each method hyperparameters are selected on validation set and the results are averaged over all combinations of fold and seed.

Method	UNI	UNI2	PLIP	H-optimus-1	Virchow	Virchow2	CONCH v1.5	Prov-GigaPath	Midnight	MUSK
ERM	0.731	0.727	0.542	0.749	0.703	0.747	0.775	0.711	0.730	0.617
CORAL	0.732	0.743	0.561	0.752	0.720	0.742	0.801	0.713	0.743	0.655
Conditional CORAL	0.723	0.726	0.571	0.753	0.709	0.738	0.800	0.714	0.735	0.640
Paired CORAL	0.727	0.722	0.578	0.672	0.720	0.742	0.801	0.719	0.735	0.641
DANN	0.731	0.745	0.557	0.767	0.704	0.756	0.815	0.707	0.741	0.635
IRM	0.731	0.728	0.551	0.761	0.706	0.762	0.809	0.721	0.749	0.652
GroupDRO	0.724	0.709	0.564	0.743	0.696	0.733	0.788	0.706	0.730	0.638
Fish	0.753	0.758	0.541	0.775	0.734	0.766	0.825	0.749	0.763	0.612
Mixup	0.722	0.718	0.568	0.744	0.702	0.734	0.786	0.705	0.739	0.645
HiLO	0.772	0.762	0.577	0.789	0.740	0.772	0.817	0.759	0.755	0.649

4.3 Comparison with DG methods - AGGC2022 dataset

We compare HiLO with CORAL (Sun & Saenko, 2016), together with its conditional and paired variants, DANN (Ganin et al., 2016), IRM (Arjovsky et al., 2019), GroupDRO (Sagawa et al., 2020), Fish (Shi et al., 2022), and Mixup (Zhang et al., 2018), each trained on the same frozen patch embeddings. Every method is given one key hyperparameter tuned on the same five-point grid. For HiLO this is the alignment weight λ_c.

Table 2 shows that the effect of DG methods on frozen patch embeddings varies across methods and FMs, with changes in OOD macro-F1 relative to ERM ranging from -0.077 for Paired CORAL on H-optimus-1 to $+0.050$ for Fish on CONCH v1.5. HiLO exceeds every DG method on 6 of 10 FMs (UNI, UNI2, H-optimus-1, Virchow, Virchow2, and Prov-GigaPath), with the largest margins over the strongest method on UNI (0.019) and H-optimus-1 (0.014). On the remaining FMs it trails the best DG method by at most 0.008, and it performs comparably to Paired CORAL on PLIP, at 0.577 against 0.578.

4.4 Combination of HiLO with DG methods - AGGC2022 and BEETLE datasets

We present results for the combination of HiLO with a few DG methods on both AGGC2022 and BEETLE datasets. We combine HiLO with Mixup and GroupDRO, and compare each combination with HiLO and the DG method alone (Table 3). For fair comparison, we fix the alignment weight at $\lambda_c = 100$.

Table 3 shows that adding HiLO to a DG method improves OOD macro-F1 over the DG method alone in 19 of 20 cases on AGGC2022, by between 0.023 and 0.060.The combination also outperforms HiLO alone in 15 of 20 cases. On CONCH v1.5, the combination with Mixup reaches 0.837, compared with 0.817 for HiLO and 0.786 for Mixup. On BEETLE, adding HiLO improves OOD macro-F1 over the DG method alone in all experiment by a value between 0.008 and 0.036. HiLO alone also outperforms both Mixup and GroupDRO on all 8 FMs and the largest gain occurs on

Table 3: OOD macro-F1 on AGGC2022 and BEETLE for HiLO combined with DG methods. The alignment weight is fixed at $\lambda_c = 100$. Results are averaged over all combinations of fold and seed. CONCH v1.5 and MUSK are not evaluated on BEETLE.

