Multi-Teacher Contrastive Distillation for Edge-Efficient Pathology Foundation Models

Tim Lenz[1†], Maurice Heide[1†], Marco Gustav[1], Nic G. Reitsam[1,2,3], and Jakob Nikolas Kather[1,4,5,6*]

[1] EKFZ for Digital Health, TU Dresden, Dresden, Germany
[2] Pathology, University of Augsburg, Augsburg, Germany
[3] Bavarian Cancer Research Center (BZKF), Augsburg, Germany
[4] Department of Medicine I, TU Dresden, Dresden, Germany
[5] Medical Oncology, NCT Heidelberg, Heidelberg, Germany
[6] Pathology & Data Analytics, University of Leeds, Leeds, United Kingdom
kather.jn@tu-dresden.de

Abstract. Computational pathology foundation models (PFMs) have advanced whole-slide image analysis. However, their size and inference cost hinder local deployment in pathology departments. We propose MuCoDi, a pretraining framework that distills frozen tile embeddings from multiple PFMs into compact edge-oriented encoders. Instead of regressing individual teacher features, MuCoDi trains lightweight MobileOne and RepViT students with a contrastive distillation objective adapted from MoCo v3, where cached Virchow2, UNI2, and H-Optimus-1 embeddings replace momentum-encoder keys. We pretrain students on 14.3M TCGA tiles from only 11.8K WSIs and evaluate frozen encoders on 23 clinically curated downstream classification tasks. RepViT-based MuCoEdge students retain near-teacher performance while reducing model size by orders of magnitude: MuCoEdge-R2.3 and MuCoEdge-R1.5 reach 71.0% external AUROC, within 0.8 percentage points of the best teacher (Virchow2, 71.8%), while MuCoEdge-R2.3 obtains the best external F1 and the second-best AUPRC (51.8% and 53.3%). MuCoEdge-R1.0 reaches 70.9% AUROC with only 6.4M parameters and 1.12 GFLOPs. On a Raspberry Pi 5, sub-million-parameter MobileOne students achieve up to 605× single-tile speedup over Virchow2 while retaining 66.5–66.9% external AUROC, demonstrating that PFM-quality pathology representations can be moved toward practical edge deployment. Code is available at https://github.com/KatherLab/MuCoDi.

Keywords: Computational pathology · Foundation models · Contrastive distillation · Knowledge distillation · Edge deployment

1 Introduction

The central promise of computational pathology is to bring AI assistance to pathologists, rather than asking pathologists to adapt to AI infrastructure. This

[†] Equal contribution, * Corresponding author

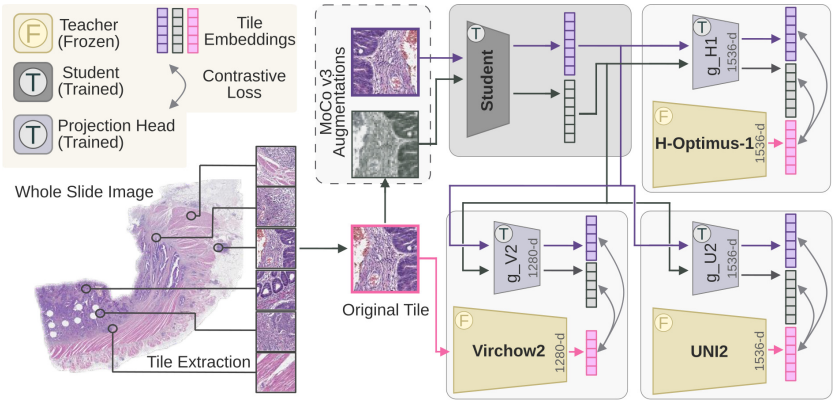


Fig. 1. Overview of MuCoDi pretraining. Frozen PFM teachers provide cached tile embeddings, while a lightweight student encoder and teacher-specific projection heads are trained with a multi-teacher contrastive distillation loss.

is increasingly urgent: global cancer incidence was close to 20 million new cases in 2022 and is projected to reach 35 million by 2050 [1]. At the same time, histopathology workforces are under pressure from rising workload, increasing diagnostic complexity and aging workforces [24]. AI systems could support screening, triage, and prescreening for therapy-relevant biomarkers, but only if they integrate into routine workflows without additional operational burden.

