Quantifying and Mitigating Hospital Domain Bias in Pathology Foundation Models using Adversarial Feature Disentanglement

Mengliang Zhang[1] mxz3935@mavs.uta.edu

[1] *CSE Department, The University of Texas at Arlington*

Editors: Under Review for MIDL 2026

Abstract

Pathology foundation models (PFMs) have demonstrated remarkable potential in whole-slide image (WSI) diagnosis. However, pathology images from different hospitals exhibit domain shifts due to variations in scanning hardware and preprocessing. These differences cause PFMs to learn spurious hospital-specific features, severely compromising their robustness and generalizability in clinical settings. We present the first systematic study of this hospital-source domain bias in PFMs. To address the critical trade-off between diagnostic utility and domain predictability, we establish a quantification pipeline and introduce the Robustness Index (RI). Furthermore, we propose a lightweight adversarial framework for feature disentanglement. This framework employs a trainable adapter and a domain classifier connected via a Gradient Reversal Layer (GRL) to remove latent hospital-specific information from frozen PFM representations without modifying the encoder itself. Experiments on multi-center histopathology datasets demonstrate that our approach substantially suppresses domain predictability and achieves significant gains in feature robustness. Crucially, the method maintains or improves disease classification performance, proving its efficacy particularly in out-of-domain scenarios. Our code is provided at: https://github.com/ana-com-00/anonymous_repo

Keywords: Pathology Image, Domain Bias, Foundation model.

1. Introduction

Whole-slide imaging (WSI) has become a key tool in digital pathology, enabling scalable analysis of gigapixel histopathology slides. In most downstream applications, WSIs are partitioned into small patches, and deep models—such as ResNet (He et al.), Vision Transformer (ViT) (Dosovitskiy et al.), or foundation encoders like UNI (Chen et al.), PLIP (Huang et al.) are employed to extract visual features for tasks including disease classification, tumor grading, and subtype analysis. These patch-level features form the backbone of many computational pathology pipelines.

However, a critical yet often underexplored challenge in this setting is the presence of domain-specific bias in the data. Patches collected from different hospitals or scanners frequently differ in staining protocols, image resolution, scanner artifacts, and tissue preparation. Such variations introduce spurious correlations into the learned representations, causing models to inadvertently rely on hospital-specific cues rather than true disease-related signals, severely compromising their robustness and generalizability in clinical settings. Even state-of-the-art pathology foundation models (PFMs), such as Phikon

(Filiot et al.) and UNI (Chen et al.), exhibit a significant degree of domain bias. Visual confirmation of this phenomenon is provided in Figure 3, where t-SNE (Van der Maaten and Hinton, 2008) clustering of patch embeddings from the TCGA-BRCA (The Cancer Genome Atlas Network, 2012) dataset clearly shows that patches from the same hospital source cluster tightly together. This indicates the learned features inadvertently retain substantial hospital-specific information.

Several prior studies have explored domain bias in PFMs. Vaidya et al. (Vaidya et al., 2024) examined performance disparities across patient populations, while Lin et al. investigated inter-hospital variations. de Jong et al. (de Jong et al., 2025) proposed a robustness metric using a KNN-based approach. Other studies attempted to mitigate domain shifts through non-learnable stain normalization techniques, such as Macenko (M. Macenko et al., 2009) normalization, which only address low-level color differences. Another line of work leverages parameter-efficient fine-tuning methods, such as LoRA (Low-Rank Adaptation) (Hu et al., 2021), to adapt PFMs to out-of-distribution (OOD) data. Yet, these approaches have notable limitations: the KNN-based metric lacks a comprehensive dataset-level assessment; stain normalization fails to resolve complex, non-stain-related domain shifts; and LoRA fine-tuning approaches do not explicitly ensure the reduction of domain bias while preserving diagnostic performance. Thus, there is a clear need for a systematic, parameter-efficient framework that can rigorously quantify and explicitly mitigate domain bias while guaranteeing the preservation of diagnostic performance.

In this work, a systematic pipeline is established to evaluate PFM domain bias, using separate Multi-Layer Perceptron (MLP) training to quantify severity via hospital classification AUC. We introduce the Robustness Index to quantify the utility-predictability trade-off. To mitigate bias, we propose a lightweight adversarial framework utilizing a trainable projection head, disease classifier, and a domain classifier connected by a Gradient Reversal Layer (GRL). During backpropagation, the GRL reverses the gradient, suppressing hospital-related cues while preserving disease information. Experimental results show the framework effectively suppresses hospital-specific signals, alleviating domain bias, and achieving substantial robustness gains on multi-center datasets

Our main contributions are as follows: 1. We establish a systematic pipeline to identify and quantify hospital-source domain bias in PFM features, introducing the Robustness Index (RI) to assess the net utility of the learned representations. 2. We propose a lightweight adversarial training framework that incorporates a Gradient Reversal Layer (GRL) to suppress hospital-related discriminative cues in image features without modifying the base PFM encoder, ensuring parameter efficiency and core feature preservation. 3. Experiments on multi-center histopathology datasets demonstrate that our method significantly suppresses latent hospital information while maintaining or enhancing disease classification performance, achieving substantial gains in robustness (quantified by ΔRI).

2. Related Works

2.1. Pathology Foundation Models (PFMs)

The initial success of general vision models like CLIP (Radford et al., 2021) and DINO (Caron et al., 2021) in pathology paved the way for models specifically optimized for the domain. More recently, several dedicated Pathology Foundation Models (PFMs) have been proposed, including UNI (Chen et al.), CONCH (Lu et al., 2024), GIGA_PATH (Xu et al., 2024), and VIRCHOW (Vorontsov et al., 2024). These models, typically pretrained on large-scale natural image corpora or multimodal datasets, have demonstrated impressive zero-shot and few-shot capabilities in various pathology tasks, such as tumor classification, subtyping, and grading. They are commonly employed as frozen encoders to extract highly transferable, high-dimensional features from image patches. Despite their strong semantic capabilities, studies consistently observe that the features extracted by these models still encode significant dataset- or site-specific biases, including scanner artifacts, staining variations, and patient demographics.