Dataset	Method	UNI	UNI2	PLIP	H-optimus-1	Virchow	Virchow2	CONCH v1.5	Prov-GigaPath	Midnight	MUSK
AGGC2022	Mixup	0.722	0.718	0.568	0.744	0.702	0.734	0.786	0.705	0.739	0.645
	GroupDRO	0.724	0.709	0.564	0.743	0.696	0.733	0.788	0.706	0.730	0.638
	HiLO	0.772	0.762	0.577	0.789	0.740	0.772	0.817	0.759	0.755	0.649
	HiLO with Mixup	0.772	0.769	0.600	0.793	0.746	0.772	0.837	0.756	0.765	0.668
	HiLO with GroupDRO	0.779	0.768	0.590	0.795	0.748	0.767	0.840	0.766	0.765	0.634
BEETLE	Mixup	0.668	0.699	0.512	0.707	0.716	0.730	–	0.651	0.713	–
	GroupDRO	0.658	0.697	0.515	0.705	0.734	0.735	–	0.649	0.725	–
	HiLO	0.669	0.713	0.531	0.728	0.737	0.739	–	0.657	0.728	–
	HiLO with Mixup	0.678	0.718	0.536	0.743	0.740	0.744	–	0.670	0.729	–
	HiLO with GroupDRO	0.686	0.729	0.539	0.731	0.763	0.762	–	0.661	0.733	–

Virchow, where the combination with GroupDRO reaches 0.763, compared with 0.740 for HiLO and 0.696 for GroupDRO.

5 Discussion

Our experiments show that the current SOTA pathology FMs encode non-biological variation alongside biological signal and existing methods address this variation only partially. The effectiveness of stain normalisation methods vary across FMs (Table 1), suggesting that the non-biological factors retained by FMs are not limited to colour and therefore cannot be completely removed by correcting the input images alone. DG methods constrain statistics or objectives defined over entire source domains and such constraints admit solutions that leave the non-biological signals intact, which explains their inconsistent gains over ERM across FMs (Table 2).

Our results clearly suggest that one-to-one coupling provide a more effective signal for correcting patch embeddings than the objectives of DG methods. Specifically, HiLO operates on individual patch embeddings, pairing each patch embedding with a patch embedding from another source domain while maintaining label-awareness. We present extensive experiments to demonstrate how HiLO contributes to the improvement of ID and OOD performance. OT methods that derive matches from mini-batch or low-rank couplings yield lower OOD macro-F1 than HiLO on all FMs. Moreover, HiLO can also be combined with DG methods, and these combinations improve over the DG methods alone.

Nevertheless, HiLO requires labeled source domains to guarantee improved alignment of patch embeddings. The pools of each pair of source domains are drawn with equal class counts, so the behaviour of HiLO under different class distributions across source domains remains to be evaluated. All experiments address patch-level classification, and the effect of HiLO on slide-level prediction remains to be explored in future research. These limitations suggest two directions for future work. Pools that preserve the class distributions of the source domains would test HiLO when some cross-class matches cannot be avoided. Applying the alignment to patch embeddings before slide-level aggregation would test whether the improvement at the patch level transfers to slide-level prediction.

6 Conclusion

We presented HiLO, a hierarchical label-aware optimal transport framework for the cross-domain alignment of patch embeddings from pathology FMs. HiLO augments each patch embedding with its label, which reduces cross-class matches while retaining the low-rank cost factorization. The resulting one-to-one couplings supervise an alignment that applies to held-out domains. On 2 publicly available WSI datasets, AGGC2022 and BEETLE, HiLO improves OOD macro-F1 over stain normalisation, other OT methods, and DG methods on all FMs. We argue that our method is a meaningful step towards clinically scalable correction of non-biological variation in pathology FM embeddings.

AI USE STATEMENT

In this work, we used generative AI tools to check the grammar and search for related literature. We have not used generative AI tools to generate synthetic data, develop theoretical frameworks, formulate or prove mathematical claims, propose hypotheses, design the research methodology, clean or reformat datasets, or support qualitative data analysis, and translation is not applicable to this work. Additionally, we used generative AI tools to edit the paper for readability, create LaTeX code for tables and figures. We have reviewed all AI-assisted work. All numbers in tables and text were checked against the original experiment logs by the authors, and the method, experimental design, and conclusions were developed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available, de-identified histopathology datasets (AGGC2022, BEETLE, and Camelyon17), which were released by their organisers with the relevant ethical approvals. No new patient data were collected, and no attempt was made to re-identify individuals. All foundation models were used under the terms of their respective licences. The methods studied here are intended for research on robustness to scanner and centre shift and are not validated for clinical use. The authors declare no conflict of interest.