Current pathology foundation models (PFMs) have changed the performance landscape but not yet the deployment landscape. Their large vision-transformer backbones and gigapixel whole-slide image (WSI) pipelines often require dedicated GPUs, local clusters, or cloud processing. This creates a workflow mismatch: instead of augmenting existing pathology practice, many methods require additional digitization, transfer, preprocessing, and compute steps. For clinical adoption, models should run where tissue is examined and where data are produced: on existing pathology workstations, ordinary clinic computers, laptops, or eventually device-integrated microscopes. Such local inference also reduces data movement, allowing sensitive images to remain inside the pathology department.

Recent work highlights the feasibility and importance of this direction. Choudhury et al. developed a low-cost, open-source microscopy platform for slide capture and computational analysis, including Raspberry-Pi-based local compute for automated histopathology evaluation [5]. LitePath similarly argues that PFM accessibility depends not only on accuracy but also on deployability, reporting large reductions in parameters, FLOPs, energy, and runtime through distilled LiteFM models and adaptive patch selection [2]. These studies motivate a practical target for pathology AI: compact encoders that can run inside existing departmental infrastructure, including low-cost workstation and microscope-connected edge setups, rather than requiring external high-performance compute.

We propose Multi-teacher Contrastive Distillation (MuCoDi), a framework for learning edge-efficient pathology encoders (Fig. 1). MuCoDi distills comple-

Table 1. Overview of feature extractors in this study. Abbreviations: #Ps denotes parameters in millions, GFLOPs denotes giga floating-point operations for one (224×224) px tile, and #WSI denotes pretraining whole-slide images in thousands. The asterisk marks the custom ViT-G/14 configuration used by UNI2.

Model	Backbone	#Ps[M]	GFLOPs	#WSI[K]
H-Optimus-1 [20]	ViT-G/14	1135.0	295.97	>1000
UNI2 [14]	ViT-G/14*	681.0	180.39	>350
Virchow2 [28]	ViT-H/14	632.0	164.59	3100
LiteFM-L [2]	ViT-B	86.6	17.58	72.3
LiteFM [2]	ViT-S	22.1	4.61	72.3
LiteFM-S [2]	ViT-T	5.7	1.26	72.3
MuCoEdge-R2.3	RepViT-M2.3	22.4	4.57	11.8
MuCoEdge-R1.5	RepViT-M1.5	13.6	2.31	11.8
MuCoEdge-Ms3	MobileOne-S3	8.0	1.89	11.8
MuCoEdge-R1.1	RepViT-M1.1	7.8	1.36	11.8
MuCoEdge-R1.0	RepViT-M1.0	6.4	1.12	11.8
MuCoEdge-Ms2	MobileOne-S2	5.8	1.30	11.8
MuCoEdge-R0.9	RepViT-M0.9	4.7	0.831	11.8
MuCoEdge-Ms1	MobileOne-S1	3.5	0.824	11.8
MuCoEdge-R0.6	RepViT-M0.6	2.2	0.394	11.8
MuCoEdge-Ms0	MobileOne-S0	1.1	0.274	11.8
MuCoEdge-Mμ2	MobileOne-μ2	0.7	0.214	11.8
MuCoEdge-Mμ1	MobileOne-μ1	0.5	0.139	11.8
MuCoEdge-Mμ0	MobileOne-μ0	0.2	0.069	11.8

mentary representations from multiple state-of-the-art PFMs into a substantially smaller student backbone designed for local inference. Instead of matching only individual teacher features, MuCoDi preserves teacher-induced relational structure through contrastive distillation, enabling a compact model to inherit complementary morphology-aware embeddings while reducing latency and hardware demands.

Our contributions are threefold. First, we introduce an efficient contrastive multi-teacher knowledge distillation pretraining framework for edge-oriented pathology representation learning using only 11.8K unlabeled TCGA WSIs for student pretraining. Second, we demonstrate extreme latency improvements while maintaining strong downstream classification performance, including sub-million-parameter students for severe edge constraints. Third, we benchmark the latency-accuracy trade-off of compact pathology encoders on low-cost edge hardware, targeting future integration into routine workstation and microscope-based pathology workflows.

