



OPEN Lightweight self supervised learning framework for domain generalization in histopathology

Abubakr Shafique¹✉, Amanda Dy¹, Xiaoli Qin¹, Najd Alshamlan², Dimitrios Androutsos¹, Susan J. Done^{3,4} & April Khademi^{1,5,6,7,8}✉

The emergence of large foundation models (FMs) in histopathology, trained on extensive image datasets using high-performance graphics processing unit (GPU) clusters, has demonstrated significant potential in advancing computational pathology. FMs have potential to overcome the domain gap between training and testing datasets, which creates more translation opportunities. However, the reliance on vast computational resources and large-scale data often limits accessibility and widespread adoption of FMs. To address this limitation, we present *HistoLite*, a lightweight self-supervised learning framework designed to enable domain-invariant representation learning in histopathology. *HistoLite* utilizes customizable auto-encoders within a self-supervised learning paradigm that learns generalized and transferable features in an efficient manner. We evaluated the proposed framework using breast Whole Slide Images (WSIs) and benchmarked performance with state-of-the-art FMs for domain generalization. A novel dataset was curated that is of the same tissue slides, scanned by two different scanning platforms, which allows for specific analysis of covariate shifts due to scanner bias. Aspects evaluated include the difference in embeddings across scanners using novel representation shift metrics, including a robustness index, and accuracy, which looks at performance on downstream tasks. The top performing models were UNI, Virchow2 and Prov-GigaPath, likely due to large model sizes and training datasets. In general, most FMs were found to be susceptible to scanner-bias, as shown by differences in embeddings and drop in performance on the held-out scanner. This has significant implications for real-world deployment of FMs in histopathology. *HistoLite* offered low representation shift in embeddings, the lowest performance drop on out-of-domain data with modest classification accuracy, indicating the smaller model may exhibit a tradeoff between accuracy and generalization.

Keywords Digital pathology, Domain generalization, Domain shift, Foundation models, Self-supervised learning, Light-weight models

Histopathology remains the gold standard for cancer diagnosis and confirming malignancies^{1,2}. Traditionally reviewed under light microscopes, pathology slides can now be digitized using whole slide imaging (WSI) for clinical sign-out, telepathology, and computer-assisted analysis²⁻⁴. While conventional computer-assisted image analysis methods have achieved moderate success, advances in artificial intelligence (AI) and deep learning are revolutionizing histopathological interpretation and computational pathology^{3,5}. AI tools have been shown to reduce subjectivity, improve accuracy, and drive workflow efficiencies that support digital pathology adoption and quality of care^{6,7}. Despite the clear benefits, computational pathology algorithms are not yet widely deployed.

Deep learning models suffer performance degradation when applied to unseen or out-of-distribution data^{8,9}. Scanner variation in histopathology images is a leading technical barrier for robust deployment of AI across pathology laboratories¹⁰⁻¹³. Different labs may support different scanner vendors, which creates differences in colour and noise distributions, among other “invisible” acquisition factors that can inadvertently affect deep learning algorithms. Figure 1 illustrates the variability in appearance when the same glass slide is scanned using two different scanners, which can lead to shifts in the model’s learned feature representations (“*representation*”).

¹Department of Electrical, Computer and Biomedical Engineering, Toronto Metropolitan University, Toronto, ON, Canada. ²King Abdullah bin Abdulaziz University Hospital, Princess Norah University, Riyadh, Saudi Arabia. ³Laboratory Medicine Program, University Health Network, Toronto, ON, Canada. ⁴Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada. ⁵Institute for Biomedical Engineering, Science Tech (iBEST), Toronto, ON, Canada. ⁶Vector Institute for Artificial Intelligence, Toronto, ON, Canada. ⁷Department of Medical Imaging, University of Toronto, Toronto, ON, Canada. ⁸Institute of Medical Science, University of Toronto, Toronto, ON, Canada. ✉email: abubakr.shafique@torontomu.ca; akhademi@torontomu.ca

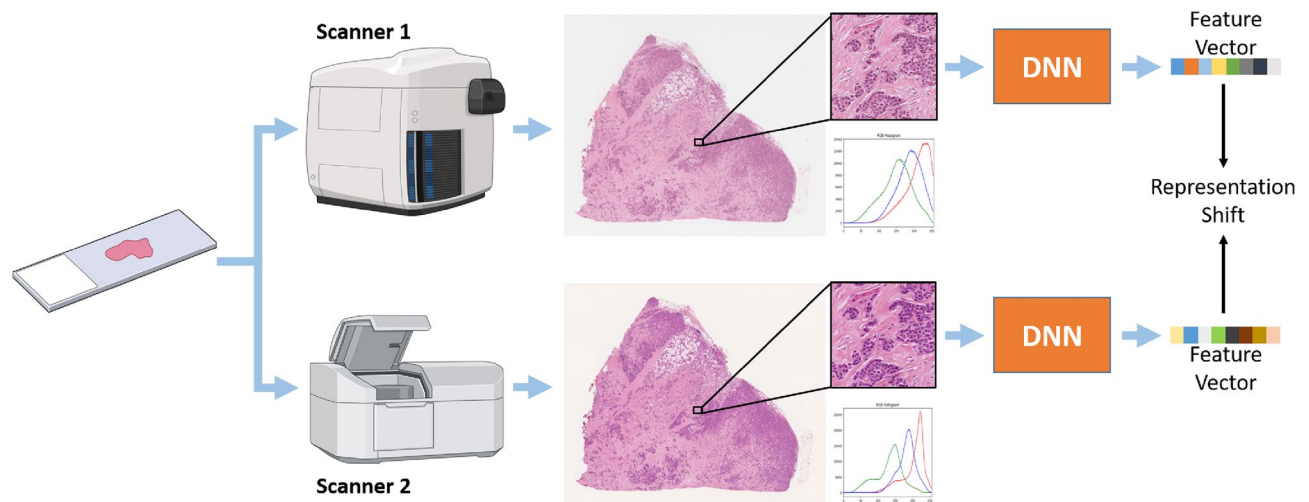


Figure 1. Representation shift. Same-slide tissue regions scanned on two devices—(Scanner 1 (Aperio) and Scanner 2 (Sakura VisionTek)—show differences in patch-level embeddings despite identical spatial locations. These differences are quantified using novel representation-shift metrics applied to deep neural network (DNN)-derived feature vectors.

shift”). These variations can undermine the consistency and reliability of model performance⁴, which would create inequities in care delivered to the patient, as the model would behave differently across different scanners and labs.

Histopathology foundation models (FMs) have emerged to improve generalization¹⁴. These models leverage large-scale, unlabeled histopathology archives to learn rich and transferable representations^{14,15}. FM generalization in pathology applications is only beginning to be explored, and to our knowledge, there are no studies that specifically assess the impact of scanner bias. A key innovation of this work lies in the evaluation of FM robustness using a novel dataset of identical histological slides scanned using two different WSI scanners. This setup enables a targeted analysis of performance variation and representational shifts arising from scanner variability.

A major drawback of FMs is their substantial demand for computational resources and large-scale clinical training datasets^{16–18}. FMs have been reported to consume up to 35 times more power than traditional task-specific models, raising concerns about environmental sustainability and long-term feasibility¹⁹. These constraints can also limit the application of AI in low-resource healthcare settings and reduce efficiency.

In response to these challenges, we propose HistoLite, a lightweight self-supervised deep learning framework designed to be trained on smaller datasets using reduced computational resources. HistoLite incorporates a novel autoencoder-based contrastive learning architecture designed to be resource-efficient and capable of learning robust, generalizable feature representations. Autoencoders represent one of the most established and effective approaches for learning deep representations in a self-supervised manner²⁰. They have been utilized successfully in a range of histopathology applications, including segmentation, detection, feature extraction, compact representation learning, and cross-modal representation learning^{21–24}.

We introduce a novel framework to assess model sensitivity to scanner bias by quantifying representation shifts in identical pathology slides digitized using two distinct scanners. Leveraging this framework, we compare HistoLite against state-of-the-art (SOTA) FMs in breast cancer pathology images.

Related work

Foundation models in histopathology

Advanced self-supervised learning (SSL) techniques^{15,25–30} have enabled the development of FMs. SSL techniques have several advantages for histopathology image analysis, where annotated data is scarce, in that large unlabeled datasets can be leveraged to obtain robust representations that can be fine-tuned for downstream tasks. SSL generates supervisory signals directly from the data itself, rather than relying on expert-provided labels^{25–28}. Recently, numerous FMs have emerged in the literature, driven by the growing popularity of SSL^{27,28}. By removing the dependency on labeled data, these models can effectively utilize large datasets obtained from gigapixel WSIs. The latest FMs for histopathology include UNI³⁰, Hibou-B³¹, Virchow²⁹, Virchow2/2G³², and Prov-GigaPath¹⁵, have utilized the DINOv2 self-supervised learning framework²⁸, leveraging their extensive in-house datasets. iBOT-Path³³ used the iBOT-SSL framework³⁴. Hierarchical Image Pyramid Transformer (HIPT)³⁵, and PathDino³⁶ used the DINO SSL framework²⁷. In addition to SSL Vision Transformer (ViT) FMs, there exists a Convolutional Neural Network (CNN) called KimiaNet³⁷, which is built upon the DenseNet-121 architecture and has been trained in a supervised manner using the entire TCGA dataset. These large-scale FMs are trained on extensive graphics processing unit (GPU) clusters with large datasets^{15,29–32}. Table 1 provides a summary of all models, including model architecture, parameters, patch sizes, number of organs, number of WSIs and patches used for training, the source of the WSIs, and the datasets from which they were obtained.

Models	Params. (M)	Architecture	Patches (M)	WSIs (K)	Patch Size	Data	Organs
HistoLite (Ours)	7	Auto-Encoder	0.5	2.7	512	Public[0]	1
KimiaNet ³⁷	7	DenseNet-121	0.3	11	1000	TCGA[2]	30
PathDino ³⁶	9	ViT-S/16[1]	6	12	512	TCGA[2]	33
HIPT ³⁵	21	ViT-S/16	104	11	224	TCGA[2]	33
iBOT-Path ³³	85	ViT-B/16	40	6	224	TCGA[2]	13
Hibou-B ³¹	85	ViT-B/14	512	1139	224	Private	12
UNI ³⁰	303	ViT-L/16	100	100	224	MGB[3]	20
Virchow ²⁹	632	ViT-H/14	2000	1,488	224	MSKCC[4]	17
Virchow2 ³²	632	ViT-H/14	2000	3,100	224	MSKCC[4]	200
Prov-GigaPath ¹⁵	1,135	ViT-G/14	1385	171	224	PHS[5]	31

Table 1. Histopathology FMs available in the literature. ⁰Public Datasets: CAMELYON16/17, HEROHE, TCGA-BRCA. ¹PathDino is a modified ViT-S/16 with only 5 blocks. ²TCGA: The Cancer Genome Atlas. ³MGB: Mass General Brigham. ⁴MSKCC: Memorial Sloan Kettering Cancer Center. ⁵PHS: Providence Health and Services

Domain shift in histopathology

A data domain is defined as the joint distribution of feature and label spaces for the source in-domain (ID) data used to train a model, and the target out-of-domain (OOD) data unseen during training⁸. Domain shift occurs when the distributions of the source and target domains differ, and may manifest as covariate, prior, posterior, or class-conditional shifts⁸. In histopathology, covariate shift (differences in image appearance due to staining, scanning, or pre-processing) is the most common³⁸, and has been widely recognized as a critical technical challenge^{39–41}. Such shifts mean that datasets acquired from different laboratories or imaging centers can yield markedly different model performance, raising the risk of healthcare inequities.

