Harnessing Adversarial Distillation to Customise Debiased, Disease-Specific Pathology Foundation Models for Breast Cancer

Zhiwei Chen[1⋆], Yang Hu[2,3,4⋆(✉)], Yuxiang Xiao[1], Yakun Ju[2], Tianyang Zhang[4], Yingxue Xu[6], Wei Li[7], Hao Chen[6], Jens Rittscher[4,5], and Kaixiang Yang[1(✉)]

[1] School of Computer Science and Engineering, South China University of Technology, China
[2] School of Computing and Mathematical Sciences, University of Leicester, UK
[3] Leicester Cancer Research Centre, University of Leicester, UK
[4] Department of Engineering Science, University of Oxford, UK
[5] Nuffield Department of Medicine, University of Oxford, UK
[6] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China
[7] ZoyMed, China

Abstract Pathology foundation models (PFMs) provide strong tissue representations and have become central to digital pathology. However, deployment in disease-specific settings is limited by 1) the high computational cost of billion-parameter PFMs and 2) distribution mismatch and non-biological bias inherited from pan-cancer, multi-centre pre-training, including site-specific signatures and imbalanced disease prevalence. These factors can encourage shortcut learning and under-emphasise subtle morphology required for reliable modelling of a specific cancer type. We present **SmartStu** (a **Smart Stu**dent), a framework to customise compact, breast-cancer-specific PFMs via distillation whilst mitigating confounding. SmartStu distils representations from multiple teacher PFMs into a lightweight student backbone. Crucially, we introduce **adversarial distillation** that leverages a dedicated noise model trained to predict nuisance, edge-dominated cues on the distillation set. Using this noise model as a counterexample, the adversarial objective encourages the student to recognise, yet suppress, features predictive of nuisance targets. We further incorporate multi-teacher ensemble distillation and an auxiliary self-supervised objective with artefact injection. We validate SmartStu on three external cohorts (Yale HER2, SLN-Breast, and BRACS) with multiple tiny backbones. SmartStu yields breast-cancer-specific PFMs that are over 30× smaller than general PFMs whilst largely preserving, and sometimes improving, downstream performance measured by balanced accuracy (bAcc) and AUC. Code is available at https://github.com/zwchen03/advDistall.

Keywords: Pathology foundation models · Active debiasing · Adversarial distillation · Breast cancer

⋆ Zhiwei Chen and Yang Hu contribute equally to this work.
Corresponding email: superhy199148@hotmail.com and yangkx@scut.edu.cn

1 Introduction

Self-supervised pre-training has driven the rapid emergence of pathology foundation models (PFMs), enabling strong and transferable representations for histology across a wide range of downstream tasks [5,20,27,6,22,11,28]. Pre-trained on large, heterogeneous collections of whole-slide images (WSIs), these models capture broad morphological priors that outperform classic histopathology feature extraction approaches limited by annotation cost and dataset scale [4,7]. Despite these advances, deploying PFMs in disease-specific clinical workflows remains challenging due to the following bottlenecks.

First, existing PFMs' parameter count and inference-time cost hinder high-throughput processing and deployment in resource-constrained clinical environments [18,23]. Scaling PFMs towards whole-slide modelling and ultra-long context [25,24] typically increases memory use and latency, restricting deployment to well-provisioned settings. Second, PFMs pre-trained on multi-centre datasets inevitably inherit non-biological biases from tissue preparation, staining protocols, and scanning pipelines. These site (collection source)-specific signatures can be exploited as shortcuts by downstream models, degrading out-of-distribution generalisation [13,21,17]. Finally, a task mismatch can arise from pan-cancer pre-training: PFMs optimised for broad transfer across multiple cancer types may under-emphasise subtle, cancer-specific morphology required for reliable modelling in the target disease.