REPRODUCIBILITY STATEMENT

All datasets and foundation models used in this work are publicly available. The leave-one-scanner-out splits, model architecture, training protocol, hyperparameter grids, and random seeds are described in all experiments. All results are averaged over all combinations of fold and seed. The full code will be released publicly upon acceptance.

REFERENCES

AGGC22 Organisers. Automated gleason grading challenge 2022 (AGGC2022). https://aggc22.grand-challenge.org/, 2022. MICCAI 2022 Challenge, National University Hospital, Singapore.

Hafsa Akebli and Vincenzo Della Mea. Domain generalization of histopathology foundation models in multicenter, multi-scanner cohorts: A comparative benchmark. *Journal of Imaging*, 12(8):381, 2026.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Peter Bandi et al. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2): 550–560, 2019.

Bioptimus. H-optimus-1: A foundation model for histology. https://huggingface.co/bioptimus/H-optimus-1, 2025.

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 *MICCAI Workshops*. Springer, 2025. doi: 10.1007/978-3-032-07845-2_5.

Richard J. Chen et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30, 2024. doi: 10.1038/s41591-024-02857-3.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

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

Kianoush Falahkheirkhah, Alex Lu, David Alvarez-Melis, and Grace Huynh. Domain adaptation using optimal transport for invariant learning using histopathology datasets. *arXiv preprint arXiv:2303.02241*, 2023.

Yaroslav Ganin et al. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations (ICLR)*, 2021.

Peter Halmos, Xinhao Liu, Julian Gold, and Benjamin J. Raphael. Low-rank optimal transport through factor relaxation with latent coupling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2411.10555.

Peter Halmos, Julian Gold, Xinhao Liu, and Benjamin J. Raphael. Hierarchical refinement: Optimal transport to infinity and beyond. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2503.03025.

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, 2023. doi: 10.1038/s41591-023-02504-3.

Mikhail Karasikov, Joost van Doorn, Nicolas Känzig, Melis Erdal Cesur, Hugo Mark Horlings, Robert Berke, Fei Tang, and Sebastian Otálora. Training state-of-the-art pathology foundation models with orders of magnitude less data. *arXiv preprint arXiv:2504.05186*, 2025.

David Krueger et al. Out-of-distribution generalization via risk extrapolation (REx). In *International Conference on Machine Learning (ICML)*, 2021.

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

Jonah Kömen, Edwin D. de Jong, Eric Marcus, Jonas Teuwen, et al. Towards robust foundation models for digital pathology. *Nature Communications*, 17(1), 2026. doi: 10.1038/s41467-026-73923-2.

Carlijn Lems, Leslie Tessier, John-Melle Bokhorst, Mart van Rijthoven, Witali Aswolinskiy, Matteo Pozzi, Natalie Klubickova, Suzanne Dintzis, Michela Campora, Maschenka Balkenhol, Peter Bult, Joey Spronck, Thomas Detone, Mattia Barbareschi, Enrico Munari, Giuseppe Bogina, Jelle Wesseling, Esther H. Lips, Francesco Ciompi, Frédérique Meeuwsen, and Jeroen van der Laak. A multicentric dataset for training and benchmarking breast cancer segmentation in H&E slides. *arXiv preprint arXiv:2510.02037*, 2025.

Ming Y. Lu et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30, 2024.

Marc Macenko et al. A method for normalizing histology slides for quantitative analysis. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2009.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. Foundations and Trends in Machine Learning, 2019.

E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

Santanu Roy, Shubhajit Panda, and Mahesh Jangid. Modified Reinhard algorithm for color normalization of colorectal cancer histopathology images. In *Proceedings of the 29th European Signal Processing Conference (EUSIPCO)*, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts. In *International Conference on Learning Representations (ICLR)*, 2020.

Meyer Scetbon, Marco Cuturi, and Gabriel Peyré. Low-rank sinkhorn factorization. In *International Conference on Machine Learning (ICML)*, 2021.

M. Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. StainGAN: Stain style transfer for digital histological images. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 953–956, 2019. doi: 10.1109/ISBI.2019.8759152.

Yuge Shi et al. Gradient matching for domain generalization. In *International Conference on Learning Representations (ICLR)*, 2022.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *ECCV Workshops*, 2016.