Domain Generalization (DG) techniques are solutions for the domain shift and use only source domain data without having access to target data^{42–44}. This is important for translation to clinical use, as models can be applied widely and robustly at new imaging centres without the need to collect data and labels or fine-tune, which can have large regulatory implications. This is in contrast to Domain Adaptation (DA) techniques, which have access to both the source and target data⁹.

In computational pathology, a variety of DG methods have been proposed and can be broadly categorized based on their underlying mechanisms. Domain alignment strategies such as stain normalization^{45,46}, the use of generative models⁴⁷, and feature alignment⁴⁸ aim to mitigate domain shift by learning feature representations that are invariant across different data distributions. Data augmentation techniques^{47,49–51} are another field of DG that enhances model robustness by artificially expanding the training dataset using perturbations of existing samples. Domain separation^{52–55} focuses on decomposing the learned feature space into domain-invariant and domain-specific components. Meta-learning approaches^{56–59} seek to learn a generalizable learning algorithm itself, enabling a model to quickly adapt to new, unseen domains with minimal data. Ensemble learning^{41,60–62} leverages the collective power of multiple models to reduce the risk of overfitting and compensate for the weaknesses of a single model. Tailored model design^{63–66} involves developing specialized architectures optimized for specific tasks, which can lead to resource savings and better regularization against overfitting. Regularization training strategies^{67–70} constrain model complexity, thereby preventing overfitting and reducing the influence of irrelevant features. While many of these methods achieved modest results, newer approaches such as FMs, which are trained on large datasets of unlabeled source domain data, are demonstrating superior performance in histopathology tasks^{15,30,32}, and have the potential to generalize due to the large datasets used for pretraining.

HistoLite: lightweight self-supervised model

This work proposes HistoLite, a lightweight self-supervised model designed to achieve strong generalization from substantially smaller datasets than those required by conventional FMs. It can be trained on a single standard GPU, prioritizing efficiency and accessibility.

HistoLite uniquely leverages a dual-stream contrastive autoencoder architecture with shared weights for improved generalization. Figure 2a shows the HistoLite architecture. In the first stream, the network learns features by reconstructing the original image, while the second stream processes an augmented version of the image that simulates realistic variations in stain, contrast, sharpness, and field-of-view (FOV). Compressed representations from the bottlenecks of both streams are aligned via a contrastive objective, encouraging the learning of domain-invariant representations. This is further supported by a novel rotation augmentation strategy called *Adaptive HistoRotate*, which dynamically adjusts rotational transformations to maximize robustness to orientation variability.

Model architecture

Autoencoders are one of the first self-supervised deep learning architectures that leverage the inherent structure of data to learn useful representations without explicit labels²⁰. Autoencoders learn deep feature representations in an unsupervised manner by defining the reconstruction of the input as the primary learning objective. Figure 2b illustrates the proposed design of the lightweight CNN-based 2D autoencoder for both contrastive learning

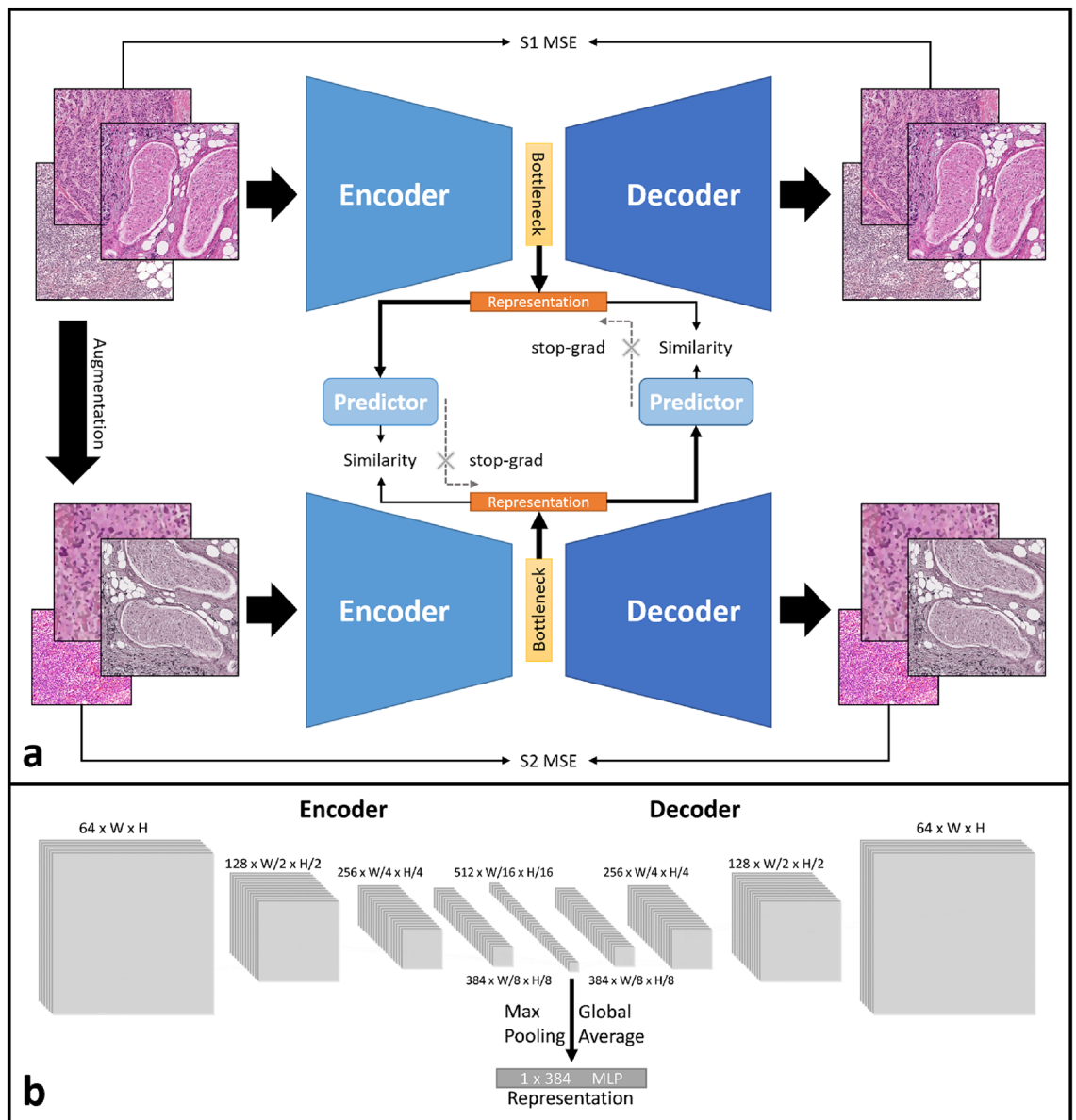


Figure 2. HistoLite SSL framework. **(a)** A dual-stream autoencoder-based self-supervised learning framework designed to achieve domain-agnostic representation learning. **(b)** HistoLite autoencoder design for both contrastive learning streams.

streams. The encoder progressively extracts 64, 128, 256, 384, and 512 2D feature maps through convolutional layers. The decoder reconstructs the input image by sequentially reducing the number of feature maps from 512 to 64. For input images of size 512×512 , HistoLite consists of 41M parameters, with the autoencoder streams and predictive modules. An individual autoencoder (Fig. 2b), the parameter count is 17.2M, with the encoder accounting for 7.6M parameters. This architecture ensures precise image reconstruction while generating a compact, robust representation suitable for downstream tasks. The dual-stream contrastive learning framework promotes feature alignment and invariance to domain-specific augmentations (see Loss Functions).

Representation vector

A compact 1D representation of the feature maps is required for the output of the encoder, which is used for feature alignment across dual contrastive learning streams, fine-tuning for downstream tasks, and for analyzing the representation shift. Similar to other small FMs, we define the model embedding to be length 384. The 512 2D feature maps are processed using a 2×2 Max Pooling followed by global average pooling. The resultant vector is then passed through a single multilayer perceptron (MLP) layer with 384 units, yielding the final encoded vectorized representation of the image. The 384-dimensional vector representation generated by the autoencoder is passed through a predictor MLP that first expands it to 1536 dimensions before reducing it back to 384 dimensions. This $4 \times$ expansion and subsequent reduction facilitate better alignment of the representations,

enabling the predictor to model more complex relationships and promote effective feature alignment. The predictor MLP comprises 3.5M trainable parameters.

Loss function

Chen et al.²⁶ introduced the concept of feature alignment through predictor similarity in a Siamese network, using a stop-gradient mechanism to improve training stability and effectiveness. Inspired by this, a similar strategy is implemented to align the Histolite's representations, as shown in Fig. 2a. A predictor is applied to the output of each autoencoder, and training minimizes the discrepancy between these predicted representations. This encourages consistency and robust feature learning while preserving the unique characteristics of each stream. The method eliminates the need for negative pairs or a momentum encoder, and remains effective with standard batch sizes, thereby eliminating the dependency on large-batch training. As a result, the approach is resource-efficient and well-suited for typical computational environments.

To ensure similarity in the features learned by the autoencoders, the Mean Squared Error (MSE) loss is applied to the decoder outputs from both autoencoders as well as to the representations generated by the corresponding bottlenecks, resulting in three MSE losses:

$$\mathcal{L}_{\text{total MSE}} = \mathcal{L}_{S1 \text{ MSE}} + \mathcal{L}_{S2 \text{ MSE}} + \mathcal{L}_{\text{Sim MSE}}, \quad (1)$$

where $\mathcal{L}_{S1 \text{ MSE}}$ represents the MSE similarity loss associated with the reconstruction of the original images in stream one, $\mathcal{L}_{S2 \text{ MSE}}$ represents the MSE similarity loss associated with the reconstruction of the augmented images in stream two, $\mathcal{L}_{\text{Sim MSE}}$ represents the MSE similarity loss associated with the representations generated from stream one and stream two, and $\mathcal{L}_{\text{total MSE}}$ represents the total MSE loss combined for backpropagation.