An intuitive direction is to customise PFMs via knowledge distillation (KD), compressing large, general teachers into compact, task-specific students by matching representations. In histopathology, distillation can yield efficient models that preserve much of the teacher performance [10]. However, standard distillation largely transfers whatever the teacher encodes and offers limited control over propagating site-related bias into the student. Domain generalisation and debiasing techniques (e.g., stain normalisation) can reduce domain shift [1,15], but may either suppress diagnostically relevant variation or leave residual centre-identifiable structure in the representation, arising from acquisition pipelines and case-mix confounding rather than superficial stain differences.

We present **SmartStu**, a framework for customising compact, disease-specific PFMs for breast cancer by distilling from multiple teacher PFMs into a lightweight student, whilst mitigating confounding inherited from centralised pre-training. SmartStu introduces adversarial distillation: we train a dedicated noise model to predict nuisance targets and capture site-related cues on the distillation set, then use it as a counterexample during KD. We optimise an adversarial objective that encourages the student to recognise biased directions yet suppress features predictive of nuisance targets. In addition, SmartStu incorporates an auxiliary self-supervised objective with multi-type artefact injection to improve stability under acquisition variability. We summarise our contributions as:

1) We propose **SmartStu**, a disease-specific and lightweight pathology foundation model framework with an adversarial distillation mechanism that actively suppresses site-related confounding during customisation.

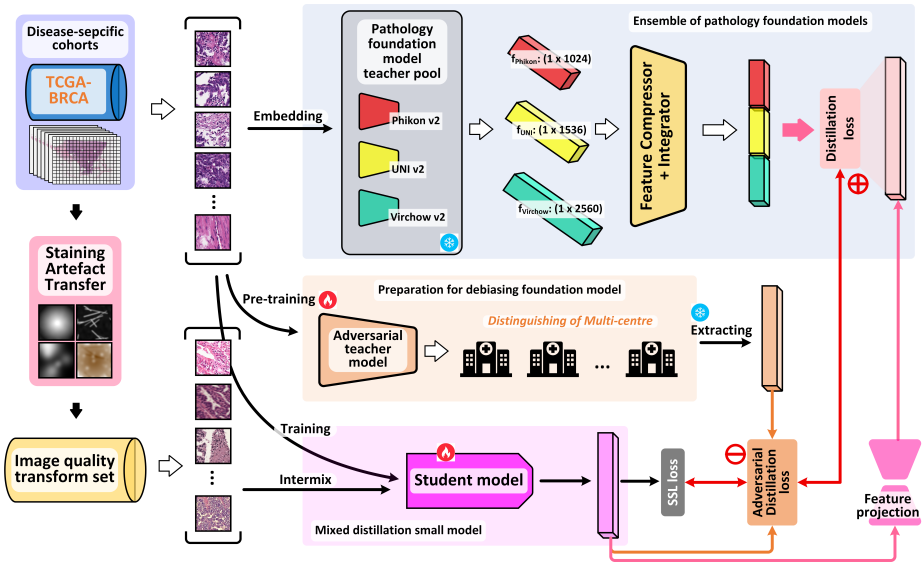

Figure 1. Overview of **SmartStu**: multi-teacher feature distillation into a compact student, augmented with **adversarial (noise) distillation** to suppress site-related confounding, and an auxiliary **self-supervised** objective with artefact injection.

2) We integrate complementary robustness components, including multi-teacher ensemble distillation and an auxiliary self-supervision with multiple artefact injection, to stabilise representation learning under realistic variability.

3) We evaluate SmartStu across 3 external breast cancer cohorts (Yale HER2, SLN-Breast, and BRACS) and various student backbones. SmartStu yields compact breast cancer PFMs that are over 30× smaller than general PFMs whilst preserving, and occasionally improving, downstream performance.

2 Methods

SmartStu is a compact tile-level disease-specific encoder and distils complementary knowledge from multiple teacher PFMs into a lightweight student while explicitly suppressing bias via adversarial distillation with a dedicated noise model (see Fig. 1). Given a whole slide image (WSI) X, the tissue region is split into M non-overlapping tiles $\mathcal{I} = \{x_i\}_{i=1}^{M}$, extracted at 20× magnification with size 256 × 256 pixels. SmartStu is trained to produce domain-specific tile embeddings; weakly supervised slide-level prediction is obtained downstream by feeding the distilled embeddings into a standard MIL aggregator [14].