Abhishek Vahadane et al. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Transactions on Medical Imaging*, 35(8):1962–1971, 2016.

Eugene Vorontsov et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 2024. doi: 10.1038/s41591-024-03141-0.

Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, et al. A vision-language foundation model for precision oncology. *Nature*, 638:769–778, 2025. doi: 10.1038/s41586-024-08378-w.

Hanwen Xu et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630, 2024. doi: 10.1038/s41586-024-07441-w.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

Yishu Zhang, Shushan Wu, Zhenzhong Zhang, Didong Li, Huaxiu Yao, Yun Li, Iain Carmichael, Katherine A. Hoadley, Hongtu Zhu, Di Wu, and Daiwei Zhang. Mitigating batch effects in histopathology via language-mediated robust embedding generation. *arXiv preprint arXiv:2606.28697*, 2026.

Eric Zimmermann et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.

A Proof of Proposition 1

Because both pools contain the same number of patch embeddings of each class, a one-to-one correspondence without cross-class matches exists. Its total transport cost is at most $4N$, since the first term of Equation (3) lies in $[0, 4]$ for every match. A one-to-one correspondence with at least one cross-class match has a total transport cost of at least $2\lambda_y^2$, which exceeds $4N$ when $\lambda_y^2 > 2N$. Every minimizer of Equation (4) therefore contains no cross-class match. □

B Datasets

AGGC2022. Patches are extracted at 256 pixels pixels and 20x. Table 4 lists the number of patches of each class.

Table 4: Number of patches of each class in AGGC2022.

	G3	G4∪G5	Normal	Stroma
Patches	202k	349k	69k	140k

BEETLE. Patches are extracted at 256 pixels and 20x from the development set. Each scanner is a domain, and we retain the five scanners whose patches contain all four classes (Table 5). Leica Aperio GT 450 DX is excluded because it contains no invasive patches, and Hamamatsu NanoZoomer 2.0-RS is excluded because it contains no necrosis patches. Three further scanners appear only in the evaluation set and are not used.

Table 5: Scanners of the BEETLE development set used as domains.

Scanner	Source	Patches
3DHistech Pannoramic P250 Flash II	RUMC	56,548
3DHistech Pannoramic 250 Flash III	SCH	18,620
3DHistech Pannoramic 1000	RUMC	11,560
Leica Aperio ScanScope XT	TCGA	10,512
Leica Aperio AT2	NKI	3,520

Camelyon17. Each of the five centres is a domain. Table 6 reports OOD macro-F1 of representative methods, which exceeds 0.992 for all methods and motivates the use of Camelyon17 for visualization only.

Table 6: OOD macro-F1 on Camelyon17.

Method	OOD macro-F1
ERM	0.992
Strongest DG method	0.994
HiLO	0.994

Table 7: Details of FMs used in the experiments.

FM	Reference	Input patch size	d
UNI	Chen et al. (2024)	256	1024
UNI2	Chen et al. (2024)	256	1536
PLIP	Huang et al. (2023)	256	512
H-optimus-1	Bioptimus (2025)	256	1536
Virchow	Vorontsov et al. (2024)	256	2560
Virchow2	Zimmermann et al. (2024)	256	2560
CONCH v1.5	Lu et al. (2024)	512	768
Prov-GigaPath	Xu et al. (2024)	256	1536
Midnight	Karasikov et al. (2025)	256	3072
MUSK	Xiang et al. (2025)	384	1024

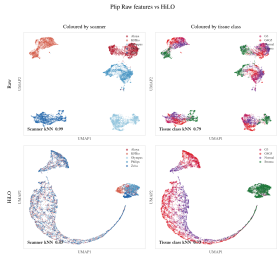
C VISUALIZATION OF PATCH EMBEDDINGS

Figures 3, 4, and 5 show UMAP projections of frozen and aligned patch embeddings on AGGC2022, BEETLE, and Camelyon17. AGGC2022 covers all 10 FMs. BEETLE and Camelyon17 cover 8 FMs, as CONCH v1.5 and MUSK are not evaluated on these datasets. All projections use the same fold, the same random subset , and the same UMAP hyperparameters. Frozen and aligned patch embeddings are projected separately because they lie in spaces of different dimensions.