2 Related work

Pathology foundation models. Large self-supervised pathology foundation models (PFMs) such as UNI [3], Virchow [23], Prov-GigaPath [27], H-optimus-0/1 [19,20],

and Virchow2 [28] have substantially improved reusable histopathology representations. However, this progress largely follows a scaling trend: recent models range from hundreds of millions to over one billion parameters and process whole-slide images through large numbers of tiles, making training and deployment costly [16].

Knowledge distillation in pathology. Knowledge distillation (KD) compresses teacher models into a student [9]. Contrastive representation distillation instead transfers relational structure in feature space [21]. In computational pathology, multi-model knowledge has been used to build stronger or more task-adapted representations: GPFM combines expert and self-KD [13], COBRA treats embeddings from multiple PFMs as feature-space augmentations for slide-level contrastive learning [11], and HistoMILKD distills multiple PFMs into a MIL-based WSI classifier [15]. These methods demonstrate the value of multi-PFM supervision, but they primarily target generalization or slide-level prediction rather than compact feature extractors for efficient edge inference.

Efficient pathology foundation models. Recent work has begun to address PFM efficiency. H0-mini is an 86M-parameter ViT-B/14 distilled from a large pathology FM for robust and efficient inference [8]. Virchow2G Mini is a 22M-parameter distillation of the 1.9B-parameter Virchow2G [28]. LitePath introduces LiteFM and reports a small aggregate benefit of three-teacher over matched Virchow2-only distillation [2]. Nevertheless, these models remain ViT-based and are not primarily designed or benchmarked for direct deployment on low-cost edge devices integrated into routine pathology hardware.

Efficient architectures. In general computer vision, mobile-first backbones explicitly optimize latency and hardware efficiency. MobileNetV3 uses hardware-aware search [10], MobileOne targets sub-millisecond mobile inference through reparameterized convolutional design [22], RepViT revisits mobile CNNs through ViT-inspired design [25], and EfficientFormer shows that transformer-like models can approach MobileNet-speed inference [12]. MuCoDi brings this design philosophy to computational pathology by distilling multiple large PFMs into edge-oriented students for efficient integration into microscope or workstation workflows.

3 Methods

3.1 MuCoDi pretraining

MuCoDi adapts the two-view contrastive learning recipe of MoCo v3 [4] to multi-teacher distillation. MoCo v3 contrasts queries against keys produced by a momentum encoder. MuCoDi keeps the contrastive objective but replaces this key encoder with cached embeddings from frozen pathology foundation models, which act as fixed teacher keys. For each training tile x, we precompute teacher features $k_t(x)$ for $T = 3$ teachers: Virchow2, UNI2, and H-Optimus-1. The teachers

remain frozen. Only the lightweight student backbone f_θ and teacher-specific linear projection heads g_t are optimized. The heads map the shared student feature to 1280 dimensions for Virchow2 and 1536 dimensions for UNI2 and H-Optimus-1. For each tile, we sample two MoCo-v3-style stochastic views x_1, x_2 resized to (224×224) px. Let $q_{t,v} = \mathrm{norm}(g_t(f_\theta(x_v)))$ be the student query for teacher t and view $v \in \{1, 2\}$. Let $K_t = \{k_t^{(i)}\}_{i=1}^{B_g}$ denote the L2-normalized teacher keys gathered across all GPUs for the current global batch of size B_g. For a query from tile j, the positive key $k_t^{+(j)}$ is the cached teacher embedding of the same underlying tile. For one teacher and view, the InfoNCE term [18] is

$$\ell_{t,v}^{(j)} = -\log \frac{\exp\left(q_{t,v}^{(j)\top} k_t^{+(j)}/\tau\right)}{\sum_{i=1}^{B_g} \exp\left(q_{t,v}^{(j)\top} k_t^{(i)}/\tau\right)}, \tag{1}$$

where τ is the temperature. Following MoCo v3, the optimized per-sample loss is scaled as $\mathcal{L}_{t,v}^{(j)} = 2\tau \ell_{t,v}^{(j)}$, which makes the gradient magnitude independent of temperature. The final MuCoDi objective is the unweighted sum over teachers and views,

$$\mathcal{L}_{\mathrm{MuCoDi}} = \sum_{t=1}^{T} \sum_{v=1}^{2} \frac{1}{B_g} \sum_{j=1}^{B_g} \mathcal{L}_{t,v}^{(j)}. \tag{2}$$

Teacher embeddings are stored with the image tiles and loaded without gradients, and negatives are provided by cross-GPU gathering of the current batch.