2.1 Multi-teacher feature distillation

For each tile $x_i \in \mathcal{X}$, the student encoder produces $z_i^s = S(x_i)$. To transfer diverse pathology semantics, we use a teacher pool of K frozen PFMs $\{T_k\}_{k=1}^{K}$.

For each $x_i \in \mathcal{X}$, teacher T_k produces an embedding $f_i^{(k)} = T_k(x_i) \in \mathbb{R}^{d_k}$. Since teacher embeddings have varying dimensions, we apply a lightweight alignment head $\phi_k(\cdot)$ to map each teacher feature into a shared dimension D. We then form the concatenated teacher target:

$$z_i^t = \mathrm{Concat}\Big(\phi_1(f_i^{(1)}), \phi_2(f_i^{(2)}), \ldots, \phi_K(f_i^{(K)})\Big) \in \mathbb{R}^{K \cdot D}. \tag{1}$$

We map student features to the teacher target space via a projection head, $h_i^{dist} = \psi_{dist}(z_i^s) \in \mathbb{R}^{K \cdot D}$, and minimise a feature-level mean-squared error between $\{h_i^{dist}\}$ and $\{z_i^t\}$:

$$\mathcal{L}_{distill} = \frac{1}{|\mathcal{X}|} \sum_{i=1}^{|\mathcal{X}|} \left\| z_i^t - h_i^{dist} \right\|_2^2. \tag{2}$$

2.2 Adversarial distillation for de-biasing

SmartStu introduces adversarial distillation by learning a noise model that captures nuisance cues and then training the student to suppress them.

Noise model pre-training (adversarial teacher). We pre-train an auxiliary adversarial teacher T_{adv} on a multi-centre cohort with site-wise nuisance labels. For a tile x with site label $y_c \in \{1, \ldots, C\}$, we extract features $v = T_{adv}(x)$ and project them to a normalised embedding z. We optimise a hybrid objective that combines contrastive learning [16] with cross-entropy ($\mathrm{CE}(\cdot)$) site classification:

$$\mathcal{L}_{ctr} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)}, \tag{3}$$

$$\mathcal{L}_{pre} = \mathrm{CE}(y_c, p_c) + \lambda \mathcal{L}_{ctr}. \tag{4}$$

Here $P(i)$ denotes positives from the same site, $A(i)$ contains all samples in the batch, τ is a temperature, and p_c is the predicted site distribution. After pre-training, T_{adv} is frozen and used as a nuisance-feature extractor.

Adversarial distillation loss. For each $x_i \in \mathcal{X}$, the frozen noise teacher outputs a nuisance signature $z_i^{adv} = T_{adv}(x_i)$. The student feature z_i^s is mapped by a feature alignment head $\psi_{adv}(\cdot)$ and then passed through a gradient reversal layer (GRL) [12], giving $h_i^{adv} = \mathcal{R}(\psi_{adv}(z_i^s))$. We use cosine similarity (bounded in $[-1, 1]$) to measure alignment between h_i^{adv} and z_i^{adv}:

$$\mathcal{L}_{adv} = 1 - \mathrm{CosSim}(z_i^{adv}, h_i^{adv}) = 1 - \frac{z_i^{adv} \cdot h_i^{adv}}{\|z_i^{adv}\|_2 \|h_i^{adv}\|_2}. \tag{5}$$

With GRL, minimising $\mathcal{L}_{adv}$ discourages the student from encoding site-related confounders in its representation.

Table 1. Performance comparison when using UNI as the only teacher. Results are reported as balanced accuracy (bAcc) and AUC (mean for 5-runs). **Bold+shaded** indicates the best result, while **bold** indicates the best record for each student backbone.