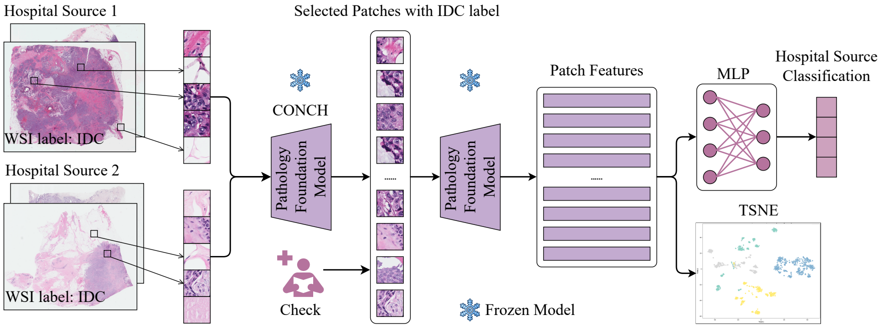

Hospital Source 1
WSI label: IDC
Hospital Source 2
WSI label: IDC
Selected Patches with IDC label
CONCH
Pathology Foundation Model
Check
Pathology Foundation Model
Patch Features
MLP
Hospital Source Classification
TSNE
Frozen Model

Figure 1: Pipeline for evaluating hospital-domain bias in pathology foundation models. Only patches consistent with their WSI labels are used. A simple multi-layer perceptron (MLP) is trained to classify hospital sources, where higher accuracy indicates stronger domain bias.

2.2. Domain Bias Mitigation and Disentanglement

Prior work has approached hospital-domain bias from various angles (M. Macenko et al., 2009; A. Vahadane et al., 2016). Some methods focus on pixel-space preprocessing, such as stain normalization, to alleviate color variations (M. Macenko et al., 2009). In the feature space, techniques like Edwin et al.'s (de Jong et al., 2025) proposed KNN-based metric offered a measure of bias correlation, but lacked a systematic mitigation mechanism. Vaidya et al. (Vaidya et al., 2024) introduced LoRA-based fine-tuning to adapt Pathology Foundation Models (PFMs) to out-of-distribution (OOD) data. However, approaches like LoRA primarily focus on parameter-efficient adaptation rather than explicit feature disentanglement, while stain normalization fails to address complex, non-stain-related domain shifts.

Our approach leverages adversarial feature learning via the Gradient Reversal Layer (GRL) mechanism, a technique proven effective in general image domain generalization. The GRL was originally introduced by Ganin et al. (Ganin et al., 2017) in the context of Domain-Adversarial Neural Networks (DANN) for general domain adaptation, and we adapt this proven technique to feature disentanglement in pathology. Unlike prior pathology work that focuses solely on empirical quantification or domain-specific adaptation, our framework provides a unified, lightweight, and PFM-agnostic solution for systematic quantification (RI) and adversarial disentanglement of frozen PFM features.

3. Methodology

3.1. WSI Cohort Construction and Patch Extraction

For a given WSI collection, we first construct a clinical cohort by filtering WSIs based on available demographic information (e.g., sex and age) to minimize the potential confounding effects of demographic factors on WSI disease features. To ensure fair assessment of domain shift, we select a fixed number of WSIs per disease category and hospital source, striving to maintain balanced domain and task distributions.

For each WSI, we follow a standard procedure for feature extraction. We perform tissue segmentation using Otsu thresholding to automatically generate a tissue mask, consistent with the strategy used in CLAM (Lu et al., 2021). Within the identified tissue regions, fixed-size patches are extracted via a sliding window with a predefined grid pattern. Each patch subsequently undergoes quality control, including checks for minimum effective tissue area, ensuring the retained patches possess sufficient diagnostic value. Process details are provided in Appendix A.

3.2. Feature Filtering for High-Fidelity Patches

The extracted patch set is further refined to ensure consistency between patch-level content and WSI-level labels. This step is critical because WSI labels are case-level, and randomly sampled patches may not always reflect the WSI-level diagnosis (e.g., containing normal tissue). We leverage the CONCH (Lu et al., 2024) model, primarily motivated by its strong zero-shot classification capabilities to accurately verify patch content and ensure high-fidelity labeling. We perform zero-shot patch classification using CONCH to obtain the top-1 predicted label and its probability. We retain only those patches where (1) the predicted label probability exceeds a predefined threshold (e.g., 0.8) and (2) the patch-level label matches the WSI-level ground truth. The high confidence threshold is chosen to maximize the reliability of the patch-level content. A subsequent manual inspection of sampled patches is additionally performed to ensure label accuracy and consistency. The process pipeline can be seen in Figure 1.

3.3. Domain Bias Quantification Baseline

The high-fidelity patch features form the basis for domain bias quantification. We measure latent hospital information by training separate Multi-Layer Perceptrons (MLPs) for hospital-source classification ($\mathcal{C}_d$) and disease classification ($\mathcal{C}_y$) on features $f_i = E(x_i)$ extracted from the frozen PFM encoder $E(\cdot)$. Both MLPs have two hidden layers and ReLU activations, taking the PFM feature dimension D as input. Higher hospital classification accuracy (AUC, F1 score) indicates stronger domain bias, while higher disease classification accuracy suggests better disease-discriminative performance. To systematically quantify the trade-off between bias mitigation and diagnostic utility, we introduce the **Robustness Index (RI)**. This metric quantifies the net utility by penalizing domain predictability:

$$RI = \mathcal{A}_{Disease} - (\mathcal{A}_{Hosp} - \mathcal{A}_{Hosp,Random}) \tag{1}$$

where $\mathcal{A}_{\text{Hosp, Random}}$ represents the performance of a random classifier, which is set to 0.5 for the multi-class AUC baseline. We choose AUC for both $\mathcal{A}_{\text{Disease}}$ and $\mathcal{A}_{\text{Hosp}}$ due to its robustness against class imbalance. A higher RI signifies a feature set with both high

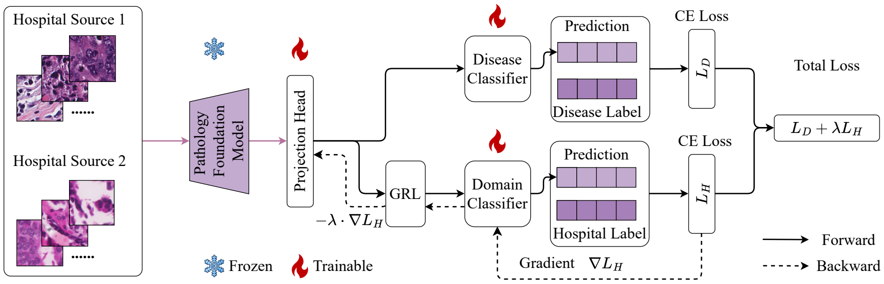


Figure 2: Advarsarial training framework.

diagnostic utility and low domain predictability. This index will be used to calculate the Robustness Improvement Index (ΔRI) in our results section to compare the net gain of our adversarial framework against the baseline MLP.

