

Do Center Biases Propagate? Robustness of Pathology Foundation Models in Whole-Slide Image Classification

Ilán Carretero¹, Pablo Meseguer¹, Rocío del Amor^{1,2}, and Valery Naranjo^{1,2}

¹ CVBLab, HumanTech, Universitat Politècnica de València (UPV), Valencia, Spain

² Artikode Intelligence S.L., Valencia, Spain

✉ {ilcarjuc, pabmees, madeam2, vnaranjo}@upv.es

How to cite: Carretero, I.; Meseguer, P.; Del Amor, R.; and Naranjo, V. (2026). Do Center Biases Propagate? Robustness of Pathology Foundation Models in Whole-Slide Image Classification. En libro de actas: *XLIV Congreso Anual de la Sociedad Española de la Ingeniería Biomédica 2026*. Valencia, 11–13 November 2026. <https://doi.org/10.4995/CASEIB2026.2026.xxxx>

Abstract

Pathology foundation models (PFMs) have transformed computational pathology through powerful representation learning from histopathological images. PFMs provide rich, discriminative representations for whole slide image (WSI) analysis, enabling tasks such as slide-level classification under multiple instance learning (MIL). However, these representations may also encode non-biological signals associated with acquisition centers, potentially introducing spurious shortcuts into downstream predictions. In this work, we evaluate center-associated robustness in WSI classification using a controlled training setting with increasing class-center correlations quantified by Cramér’s V . We benchmark six PFMs across four datasets and two MIL aggregators, while evaluating ComBat as a robustification strategy. We further introduce the Area Under the Cramér’s V Curve (AUCC) to jointly capture absolute classification performance and its degradation as spurious correlation increases. Results show that center-related information encoded by PFMs propagates to WSI-level predictions, with robustness depending on both the PFM representation and MIL aggregation strategy. Additionally, ComBat harmonization does not provide consistent robustness gains across datasets.

Keywords: Pathology Foundation Models, Center Robustness, Whole Slide Images, Spurious Correlations.

1. Introduction

The emergence of pathology foundation models (PFMs) have notably enhanced AI-based workflows in computational pathology (CPath) [1]. Due to high-capacity model architectures and large-scale, self-supervised, in-domain pre-training, PFMs extract powerful data representations from histopathological images. Their enhanced representational learning has enhanced CPath algorithms for tasks such as whole slide image (WSI) classification. Given the gigapixel scale of WSIs, multiple instance learning (MIL) has become the most prominent approach for WSI classification, requiring only WSI-level annotations [2].

Beyond disease-related biological patterns, PFMs have been shown to encode signals associated with center-specific

procedures [3]. Across pathology departments, differences in tissue processing, staining protocols, and digitization scanners can introduce variations in color, intensity, and other image characteristics. These technical signatures may act as spurious shortcuts, causing downstream models to rely on center-related information rather than biological signal and limiting generalization across centers. Although large-scale in-domain pretraining may improve robustness to such shifts, its impact on clinically relevant downstream predictions remains an important consideration.

In this work, we evaluate center-associated robustness in WSI-level prediction using a weakly supervised MIL framework. We construct controlled training settings with increasing class–center correlations quantified by Cramér’s V and assess how these spurious associations propagate to downstream predictions across PFMs and MIL aggregators. We further introduce the Area Under the Cramér’s V Curve (AUCC), which jointly captures absolute classification performance and performance degradation as center–label correlation increases. Finally, we evaluate ComBat harmonization [4] as a robustification strategy for mitigating center-associated variation in PFM representations.

2. Methodology

This section presents our framework for evaluating center-associated robustness of PFMs under spurious correlations. Figure 1 summarizes the overall approach and its main components are detailed below.

2.1. Multiple Instance Learning (MIL)

In MIL, each WSI is represented as a bag $X = \{x_n\}_{n=1}^N$ of N image patches and assigned a slide-level label $y \in Y$, where Y denotes the set of target classes. A PFM encoder maps each patch into a feature representation in c , which is subsequently aggregated by a trainable MIL module into a slide-level embedding z . Finally, an MLP maps z to the prediction $\hat{y}$, and the downstream model is optimized using the cross-entropy loss $\mathcal{L}_{CE}$.