		Yale HER2		SLN-Breast		BRACS	
Model	Mode	bAcc	AUC	bAcc	AUC	bAcc	AUC
UNI-v2 (681M)		**0.9032**	**0.9500**	**0.9323**	**0.9805**	**0.5083**	**0.8699**
ResNet18 (11.18M)	Default	0.7642	0.8479	0.7707	0.8722	0.3281	0.7455
	Sig-SmartStu	0.8208	0.9105	0.8647	0.9143	0.3697	0.8083
	Adv-SmartStu	0.8361	0.9121	0.8842	0.9308	0.3984	0.8163
	AdvS-SmartStu	**0.8468**	**0.9226**	**0.8985**	**0.9534**	**0.4378**	**0.8180**
TinyViT-5M (5.07M)	Default	0.7513	0.8553	0.8323	0.8917	0.3497	0.7300
	Sig-SmartStu	0.8450	0.9192	0.9038	0.9188	0.4002	0.8016
	Adv-SmartStu	**0.8553**	0.9247	**0.9090**	0.9308	0.4088	0.8039
	AdvS-SmartStu	0.8453	**0.9268**	0.9060	**0.9398**	**0.4119**	**0.8198**
MobileNetV3-Small (0.93M)	Default	0.7582	0.8474	0.8263	0.8827	0.3313	0.7414
	Sig-SmartStu	0.8103	0.9011	0.8752	0.9218	0.3618	0.7529
	Adv-SmartStu	**0.8258**	**0.9079**	0.8752	**0.9233**	**0.3666**	0.7791
	AdvS-SmartStu	0.8205	0.9063	**0.9038**	0.9203	0.3622	**0.8030**

2.3 Auxiliary self-supervised adaptation & overall optimisation

Training batch with artefact injection. We apply an artefact injection operator $\mathcal{A}(\cdot)$ during training, including 1) brown colour cast to mimic stain shift, 2) exposure/illumination jitter, 3) resolution degradation via downsampling with lossy compression, and 4) additive Gaussian noise. For each mini-batch $\mathcal{B}$, we form $\mathcal{X} = \mathcal{B} \uplus \mathcal{A}(\mathcal{B})$ and optimise all objectives on $\mathcal{X}$.

Self-supervised adaptation (SSL adaptation). On $\mathcal{X}$, we add a DINO-style self-supervised term [5]. For each tile, we sample two global crops and two local crops (after $\mathcal{A}(\cdot)$), and compute their predicted distributions via the student projection head and softmax. Denoting the predictions of global crops as P^g, P^g_{aug} and those of local crops as P^l, P^l_{aug}, we minimise the symmetric cross-entropy:

$$\mathcal{L}_{ssl} = \frac{1}{2}\Big(\mathrm{CE}(P^g, P^l_{aug}) + \mathrm{CE}(P^g_{aug}, P^l)\Big). \tag{6}$$

Overall optimisation objective. The final training objective is a weighted combination of multi-teacher distillation, adversarial (noise) distillation, and self-supervised adaptation:

$$\mathcal{L}_{all} = \mathcal{L}_{distill} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{ssl}\mathcal{L}_{ssl}, \tag{7}$$

where λ_{adv} and λ_{ssl} balance the contribution of de-biasing and SSL adaptation.

3 Experiments and Results

3.1 Datasets and implementation details

Datasets. We pre-train SmartStu via distillation on TCGA-BRCA (n = 1,133; 40 sites)[1]. Generalisation is evaluated on three external breast cancer cohorts:

[1] https://portal.gdc.cancer.gov/projects/TCGA-BRCA

Table 2. Performance comparison when distilling from an ensemble of three teacher foundation models (UNI-v2, Phikon-v2, and Virchow-v2).