Data augmentation

To capture the inherent variability of histopathology data, we applied a set of augmentations encompassing changes in orientation, staining, contrast, sharpness, FOV, and magnification. The objective is to perturb the data realistically to encourage learning of domain-invariant features across the two streams.

Rotation augmentation was applied to the inputs of both streams. Previously, Alfasy et al.³⁶ proposed *HistoRotate*, a 360° rotation augmentation in which a large image is randomly rotated and then center-cropped to form the network input. Although this method improves rotational invariance and robustness⁷¹, it can generate repeated versions of the same patch. We address this limitation with *Adaptive HistoRotate*, which first performs a random crop from the larger image, followed by rotation, ensuring that a distinct patch is selected at each iteration (Fig. 3).

Stream two was further subjected to additional perturbations to enhance representation diversity and promote domain-invariant contrastive learning, analogous to Siamese²⁶ and SimCLR⁷² frameworks. Specifically, **Color Jitter** was applied with brightness, contrast, saturation, and hue factors of 0.3, 0.3, 0.15, and 0.05, respectively, each with a probability of 1.0. **Gaussian Blur** used a 7×7 kernel with $\sigma \in [0.1, 3.0]$ and a probability of 0.3. **Sharpness** augmentation was applied with a factor of 2.0 with a probability of 0.3. **Horizontal and vertical flips** were applied independently with a probability of 0.3. For size variation, **Magnification** augmentation was implemented by random cropping to 448^2 , 480^2 , or 512^2 pixels. The augmentation parameters are determined based on empirical observation.

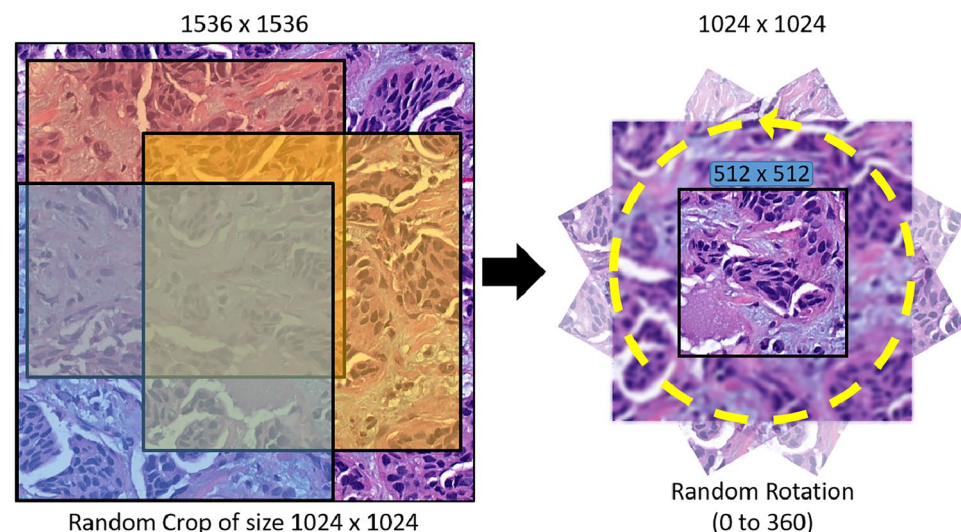


Figure 3. Proposed orientation augmentation method that incorporates the surrounding context and tissue patterns. A random region of size 1024×1024 is initially selected from a larger 1536×1536 image. This region is then subjected to a random 360-degree rotation. Finally, a 512×512 region is cropped from the center of the randomly rotated 1024×1024 region.

Experimental design

Generalization performance of HistoLite and SOTA FMs through two experiments. First, zero-shot feature representations were used to quantify *representation shift*, the change in embeddings across scanners (covariate shift) as shown in Fig. 1. Second, we assessed downstream performance on automated tumour versus non-tumour patch classification using ground truth labels from slides scanned on different devices.

Datasets

Model evaluation was conducted with 111 FFPE H&E-stained breast cancer slides from the Ontario Institute for Cancer Research (OICR), comprising invasive ductal carcinoma, ductal carcinoma in situ, encapsulated papillary carcinoma, and invasive lobular carcinoma. Each slide was scanned on two devices—Aperio AT2 (40×, 0.25 mpp) and Sakura VisionTek (20×, 0.27 mpp)—yielding 222 WSIs. VisionTek images were upsampled to match Aperio resolution, and affine registration¹ was performed using Aperio as reference (Fig. 1).

An expert pathologist annotated ten tumour and ten non-tumour HPFs (212,000 μm^2 each) on Aperio WSIs; the same regions were mapped to VisionTek images via registration. From these, 512×512 pixel patches at 0.5 mpp were extracted at identical coordinates across scanners, producing 4,952 patches per scanner (2,959 tumour; 1,993 non-tumour), for a total of 9,904 aligned patches. This design enables controlled analysis of scanner-induced domain shift and model robustness.

HistoLite was trained on 2,761 public breast cancer WSIs: CAMELYON16⁷³ (270 WSIs, 46,239 patches), CAMELYON17⁷⁴ (997 WSIs, 136,837 patches), HEROHE⁷⁵ (360 WSIs, 108,490 patches), and TCGA⁷⁶ (1,134 WSIs, 253,456 patches). Patches (1536×1536 pixels, 20×) were selected to contain $\geq 80\%$ tissue while excluding artifacts and blur, yielding 545,022 training patches.

HistoLite training

The model was trained for 14 epochs using the proposed SSL framework, with training halted once the loss reached a plateau. The training process used the Adam optimizer with a learning rate and weight decay, both set to $1e^{-4}$. A batch size of four images was used, effectively doubled to eight images after data augmentation, as both the original and augmented images were input into the two network streams. Figure 4a presents the loss curves observed during training, which include the MSE loss for the first stream, representing the reconstruction error of the original input image; the MSE loss for the second stream, representing the reconstruction error of the augmented input image; and the similarity MSE loss, which enforces alignment between the representations generated by the bottlenecks of both streams to ensure consistency despite domain variations. These individual losses are combined to form a unified loss function used for backpropagation, enabling the network to optimize effectively and learn robust features.

Zero-shot feature representations across scanners

To analyze the generalization properties of HistoLite and the SOTA FMs, we conducted a novel assessment of the same tissue across scanners. The same tissue sample scanned by two different scanners is used as an input to the pre-trained models for inference. All models are used as is, i.e. zero-shot with no additional training or fine-tuning. The embeddings produced by each network serve as the model representations, which are subsequently compared pairwise with the embeddings of the patches from the two scanners to measure representation shift.

To quantify the representation shift, vector-, histogram- and cluster-based measures are proposed. Vector-based measures use the z-score normalized embeddings to standardize for different embedding lengths and ranges. Representation shift is measured using mean absolute error (MAE) and cosine distance between paired patches across scanners. The Kullback-Leibler (KL) divergence is used to measure the differences in the embedding probability distributions of paired patches from two scanners.

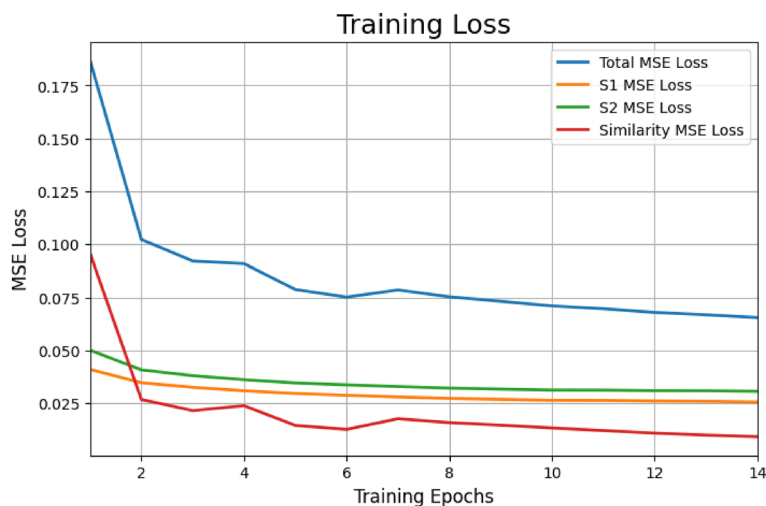


Figure 4. Training losses. Illustrate the training loss curves for HistoLite over 14 epochs.

Cluster-based metrics consider proximity of samples in a high-dimensional feature space and are used to investigate the compactness and separability of the embeddings across classes. In this work, the Calinski-Harabasz (CH) Index is used, which measures the ratio of the weighted sum of squared Euclidean distances between the embeddings and their respective cluster centroid (within-class dispersion) to the sum of squared Euclidean distances of each class centroid to the overall centroid of the data (between-cluster dispersion). The centroid (mean) and classes are based on scanner type and tissue class. This results in the following comparisons (with ↓ and ↑ denoting the desirable outcome, i.e. small or large):

- Tumour CH ↓: Compares tumour patches from the Aperio scanner to tumour patches from the VisionTek scanner (Tumour AP vs. Tumour VT).
- Non-tumour CH ↓: Compares non-tumour patches from Aperio to non-tumour patches from VisionTek (Non-Tumour AP vs. Non-Tumour VT).
- Scanner CH ↓: Compares all patches from Aperio to all patches from VisionTek.
- Tissue CH ↑: Compares all tumour patches (Tumour AP + Tumour VT combined) to all non-tumour patches (Non-Tumour AP + Non-Tumour VT).

To summarize the similarity and differences in the embeddings across tissues and scanners, we propose the Robustness Index (RI) as:

$$\text{Robustness Index (RI)} = \frac{2 \times (\text{Tissue CH})}{\text{Tumour CH} + \text{Non-Tumour CH}}, \quad (2)$$

the numerator represents the inter-class separation, specifically the difference between the centroids of the tumor and non-tumour clusters (tissue discrimination), and the denominator accounts for intra-class compactness of each class across scanners (tissue-specific scanner discrimination), as the average CH for the tumor and non-tumour clusters.

Breast cancer classification across scanners

To evaluate the models on a downstream task, the network backbone was frozen for all models, and a two-class classification head (tumor vs. non-tumour) was added on top of the frozen feature extractor. This classification head was trained for 20 epochs using a cross-entropy loss function, the Adam optimizer, with a learning rate and weight decay, both set to $1e^{-4}$. Data augmentation strategies were applied consistently across all linear probing models. These included color jitter, Gaussian blur, random sharpness, autocontrast, horizontal and vertical flips, as defined in “Data augmentation”. No augmentations were applied to the validation or test sets. For models pretrained on 224×224 resolution (e.g., most ViT-based FMs), images were center-cropped to match the input size. In contrast, models such as PathDino, KimiaNet, and HistoLite were evaluated using 512×512 input patches.