Training details. In the student sweep, all students are pretrained for 10 epochs with AdamW using PyTorch default betas, base learning rate 2×10^{-2}, weight decay 10^{-6}, global batch size 2048, and temperature $\tau = 0.2$. The learning rate is linearly warmed up for the first 5% of optimizer steps and then cosine-decayed to zero. Gradients are clipped to a maximum norm of 1.0. Training uses bfloat16 automatic mixed precision, SyncBatchNorm, and distributed data parallelism on four GPUs. The augmentation pipeline follows the MoCo v3 two-crop scheme with random resized crops, color jitter, grayscale conversion, horizontal flips, asymmetric Gaussian blur, and solarization.

3.2 Data and task selection

MuCoDi is pretrained on TCGA tissue tiles with cached teacher features [26]. The pretraining set comprises 11,803 FFPE diagnostic WSIs and 14,280,892 tissue tiles. As shown in Tab. 1, this student distillation corpus uses approximately 6× fewer WSIs than LiteFM and over 260× fewer than Virchow2. Pretraining and evaluation use the same 2.0 μm/px (5×) physical resolution, which reduces tile count and compute [17]. Downstream evaluation follows the weakly supervised tile-feature benchmarking paradigm used in recent slide-representation and pathology foundation-model studies [11,16]. Task heads are trained on TCGA and deployed unchanged on matched CPTAC cohorts [7]. Evaluation uses the matched TCGA

set (3,933 patients, 4,633 slides) and matched CPTAC set (1,307 patients, 3,736 deployable slides), covering 23 clinically curated binary endpoints with at least 20 minority-class patients. All three teacher PFMs achieve at least 0.65 AUROC on internal TCGA evaluation for each endpoint, ensuring that each endpoint has a morphology-predictable signal before external deployment. The selected tasks span BRCA (*TP53*, *ESR1*, and *PGR*), CRC (*TP53* and MSI), GBM (*NF1*), KIRC (*SETD2*, grade, stage, and T stage), LUAD (*EGFR*, *STK11*, *TP53*, and T stage), NSCLC subtype, and UCEC (*BRD4*, *KMT2A*, *MGA*, *NF1*, *PTEN*, *ROS1*, *TP53*, and grade).

3.3 Evaluation protocol

Downstream performance is measured by freezing the student as a tile encoder, extracting one feature vector per tile, and training slide-level multiple-instance learning (MIL) heads with the STAMP workflow [6] on TCGA using patient-level stratified 5-fold cross-validation. External generalization is evaluated by applying each TCGA-trained fold checkpoint unchanged to the matched CPTAC cohort.

Slide-level heads and metrics. Main results use the STAMP transformer MIL head. Performance is reported as AUROC, F1 score, and AUPRC with 95% confidence-interval half-widths.

Edge inference benchmark. Deployment efficiency is measured on a Raspberry Pi 5 Model B with a quad-core Cortex-A76 CPU, 16 GB RAM, and no inference accelerator. Latency is measured in single-threaded float32 CPU mode without quantization and with batch size 1. MobileOne and RepViT models are first converted to their fused inference-time forms. For each model, we report parameters, GFLOPs for one (224×224) px RGB tile, and median single-tile latency from 100 timed forward passes after 20 warmup passes.