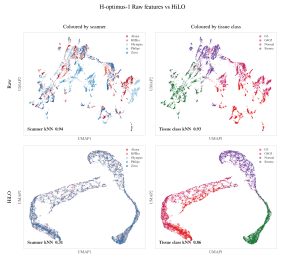


(a) UNI

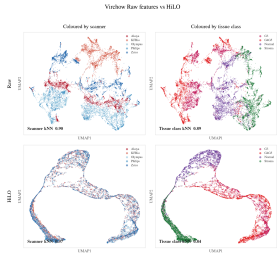
(b) UNI2



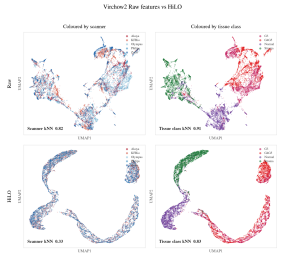
(c) PLIP



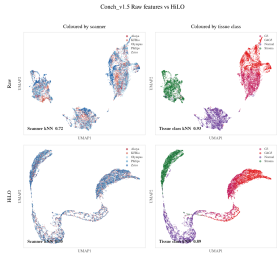
(d) H-optimus-1



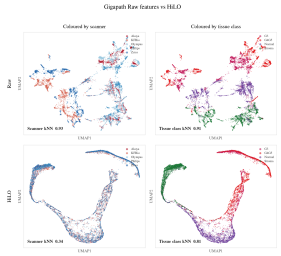
(e) Virchow



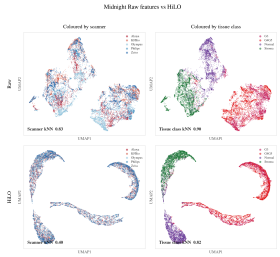
(f) Virchow2



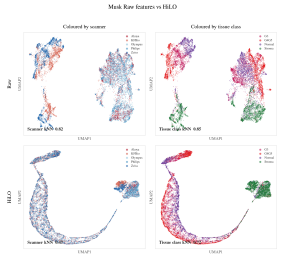
(g) CONCH v1.5



(h) Prov-GigaPath



(i) Midnight



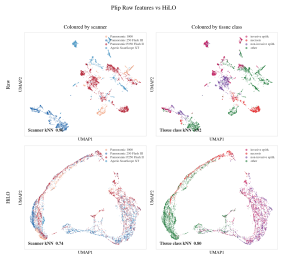
(j) MUSK

Figure 3: UMAP of frozen (left) and aligned (right) patch embeddings on AGGC2022 with KFBio as the held-out scanner. Points are coloured by scanner and tissue class.