The models were first trained on Aperio data and evaluated on both the Aperio test set (in-domain, ID) and the paired VisionTek set (out-of-domain, OOD) containing the same matched patches. In the second experiment, models were trained on VisionTek data and tested on the same paired datasets, with VisionTek as ID and Aperio as OOD.

Classification performance of the models is evaluated over five folds with an 80/20 training/testing split to predict tumour or non-tumour labels for each patch. The True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), and False Negative Rate (FNR) are measured. Accuracy is calculated as the ratio of correctly classified instances (true positives, correctly identified tumour patches, and true negatives, correctly identified non-tumour patches) to the total number of instances in the test dataset.

The recently proposed generalization test for histopathology algorithms using Two One-Sided Test (TOST) was conducted⁷⁷. TOST analysis is a statistical approach to assess performance equivalence by verifying whether performance differences between ID and OOD datasets fall within an automatically derived equivalence margin. This margin is determined using the standard error to compute the 95% confidence interval⁷⁷.

To determine the relationship between performance consistency and feature similarity, we introduce a novel comparative analysis that examines the change in ID and OOD performance of the models compared to the representation shift. To further understand the trade-offs in model designs, we also take into account the model size, considering its impact on both computational efficiency and generalization ability.

Results

Experiments were conducted on a Windows PC equipped with an Intel i9-12900K CPU, 64GB of RAM, and an NVIDIA RTX 3090 Ti GPU with 24GB of VRAM. Python version 3.9.19, PyTorch version 2.2.2, and CUDA version 11.8 were used.

Zero-shot feature representations across scanners

The same patches from the two scanners were processed by HistoLite and the comparative FMs using zero-shot feature representations, and the difference in embeddings are analyzed using the proposed metrics. Table 2 presents the mean and standard deviation of vector- and histogram-based similarity metrics, including MAE, cosine distance, and KL divergence. The distributions for MAE, cosine distance, and KL divergence are included in the supplementary material (see Figs. S1, S2 and S3, respectively). HistoLite achieves the lowest KL divergence, which indicates that the distributions of the embeddings are more similar across scanners, followed closely by HIPT. HIPT achieves the lowest MAE and Cosine Distance, followed closely by HistoLite. These results suggest

Models	Patch size	Embedding vector	Mean \pm Std MAE	Mean \pm Std Cosine Dist.	Mean \pm Std KL Div.
HistoLite (ours)	512	384	0.51 \pm 0.18	0.23 \pm 0.14	0.21 \pm 0.18
KimiaNet	512	1024	0.79 \pm 0.16	0.52 \pm 0.13	0.54 \pm 0.26
PathDino	512	384	0.63 \pm 0.14	0.33 \pm 0.14	0.33 \pm 0.16
HIPT	224	384	0.49 \pm 0.14	0.22 \pm 0.11	0.22 \pm 0.15
iBot-Path	224	768	0.73 \pm 0.12	0.44 \pm 0.12	0.43 \pm 0.15
Hibou-B	224	768	0.79 \pm 0.13	0.50 \pm 0.11	0.50 \pm 0.17
UNI	224	1024	0.70 \pm 0.10	0.41 \pm 0.11	0.42 \pm 0.13
Virchow	224	2560	0.68 \pm 0.14	0.42 \pm 0.11	0.42 \pm 0.20
Virchow2	224	2560	0.66 \pm 0.17	0.37 \pm 0.16	0.37 \pm 0.20
Prov-Gigapath	224	1536	0.66 \pm 0.11	0.36 \pm 0.10	0.35 \pm 0.12

Table 2. Representation shift metrics MAE, cosine distance, and KL divergence to quantify the similarity between the embeddings of corresponding OICR patches from two different scanners.

Models	Tumour CH \downarrow	Non-Tumour CH \downarrow	Scanner CH \downarrow	Tissue CH \uparrow	RI
HistoLite (ours)	1567.1	617.0	1709.4	1204.3	1.10
KimiaNet	1005.2	601.6	1426.2	309.2	0.38
PathDino	1249.6	836.1	1814.0	768.6	0.73
HIPT	112.2	47.6	128.8	1802.5	22.57
iBOT-Path	755.5	567.8	1105.6	838.7	1.26
Hibou-B	1507.3	810.6	1977.2	434.5	0.37
UNI	497.8	317.4	714.4	353.4	0.86
Virchow	2176.5	1212.1	2956.9	464.4	0.27
Virchow2	615.1	305.5	774.9	584.8	1.27
Prov-GigaPath	311.6	208.3	433.8	522.8	2.01

Table 3. Calinski–Harabasz (CH) Index, and Robustness Index (RI) for different scanners and tissue types for all models.

that HistoLite and HIPT are learning more scanner-invariant features (most similar embedding magnitudes across scanners).

The cluster-based CH metric is reported in Table 3, which considers the degree of overlap in embeddings in a high-dimensional space with respect to tissue or scanner classes. HIPT consistently has the lowest CH over the Non-Tumour CH, Tumour CH and Scanner CH, and the highest Tissue CH indicating good discrimination over tissue classes and feature overlap across scanners. HistoLite has moderate CH metrics across all categories, with good separation between tumour and non-tumour in Tissue CH. Consistently, Prov-GigaPath, Virchow2, UNI and i-BOT-Path, all demonstrate strong performance in the clustering metrics, indicating good generalization across scanners and good separability between tissue classes.

The RI shown in Fig. 6e and Table 3 indicates that HIPT has the highest robustness, followed by Prov-Gigapath, Virchow2 then, iBOT-Path. The RI of HistoLite is comparable to iBOT-path. It shows that there is a tradeoff in performance between tissue discriminability and scanner-variance. From the RI, HistoLite is outperforming KimiaNet, PathDino, Hibou-B, UNI and Virchow. In addition, the embeddings are visualized using the t-distributed Stochastic Neighbor Embedding (t-SNE) in Fig. 5. The t-SNE plots support the cluster-based metrics in feature separation and overlap for different classes and scanners.

Breast cancer classification across scanners

Classification as a downstream task was evaluated using fine-tuned models to classify tumour and non-tumour tissue patches across scanners. The 4,952 patches from each scanner were split into five folds for cross-validation, with each network performing 10 classification experiments in total (five folds per scanner). Table 4 presents the mean classification accuracy for both ID and OOD scenarios. Table 5 reports the performance differences in ID and OOD data for both scanners. The corresponding mean accuracy and performance drop are also illustrated in Fig. 7. To enable a more comprehensive evaluation of model performance, we report additional metrics beyond accuracy. Specifically, fold-wise accuracy along with the average TPR, TNR, FPR, and FNR computed across the five cross-validation folds, are provided in the supplementary material (Tables S1 and S2, respectively).

The best performing models are UNI, Virchow2, and Prov-GigaPath possibly due to the large model size and training datasets. The drop in mean classification accuracy of these models are some of the lowest as well, indicating they are performing well across scanners in downstream tasks. The proposed light-weight model, HistoLite, had an average classification accuracy of 91.8% across both ID and OOD datasets, which

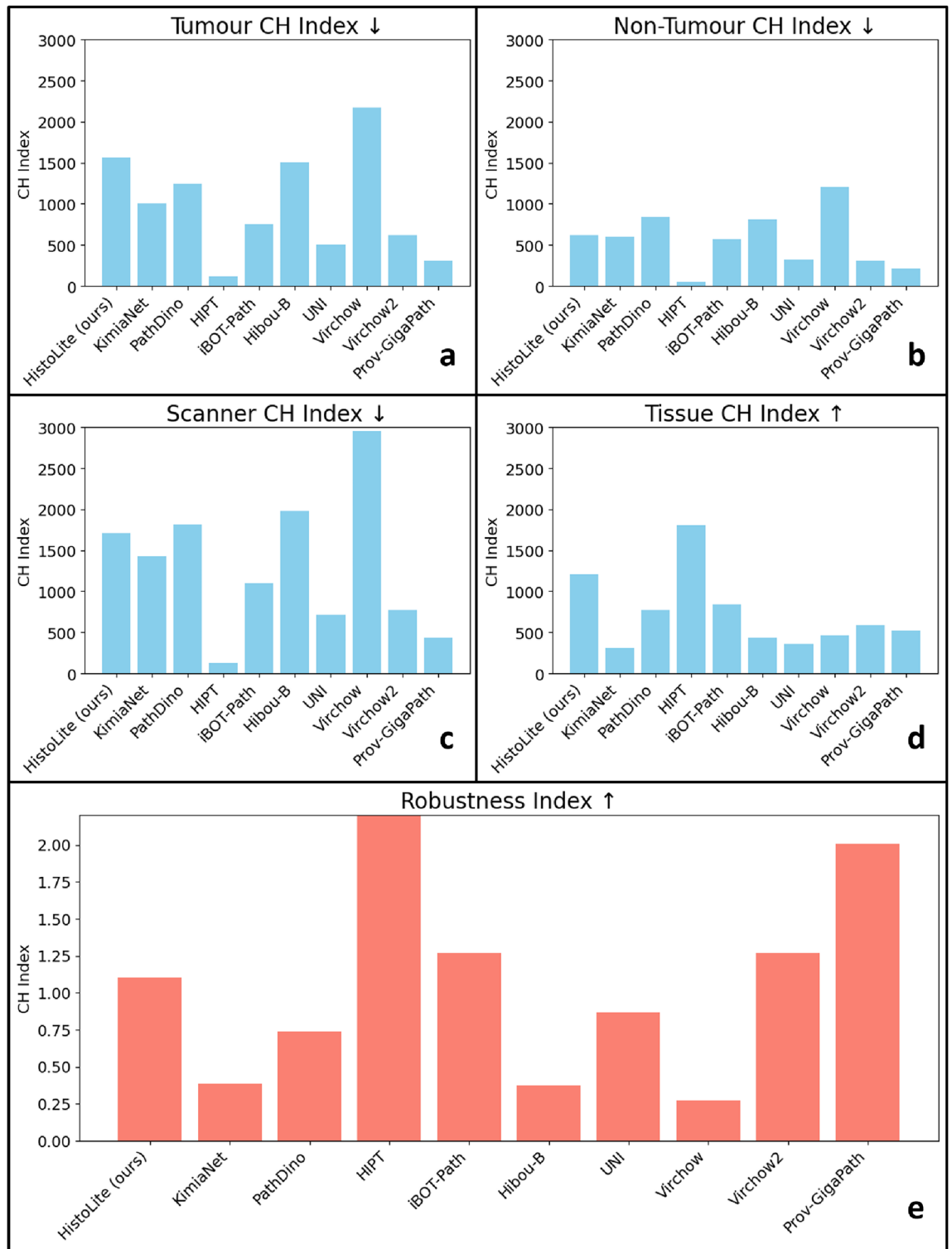


Figure 6. CH and robustness index. The CH Indexes are illustrated as follows: (a) Tumour CH Index, (b) Non-Tumour CH Index, (c) Scanner CH Index, and (d) Tissue CH Index. Additionally, the Robustness Index is shown in (e). This figure corresponds to Table 3.