Model	Mode	Yale HER2		SLN-Breast		BRACS	
		bAcc	AUC	bAcc	AUC	bAcc	AUC
Original PFMs (300M ∼ 680M)	UNI-v2	0.9032	0.9500	**0.9323**	0.9805	**0.5083**	0.8699
	Phikon-v2	**0.9082**	**0.9505**	0.9271	0.9609	0.4527	0.8482
	Virchow-v2	0.8629	0.9416	0.9203	**0.9865**	0.5059	**0.8747**
ResNet18 (11.18M)	Default	0.7642	0.8479	0.7707	0.8722	0.3281	0.7455
	Itg-SmartStu	0.8366	0.9089	0.8895	0.9248	0.4298	0.8483
	Adv-SmartStu	0.8582	0.9153	0.9075	0.9383	0.4488	**0.8573**
	AdvS-SmartStu	**0.8674**	**0.9189**	**0.9128**	**0.9549**	**0.4576**	0.8489
ResNet34 (21.28M)	Default	0.7787	0.8405	0.8030	0.8977	0.3471	0.7638
	Itg-SmartStu	0.8447	0.9337	0.8865	0.9519	0.4524	0.8553
	Adv-SmartStu	0.8605	**0.9384**	**0.9113**	**0.9684**	0.4911	0.8574
	AdvS-SmartStu	**0.8708**	0.9363	0.8955	0.9474	**0.5085**	**0.8596**
ResNet50 (23.51M)	Default	0.8000	0.8758	0.8504	0.9098	0.3396	0.7150
	Itg-SmartStu	0.8616	0.9379	0.9090	0.9594	0.4699	0.8490
	Adv-SmartStu	0.8929	0.9384	0.9180	0.9609	0.4939	0.8482
	AdvS-SmartStu	**0.8976**	**0.9411**	**0.9233**	**0.9820**	**0.5109**	**0.8591**
TinyViT-5M (5.07M)	Default	0.7513	0.8553	0.8323	0.8917	0.3497	0.7300
	Itg-SmartStu	0.8561	0.9326	0.9090	0.9293	0.4115	0.7922
	Adv-SmartStu	**0.8666**	0.9321	0.9143	0.9459	0.4326	0.8076
	AdvS-SmartStu	0.8550	**0.9342**	**0.9180**	**0.9504**	**0.4434**	**0.8118**
TinyViT-11M (10.55M)	Default	0.7871	0.8511	0.8271	0.8857	0.3664	0.7833
	Itg-SmartStu	0.8816	0.9332	0.9143	0.9564	0.4153	0.8328
	Adv-SmartStu	**0.8916**	0.9395	0.9218	0.9429	0.4405	0.8361
	AdvS-SmartStu	**0.8916**	**0.9432**	**0.9271**	**0.9699**	**0.4902**	**0.8477**
TinyViT-21M (20.62M)	Default	0.7779	0.8268	0.8752	0.9398	0.3836	0.8028
	Itg-SmartStu	0.8924	0.9468	0.9180	0.9684	0.4367	0.8272
	Adv-SmartStu	**0.8974**	0.9489	0.9233	0.9684	0.4607	0.8359
	AdvS-SmartStu	**0.8974**	**0.9505**	**0.9323**	**0.9940**	**0.4958**	**0.8410**
MobileNetV3-Small (0.93M)	Default	0.7582	0.8474	0.8263	0.8827	0.3313	0.7414
	Itg-SmartStu	0.8147	0.8989	0.8805	0.9293	0.3657	0.7923
	Adv-SmartStu	**0.8303**	**0.9079**	0.8805	0.9489	**0.3905**	0.7914
	AdvS-SmartStu	0.8253	0.9061	**0.9090**	**0.9639**	0.3886	**0.8065**
MobileNetV3-Large (2.97M)	Default	0.7908	0.8579	0.8436	0.8782	0.3443	0.7511
	Itg-SmartStu	0.8205	0.9142	0.8985	0.9459	0.3978	0.8282
	Adv-SmartStu	**0.8671**	**0.9189**	**0.9271**	**0.9474**	**0.4162**	**0.8312**
	AdvS-SmartStu	0.8363	0.9137	0.9090	0.9459	0.4062	0.8258

Yale-HER2 ($n = 192$) [8,9], SLN-Breast ($n = 130$) [3,4], and BRACS ($n = 547$) [2]. Yale-HER2 and SLN-Breast are binary classification tasks with a 0.8/0.2 train/test split. BRACS is a 7-class classification task, and we follow the dataset-default train/test split (0.84/0.16).