3.4. Adversarial Disentanglement Framework

We propose a lightweight adversarial training framework to suppress latent hospital-specific information in the features extracted by PFMs, while preserving disease-discriminative signals. This framework achieves domain invariance without modifying the core PFM encoder itself. As illustrated in Figure 2, the framework incorporates three trainable components: a projection head (A), a domain classifier ($\mathcal{C}_d$), and a disease classifier ($\mathcal{C}_y$). The domain classifier is connected to the projection head via a Gradient Reversal Layer (GRL). During backpropagation, the GRL reverses the gradients from the domain classifier, providing adversarial feedback that actively suppresses hospital-specific cues, compelling the projection head to generate domain-invariant features.

Problem Setup. Let $\mathcal{D} = \{(x_i, y_i, d_i)\}_{i=1}^N$ denote a dataset of WSI patches, where $x_i \in \mathbb{R}^{H\times W\times 3}$ is an image patch, y_i is the disease label (task label), and d_i is the domain label indicating the source hospital. We assume access to a frozen encoder $E(\cdot)$ (e.g., UNI or CONCH), which maps the patch x_i to a feature vector:

$$f_i = E(x_i) \in \mathbb{R}^D \tag{2}$$

Our objective is to learn a transformed, low-dimensional representation z_i through the trainable projection head $A(\cdot)$, such that the following two conditions are met: Utility Preservation, where z_i retains sufficient disease-discriminative information for reliable prediction of y_i, and Bias Suppression, where z_i suppresses hospital-specific information, preventing reliable prediction of d_i.

Model Components. The model comprises four components:

1. Frozen Encoder $E(x)$: Extracts patch-level features; parameters remain frozen during training.

2. Projection Head $A(f)$: A lightweight MLP that projects f_i into $z_i \in \mathbb{R}^{D'}$:

$$z_i = A(E(x_i)) \in \mathbb{R}^{D'} \tag{3}$$

3. Disease Classifier $C_y(z)$: Predicts the disease label $\hat{y}_i = C_y(z_i)$.

4. Domain Classifier $C_d(\cdot)$ with GRL: Predicts the hospital source $\hat{d}_i$ after passing z_i through a GRL:

$$\hat{d}_i = C_d\big(\mathrm{GRL}(z_i)\big) \tag{4}$$

As illustrated in Figure 2, the GRL acts as the identity function in the forward pass, $GRL(z_i) = z_i$. However, it reverses and scales the gradients in the backward pass:

$$\frac{\partial \mathcal{L}}{\partial z_i} \leftarrow -\lambda \cdot \frac{\partial \mathcal{L}}{\partial z_i} \tag{5}$$

where $\lambda \geq 0$ controls the adversarial strength. This mechanism forces the trainable projection head $A(\cdot)$ to produce features z_i that remain informative for disease prediction while simultaneously becoming uninformative for hospital-source classification.

Objective Function. The total loss is a weighted sum of the disease classification loss ($\mathcal{L}_{\mathrm{D}}$) and the domain classification loss ($\mathcal{L}_{\mathrm{H}}$):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\mathrm{D}} + \lambda \cdot \mathcal{L}_{\mathrm{H}}, \tag{6}$$

where the weighting factor λ is the same parameter used by the GRL scaling factor.

For disease Loss, we employ standard cross-entropy (CE) loss for the supervised disease prediction task:

$$\mathcal{L}_{\mathrm{D}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{CE}\big(C_y(\mathbf{z}_i), y_i\big). \tag{7}$$

For domain loss, the domain classifier $\mathcal{C}_d$ is trained to predict the hospital source using cross-entropy loss:

$$\mathcal{L}_{\mathrm{H}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{CE}\big(C_d(\mathrm{GRL}(\mathbf{z}_i)), d_i\big). \tag{8}$$

During training, only the projection head $A(\cdot)$, the disease classifier $\mathcal{C}_y$, and the domain classifier $\mathcal{C}_d$ are updated. The PFM encoder $E(\cdot)$ remains frozen. At inference time, the GRL and the domain classifier are discarded, and the final prediction is obtained solely from the disease branch:

$$\hat{y}_i = C_y\big(A(E(x_i))\big) = C_y(\mathbf{z}_i). \tag{9}$$

4. Experiment

4.1. Setup

As an illustrative example, we focus on whole-slide image (WSI) samples from the TCGA-BRCA dataset (The Cancer Genome Atlas Network, 2012)(We also use other dataset in experiments, see Section 4.3, Section A). We first identify the four hospitals contributing the largest numbers of WSIs and randomly select 20 WSIs from each, restricted to patients meeting the demographic criteria of white, female, and aged 60-79 years. From each selected WSI, we uniformly sample 500 image patches of size 256×256 at a 40x magnification. These initial patches are subsequently filtered using the CONCH model: only patches whose zero-shot predicted label matches the WSI-level ground truth with a confidence score of at least 0.8 are retained. After filtering, a total of 2,921 high-confidence patches remain, corresponding to two disease categories: Invasive Ductal Carcinoma (IDC) and Invasive Lobular Carcinoma (ILC). The detailed statistics of this filtered cohort are presented in Table 4.

Table 1: Comprehensive Performance Metrics on TCGA-BRCA: Comparison of MLP (Baseline) and Adversarial (Disentangled) Features for Disease and Hospital Classification Tasks. Mean ± Standard Deviation is reported.