is higher than KimiaNet and HIPT. While HIPT and HistoLite had the smallest vector and histogram-based representation shift, they exhibited lower mean accuracy in classification. This means the embedding vectors had similar magnitudes and distributions in zero-shot inference, but the features are less robust at differentiating tissue classes when fine-tuned for classification. Considering the performance drop across datasets, HistoLite had the smallest difference between ID and OOD, followed by HIPT. Therefore, although these models did not

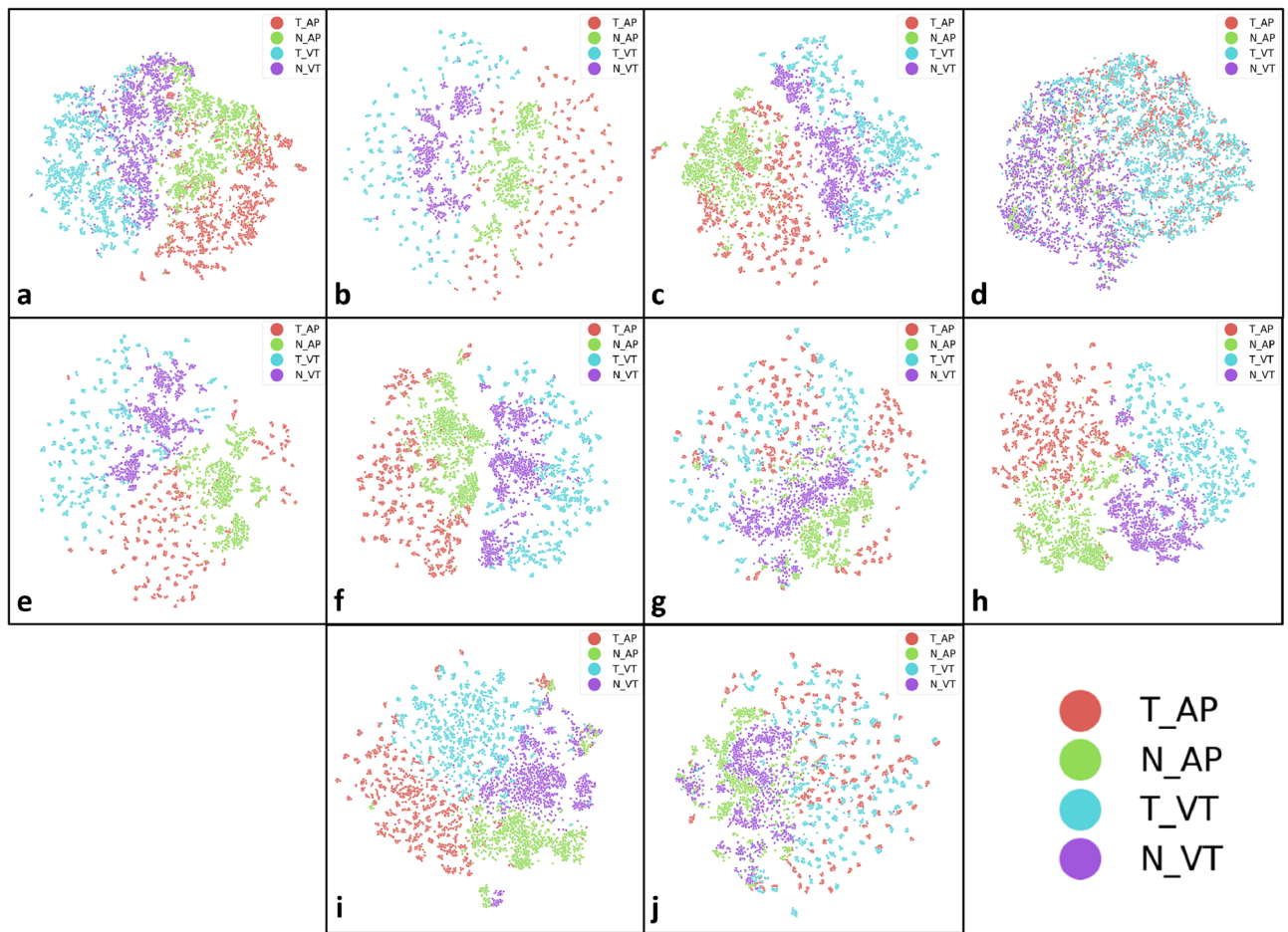


Figure 5. t-SNE plots. Combined t-SNE projections of tumorous and non-tumorous (normal) patch embeddings from two scanners, Aperio (AP) and VisionTek (VT), are shown. Subfigures (a–j) correspond to the models: (a) HistoLite (ours), (b) KimiaNet, (c) PathDino, (d) HIPT, (e) iBOT-Path, (f) Hibou-B, (g) UNI, (h) Virchow, (i) Virchow2, and (j) Prov-Gigapath. All the plots are at the same scale.

Models	AP as ID, VT as OOD		VT as ID, AP as OOD		Mean
	ID \pm Std	OOD \pm Std	ID \pm Std	OOD \pm Std	
HistoLite (ours)	92.4 \pm 1.0	92.1 \pm 0.6	92.3 \pm 0.6	90.2 \pm 1.2	91.8
KimiaNet	94.7 \pm 0.6	79.6 \pm 1.7	92.7 \pm 1.1	85.3 \pm 5.6	88.1
PathDino	97.0 \pm 1.4	88.5 \pm 5.2	96.6 \pm 1.1	92.4 \pm 3.3	93.6
HIPT	90.4 \pm 1.2	89.6 \pm 0.8	90.6 \pm 1.1	88.4 \pm 0.8	89.8
IBOT-Path	96.5 \pm 0.8	95.2 \pm 1.2	96.4 \pm 1.1	92.4 \pm 4.1	95.1
Hibou-B	95.8 \pm 1.2	90.9 \pm 1.2	94.8 \pm 1.4	92.9 \pm 0.8	93.6
UNI	96.9 \pm 1.2	95.3 \pm 2.4	96.4 \pm 1.3	94.9 \pm 1.8	95.9
Virchow	96.5 \pm 1.0	92.5 \pm 3.6	96.2 \pm 0.6	91.7 \pm 1.1	94.2
Virchow2	97.4 \pm 1.0	95.4 \pm 1.1	96.5 \pm 1.1	94.5 \pm 2.6	95.9
Prov-GigaPath	96.6 \pm 0.6	94.5 \pm 0.5	96.3 \pm 0.6	95.0 \pm 1.3	95.6

Table 4. Mean classification accuracy over five folds: (1) Aperio training data (ID) with VisionTek as OOD data and (2) VisionTek training data (ID) with Aperio as OOD data. The table also reports the overall mean accuracy across all ID and OOD scenarios.

achieve the highest performance, the performance across ID and OOD datasets is more consistent, suggesting higher model reliability.

TOST was completed to investigate whether the performance on ID and OOD datasets were statistically equivalent which may be used as a proxy for generalization. The equivalence margin is automatically computed

	AP as ID, VT as OOD	VT as ID, AP as OOD	
Models	ID - OOD	ID - OOD	Mean
HistoLite (ours)	0.35	2.14	1.25
KimiaNet	15.07	7.41	11.24
PathDino	8.50	4.16	6.33
HIPT	0.78	2.20	1.49
IBOT-Path	1.32	3.96	2.64
Hibou-B	4.88	1.93	3.41
UNI	1.64	1.47	1.55
Virchow	3.99	4.55	4.27
Virchow2	1.99	1.99	1.99
Prov-GigaPath	2.09	1.38	1.74

Table 5. Difference in mean classification accuracy between ID and OOD: (1) Aperio training data (ID) with VisionTek as OOD, and (2) VisionTek training data (ID) with Aperio as OOD. The table also reports the mean difference across both scenarios.

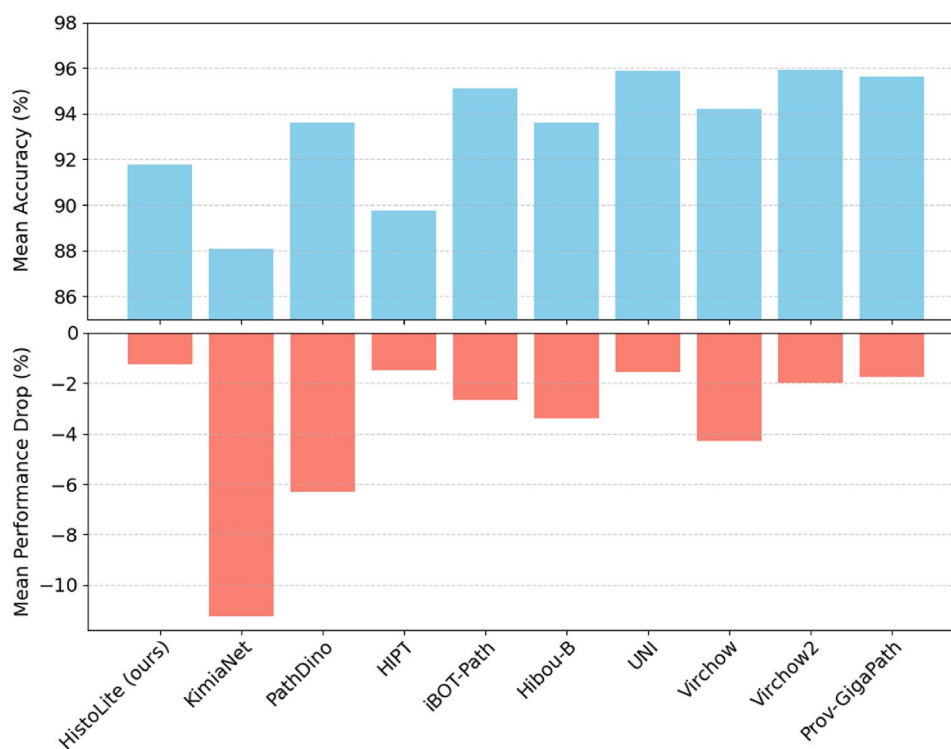


Figure 7. Mean accuracy and performance drop. Mean accuracies across all ID and OOD evaluation scenarios, along with the mean performance drop for the corresponding OOD datasets. The top bar chart corresponds to the mean values in Table 4, and the bottom bar chart corresponds to the mean values in Table 5.

as proposed by Varnava et al.⁷⁷. The tissue classification accuracy for both training setups, as well as the TOST analysis with the data-driven bounds are shown in Fig. 8. HistoLite has low mean differences and variance, and also is contained within the equivalence bounds indicating that HistoLite is generalizing across scanners. In contrast, SOTA FMs such as KimiaNet, PathDino, and Virchow exceed the bound in both training scenarios, indicating these models experience significant performance drops on the OOD dataset.