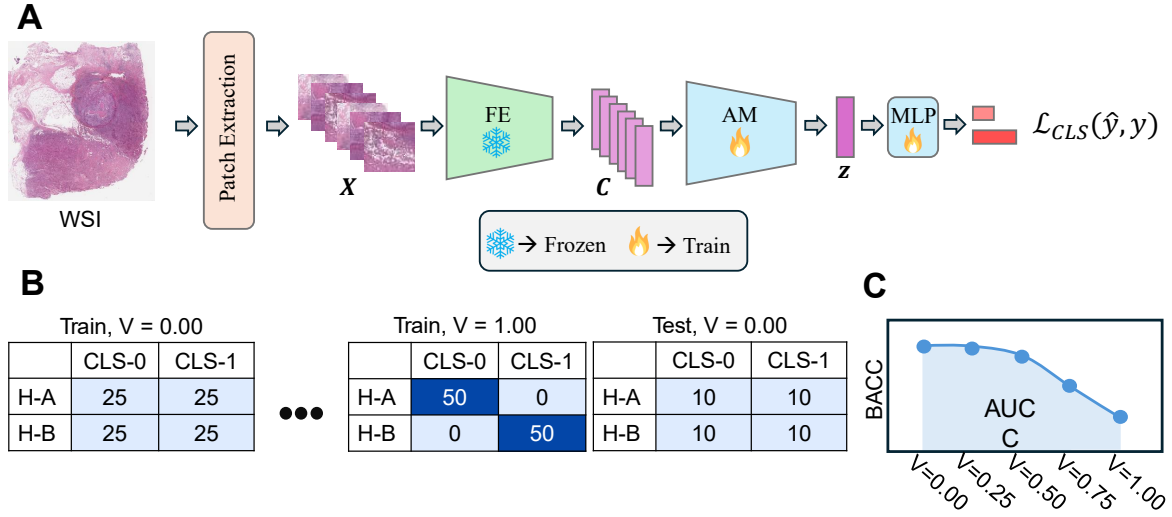


Figure 1. Framework overview: (A) MIL-based WSI classification, (B) center-label spurious correlations, and (C) Robustness quantification.

2.2. Center-Label Spurious Correlations

To assess reliance on center-related shortcuts, we construct training sets with progressively stronger associations between the class label $y \in Y$ and acquisition center $h \in H$, while keeping the number of WSIs fixed. The association is quantified through Cramér’s V ,

$$V = \sqrt{\frac{\chi^2}{M \min(|Y| - 1, |H| - 1)}}, \quad \chi^2 = \sum_{y \in Y} \sum_{h \in H} \frac{(O_{yh} - E_{yh})^2}{E_{yh}},$$

where M is the number of training WSIs, and O_{yh} and E_{yh} are the observed and expected counts for each class-center pair. As shown in Figure 1 B, Cramér’s $V \in [0, 1]$ ranges from independence ($V = 0$) to maximal association ($V = 1$). The test set is always kept uncorrelated ($V = 0$), such that performance degradation reflects reliance on the spurious center-label association.

2.3. Robustness Quantification with AUCC

To summarize downstream robustness across the different center-label correlations, we propose the Area Under the Cramér’s V Curve (AUCC) score. Given the ordered configurations $\{V_t\}_{t=0}^T$, we compute the area under the balanced-accuracy curve using trapezoidal integration,

$$\text{AUCC} = \sum_{t=1}^T \frac{\text{BACC}(V_t) + \text{BACC}(V_{t-1})}{2} (V_t - V_{t-1}),$$

where $\text{BACC}(V_t)$ denotes the balanced accuracy obtained when training under correlation level V_t , and $T + 1$ is the number of evaluated configurations. Unlike measuring performance degradation at a single correlation level, AUCC jointly captures absolute predictive performance and its stability as the spurious association increases. Thus, higher AUCC indicates a stronger performance-robustness trade-off across the complete range of center-label correlations.

3. Experimental design

3.1. Datasets and Preprocessing

We evaluate four binary WSI classification tasks spanning skin (AI4SKIN [5]), breast (CAMELYON16 [6], TCGA-BRCA) and lung (TCGA-NSCLC) cancer. For each dataset, we retain WSIs from two acquisition centers, enabling controlled manipulation of center-label correlations. Table 1 summarizes the resulting class and center distributions.