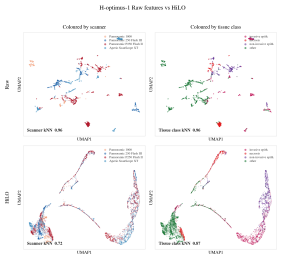


(a) UNI

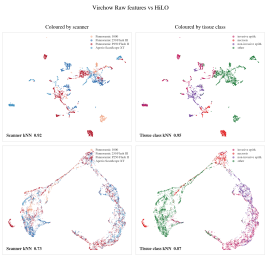
(b) UNI2



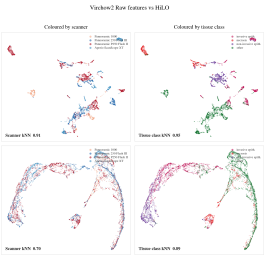
(c) PLIP



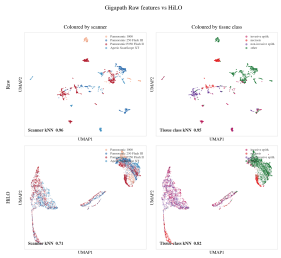
(d) H-optimus-1



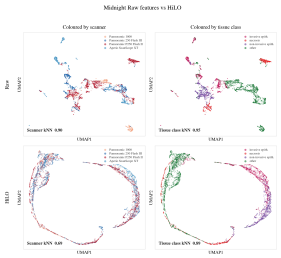
(e) Virchow



(f) Virchow2

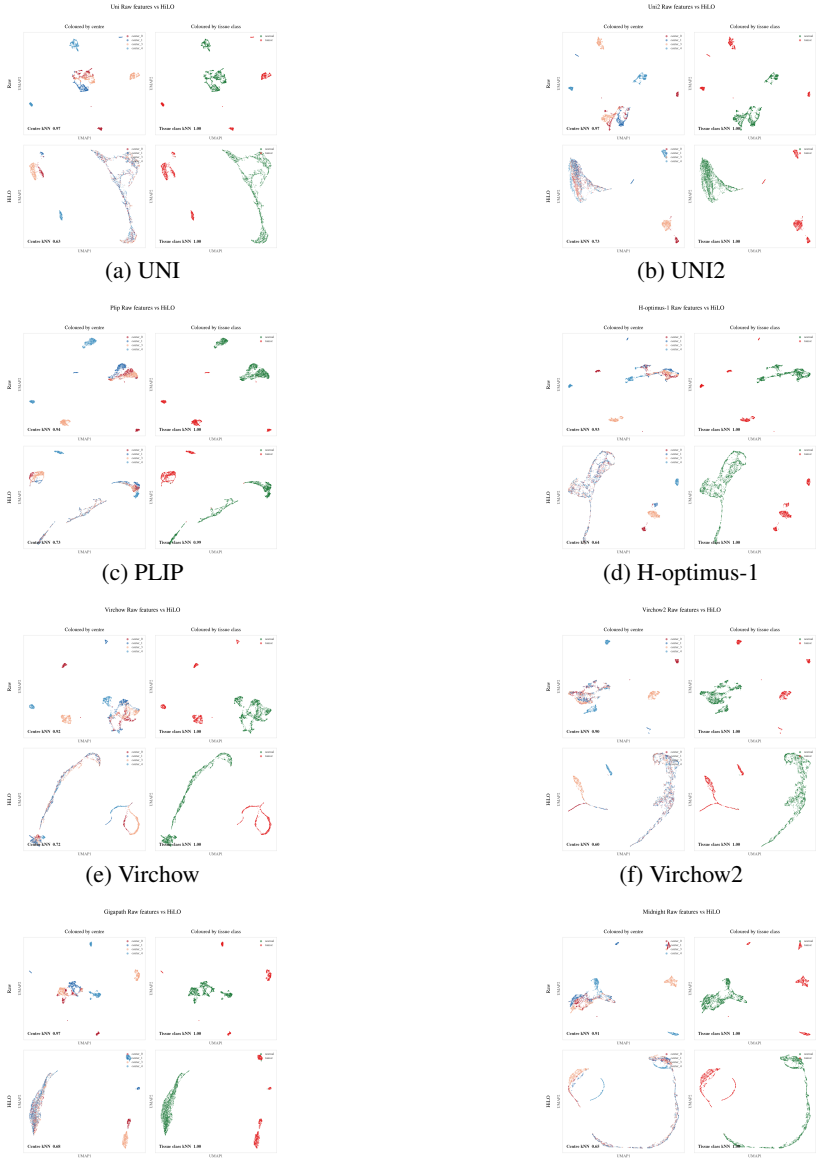


(g) Prov-GigaPath



(h) Midnight

Figure 4: UMAP of frozen (left) and aligned (right) patch embeddings on BEETLE with Leica Aperio AT2 as the held-out (scanner) domain. Points are coloured by scanner and tissue class.



(a) UNI (b) UNI2

(c) PLIP (d) H-optimus-1

(e) Virchow (f) Virchow2

(g) Prov-GigaPath (h) Midnight

Figure 5: UMAP of frozen (left) and aligned (right) patch embeddings on Camelyon17 with Center 1 as the held-out centre. Points are coloured by scanner and tissue class.

Table 8: Six Folds on AGGC2022 and the number of matches for UNI.

Held-out domain	Validation domain	Source domains	Matches
Akoya	KFBio	Leica, Olympus, Philips, Zeiss	176,128
KFBio	Leica	Akoya, Olympus, Philips, Zeiss	184,320
Leica	Olympus	Akoya, KFBio, Philips, Zeiss	196,608
Olympus	Philips	Akoya, KFBio, Leica, Zeiss	184,320
Philips	Akoya	KFBio, Leica, Olympus, Zeiss	176,128
Zeiss	Akoya	KFBio, Leica, Olympus, Philips	176,128

Table 9: Scanner accuracy and class accuracy of k-nearest-neighbour classifiers on frozen and aligned patch embeddings on AGGC2022, with KFBio as the held-out domain. Chance levels are 0.20 for the scanner and 0.25 for the class.