To investigate the relationship between model performance and differences in embeddings, classification accuracy is correlated to MAE representation shift and RI, with circle sizes representing the model size, as shown in Fig. 9. The performance difference between ID and OOD data widens as the MAE representation shift increases over all models. Despite being the smallest model, HistoLite exhibits a minimal performance drop with minimal representation shift across scanners (and modest classification performance). HIPT, which has approximately three times the parameters of HistoLite and is trained on 100 million more patches, exhibits comparable results. However, considering the downstream task of tissue classification, HistoLite outperforms HIPT. Other FMs, which have considerably larger MAE and performance differences (but with top classification accuracy), include

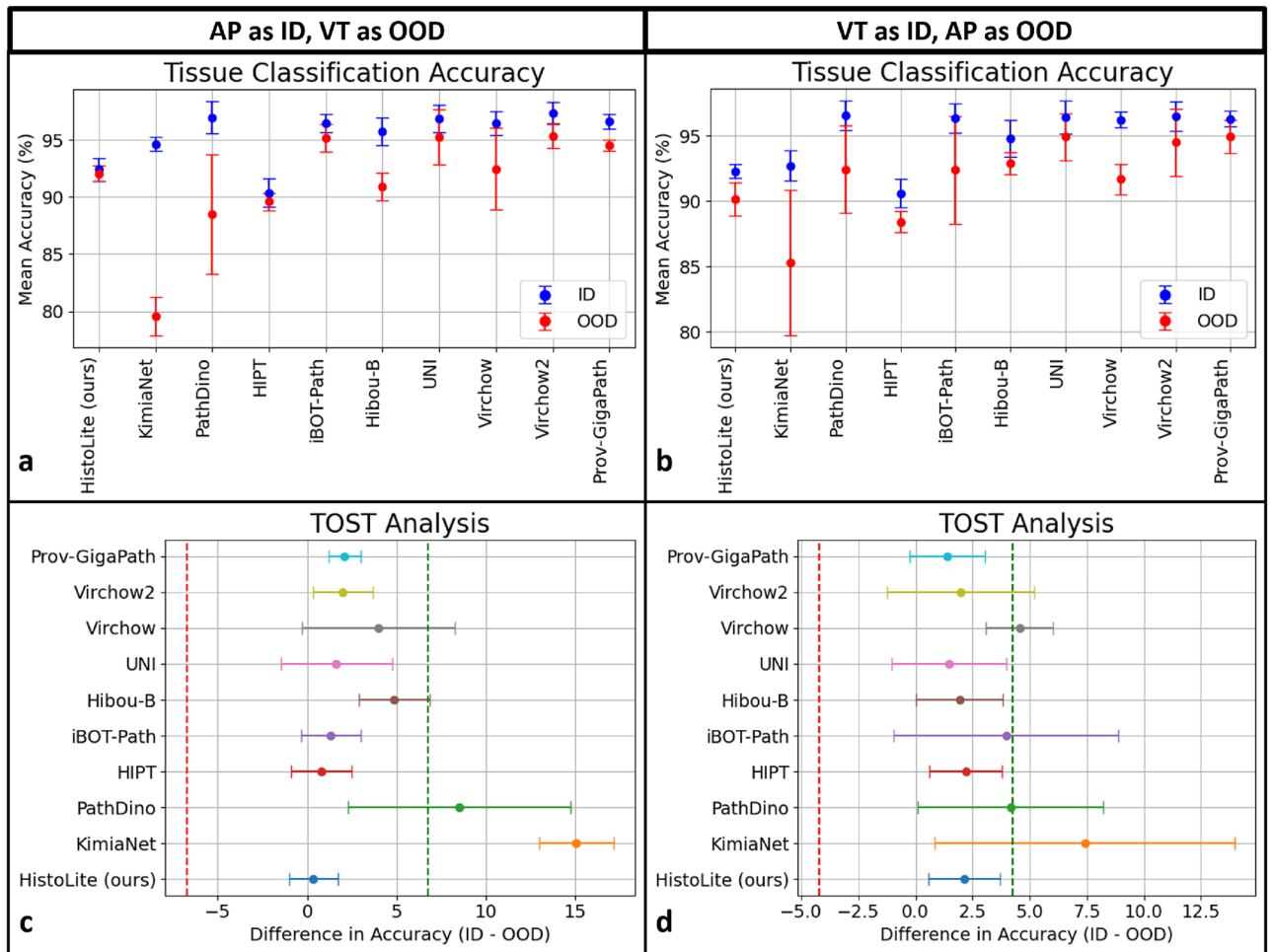


Figure 8. Cross-validation. Five-fold cross-validation accuracy results for all the models. (a) and (b) present the mean accuracy and standard deviation when Aperio and VisionTek were used as the ID datasets, respectively. (c) and (d) present the TOST analysis.

Prov-GigaPath, Virchow2, UNI and iBOT-Path. KimiaNet has both the largest performance drop and largest MAE representation shift. Additionally, Fig. S4 in the supplementary file illustrates the relationship between performance differences and both cosine distance and KL divergence, along with trend lines for tumor, normal, and mixed tumor-normal cases.

Figure 9c,d demonstrates the relationship between the ID to OOD accuracy difference, to the RI. Prov-GigaPath has the highest robustness across both Aperio and VisionTek scanners, with low drop in performance. HistoLite is operating with a moderate RI and low performance drop. KimiaNet has low RI and the highest accuracy difference indicating poor generalization. HistoLite, positioned in the middle, demonstrates a balanced approach reflecting its ability to maintain a stable latent representation and consistency between ID and OOD data, with modest classification performance but greater efficiency. HIPT is not included in this figure due to its exceptionally high RI. However, this highlights that a single metric does not capture the complete picture, and it is essential to consider additional metrics for a comprehensive evaluation. Cosine distance and KL divergence as metrics for representation shift, are provided in Fig. S4 of the n materials.

Discussion

Domain variation in digital pathology, arising from differences in scanners can lead to performance degradation and poses a significant challenge to the generalization of machine learning models. Recent large-scale FMs have tried to solve this but a recent study conducted by Mulliqi et al. raises concerns regarding the purported generalizability benefits of the FMs.¹⁹ It was also found that FMs consume 35 times more energy and significantly more data compared to conventional task-specific models.

In this study, we present HistoLite, a lightweight SSL framework designed to achieve domain-generalized representation learning in histopathology. Unlike resource-intensive FMs, HistoLite is optimized to operate efficiently on modest computational setups, such as a personal GPU computer. The framework incorporates a customizable autoencoder architecture, which can be tailored to fit the computational resources available, enhancing its adaptability.

Accuracy Difference (ID – OOD) vs Representation Shift

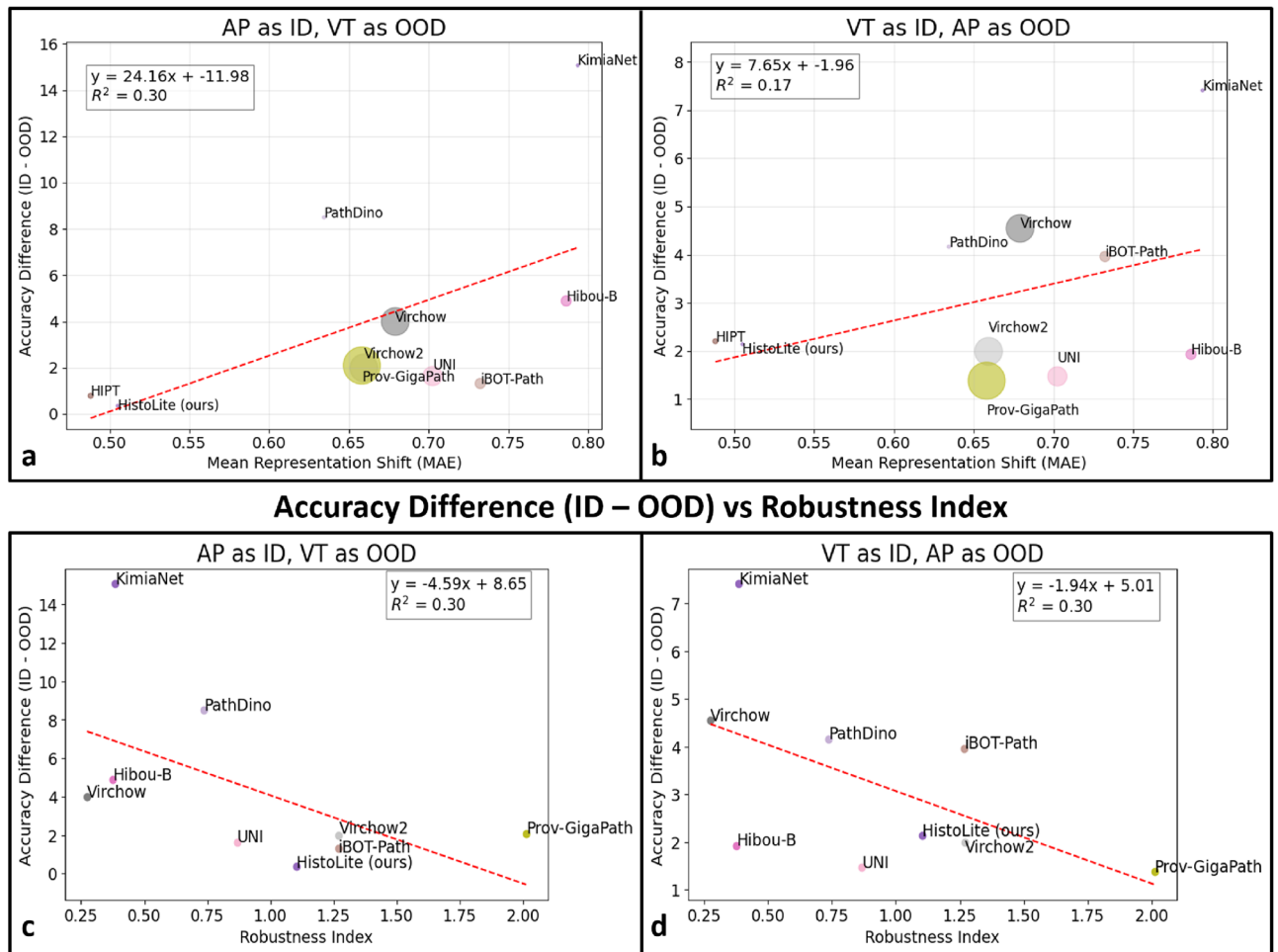


Figure 9. Performance vs representation shift/robustness index. Relationship between the average performance difference between ID and OOD against the representation shift is illustrated in panels (a) and (b), while the relationship with the robustness index is depicted in panels (c) and (d). In (a) and (b) circle size represents the model's parameter count, providing a visual comparison of model scale and the red line shows the trend.