Implementation. WSIs are tiled into 256×256 patches from tissue regions. Preprocessing is performed with Trident toolbox [26]. Training has two stages: 1) adversarial teacher training with learning rate 1×10^{-4}, weight decay 0.05, $\tau = 0.07$, and $\lambda = 1$; 2) student distillation with learning rate 1×10^{-4}, weight decay 0.05, and loss weights $\lambda_{adv} = 10^{-4}$, $\lambda_{ssl} = 10^{-4}$.

For downstream evaluation, we employ ABMIL [14] as the MIL aggregator. Our student models are compared against several PFMs, including UNI-V2 [6], Phikon-V2 [11], and Virchow-V2 [28]. We evaluate a series of lightweight backbones for the student model, specifically ResNet (18, 34, 50), TinyViT (5M, 11M, 21M), and MobileNet-v3 (Small, Large). All experiments are conducted

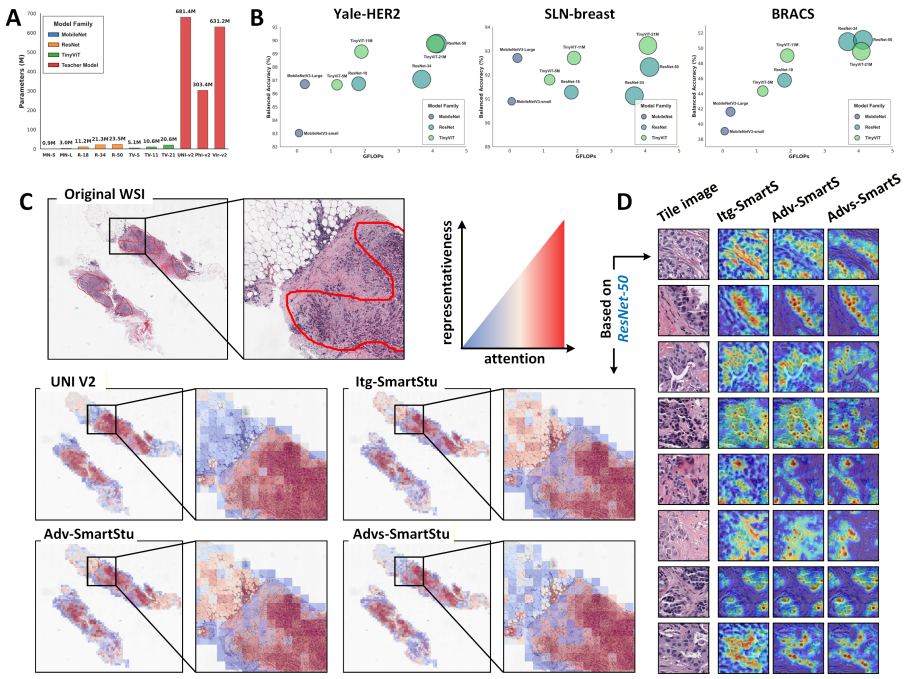

Figure 2. Efficiency and interpretability analysis of the proposed method. **(A)** Parameter count comparison between student and teacher models. Abbreviations: MN (MobileNet-V3), R (ResNet), and TV (TinyViT). **(B)** Performance-efficiency trade-off (Balanced Accuracy vs. GFLOPs), where bubble size is proportional to the parameter count. **(C)** Slide-level attention heatmaps comparison on a sample from the Yale HER2 test set. **(D)** Fine-grained interpretability visualization using tile-level Eigen-CAM.

with 5 runs, with the average performance reported. We employ balanced accuracy (BACC) and the area under the receiver operating characteristic curve (AUC) as the evaluation metrics.

In this context, **Default** denotes the model initialised with ImageNet[2] pre-trained weights. **Sig-SmartStu** and **Itg-SmartStu** denote single- and multi-teacher variants trained with distillation only (i.e., $\mathcal{L}_{distill}$). **Adv-SmartStu** is trained with $\mathcal{L}_{distill} + \lambda_{adv}\mathcal{L}_{adv}$. Finally, **AdvS-SmartStu** is trained with the full objective $\mathcal{L}_{all}$.