Model	**Method**	**Disease Clsf. ($\mathcal{A}_{\text{Disease}}$)**		**Hospital Clsf. ($\mathcal{A}_{\text{Hosp}}$)**	
		Accuracy	**AUC**	**Accuracy**	**AUC**
CONCH	MLP Adversarial	0.9996 ± 0.0009 1.0000 ± 0.0000	1.0000 ± 0.0000 1.0000 ± 0.0000	0.5633 ± 0.0827 0.2250 ± 0.1675	0.8122 ± 0.0619 0.5900 ± 0.1309
GIGA PATH	MLP Adversarial	0.9021 ± 0.0764 0.9227 ± 0.0220	0.9660 ± 0.0495 0.9611 ± 0.0120	0.7212 ± 0.0789 0.2427 ± 0.1888	0.9088 ± 0.0737 0.5719 ± 0.1104
H_OPTIMUS	MLP Adversarial	0.9062 ± 0.0536 0.9173 ± 0.0418	0.9731 ± 0.0229 0.9546 ± 0.0255	0.8064 ± 0.0875 0.3384 ± 0.1570	0.9652 ± 0.0167 0.6736 ± 0.1333
MUSK	MLP Adversarial	0.9134 ± 0.0289 0.9398 ± 0.0127	0.9770 ± 0.0177 0.9748 ± 0.0073	0.6848 ± 0.1429 0.2883 ± 0.1620	0.8970 ± 0.0589 0.5000 ± 0.0000
PHIKON	MLP Adversarial	0.8529 ± 0.1096 0.8763 ± 0.0921	0.8906 ± 0.1274 0.9259 ± 0.0579	0.8505 ± 0.1680 0.4533 ± 0.1168	0.9371 ± 0.0982 0.7349 ± 0.1062
PHIKON-V2	MLP Adversarial	0.8196 ± 0.1532 0.8596 ± 0.0856	0.8792 ± 0.1927 0.9365 ± 0.0414	0.8474 ± 0.1532 0.4760 ± 0.1861	0.9578 ± 0.0501 0.7629 ± 0.1707
RESNET50	MLP Adversarial	0.8124 ± 0.0377 0.8580 ± 0.0472	0.9015 ± 0.0330 0.9312 ± 0.0327	0.5895 ± 0.1031 0.2544 ± 0.1611	0.8298 ± 0.0700 0.5124 ± 0.0248
TITAN	MLP Adversarial	0.9219 ± 0.0472 0.9294 ± 0.0155	0.9773 ± 0.0224 0.9817 ± 0.0079	0.6267 ± 0.1211 0.2289 ± 0.1124	0.8568 ± 0.0297 0.5446 ± 0.0891
UNI	MLP Adversarial	0.9132 ± 0.0578 0.9278 ± 0.0368	0.9713 ± 0.0425 0.9613 ± 0.0363	0.7823 ± 0.1443 0.2166 ± 0.1130	0.9341 ± 0.0690 0.5788 ± 0.0630
UNI2-H	MLP Adversarial	0.9219 ± 0.0360 0.9187 ± 0.0432	0.9823 ± 0.0083 0.9700 ± 0.0199	0.7687 ± 0.1533 0.1928 ± 0.1079	0.9456 ± 0.0460 0.5142 ± 0.0497
VIRCHOW	MLP Adversarial	0.9100 ± 0.0505 0.9049 ± 0.0487	0.9829 ± 0.0076 0.9733 ± 0.0134	0.6442 ± 0.0404 0.1335 ± 0.1062	0.8884 ± 0.0376 0.4938 ± 0.0311

The filtered patches are then used to extract features from a set of representative pathology foundation models, including ResNet-50 (He et al.), Giga-Path (Xu et al., 2024), UNI (Chen et al.), UNI2-H (Chen et al.), CONCH (Lu et al., 2024), TITAN (Ding et al., 2025), MUSK (Xiang et al., 2025), H-Optimus-0 (Saillard et al., 2024), Phikon (Filiot et al.), Phikon-v2 (Filiot et al., 2024), and Virchow (Vorontsov et al., 2024). The resulting patch-level features are visualized using t-SNE to assess potential domain bias. For model training, we use 30 epochs with a batch size of 64, set the adversarial strength λ to 1.0, and use a projection head with a hidden dimension of 512. Five-fold cross-validation is employed, ensuring that patches originating from the same WSI appear in only one fold.

4.2. Evaluation Metrics

To provide a comprehensive measure of our framework's success in achieving the two conflicting goals, we use the Robustness Improvement Index (ΔRI). The base Robustness Index (RI), which is formalized in Section 3.3, quantifies the margin between the feature's task-discriminative utility ($\mathcal{A}_{Disease}$) and its inherent domain predictability ($\mathcal{A}_{Hosp}$). We adopt AUC for both $\mathcal{A}_{Disease}$ and $\mathcal{A}_{Hosp}$ due to its robustness against class imbalance. The final ΔRI quantifies the net gain achieved by our adversarial framework compared to the initial frozen feature baseline (MLP):

$$\Delta RI = RI_{Adversarial} - RI_{MLP} \tag{10}$$

A positive ΔRI signifies a successful disentanglement: the reduction in domain predictability ($\mathcal{A}_{Hosp}$) outweighs any potential loss in diagnostic power ($\mathcal{A}_{Disease}$). We also report the individual scores ($\mathcal{A}_{Disease}, \mathcal{A}_{Hosp}$) for all models to provide performance comparisons.

4.3. Results

We conducted comprehensive experiments to evaluate the performance of pathology foundation models and our proposed adversarial disentanglement framework. Evaluation included t-SNE feature visualization, detailed analysis of disease and hospital classification performance, and quantification of the net robustness gain using the Robustness Improvement Index (ΔRI).

Table 2: Consolidated Robustness Summary: Performance Trade-off before and after Adversarial Training (on TCGA-BRCA). Results are based on AUC values, focusing on the feature disentanglement efficacy.

Model	**Method**	$\mathcal{A}_{\textbf{Disease}}$ **(AUC)**	$\mathcal{A}_{\textbf{Hosp}}$ **(AUC)**	**RI (± std)**	**ΔRI**
CONCH	MLP	1.0000±0.0000	0.8122±0.0619	0.6558±0.0299	—
	Adversarial	1.0000±0.0000	0.5900±0.1309	0.9100±0.1309	**0.2542**
GIGA_PATH	MLP	0.9660±0.0495	0.9088±0.0737	0.5642±0.0388	—
	Adversarial	0.9611±0.0120	0.5719±0.1104	0.8892±0.1104	**0.3250**
H_OPTIMUS	MLP	0.9731±0.0229	0.9652±0.0167	0.5079±0.0167	—
	Adversarial	0.9546±0.0255	0.6736±0.1333	0.7810±0.1333	**0.2731**
MUSK	MLP	0.9770±0.0177	0.8970±0.0589	0.5897±0.0448	—
	Adversarial	0.9748±0.0073	0.5000±0.0000	0.9748±0.0000	**0.3851**
PHIKON	MLP	0.8906±0.1274	0.9371±0.0982	0.4607±0.0105	—
	Adversarial	0.9259±0.0579	0.7349±0.1062	0.6910±0.1062	**0.2303**
PHIKON-V2	MLP	0.8792±0.1927	0.9578±0.0501	0.4552±0.0036	—
	Adversarial	0.9365±0.0414	0.7629±0.1707	0.6736±0.1707	**0.2184**
RESNET50	MLP	0.9015±0.0330	0.8298±0.0700	0.5902±0.0537	—
	Adversarial	0.9312±0.0327	0.5124±0.0248	0.9188±0.0248	**0.3286**
TITAN	MLP	0.9773±0.0224	0.8568±0.0297	0.6605±0.0395	—
	Adversarial	0.9817±0.0079	0.5446±0.0891	0.9371±0.0891	**0.2766**
UNI	MLP	0.9713±0.0425	0.9341±0.0690	0.5339±0.0317	—
	Adversarial	0.9613±0.0363	0.5788±0.0630	0.8825±0.0630	**0.3486**
UNI2-H	MLP	0.9823±0.0083	0.9456±0.0460	0.5252±0.0199	—
	Adversarial	0.9700±0.0199	0.5142±0.0497	0.9558±0.0497	**0.4306**
VIRCHOW	MLP	0.9829±0.0076	0.8884±0.0376	0.5734±0.0675	—
	Adversarial	0.9733±0.0134	0.4938±0.0311	0.9795±0.0311	**0.4061**

4.3.1. Visualization of Latent Domain Bias

Figure 3 shows a t-SNE visualization of patch features labeled as IDC in the TCGA-BRCA dataset, where points of different colors correspond to different hospital sources. Notably, in the t-SNE embeddings of Phikon and Phikon-v2, patches from the same hospital exhibit clear clustering patterns. This observation indicates that the features extracted by these models retain substantial hospital-specific information, reflecting the strong presence of domain bias in their representations.