A dataset was created from the same tissue slides from two different scanners to study the variability arising from scanner differences (covariate shift). Representation shift was measured at the embedding level across scanners, which is a novel approach, and the first work of its kind. Results can further our understanding of model generalizability. Model performance was further validated through a downstream classification task and correlated to representation shift. HistoLite exhibits the smallest (ID – OOD) performance drop, the second-lowest representation shift, and moderate classification accuracy. This performance profile is likely due to its smaller model size, which provides a favorable balance between accuracy and generalization. Although its downstream classification accuracy is moderate, HistoLite achieves the strongest cross-scanner generalization, as indicated by the reduced (ID – OOD) performance gap and further supported by the TOST analysis results.

It was shown that HIPT has great similarity in embeddings across scanners, but classification performance using the fine-tuned model is one of the lowest. The clustering metrics showed good overlap across scanners, and separation between tissue classes, but on the downstream task it did not perform as well. However, the generalization analysis using TOST showed that while the performance was lower, it was consistent across scanners. These results for HistoLite and HIPT suggest that perhaps there is a tradeoff between accuracy and generalization in smaller models. It may be possible to generate robust representations that are more consistent across scanners, but this may come with a reduction of performance.

The top-performing models in terms of classification accuracy are Prov-GigaPath, Virchow2, UNI, and iBOT-Path, each outperforming HistoLite by at least 4%. Among these, UNI achieves the smallest (ID – OOD) performance difference (1.55%), followed by Prov-GigaPath (1.74%) and Virchow2 (1.99%), while iBOT-Path exhibits the largest gap among the top performers (2.64%). These are all ViT-based models, trained on multiple organs and large datasets, which could indicate these configurations are more optimal for top classification accuracy and generalization. In contrast, KimiaNet demonstrates the lowest accuracy and the poorest

generalization, with a mean accuracy of 88.1% and the largest mean performance drop of 11.24%. KimiaNet is a relatively small CNN-based model trained using supervision learning. These factors may contribute to the reduced generalization and classification performance.

Our analysis reveals several important insights into the relationship between model characteristics, representation similarity, and cross-scanner performance. First, we observe that a small representation shift (i.e., high similarity between ID and OOD embeddings) does not necessarily guarantee superior classification accuracy; rather, it tends to be associated with a smaller performance gap between ID and OOD evaluations. This is an important observation as it may suggest that there is a trade-off between representation robustness, and task-specific accuracy. This trend is illustrated in Fig. 5, which correlates representation shift with performance drop. The models with the smallest performance drop also had small representation shift but not necessarily the top classification accuracy. The larger and top performing models had modest performance drop and representation shift. Similarly, a high RI does not inherently translate to the best performance. For example, while HIPT achieves the highest RI, it does not yield the top accuracy. These findings suggest that consistent performance is achieved through better representation similarity and RI, but more investigation is required to ensure that these models also perform well on downstream tasks. The relationship between accuracy and generalization should be studied further, and models that can achieve both are an interesting avenue of future research.

Contrary to the common assumption that larger models inherently deliver better performance and generalization, our results show this is not always the case. For instance, UNI, which is smaller than Virchow/Virchow2 and outperforms Virchow in terms of generalizability, representation shift, and performance drop. Likewise, Virchow and Virchow2 share the same architecture, yet Virchow2 achieves better generalization and classification accuracy, which may be due to the multi-resolution training and broader diversity of organ types in its training data. Interestingly, Prov-GigaPath, the largest model in our evaluation, performs comparably to smaller models such as Virchow2 and UNI, and shows only a slight advantage in terms of representation shift, robustness index and accuracy.

Model comparisons further support these observations. Hibou-B and iBOT-Path are similar in size, yet iBOT-Path exhibits stronger generalization, as evidenced by its smaller performance drop and reduced representation shift. Among smaller models, HistoLite and KimiaNet have comparable sizes, while PathDino is slightly larger. HistoLite demonstrates superior cross-scanner generalization, as indicated by both its lower representation shift and smaller performance drop, whereas PathDino achieves higher classification accuracy (by approximately 1%) but exhibits a larger performance drop (6.33%). Therefore, this slight gain in classification accuracy is coming at the expense of scanner-robustness. In contrast, KimiaNet shows the weakest performance overall, likely due to its supervised training paradigm and CNN-based architecture, compared to the self-supervised, transformer-based approaches used by the other evaluated models. Notably, while HistoLite also employs a CNN-based encoder, its self-supervised training enables it to outperform KimiaNet.

A limitation of this work may be the differences in input patch sizes between HistoLite and the evaluated FMs. HistoLite operates on 512×512 patches, while most FMs, including ViT-based architectures, use 224×224 patches at $20\times$ magnification. We chose to employ these models in their off-the-shelf form without fine-tuning or altering their patch size. It was not our intention to optimize the performance of each model individually, but rather to apply them as they were designed. To ensure a fair comparison across models, the pixel resolution of all patches was the same for all models (which may be just as important as the patch-size). Harmonizing patch sizes in the future could lead to more standardized results for isolating further architectural differences.

Going forward, we aim to explore the proposed HistoLite framework with alternative architectures, such as ViTs, to investigate how the attention mechanisms in ViTs can learn domain-invariant features using the proposed SSL framework. This could provide insights into whether attention-based models further enhance the framework's ability to generalize across domains. To expand the findings more broadly, the framework could be evaluated on datasets from various organs as this study focused exclusively on breast tissue and may not fully represent other anatomies. However, since the same tissue slides were acquired by both scanners, perhaps the remaining variability was related to the scanner-bias alone. Expanding the evaluation to include datasets from organs such as lung, colon, and prostate will enable a more comprehensive assessment of generalizability, robustness, and fairness across diverse biological and clinical contexts.

Another potential future direction involves refining the representation alignment approach. Specifically, we propose replacing the current predictor, which is based on a Siamese network²⁶, with a DINO head^{27,28}. By incorporating the DINO head, it would be possible to examine how it aligns features and whether it enhances the performance of the framework. However, this approach would introduce a higher number of learnable parameters, which could impact the computational efficiency and may require careful resource management during training.

Conclusion

Developing models that generalize across scanners is critical for translation of computational pathology algorithms. FMs have been developed as tools to overcome these challenges. However, the training of large FMs is often inaccessible due to limitations in data availability and computational resources. Additionally, the generalization abilities of the models across different scanning platforms is largely understudied. To address these challenges, we propose HistoLite: a lightweight, self-supervised representation learning framework that is resource-efficient, customizable and provides robustness to covariate shifts. HistoLite is designed to be trained on a personal GPU, making it accessible for researchers and institutions with limited resources. Due to the feature alignment of augmented images, and dual-stream autoencoder, the framework is designed to be robust to data variability.

The experimental results demonstrate that HistoLite achieves performance comparable to larger models trained on millions of patches using extensive GPU clusters. Moreover, HistoLite has consistent performance across scanners, highlighting its robustness. This framework offers a practical solution for smaller healthcare facilities or research units with limited data and computational capabilities. Depending on the available resources and domain requirements, one must carefully consider whether to develop a lightweight, domain-specific model with scanner robustness or opt for a larger model with slightly higher accuracy that demands significantly greater resources and data.

Data availability

CAMELYON16/17 data is available at: “<https://camelyon17.grand-challenge.org/Data/>”. TCGA-BC data is available at: “<https://portal.gdc.cancer.gov/>”. HEROHE data is available at: “<https://ecdp2020.grand-challenge.org/Dataset/>”. The private datasets generated during and/or analyzed during the current study are not publicly available due to privacy concerns and ethical considerations. The source code for the HistoLite framework is publicly accessible at <https://github.com/IAMLAB-Ryerson/HistoLite>.

Received: 4 June 2025; Accepted: 15 September 2025

Published online: 21 October 2025

References

- Shafique, A., Babaie, M., Sajadi, M., Batten, A., & Skdar, S. Automatic multi-stain registration of whole slide images in histopathology. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3622–3625 (2021). IEEE
- Haan, K. et al. Deep learning-based transformation of h & e stained tissues into special stains. *Nat. Commun.* **12**(1), 4884 (2021).
- Zhang, D.Y., Venkat, A., Khasawneh, H., Sali, R., Zhang, V., & Pei, Z. Implementation of digital pathology and artificial intelligence in routine pathology practice. *Lab. Investig.* 102111 (2024).
- Dunn, C., Brettle, D., Hodgson, C., Hughes, R., & Treanor, D. An international study of stain variability in histopathology using qualitative and quantitative analysis. *J. Pathol. Inf.* 100423 (2025).
- Cui, M. & Zhang, D. Y. Artificial intelligence and computational pathology. *Lab. Invest.* **101**(4), 412–422 (2021).
- Dy, A. et al. Ai improves accuracy, agreement and efficiency of pathologists for ki67 assessments in breast cancer. *Sci. Rep.* **14**(1), 1283 (2024).
- Angeloni, M. et al. Closing the gap in the clinical adoption of computational pathology: a standardized, open-source framework to integrate deep-learning models into the laboratory information system. *Genome Med.* **17**(1), 60 (2025).
- Zamanitajeddin, N., Jahanifar, M., Xu, K., Siraj, F., & Rajpoot, N. Benchmarking domain generalization algorithms in computational pathology. arXiv preprint [arXiv:2409.17063](https://arxiv.org/abs/2409.17063) (2024).
- Jahanifar, M., Raza, M., Xu, K., Vuong, T.T.L., Jewsbury, R., Shephard, A., Zamanitajeddin, N., Kwak, J.T., Raza, S.E.A., Minhas, F., et al. Domain generalization in computational pathology: Survey and guidelines. *ACM Comput. Surveys* (2023).
- Stacke, K., Eilertsen, G., Unger, J. & Lundström, C. Measuring domain shift for deep learning in histopathology. *IEEE J. Biomed. Health Inform.* **25**(2), 325–336 (2020).
- Shen, Y., Luo, Y., Shen, D., Ke, J.: 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*, pp. 212–221. Springer (2022).
- Hosseini, S.M., Shafique, A., & Babaie, M. Class-imbalanced unsupervised and semi-supervised domain adaptation for histopathology images. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–7 (2023). IEEE.
- Asadi-Aghbolaghi, M. et al. Learning generalizable ai models for multi-center histopathology image classification. *NPJ Precis. Oncol.* **8**(1), 151 (2024).
- Waqas, A. et al. Revolutionizing digital pathology with the power of generative artificial intelligence and foundation models. *Lab. Invest.* **103**(11), 100255 (2023).
- Xu, H., Suyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* 1–8 (2024).
- Tizhoosh, H. R. & Pantanowitz, L. Artificial intelligence and digital pathology: challenges and opportunities. *J. Pathol. Inf.* **9**(1), 38 (2018).
- Alfasly, S. et al. Foundation models for histopathology—fanfare or flair. *Mayo Clin. Proc. Digit. Health* **2**(1), 165–174 (2024).
- Alfasly, S., Nejat, P., Hemati, S., Khan, J., Lahr, I., Alsaafin, A., Shafique, A., Comfere, N., Murphree, D., Meroueh, C., et al. When is a foundation model a foundation model. arXiv preprint [arXiv:2309.11510](https://arxiv.org/abs/2309.11510) (2023).
- Mulliqi, N., Blilie, A., Ji, X., Szolnoky, K., Olsson, H., Boman, S.E., Titus, M., Gonzalez, G.M., Mielcarz, J.A., Valkonen, M., et al. Foundation models—a panacea for artificial intelligence in pathology? arXiv preprint [arXiv:2502.21264](https://arxiv.org/abs/2502.21264) (2025).
- Li, P., Pei, Y. & Li, J. A comprehensive survey on design and application of autoencoder in deep learning. *Appl. Soft Comput.* **138**, 110176 (2023).
- Roy, M. et al. Convolutional autoencoder based model histocae for segmentation of viable tumor regions in liver whole-slide images. *Sci. Rep.* **11**(1), 139 (2021).
- Xu, J. et al. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *IEEE Trans. Med. Imaging* **35**(1), 119–130 (2015).
- Ashraf, A., Nawari, N.M., Shahzad, T., Aamir, M., Khan, M.A., & Ouahada, K. Dimension reduction using dual-featured auto-encoder for the histological classification of human lungs tissues. *IEEE Access* (2024).
- Alsaafin, A., Shafique, A., Alfasly, S., & Tizhoosh, H. Personalized 2d binary patient codes of tissue images and immunogenomic data through multimodal self-supervised fusion. arXiv preprint [arXiv:2409.13115](https://arxiv.org/abs/2409.13115) (2024).
- Chen, T., Kornblith, S., Swersky, K., Norouzi, M. & Hinton, G. E. Big self-supervised models are strong semi-supervised learners. *Adv. Neural. Inf. Process. Syst.* **33**, 22243–22255 (2020).
- Chen, X., & He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758 (2021).
- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660 (2021).
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. arXiv preprint [arXiv:2304.07193](https://arxiv.org/abs/2304.07193) (2023).
- Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholz, N., Fusi, N., et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med.* 1–12 (2024).
- Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**(3), 850–862 (2024).