3.2 Distillation performance results

Single-teacher setting. We first distil from UNI-v2 as the sole teacher (Table 1). Compared with *Default* (ImageNet pre-trained), distillation improves downstream performance, confirming successful knowledge transfer to compact back-

[2] https://www.image-net.org

bones. Although these students generally remain below using frozen UNI features directly, adding adversarial distillation and then AdvS-SmartStu with the full objective still produces a clear, stepwise improvement, indicating that each component contributes to stronger cross-cohort generalisation.

Multi-teacher setting. We next distil from an ensemble of three teachers (UNI-v2, Phikon-v2 and Virchow-v2). Table 2 shows a consistent progression: Itg-SmartStu (multi-teacher distillation only) improves over Default, and Adv-SmartStu further improves by explicitly discouraging site-related confounders. AdvS-SmartStu is commonly best or competitive across student families, suggesting that noise-guided de-biasing and artefact-augmented self-supervision encourage features that are less sensitive to acquisition-specific signatures while retaining diagnostically relevant morphology. However, very small backbones do not always benefit from the auxiliary self-supervised term, which is consistent with limited capacity and stronger regularisation pressure.

Across external cohorts (Table 2), compact students can match or slightly exceed teachers in various backbone settings. This indicates that SmartStu can consolidate multiple teachers' knowledge into a lightweight encoder while suppressing site-related bias via adversarial distillation; the auxiliary artefact-injection self-supervision further stabilises the learned features. Overall, these results support the portability of SmartStu under different deployment constraints.

3.3 Efficiency and interpretable visualisation

Figure 2-A shows that SmartStu students require substantially fewer parameters than the teacher PFMs, improving deployability in resource-constrained circumstances. Figure 2-B illustrates the relationship between performance and compute (BACC vs. GFLOPs). Overall, SmartStu concentrates strong performance in the low-compute regime, and several student architectures (notably ResNet-50 and TinyViT-21M) remain competitive across cohorts whilst operating at tens-fold lower computational cost than the teachers.

Figure 2-C compares tissue-level attention on a test WSI, where red contours mark tumour region of interest (ROI) annotations. In this example, the teacher PFM exhibits spurious activation outside the annotated ROI. SmartStu suppresses these false responses and concentrates attention on the ROI, improving localisation. AdvS-SmartStu further reduces background activation, producing a cleaner and more specific attention map on tumour regions.

Figure 2-D shows tile-level interpretation visualised by Eigen-CAM [19]. Activations become progressively more localised from Itg-SmartStu to AdvS-SmartStu, aligning better with salient morphology (e.g., nuclear boundaries and histological textures) whilst suppressing responses in less informative regions.

4 Conclusion

We propose **SmartStu**, a framework for customising compact, breast-cancer pathology foundation models. SmartStu introduces adversarial distillation to

suppress site-related bias whilst retaining disease-relevant morphology, and incorporates multi-teacher distillation with artefact-injected self-supervision to stabilise tile representations. Across external cohorts, SmartStu substantially reduces model size and compute whilst maintaining strong downstream performance. More broadly, adversarial distillation offers a controllable customisation paradigm for PFMs: by varying or combining nuisance targets, one can discourage specific non-biological signatures during knowledge transfer to better match personalised deployment. Future work will explore richer nuisance targets and adaptive combinations of adversarial objectives across clinical settings.

Acknowledgments. This work was supported by the following grants. JR is supported by the NIHR Oxford Biomedical Research Centre. HC is supported by the HKUST 30 for 30 Research Initiative Scheme (Project No. FS111). KY is supported by the National Natural Science Foundation of China (No. 62476101) and the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515140137). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Asadi-Aghbolaghi, M., Darbandsari, A., Zhang, A., Contreras-Sanz, A., Boschman, J., Ahmadvand, P., Köbel, M., Farnell, D., Huntsman, D.G., Churg, A., et al.: Learning generalizable ai models for multi-center histopathology image classification. NPJ Precision Oncology **8**(1), 151 (2024)
2. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., et al.: Bracs: A dataset for breast carcinoma subtyping in h&e histology images. Database **2022**, baac093 (2022)
3. Campanella, G., Hanna, M.G., Brogi, E., Fuchs, T.J.: Breast metastases to axillary lymph nodes. The Cancer Imaging Archive (2019), dOI: 10.7937/TCIA.2019.3XBN2JCC
4. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)
5. 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)
6. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024)
7. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A.: Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature medicine **24**(10), 1559–1567 (2018)