Dataset	Class	Center		Total
		H-A	H-B	
AI4SKIN [5]	Benign	133	166	299
	Malignant	136	191	327
CAMELYON16 [6]	Negative	150	89	239
	Positive	95	65	160
TCGA-BRCA	IDC	91	74	165
	ILC	18	21	39
TCGA-NSCLC	LUAD	84	91	175
	LUSC	50	41	91

Table 1. WSI distribution by class and acquisition center. IDC: invasive ductal carcinoma; ILC: invasive lobular carcinoma; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma.

WSIs undergo tissue detection, downsampling to $10\times$ magnification (1 MPP), and tiling into 512×512 patches. We deliberately avoid stain normalization to preserve center-associated appearance and apply the input transformations prescribed by each PFM.

3.2. Models and Evaluation Protocol

We evaluate six frozen PFMs: CONCH [7] and KEEP [12] as vision-language models, and VIRCHOW-2 [11], H-Optimus-1 [14], KAIKO [10], and UNI-2 [13] as vision-only models. Slide-level aggregation uses ABMIL [8] and TransMIL [9], trained for 20 epochs with AdamW and learning rates of 10^{-4} and 10^{-5} , respectively.

We use 5-fold cross-validation and evaluate five levels of Cramér’s V , with $V \in \{0, 0.25, 0.5, 0.75, 1\}$. Each config-

PFM	Aggregator	ComBat [4]	AUCC ($\uparrow$)				
			AI4SKIN [5]	CAMELYON16 [6]	TCGA-BRCA	TCGA-NSCLC	Average
CONCH Nat. Med.'24 [7]	ABMIL [8]	$\times$	88.42	79.06	88.77	83.76	85.00
		$\checkmark$	86.71 $\downarrow$ 1.71	81.13 $\downarrow$ 2.07	84.75 $\downarrow$ 4.02	83.37 $\downarrow$ 0.39	83.99 $\downarrow$ 1.01
	TransMIL [9]	$\checkmark$	87.54	66.19	90.52	81.37	81.40
KAIKO arXiv'24 [10]	ABMIL	$\times$	82.80	61.57	61.69	63.02	67.27
		$\checkmark$	81.95 $\downarrow$ 0.85	62.49 $\downarrow$ 0.92	60.96 $\downarrow$ 0.73	63.05 $\downarrow$ 0.03	67.11 $\downarrow$ 0.16
	TransMIL	$\checkmark$	82.21	57.49	59.21	62.26	65.29
VIRCHOW2 arXiv'24 [11]	ABMIL	$\times$	88.63	65.49	66.58	68.01	72.18
		$\checkmark$	87.16 $\downarrow$ 1.47	63.36 $\downarrow$ 2.13	65.21 $\downarrow$ 1.37	66.79 $\downarrow$ 1.22	70.63 $\downarrow$ 1.55
	TransMIL	$\checkmark$	87.13	60.90	66.88	67.09	70.50
KEEP Cancer Cell'24 [12]	ABMIL	$\times$	81.99	54.76	85.69	83.35	76.45
		$\checkmark$	81.08 $\downarrow$ 0.91	59.82 $\downarrow$ 5.06	77.21 $\downarrow$ 8.48	79.56 $\downarrow$ 3.79	74.42 $\downarrow$ 2.03
	TransMIL	$\checkmark$	87.58	57.89	88.50	81.36	78.83
UNI2 Nat. Med.'25 [13]	ABMIL	$\times$	87.55	63.58	67.23	69.05	71.85
		$\checkmark$	84.59 $\downarrow$ 2.96	63.39 $\downarrow$ 0.19	65.90 $\downarrow$ 1.33	68.80 $\downarrow$ 0.25	70.67 $\downarrow$ 1.18
	TransMIL	$\checkmark$	87.49	62.11	65.27	66.39	70.31
HOptimus1 AACR'25 [14]	ABMIL	$\times$	81.89	56.16	65.21	67.41	67.67
		$\checkmark$	81.14 $\downarrow$ 0.75	56.03 $\downarrow$ 0.13	63.17 $\downarrow$ 2.04	66.44 $\downarrow$ 0.97	66.69 $\downarrow$ 0.98
	TransMIL	$\checkmark$	82.57	54.70	63.00	67.71	67.00
	ABMIL	$\times$	82.92	55.07	62.94	67.10	67.01
		$\checkmark$	82.92 $\downarrow$ 0.35	55.07 $\downarrow$ 0.37	62.94 $\downarrow$ 0.06	67.10 $\downarrow$ 0.61	67.01 $\downarrow$ 0.01
	TransMIL	$\checkmark$	82.92	55.07	62.94	67.10	67.01

Table 2. AUCC robustness of patch-level pathology FMs across datasets and MIL aggregators. Arrows denote the absolute percentage-point change after ComBat correction. Model years indicate the initial public release of each PFM, rather than the publication year of the associated reference.

uration is repeated over 10 random training-set samplings while keeping the test fold fixed and uncorrelated ($V = 0$), with results averaged across folds and samplings.