4 Results

External validation. We compare MuCoEdge students against their teacher PFMs and LiteFM competitors on the 23 selected downstream classification tasks (Tab. 2). Teacher PFMs retain the highest external AUROC, with Virchow2 reaching 71.8% and UNI2 71.7%. RepViT-based students close most of this gap at much smaller scale: MuCoEdge-R2.3 and MuCoEdge-R1.5 reach 71.0% AUROC, while MuCoEdge-R1.1 and MuCoEdge-R1.0 reach 70.9%. The strongest edge trade-off is achieved by smaller RepViT students. MuCoEdge-R0.9 reaches 70.5% AUROC, 51.3% F1, and 52.4% AUPRC with 4.7M parameters, outperforming LiteFM-L on all three external metrics despite using 18× fewer parameters and 21× fewer GFLOPs. MuCoEdge-R0.6 remains competitive with LiteFM while using only 2.2M parameters. These results are obtained from only 11.8K student-pretraining WSIs, approximately 6× fewer than LiteFM and over 260× fewer than Virchow2 (Tab. 1). LiteFM also includes both TCGA and CPTAC in its

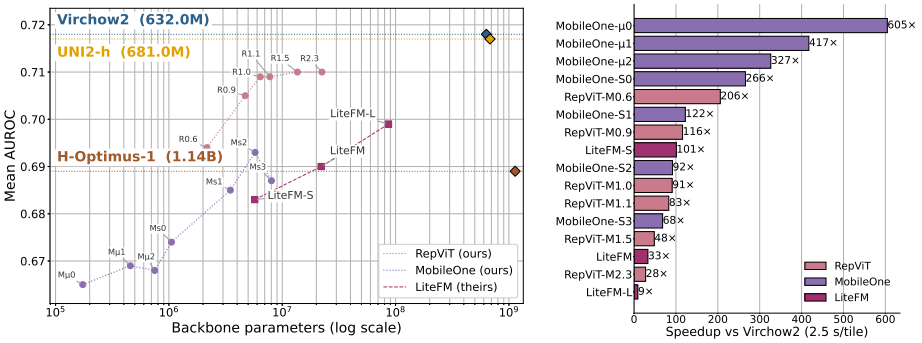
Model	Params	Type	TCGA (internal)			CPTAC (external)		
			AUC	F1	AUPRC	AUC	F1	AUPRC
H-Optimus-1	1135.0M	Theirs	$74.1_{7.7}$	$53.4_{8.1}$	$55.4_{8.7}$	$68.9_{3.6}$	$48.1_{5.7}$	$51.1_{3.0}$
UNI2	681.0M	Theirs	$76.4_{7.1}$	$55.3_{7.9}$	$\mathbf{58.8_{8.9}}$	$\underline{71.7_{3.3}}$	$50.6_{6.1}$	$\mathbf{53.4_{3.2}}$
Virchow2	632.0M	Theirs	$76.8_{7.8}$	$\mathbf{56.7_{7.8}}$	$58.7_{8.5}$	$\mathbf{71.8_{3.6}}$	$50.7_{6.8}$	$52.9_{3.8}$
LiteFM-L	86.6M	Theirs	$75.0_{6.7}$	$53.9_{7.6}$	$56.8_{8.1}$	$69.9_{3.1}$	$48.8_{6.3}$	$51.7_{3.1}$