8. Farahmand, S., Fernandez, A.I., Ahmed, F.S., Rimm, D.L., Chuang, J.H., Reisenbichler, E., Zarringhalam, K.: Deep learning trained on hematoxylin and eosin tumor region of interest predicts her2 status and trastuzumab treatment response in her2+ breast cancer. Modern Pathology **35**(1), 44–51 (2022)
9. Farahmand, S., Fernandez, A.I., Ahmed, F.S., Rimm, D.L., Chuang, J.H., Reisenbichler, E., Zarringhalam, K.: HER2 and trastuzumab treatment response H&E slides with tumour ROI annotations (version 3). The Cancer Imaging Archive (2022), dOI: 10.7937/E65C-AM96
10. Filiot, A., Dop, N., Tchita, O., Riou, A., Dubois, R., Peeters, T., Valter, D., Scalbert, M., Saillard, C., Robin, G., et al.: Distilling foundation models for robust and efficient models in digital pathology. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2025. pp. 162–172. Springer (2025)
11. Filiot, A., Jacob, P., Mac Kain, A., Saillard, C.: Phikon-v2: A large, public feature extractor for biomarker prediction. arXiv preprint arXiv:2409.09173 (2024)
12. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of machine learning research **17**(59), 1–35 (2016)
13. Howard, F.M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O.I., Kather, J.N., et al.: The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nature communications **12**(1), 4423 (2021)
14. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
15. 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 Computing Surveys **57**(11), 1–37 (2025)
16. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
17. Kömen, J., Marienwald, H., Dippel, J., Hense, J.: Do histopathological foundation models eliminate batch effects? a comparative study. arXiv preprint arXiv:2411.05489 (2024)
18. Li, D., Wan, G., Wu, X., Wu, X., Nirmal, A.J., Lian, C.G., Sorger, P.K., Semenov, Y.R., Zhao, C.: A survey on computational pathology foundation models: Datasets, adaptation strategies, and evaluation tasks. arXiv preprint arXiv:2501.15724 (2025)
19. Muhammad, M.B., Yeasin, M.: Eigen-cam: Class activation map using principal components. In: 2020 international joint conference on neural networks (IJCNN). pp. 1–7. IEEE (2020)
20. 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 (2023)
21. Veta, M., Pluim, J.P., Van Diest, P.J., Viergever, M.A.: Breast cancer histopathology image analysis: A review. IEEE transactions on biomedical engineering **61**(5), 1400–1411 (2014)
22. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., et al.: A foundation model for clinical-grade computational pathology and rare cancers detection. Nature medicine **30**(10), 2924–2935 (2024)

23. Xiong, C., Chen, H., Sung, J.J.: A survey of pathology foundation model: Progress and future directions. arXiv preprint arXiv:2504.04045 (2025)
24. Xu, H., Usuyama, 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 **630**(8015), 181–188 (2024)
25. Xu, Y., Wang, Y., Zhou, F., Ma, J., Jin, C., Yang, S., Li, J., Zhang, Z., Zhao, C., Zhou, H., et al.: A multimodal knowledge-enhanced whole-slide pathology foundation model. Nature Communications (2025)
26. Zhang, A., Jaume, G., Vaidya, A., Ding, T., Mahmood, F.: Accelerating data processing and benchmarking of ai models for pathology. arXiv preprint arXiv:2502.06750 (2025)
27. 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 (2021)
28. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Fuchs, T., Fusi, N., Liu, S., Severson, K.: Virchow2: Scaling self-supervised mixed-magnification models in pathology. arXiv preprint arXiv:2408.00738 (2024)