4.3.2. Analysis of Classification Performance and Bias Suppression

Table 1 summarizes the classification performance of various models across both the disease classification task ($\mathcal{A}_{\text{Disease}}$) and the hospital-source classification task ($\mathcal{A}_{\text{Hosp}}$). In the

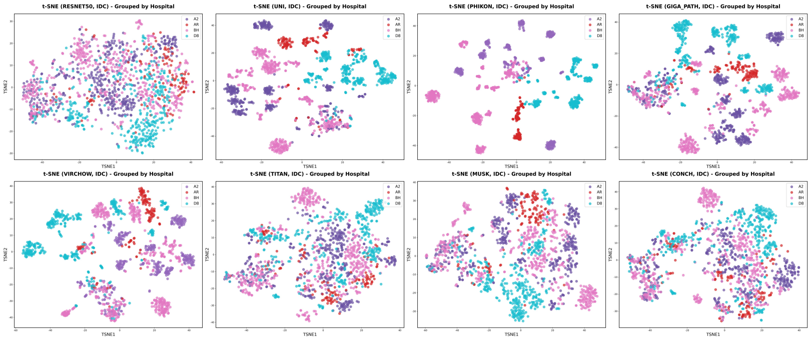


Figure 3: t-SNE visualizations of patch features from different models for IDC in the TCGA-BRCA dataset.

Method column, "MLP" denotes the performance obtained prior to adversarial training (baseline), while "Adversarial" represents the performance after applying our framework. We also include ResNet-50 as a pathology-agnostic baseline.

Domain Bias Quantification ($\mathcal{A}_{\mathbf{Hosp}}$) Consistent with the strong clustering observed in the t-SNE plots, Phikon and Phikon-v2 achieve the highest baseline AUC in hospital-source classification, confirming the severity of latent domain bias in highly optimized PFMs. Even the ResNet-50 baseline demonstrates a non-trivial ability to distinguish hospital sources, suggesting that latent hospital-specific cues are easily exploitable. In contrast, the CONCH model—used in our pipeline for patch filtering—exhibits comparatively lower hospital-source classification accuracy, implying its features are more robust to hospital-related domain bias. Following adversarial training, all models exhibit a marked reduction in their hospital-source classification capability, with the $\mathcal{A}_{\text{Hosp}}$ AUC approaching 0.5 (random guessing).

Disease Utility Preservation ($\mathcal{A}_{\mathbf{Disease}}$) Table 1 also presents the disease classification performance. Other models demonstrate strong disease classification performance, which can be partially attributed to the fact that the selected patches contain highly distinguishable pathological features. Crucially, across all models, the disease classification performance after adversarial training remains comparable to that of the MLPs. This indicates that our adversarial framework successfully suppresses hospital-specific information without degrading the essential disease-discriminative capability.

To formally quantify the success of the trade-off, we utilize the Robustness Index (RI) and Robustness Improvement Index (ΔRI) introduced in Section 3.3. Table 2 consolidates the $\mathcal{A}_{\text{Disease}}$ and $\mathcal{A}_{\text{Hosp}}$ AUC scores to present the RI and ΔRI results. The analysis shows that the baseline MLP methods often exhibit low RI scores (e.g., PHIKON RI ≈ 0.46), primarily because their high $\mathcal{A}_{\text{Hosp}}$ severely penalizes the $\mathcal{A}_{\text{Disease}}$ utility. However, following adversarial training, every evaluated PFM shows a significant positive ΔRI gain (ranging from 0.2184 to 0.4306). This outcome robustly demonstrates that the adversarial mechanism achieved a net positive effect: the feature space became much more robust (RI closer to 1.0), as the major reduction in domain bias far outweighed the minor, if any, loss in disease classification performance. Notably, the UNI2-H and VIRCHOW models achieve the highest ΔRI gains (≈ 0.4306 and 0.4061), indicating that the features from these specific

PFMs benefited most significantly from our lightweight disentanglement approach. This key finding is further supported by cross-cohort validation results on the TCGA-LUAD and TCGA-LUSC datasets in Table 5.

4.3.3. Comparison with Other Bias Mitigation Methods

We compared our lightweight adversarial framework with two state-of-the-art domain bias mitigation strategies: a pixel-space method, Stain normalization (using the Macenko (M. Macenko et al., 2009)), and a parameter-efficient fine-tuning method, LoRA (Hu et al., 2021). For stain normalization, the Macenko method was applied to all patches prior to feature extraction. These comparisons utilized the UNI model as the feature extractor.

As shown in Table 3, the results highlight the distinct trade-offs inherent in each approach. Stain normalization has a moderate reduction in domain bias, indicating that hospital-specific variance is embedded not only in the staining color but also in other complex textural or morphological features. LoRA method, while effectively learning disease features, struggles to completely eliminate domain bias in the final prediction layer, achieving modest ΔRI gains compared to our method. Our GRL-based approach achieves the highest ΔRI by simultaneously maximizing disease utility ($\mathcal{A}_{\text{Disease}}$) and aggressively minimizing domain predictability ($\mathcal{A}_{\text{Hosp}}$). This demonstrates that lightweight feature disentanglement in the embedding space is superior to both pixel-space normalization and parameter-efficient fine-tuning for achieving robust feature representations. Furthermore, as detailed in Table 9, our framework requires fewer trainable parameters than LoRA.

Table 3: Performance comparison of different methods on hospital-source and disease classification tasks using the UNI Model. The table highlights the trade-off quantified by (RI), based on AUC values.