LiteFM	22.1M	Theirs	$74.0_{8.6}$	$53.1_{7.7}$	$55.8_{8.7}$	$69.0_{3.4}$	$47.4_{5.8}$	$50.9_{3.4}$
LiteFM-S	5.7M	Theirs	$73.4_{8.5}$	$52.8_{8.2}$	$54.8_{8.2}$	$68.3_{3.9}$	$46.7_{6.5}$	$50.2_{3.7}$
MuCoEdge-R2.3	22.4M	Ours	$\mathbf{77.1_{7.3}}$	$55.6_{6.3}$	$58.5_{7.8}$	$71.0_{3.2}$	$\mathbf{51.8_{4.8}}$	$\underline{53.3_{3.4}}$
MuCoEdge-R1.5	13.6M	Ours	$76.7_{7.3}$	$55.4_{7.9}$	$57.8_{8.7}$	$71.0_{2.9}$	$51.1_{5.7}$	$52.8_{3.1}$
MuCoEdge-R1.1	7.8M	Ours	$76.3_{7.6}$	$55.2_{7.7}$	$58.3_{9.1}$	$70.9_{3.3}$	$50.5_{5.4}$	$52.2_{3.4}$
MuCoEdge-R1.0	6.4M	Ours	$\underline{76.9_{7.7}}$	$55.6_{8.0}$	$\underline{58.7_{9.4}}$	$70.9_{2.9}$	$50.2_{6.3}$	$52.0_{3.4}$
MuCoEdge-R0.9	4.7M	Ours	$76.6_{8.0}$	$\underline{56.2_{7.4}}$	$58.6_{9.6}$	$70.5_{3.3}$	$\underline{51.3_{5.0}}$	$52.4_{3.6}$
MuCoEdge-R0.6	2.2M	Ours	$75.2_{8.0}$	$54.0_{7.2}$	$57.3_{9.3}$	$\underline{69.4_{3.7}}$	$48.8_{5.8}$	$51.1_{4.0}$
MuCoEdge-Ms3	8.0M	Ours	$74.6_{7.6}$	$54.0_{7.9}$	$56.6_{8.8}$	$68.7_{3.6}$	$46.4_{5.1}$	$50.9_{3.8}$
MuCoEdge-Ms2	5.8M	Ours	$74.9_{7.7}$	$53.6_{7.6}$	$56.7_{9.2}$	$69.3_{3.6}$	$46.8_{6.2}$	$51.1_{3.6}$
MuCoEdge-Ms1	3.5M	Ours	$74.6_{7.1}$	$53.7_{8.1}$	$55.8_{7.7}$	$68.5_{3.9}$	$46.0_{5.7}$	$50.7_{3.7}$
MuCoEdge-Ms0	1.1M	Ours	$73.2_{9.2}$	$52.6_{8.2}$	$54.9_{8.3}$	$67.4_{4.3}$	$45.2_{7.6}$	$49.7_{3.8}$
MuCoEdge-Mμ2	0.7M	Ours	$73.1_{8.7}$	$52.5_{8.5}$	$54.8_{8.2}$	$66.8_{4.2}$	$44.3_{6.2}$	$48.9_{3.8}$
MuCoEdge-Mμ1	0.5M	Ours	$72.9_{8.4}$	$52.2_{7.1}$	$54.4_{8.0}$	$66.9_{3.9}$	$45.7_{5.6}$	$48.6_{3.5}$
MuCoEdge-Mμ0	0.2M	Ours	$72.1_{8.1}$	$51.6_{8.0}$	$53.7_{7.8}$	$66.5_{4.5}$	$44.5_{6.5}$	$48.4_{3.4}$