	Scanner accuracy ↓		Class accuracy ↑	
FM	Frozen	Aligned	Frozen	Aligned
UNI	0.969	0.340	0.926	0.814
UNI2	0.972	0.395	0.928	0.842
PLIP	0.993	0.428	0.793	0.726
H-optimus-1	0.935	0.311	0.926	0.857
Virchow	0.896	0.368	0.893	0.841
Virchow2	0.821	0.332	0.909	0.832
CONCH v1.5	0.723	0.351	0.931	0.886
Prov-GigaPath	0.933	0.339	0.905	0.805
Midnight	0.834	0.401	0.897	0.818
MUSK	0.819	0.471	0.849	0.721
Mean	0.889	0.374	0.896	0.814

Table 10: ID macro-F1 on AGGC2022 for stain normalisation and OT methods across 10 FMs, evaluated on held-out specimens from the four training scanners. For each method hyperparameters are selected on validation set and the results are averaged over all combinations of fold and seed.

Method	UNI	UNI2	PLIP	H-optimus-1	Virchow	Virchow2	CONCH v1.5	Prov-GigaPath	Midnight	MUSK
ERM	0.738	0.736	0.575	0.745	0.708	0.750	0.786	0.713	0.729	0.654
Reinhard	0.725	0.747	0.635	0.749	0.714	0.744	0.773	0.729	0.741	0.651
Modified Reinhard	0.744	0.745	0.635	0.743	0.709	0.757	0.785	0.724	0.748	0.653
Macenko	0.732	0.750	0.619	0.754	0.682	0.734	0.795	0.729	0.724	0.661
Vahadane	0.720	0.761	0.656	0.731	0.702	0.721	0.807	0.719	0.719	0.722
Mini-batch Sinkhorn	0.739	0.712	0.637	0.719	0.725	0.720	0.803	0.722	0.767	0.684
Low-rank OT	0.757	0.740	0.642	0.787	0.753	0.763	0.812	0.744	0.749	0.682
FRLC	0.738	0.732	0.633	0.771	0.729	0.771	0.805	0.733	0.751	0.686
HiLO ($\lambda_y = 0$)	0.771	0.766	0.647	0.782	0.742	0.767	0.813	0.753	0.760	0.677
HiLO	0.778	0.774	0.649	0.785	0.748	0.772	0.816	0.765	0.762	0.670

Table 11: ID macro-F1 on AGGC2022 for domain generalisation methods across 10 FMs, evaluated on held-out specimens from the four training scanners. For each method hyperparameters are selected on validation set and the results are averaged over all combinations of fold and seed.

Method	UNI	UNI2	PLIP	H-optimus-1	Virchow	Virchow2	CONCH v1.5	Prov-GigaPath	Midnight	MUSK
ERM	0.738	0.736	0.575	0.745	0.708	0.750	0.786	0.713	0.729	0.654
CORAL	0.740	0.745	0.641	0.747	0.726	0.749	0.800	0.725	0.756	0.700
Conditional CORAL	0.737	0.748	0.636	0.750	0.722	0.748	0.798	0.720	0.746	0.679
Paired CORAL	0.740	0.745	0.641	0.747	0.726	0.749	0.800	0.723	0.746	0.678
DANN	0.742	0.767	0.660	0.778	0.731	0.773	0.818	0.726	0.749	0.697
IRM	0.754	0.736	0.667	0.771	0.734	0.768	0.814	0.733	0.767	0.699
GroupDRO	0.730	0.733	0.639	0.739	0.715	0.735	0.796	0.708	0.738	0.677
Fish	0.767	0.771	0.621	0.786	0.747	0.788	0.825	0.752	0.758	0.647
Mixup	0.739	0.737	0.641	0.755	0.726	0.741	0.794	0.724	0.745	0.681
HiLO	0.778	0.774	0.649	0.785	0.748	0.772	0.816	0.765	0.762	0.670