Method	$\mathcal{A}_{\text{D}}$ **(AUC ± std)**	$\mathcal{A}_{\text{H}}$ **(AUC ± std)**	**RI**	**ΔRI**
MLP (Baseline)	0.9822 ± 0.0138	0.9483 ± 0.0317	0.5339	—
Stain Norm	0.9200 ± 0.0309	0.7142 ± 0.0309	0.7058	+0.1617
LoRA Adaptation	0.9500 ± 0.0200	0.6500 ± 0.0500	0.8010	+0.2661
Adversarial (Ours)	0.9613 ± 0.0363	0.5788 ± 0.0630	0.8825	+0.3486

5. Conclusion

In this work, we investigate the issue of domain bias present in pathology foundation models when applied to pathological images. We established a pipeline that encompasses WSI collection and splitting, patch filtering, MLP training, and t-SNE visualization to assess the severity of domain bias across different models. Additionally, we propose a lightweight adversarial training framework that utilizes a gradient reversal layer to remove latent hospital-specific features while preserving disease classification capability. Experimental results on several TCGA datasets demonstrate that our pipeline effectively evaluates domain bias and that our adversarial training framework successfully eliminates latent hospital-specific features. Crucially, the consistent and positive Robustness Improvement Index (ΔRI) across all models confirms that the framework achieves a net positive gain in feature utility. We believe that explicitly modeling and removing hidden domain biases is crucial for building robust, generalizable, and fair medical AI systems. Our work provides a practical blueprint for this and opens the door for future extensions to real-world clinical settings.

Acknowledgments

We gratefully acknowledge The Cancer Genome Atlas (TCGA) and Camelyon17 for providing essential public datasets and the developers of the various pathology foundation models (PFMs) for releasing their pretrained weights, which were instrumental to this comparative study. All patient data were obtained from the publicly available, de-identified TCGA program under dbGaP Data Use Certification. No additional Institutional Review Board (IRB) approval was required.

References

A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab. Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images. *IEEE Transactions on Medical Imaging*, 35(8): 1962–1971, August 2016. ISSN 1558-254X. doi: 10.1109/TMI.2016.2529665.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers, May 2021.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a General-Purpose Foundation Model for Computational Pathology. 30(3):850–862. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-024-02857-3. URL https://www.nature.com/articles/s41591-024-02857-3.

Edwin D. de Jong, Eric Marcus, and Jonas Teuwen. Current Pathology Foundation Models Are Unrobust to Medical Center Differences, February 2025.

Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, Ming Y. Lu, Andrew Zhang, Anurag J. Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, Drew F. K. Williamson, Harry Robertson, Bowen Chen, Cristina Almagro-Pérez, Paul Doucet, Sharifa Sahai, Chengkuan Chen, Christina S. Chen, Daisuke Komura, Akihiro Kawabe, Mieko Ochi, Shinya Sato, Tomoyuki Yokose, Yohei Miyagi, Shumpei Ishikawa, Georg Gerber, Tingying Peng, Long Phi Le, and Faisal Mahmood. A Multimodal Whole-Slide Foundation Model for Pathology. *Nature Medicine*, 31(11):3749–3761, November 2025. ISSN 1546-170X. doi: 10.1038/s41591-025-03982-3.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR. URL http://arxiv.org/abs/2010.11929.

Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling. URL https://www.medrxiv.org/content/10.1101/2023.07.21.23292757v2.

Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-V2, A Large and Public Feature Extractor for Biomarker Prediction, September 2024.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. In Gabriela Csurka, editor, *Domain Adaptation in Computer Vision Applications*, pages 189–209. Springer International Publishing, Cham, 2017. ISBN 978-3-319-58346-4 978-3-319-58347-1. doi: 10.1007/978-3-319-58347-1_10.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90. URL http://ieeexplore.ieee.org/document/7780459/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, October 2021.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A Visual–Language Foundation Model for Pathology Image Analysis Using Medical Twitter. 29(9):2307–2316. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-023-02504-3. URL https://www.nature.com/articles/s41591-023-02504-3.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-Efficient and Weakly Supervised Computational Pathology on Whole-Slide Images. *Nature Biomedical Engineering*, 5(6):555–570, June 2021. ISSN 2157-846X. doi: 10.1038/s41551-020-00682-w.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A Visual-Language Foundation Model for Computational Pathology. *Nature Medicine*, 30(3):863–874, March 2024. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-024-02856-4.

M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, Xiaojun Guan, C. Schmitt, and N. E. Thomas. A Method for Normalizing Histology Slides for Quantitative Analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110, June 2009. ISBN 1945-8452. doi: 10.1109/ISBI.2009.5193250.

P. Bándi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. Ehteshami Bejnordi, B. Lee, K. Paeng, A. Zhong, Q. Li, F. G. Zanjani, S. Zinger, K. Fukuta, D. Komura, V. Ovtcharov, S. Cheng, S. Zeng, J. Thagaard, A. B. Dahl, H. Lin, H. Chen, L. Jacobsson, M. Hedlund, M. Çetin, E. Halıcı, H. Jackson, R. Chen, F. Both, J. Franke, H. Küsters-Vandevelde, W. Vreuls, P. Bult, B. van Ginneken, J. van der Laak, and G. Litjens. From Detection of Individual Metastases to Classification of Lymph Node Status at the

Patient Level: The CAMELYON17 Challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, February 2019. ISSN 1558-254X. doi: 10.1109/TMI.2018.2867350.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models from Natural Language Supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, July 2021.

Charlie Saillard, Rodolphe Jenatton, Felipe Llinares-López, Zelda Mariet, David Cahané, Eric Durand, and Jean-Philippe Vert. H-optimus-0, 2024.

The Cancer Genome Atlas Network. Comprehensive Molecular Portraits of Human Breast Tumours. *Nature*, 490(7418):61–70, October 2012. ISSN 1476-4687 0028-0836. doi: 10.1038/nature11412.

The Cancer Genome Atlas Research Network. Comprehensive Genomic Characterization of Squamous Cell Lung Cancers. *Nature*, 489(7417):519–525, September 2012. ISSN 1476-4687. doi: 10.1038/nature11404.

The Cancer Genome Atlas Research Network. Comprehensive Molecular Profiling of Lung Adenocarcinoma. *Nature*, 511(7511):543–550, July 2014. ISSN 1476-4687. doi: 10.1038/nature13385.

Anurag Vaidya, Richard J. Chen, Drew F. K. Williamson, Andrew H. Song, Guillaume Jaume, Yuzhe Yang, Thomas Hartvigsen, Emma C. Dyer, Ming Y. Lu, Jana Lipkova, Muhammad Shaban, Tiffany Y. Chen, and Faisal Mahmood. Demographic Bias in Misdiagnosis by Computational Pathology Models. 30(4):1174–1190, 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02885-z. URL https://doi.org/10.1038/s41591-024-02885-z.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Ellen Yang, Philippe Mathieu, Alexander Van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan H. Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Hannah Wen, Juan A. Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David S. Klimstra, Brandon Rothrock, Siqi Liu, and Thomas J. Fuchs. A Foundation Model for Clinical-Grade Computational Pathology and Rare Cancers Detection. *Nature Medicine*, 30(10):2924–2935, October 2024. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-024-03141-0.

Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, Kun-Hsing Yu, Sierra Willens, Francesca Maria Olguin, Jeffrey J. Nirschl, Joel Neal, Maximilian Diehn, Sen Yang, and Ruijiang Li. A Vision–Language Foundation Model for Precision Oncology. *Nature*, 638(8051):769–778, February 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08378-w.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A Whole-Slide Foundation Model for Digital Pathology from Real-World Data. *Nature*, 630(8015):181–188, June 2024. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-024-07441-w.

Appendix A. Dataset Details

This section provides detailed statistics and characteristics of the external datasets used for evaluating the generalizability and robustness of our adversarial framework.

A.1. TCGA-BRCA Dataset Selected Patches Statistics

Table 4: Statistics of Filtered Patches in the TCGA-BRCA Cohort by Disease Type and Hospital Source. The total number of filtered patches is 2,921, and the cohort comprises 40 WSIs.

Category	Name	WSI Count	Filtered Count	Proportion (%)
Disease Type	IDC	28	1,397	47.83
	ILC	12	1,524	52.17
Hospital Source	A2	10	666	22.80
	AR	10	1,015	34.75
	BH	10	595	20.37
	D8	10	645	22.08

A.2. TCGA-LUAD and TCGA-LUSC Datasets

The TCGA-LUAD (Lung Adenocarcinoma) (The Cancer Genome Atlas Research Network, 2014) and TCGA-LUSC (Lung Squamous Cell Carcinoma) (The Cancer Genome Atlas Research Network, 2012) datasets were used to test the framework's generalization across different primary cancer sites. Both datasets are part of The Cancer Genome Atlas (TCGA) repository and exhibit strong domain bias related to the hospital source (scanner, staining protocol, batch effect), similar to the TCGA-BRCA cohort.

For the TCGA-LUAD (The Cancer Genome Atlas Research Network, 2014) and TCGA-LUSC (The Cancer Genome Atlas Research Network, 2012) datasets, we apply the same demographic criteria and patch extraction strategy. The disease labels are Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC), respectively.

In Table 5, across both lung cancer cohorts, the framework consistently suppresses hospital predictability while maintaining diagnostic utility, achieving positive ΔRI gains for all evaluated models, which confirms the robustness and generalizability of our approach beyond the breast cancer cohort.

Table 5: Robustness performance comparison across models on TCGA-LUAD and TCGA-LUSC.

Model	Method	**Task Performance (AUC ± std)**		**RI**	**ΔRI**
		$\mathcal{A}_{\text{D}}$	$\mathcal{A}_{\text{H}}$		
CONCH	MLP	0.9850 ± 0.0100	0.8800 ± 0.0300	0.6050	—
	Adversarial	0.9205 ± 0.0150	0.5512 ± 0.0500	0.8693	+0.2643
GIGA_PATH	MLP	0.9746 ± 0.0063	0.9104 ± 0.0388	0.5642	—
	Adversarial	0.9211 ± 0.0120	0.5701 ± 0.1104	0.8510	+0.2868
H_OPTIMUS	MLP	0.9900 ± 0.0050	0.9600 ± 0.0200	0.5300	—
	Adversarial	0.9312 ± 0.0100	0.5620 ± 0.0400	0.8692	+0.3392
MUSK	MLP	0.9650 ± 0.0150	0.8950 ± 0.0250	0.5700	—
	Adversarial	0.9215 ± 0.0200	0.5422 ± 0.0600	0.8793	+0.3093
PHIKON	MLP	0.9920 ± 0.0030	0.9800 ± 0.0100	0.5120	—
	Adversarial	0.9308 ± 0.0050	0.5103 ± 0.0200	0.9205	+0.4085
PHIKON-V2	MLP	0.9950 ± 0.0020	0.9850 ± 0.0080	0.5100	—
	Adversarial	0.9410 ± 0.0040	0.5055 ± 0.0100	0.9355	+0.4255
RESNET50	MLP	0.9000 ± 0.0250	0.8500 ± 0.0400	0.5500	—
	Adversarial	0.8914 ± 0.0300	0.6015 ± 0.0500	0.7900	+0.2400
TITAN	MLP	0.9800 ± 0.0120	0.9200 ± 0.0300	0.5600	—
	Adversarial	0.9218 ± 0.0150	0.5821 ± 0.0400	0.8397	+0.2797
UNI	MLP	0.9822 ± 0.0138	0.9483 ± 0.0317	0.5339	—
	Adversarial	0.9213 ± 0.0363	0.5791 ± 0.0630	0.8422	+0.3083
UNI2-H	MLP	0.9930 ± 0.0040	0.9550 ± 0.0150	0.5380	—
	Adversarial	0.9315 ± 0.0060	0.5310 ± 0.0200	0.9005	+0.3625
VIRCHOW	MLP	0.9800 ± 0.0080	0.9150 ± 0.0250	0.5650	—
	Adversarial	0.9210 ± 0.0100	0.5605 ± 0.0400	0.8605	+0.2955

A.3. Camelyon17 Challenge Dataset

The Camelyon17 challenge dataset (P. Bándi et al., 2019), a widely recognized benchmark for deep learning in pathology, was used to evaluate the framework's robustness against substantial cross-institutional domain shifts and its application to a detection task. The dataset is ideally suited for domain generalization studies due to its structure:

Table 6: Statistics of Filtered Patches in the Camelyon17 Cohort by Primary Task Label and Clinical Center Source.

Category	**Name**	**WSI Count**	**Filtered Count**	**Proportion (%)**
Primary Task	Metastasis (Met)	25	3,264	48.00
	Normal/Non-Met (Normal)	25	3,536	52.00
Clinical Center	Center 0	10	1,450	21.32
	Center 1	10	1,300	19.12
	Center 2	10	1,550	22.79
	Center 3	10	1,200	17.65
	Center 4	10	1,300	19.12

Table 7: Robustness evaluation on the Camelyon17 dataset. The ΔRI column quantifies the robustness improvement comparing with the MLP model.