4. Results

4.1. Representation-Level Center Encoding

Representation-level center encoding is assessed on mean-pooled slide embeddings using FM-SI [15] and RI [3]. FM-SI measures center-wise clustering (lower is better), whereas RI contrasts biological- and center-driven neighborhood structure (higher is better).

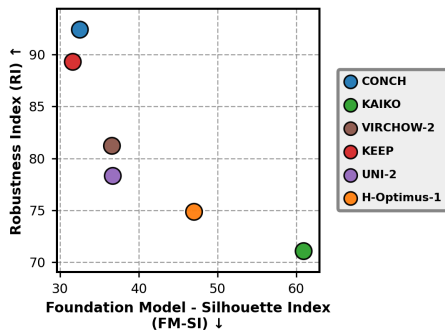


Figure 2. Slide-level robustness of patch-level pathology foundation models across center-related metrics.

Figure 2 reveals substantial differences in center robustness across PFMs, with lower FM-SI consistently associated with higher RI. CONCH and KEEP occupy the most robust region of the representation space, whereas KAIKO shows the

strongest center-related encoding, followed by H-Optimus-1. Notably, both vision-language PFMs outperform all vision-only models on the two complementary metrics, suggesting that, within our benchmark, multimodal pretraining yields representations less dominated by acquisition-center information.

4.2. Robustness to Spurious Correlations

Downstream performance is measured by BACC, while AUCC summarizes performance and robustness across the full range of Cramér’s V .

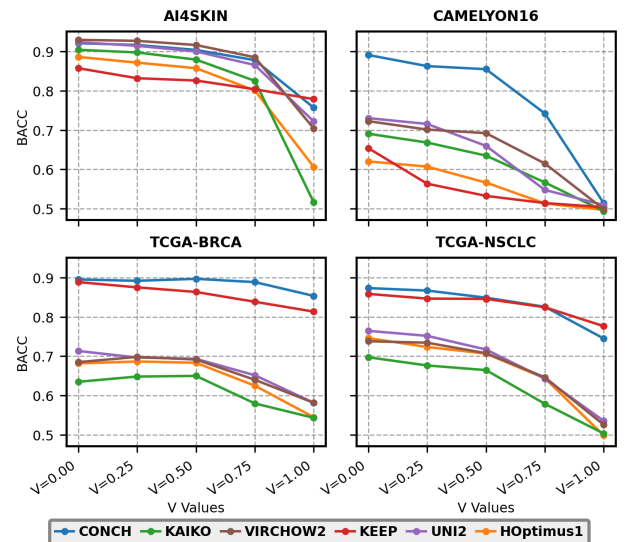


Figure 3. BACC under increasing spurious center-label correlation.

Table 2 shows that, without robustification, ABMIL achieves higher average AUCC for five out of six PFMs, outperforming TransMIL by 1.18 percentage points on average. This indicates that the aggregation strategy itself influences downstream robustness, with ABMIL generally preserving performance better under spurious center-label correlations. Figure 3 further reveals a consistent performance degradation as the correlation increases across models and datasets. CONCH and KEEP remain notably robust on TCGA-BRCA at high V , whereas most configurations deteriorate substantially, showing that center information encoded in PFM representations can propagate to slide-level predictions through spurious shortcuts.

4.3. Effect of ComBat Harmonization

As a widely used harmonization strategy, we apply ComBat [4] to mitigate center-associated bias in PFM embeddings. ComBat estimates and removes center-specific additive and scale effects while preserving biological variation.

Despite this correction, ComBat does not provide consistent robustness gains. Across the 48 PFM-aggregator-dataset configurations, AUCC decreases in 34 cases and improves in only 14, yielding an average change of -0.63 percentage points. Moreover, its effect is strongly dataset-dependent, with frequent improvements on CAMELYON16 but almost systematic degradation on TCGA-BRCA. These results suggest that generic feature harmonization is insufficient to reliably mitigate center-related shortcuts and may compromise discriminative information in some settings.