31. Nechaev, D., Pchel'nikov, A., & Ivanova, E. Hibou: A family of foundational vision transformers for pathology. arXiv preprint [arXiv:2406.05074](https://arxiv.org/abs/2406.05074) (2024).
32. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Fuchs, T., Fusi, N., et al. Virchow 2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint [arXiv:2408.00738](https://arxiv.org/abs/2408.00738) (2024).
33. Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Mac Kain, A., Saillard, C., & Schiratti, J.-B. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023–07 (2023).
34. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., & Kong, T. ibot: Image bert pre-training with online tokenizer. arXiv preprint [arXiv:2111.07832](https://arxiv.org/abs/2111.07832) (2021).
35. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., & Mahmood, F. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16144–16155 (2022).
36. Alfasly, S., Shafique, A., Nejat, P., Khan, J., Alsaafin, A., & Alabtah, G. Rotation-agnostic image representation learning for digital pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11683–11693 (2024).
37. Riasatian, A. et al. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Med. Image Anal.* **70**, 102032 (2021).
38. Faryna, K., Laak, J. & Litjens, G. Automatic data augmentation to improve generalization of deep learning in h & e stained histopathology. *Comput. Biol. Med.* **170**, 108018 (2024).
39. Koch, L. M., Baumgartner, C. F. & Berens, P. Distribution shift detection for the postmarket surveillance of medical ai algorithms: a retrospective simulation study. *NPJ Digit. Med.* **7**(1), 120 (2024).
40. Wang, J. et al. Generalizing to unseen domains: A survey on domain generalization. *IEEE Trans. Knowl. Data Eng.* **35**(8), 8052–8072 (2022).
41. Zhou, K., Liu, Z., Qiao, Y., Xiang, T. & Loy, C. C. Domain generalization: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(4), 4396–4415 (2022).
42. Guan, H. & Liu, M. Domain adaptation for medical image analysis: a survey. *IEEE Trans. Biomed. Eng.* **69**(3), 1173–1185 (2021).
43. Blanchard, G., Lee, G., & Scott, C. Generalizing from several related classification tasks to a new unlabeled sample. *Adv. Neural Inf. Process. Syst.* **24** (2011).
44. Muandet, K., Balduzzi, D., & Schölkopf, B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR (2013).
45. Vahadane, A. et al. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans. Med. Imaging* **35**(8), 1962–1971 (2016).
46. Khan, A. M., Rajpoot, N., Treanor, D. & Magee, D. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Trans. Biomed. Eng.* **61**(6), 1729–1738 (2014).
47. Goodfellow, I. et al. Generative adversarial networks. *Commun. ACM* **63**(11), 139–144 (2020).
48. Mahmood, F. et al. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Trans. Med. Imaging* **39**(11), 3257–3267 (2019).
49. Chlap, P. et al. A review of medical image data augmentation techniques for deep learning applications. *J. Med. Imaging Radiat. Oncol.* **65**(5), 545–563 (2021).
50. Li, P., Li, D., Li, W., Gong, S., Fu, Y., & Hospedales, T.M. A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8886–8895 (2021).
51. Zaffar, I., Jaume, G., Rajpoot, N., & Mahmood, F. Embedding space augmentation for weakly supervised learning in whole-slide images. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4 (2023). IEEE.
52. Chikontwe, P., Nam, S. J., Go, H., Kim, M., Sung, H. J., & Park, S. H. Feature re-calibration based multiple instance learning for whole slide image classification. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 420–430 (2022). Springer.
53. Ilse, M., Tomczak, J. M., Louizos, C., & Welling, M. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pp. 322–348 (2020). PMLR.
54. Wagner, S. J., Khalili, N., Sharma, R., Boxberg, M., Marr, C., De Back, W., & Peng, T. Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 257–266 (2021). Springer.
55. Zhang, B., Knudsen, B., Sirohi, D., Ferrero, A., & Tasdizen, T. Stain based contrastive co-training for histopathological image analysis. In *Workshop on Medical Image Learning with Limited and Noisy Data*, pp. 106–116 (2022). Springer.
56. Finn, C., Abbeel, P., & Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135 (2017). PMLR.
57. Hospedales, T., Antoniou, A., Micaelli, P. & Storkey, A. Meta-learning in neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(9), 5149–5169 (2021).
58. Han, C. et al. Meta multi-task nuclei segmentation with fewer training samples. *Med. Image Anal.* **80**, 102481 (2022).
59. Liu, Y., Wang, W., Ren, C.-X., & Dai, D.-Q. Metacon: meta contrastive learning for microsatellite instability detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 267–276 (2021). Springer.
60. Peyret, R., Bouridane, A., Khelifi, F., Tahir, M. A. & Al-Maadeed, S. Automatic classification of colorectal and prostatic histologic tumor images using multiscale multispectral local binary pattern texture features and stacked generalization. *Neurocomputing* **275**, 83–93 (2018).
61. Kotte, S., Saipradeep, V., Sivadasan, N., Joseph, T., Sharma, H., Walia, V., Varma, B., & Mukherjee, G. A deep learning based ensemble model for generalized mitosis detection in h & e stained whole slide images. In *MICCAI Challenge on Mitosis Domain Generalization*, pp. 221–225. Springer (2022).
62. Zolfaghari, B., Mirsadeghi, L., Bibak, K. & Kavousi, K. Cancer prognosis and diagnosis methods based on ensemble learning. *ACM Comput. Surv.* **55**(12), 1–34 (2023).
63. Silva, J. et al. Boundary-aware glomerulus segmentation: toward one-to-many stain generalization. *Comput. Med. Imaging Graph.* **100**, 102104 (2022).
64. Javed, S., Mahmood, A., Dias, J., Werghi, N. & Rajpoot, N. Spatially constrained context-aware hierarchical deep correlation filters for nucleus detection in histology images. *Med. Image Anal.* **72**, 102104 (2021).
65. Tang, S., Hosseini, M.S., Chen, L., Varma, S., Rowsell, C., Damaskinos, S., Plataniotis, K.N., & Wang, Z. Probeable darts with application to computational pathology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 572–581 (2021).
66. Anand, D., Patel, G., Dang, Y., & Sethi, A. Switching loss for generalized nucleus detection in histopathology. arXiv preprint [arXiv:2008.03750](https://arxiv.org/abs/2008.03750) (2020).
67. Santos, C. F. G. D. & Papa, J. P. Avoiding overfitting: A survey on regularization methods for convolutional neural networks. *ACM Comput. Surveys (Csur)* **54**(10s), 1–25 (2022).
68. Minhas, F., Toss, M.S., Wahab, N., Rakha, E., & Rajpoot, N.M. L1-regularized neural ranking for risk stratification and its application to prediction of time to distant metastasis in luminal node negative chemotherapy naïve breast cancer patients. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 390–400 (2021). Springer.
69. Liang, M. et al. Interpretable classification of pathology whole-slide images using attention based context-aware graph convolutional neural network. *Comput. Methods Programs Biomed.* **229**, 107268 (2023).

70. Huang, Z., Wang, H., Xing, E. P., & Huang, D. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pp. 124–140 (2020). Springer.
71. Elphick, M., Turajlic, S., & Yang, G. Are the latent representations of foundation models for pathology invariant to rotation? arXiv preprint [arXiv:2412.11938](https://arxiv.org/abs/2412.11938) (2024).
72. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607 (2020). PmlR.
73. Bejnordi, B. E. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**(22), 2199–2210 (2017).
74. Litjens, G. et al. 1399 h &e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience* **7**(6), 065 (2018).
75. Conde-Sousa, E. et al. Herohe challenge: predicting her2 status in breast cancer from hematoxylin-eosin whole-slide imaging. *J. Imaging* **8**(8), 213 (2022).
76. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemp. Oncol.* **2015**(1), 68–77 (2015).
77. Varnava, Y. et al. Out-of-distribution generalization for segmentation of lymph node metastasis in breast cancer. *Sci. Rep.* **15**(1), 1127 (2025).

Author contributions

The study represents a collaborative effort among seven coauthors: Abubakr Shafique, Amanda Dy, Xiaoli Qin, Najd Alshamlan, Dimitrios Androustos, Susan J. Done, and April Khademi. Abubakr Shafique led the conceptualization, methodology development, implementation, and validation of the research. Amanda Dy, Xiaoli Qin, Dimitrios Androustos, and April Khademi supported data acquisition and validation. Najd Alshamlan and Susan J. Done curated the histopathological datasets. The manuscript was initially drafted by Abubakr Shafique, with editorial guidance and final revisions led by April Khademi, who also supervised the overall project.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-20411-0>.

Correspondence and requests for materials should be addressed to A.S. or A.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025