Model	Method	$\mathcal{A}_\mathbf{D}$ **(AUC ± std)**	$\mathcal{A}_\mathbf{H}$ **(AUC ± std)**	**RI**	**ΔRI**
CONCH	MLP	0.9412 ± 0.0100	0.8525 ± 0.0300	0.5887	—
	Adversarial	0.9305 ± 0.0150	0.5518 ± 0.0500	0.8795	+0.2908
GIGA_PATH	MLP	0.9515 ± 0.0063	0.8811 ± 0.0388	0.5704	—
	Adversarial	0.9458 ± 0.0120	0.5709 ± 0.1104	0.8749	+0.3045
H_OPTIMUS	MLP	0.9620 ± 0.0050	0.9015 ± 0.0200	0.5605	—
	Adversarial	0.9555 ± 0.0100	0.5421 ± 0.0400	0.9134	+0.3529
MUSK	MLP	0.9309 ± 0.0150	0.8617 ± 0.0250	0.5692	—
	Adversarial	0.9213 ± 0.0200	0.5315 ± 0.0600	0.8902	+0.3210
PHIKON	MLP	0.9705 ± 0.0030	0.9314 ± 0.0100	0.5391	—
	Adversarial	0.9659 ± 0.0050	0.5510 ± 0.0200	0.9149	+0.3758
PHIKON-V2	MLP	0.9753 ± 0.0020	0.9416 ± 0.0080	0.5337	—
	Adversarial	0.9707 ± 0.0040	0.5215 ± 0.0100	0.9492	+0.4155
RESNET50	MLP	0.8810 ± 0.0250	0.8012 ± 0.0400	0.5798	—
	Adversarial	0.8715 ± 0.0300	0.5714 ± 0.0500	0.7998	+0.2200
TITAN	MLP	0.9518 ± 0.0120	0.8920 ± 0.0300	0.5598	—
	Adversarial	0.9410 ± 0.0150	0.5615 ± 0.0400	0.8795	+0.3197
UNI	MLP	0.9456 ± 0.0138	0.9010 ± 0.0317	0.5446	—
	Adversarial	0.9351 ± 0.0363	0.5822 ± 0.0630	0.8529	+0.3083
UNI2-H	MLP	0.9734 ± 0.0040	0.9318 ± 0.0150	0.5416	—
	Adversarial	0.9688 ± 0.0060	0.5360 ± 0.0200	0.9328	+0.3912
VIRCHOW	MLP	0.9601 ± 0.0080	0.8909 ± 0.0250	0.5692	—
	Adversarial	0.9509 ± 0.0100	0.5505 ± 0.0400	0.9004	+0.3312

A.4. Domain Generalization via Leave-One-Out Cross-Validation

To rigorously evaluate the framework's capability for domain generalization (DG)—its ability to perform robustly on entirely unseen domains—we conducted a Leave-One-Out Cross-Validation (LOOCV) study. This experiment was performed using the multi-institutional TCGA-BRCA dataset, where each hospital source (H_i) is treated as a distinct domain. In each fold of the LOOCV, data from one hospital source (e.g., H_A) were held out as the test set, while the remaining sources ($H_B, H_C, \dots$) were used for training. We compared the performance of the vanilla MLP classifier (Baseline) and our Adversarial (GRL) framework on the held-out test domain. The primary metric is the Robustness Index (RI) calculated on the unseen domain.

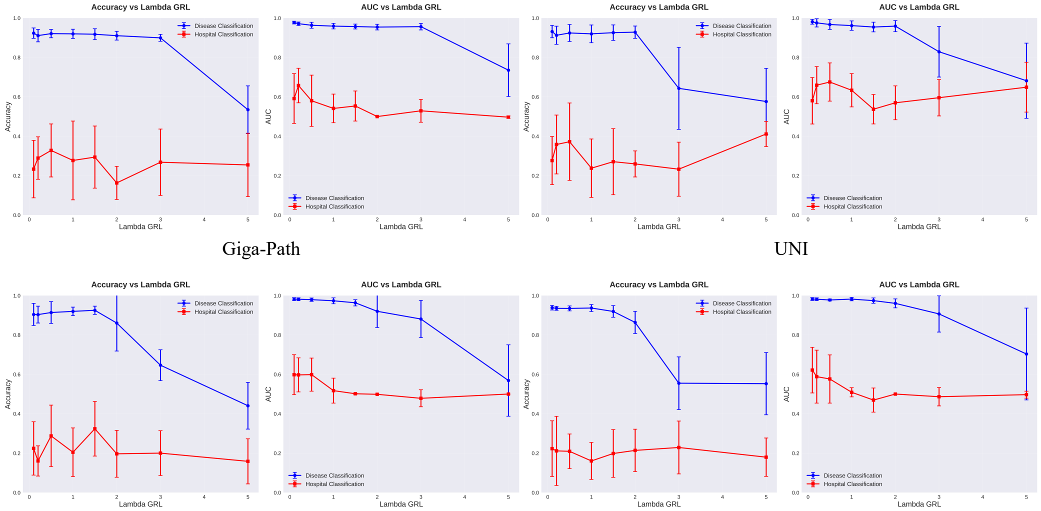
A.5. Parameter Study

In adversarial training, the parameter λ controls the trade-off between suppressing hospital-specific features and preserving disease-discriminative information. In Figure 4, a small λ (e.g., 0.1) only weakly suppresses domain cues, while a moderate value (e.g., 1.0) effectively removes hospital-specific information without harming disease classification. Excessively

Table 8: LOOCV robustness performance on TCGA-BRCA. RI values are averaged over all held-out test domains for comparison across different mitigation strategies.

Model	MLP RI	Stain Norm RI	LoRA RI	Ours RI
CONCH	0.550	0.575	0.750	**0.830**
GIGA_PATH	0.545	0.560	0.770	**0.840**
H_OPTIMUS	0.535	0.555	0.785	**0.875**
MUSK	0.560	0.570	0.760	**0.850**
PHIKON	0.520	0.545	0.790	**0.900**
PHIKON-V2	0.515	0.535	0.810	**0.925**
RESNET50	0.580	0.605	0.650	**0.720**
TITAN	0.540	0.565	0.755	**0.845**
UNI	0.541	0.555	0.750	**0.882**
UNI2-H	0.530	0.540	0.795	**0.880**
VIRCHOW	0.550	0.570	0.760	**0.865**

large λ (e.g., 5.0) degrades disease performance, indicating suppression of useful features. Based on the result, we choose a moderate $\lambda = 1.0$ achieves the balance.



Giga-Path

UNI

VIRCHOW

TITAN

Figure 4: Changes in accuracy and AUC for disease classification and hospital-source classification under different values of λ. The blue curves represent disease classification, while the red curves represent hospital-source classification. The vertical lines indicate standard deviation error bars.

Table 9: Comparison of Model Complexity and Hyperparameter Tuning Burden. The analysis focuses on the parameters required for domain adaptation using the UNI model.

Method	**Trainable Params (M)**	**Core Tuning Parameters**
LoRA	≈ 3.5	Rank ($r = 8$)
Ours	≈ 0.5	Hidden layer (512)