5. Conclusion

This work provides a controlled evaluation of center-related robustness in pathology foundation models, showing that center information encoded at the representation level can propagate to downstream WSI classification under spurious center-label correlations. Across six PFMs, four datasets, and two MIL aggregators, the proposed AUCC enables a unified assessment of predictive performance and robustness, while ComBat provides limited and inconsistent mitigation of center effects. Our study is nevertheless constrained by the controlled Cramér’s V setting, which restricts the available training samples and focuses on binary tasks with two acquisition centers, as well as by the use of two MIL aggregators and patch-level PFMs. Future work will therefore extend this analysis to broader clinical settings and slide-level foundation models, while developing center-robust foundation models and plug-in robustification strategies that can mitigate acquisition-related shortcuts without compromising discriminative information.

Acknowledgments

This work has been supported by the Generalitat Valenciana (GVA) through the project CIPROM/2022/20 (PROMETEO) and the Spanish Ministry of Science and Innovation under PID2022-140189OB-C21 (ASSIST).

References

- [1] M. Ochi, D. Komura, y S. Ishikawa, “Pathology foundation models”, *JMA journal*, 8no., 1, pp. 121–130, 2025.
- [2] G. Campanella, M. G. Hanna, L. Geneslaw, A. Mirafior, y cols., “Clinical-grade computational pathology using weakly supervised deep learning on whole slide images”, *Nature medicine*, 25no., 8, pp. 1301–1309, 2019.
- [3] J. Kömen, E. D. de Jong, J. Hense, H. Marienwald, y cols., “Towards robust foundation models for digital pathology”, *Nature Communications*, 17no., 1, pp. 5218, 2026.
- [4] W. E. Johnson, C. Li, y A. Rabinovic, “Adjusting batch effects in microarray expression data using empirical Bayes methods”, *Biostatistics*, 8no., 1, pp. 118–127, 2007.
- [5] R. Del Amor, M. López-Pérez, P. Meseguer, S. Morales, y cols., “A fusocelular skin dataset with whole slide images for deep learning models”, *Scientific Data*, 12no., 1, pp. 788, 2025.
- [6] G. Litjens, P. Bandi, B. Ehteshami Bejnordi, O. Geessink, y cols., “1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset”, *Giga-Science*, 7no., 6, pp. giv065, 2018.
- [7] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, y cols., “A visual-language foundation model for computational pathology”, *Nature medicine*, 30no., 3, pp. 863–874, 2024.
- [8] M. Ilse, J. Tomczak, y M. Welling, “Attention-based deep multiple instance learning”, *International conference on machine learning*, 2018, pp. 2127–2136.
- [9] Z. Shao, H. Bian, Y. Chen, Y. Wang, y cols., “Transmil: Transformer based correlated multiple instance learning for whole slide image classification”, *Advances in neural information processing systems*, 34, pp. 2136–2147, 2021.
- [10] N. Aben, E. D. de Jong, I. Gatopoulos, N. Känzig, y cols., “Towards large-scale training of pathology foundation models”, *arXiv preprint arXiv:2404.15217*, 2024.
- [11] E. Zimmermann, E. Vorontsov, J. Viret, A. Casson, y cols., “Virchow2: Scaling self-supervised mixed magnification models in pathology”, *arXiv preprint arXiv:2408.00738*, 2024.
- [12] X. Zhou, L. Sun, D. He, W. Guan, y cols., “Knowledge-enhanced pretraining for vision-language pathology foundation model on cancer diagnosis”, *Cancer Cell*, 44no., 4, pp. 777–791, 2026.
- [13] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, y cols., “Towards a General-Purpose Foundation Model for Computational Pathology”, *Nature Medicine*, 2024.
- [14] M. Scalbert, C. Saillard, T. Peeters, L. Gonzalez, y cols., “H-optimus-1: A foundation model for computational histopathology”, *Proceedings of the American Association for Cancer Research Annual Meeting 2026; Part 2 (Late-Breaking, Clinical Trial, and Invited Abstracts)*, 2026, pp. LB174.
- [15] P. Meseguer, R. del Amor, y V. Naranjo, “Benchmarking histopathology foundation models in a multi-center dataset for skin cancer subtyping”, *Annual Conference on Medical Image Understanding and Analysis*, 2025, pp. 16–28.