Table 2. Aggregated results for TCGA (internal) and CPTAC (external) evaluation using the STAMP transformer MIL head. Values are means across the 23 selected comparison tasks and are reported in %. Subscripts indicate the mean 95% t-interval CI half-width across the five cross-validation folds. **Bold** and underlining denote the best and second-best values in each column, respectively.

Model	Breast		Colon		Brain		Kidney		Lung		Uterus		**Avg**	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
H-Optimus-1	78.7	<u>72.6</u>	73.0	50.6	57.5	26.1	63.5	44.4	77.6	57.5	63.0	37.0	68.9	48.1
UNI2	<u>81.2</u>	69.5	<u>77.0</u>	59.9	63.1	32.3	<u>66.4</u>	**47.9**	77.4	57.8	66.9	40.2	<u>71.7</u>	50.6
Virchow2	**83.3**	**76.4**	**78.0**	60.6	60.2	28.4	**66.9**	44.9	75.9	56.8	<u>67.3</u>	40.4	**71.8**	50.7
LiteFM-L	80.1	72.4	73.6	58.4	62.1	**34.7**	65.1	42.3	76.6	57.2	64.3	37.5	69.9	48.8
LiteFM	79.8	71.3	71.9	52.9	62.3	32.7	64.5	41.9	74.6	55.6	63.9	36.6	69.0	47.4
LiteFM-S	77.5	70.3	68.8	45.3	61.1	28.0	64.3	42.0	73.9	54.6	64.2	37.8	68.3	46.7
MuCoEdge-R2.3	78.1	72.3	76.3	**67.8**	60.9	30.5	66.4	45.9	<u>78.0</u>	58.5	66.1	<u>41.5</u>	71.0	**51.8**
MuCoEdge-R1.5	77.7	71.5	73.5	60.1	62.0	30.6	65.9	45.3	**78.4**	58.3	67.0	**42.3**	71.0	51.1
MuCoEdge-R1.1	76.0	69.8	76.3	62.5	62.8	28.3	65.8	46.6	77.2	<u>58.7</u>	**67.3**	40.0	70.9	50.5
MuCoEdge-R1.0	77.8	70.6	74.8	55.8	60.9	<u>33.8</u>	66.2	45.4	77.0	57.4	67.1	41.1	70.9	50.2
MuCoEdge-R0.9	77.0	71.7	74.4	<u>63.3</u>	60.3	32.8	65.8	<u>47.3</u>	77.6	**59.0**	66.4	40.1	70.5	<u>51.3</u>
MuCoEdge-R0.6	73.8	67.0	73.1	54.6	60.6	31.2	65.0	45.4	76.5	57.4	65.6	39.1	<u>69.4</u>	48.8
MuCoEdge-Ms3	77.9	68.6	70.3	40.7	63.2	31.1	64.1	43.8	74.6	54.8	64.2	37.4	68.7	46.4
MuCoEdge-Ms2	77.7	64.9	73.4	52.3	62.8	28.2	65.1	43.0	74.8	55.1	64.6	37.6	69.3	46.8
MuCoEdge-Ms1	78.1	66.6	70.9	48.8	62.2	31.0	65.2	42.4	74.2	54.6	63.1	36.0	68.5	46.0
MuCoEdge-Ms0	75.7	58.8	71.0	54.5	62.2	30.3	64.1	41.9	72.5	52.6	62.6	36.5	67.4	45.2
MuCoEdge-Mμ2	75.1	61.0	67.1	42.1	**63.9**	29.3	64.2	43.0	70.7	51.9	62.8	36.3	66.8	44.3
MuCoEdge-Mμ1	74.9	61.4	67.8	53.4	<u>63.8</u>	31.6	64.0	44.8	71.7	53.8	62.3	35.2	66.9	45.7
MuCoEdge-Mμ0	73.0	58.4	68.4	51.1	61.4	28.3	64.4	43.1	71.0	51.9	62.4	35.8	66.5	44.5

Table 3. Per-organ AUC and F1, CPTAC (external validation), STAMP transformer MIL head. Each cell is the mean over the organ's tasks (Lung pools LUAD and NSCLC); values in %. **Bold** and underline mark the best and second-best value per column.

pretraining corpus. At the extreme edge end, sub-million MobileOne variants retain 66.5–66.9% external AUROC with only 0.2–0.7M parameters. Internal TCGA evaluation shows the same trend: MuCoEdge-R2.3 obtains the highest mean AUROC (77.1%), and several students remain within the teacher range. Per-organ results (Tab. 3) show that performance is not driven by a single tissue type. MuCoEdge variants obtain the best average F1 and are competitive across colon, kidney, lung, and uterine tasks.



(a) Performance-efficiency trade-off. **(b)** Edge speedup on Raspberry Pi 5.

Fig. 2. Efficiency and edge deployment. (a) MuCoEdge models compared with LiteFM variants and teacher foundation models on CPTAC external validation. **(b)** Single-tile inference speedup over Virchow2 on Raspberry Pi 5.

Efficiency and edge deployment. The efficiency gains are substantial (Tab. 1, Fig. 2a). Speedups are reported against Virchow2, the most efficient high-performing teacher reference in our benchmark. MuCoEdge-R0.9 improves over LiteFM-L while using 18× fewer parameters, 21× fewer GFLOPs, and 12.9× faster Raspberry-Pi inference. MuCoEdge-R0.6 reaches 69.4% AUROC with 2.2M parameters and runs 6.2× faster than LiteFM and 2.0× faster than LiteFM-S. The sub-million MobileOne variants provide the most aggressive latency reduction, achieving 327–605× single-tile speedups over Virchow2 while retaining 66.5–66.9% external AUROC. Together, these results define a practical edge-deployment spectrum, from higher-accuracy RepViT models to sub-million-parameter MobileOne models for strict latency budgets.

5 Conclusion

MuCoDi transfers multiple PFM feature spaces into compact, edge-oriented pathology encoders. In broad internal TCGA and external CPTAC evaluation, MuCoEdge-R0.9 improves over LiteFM-L with 18× fewer parameters, MuCoEdge-R0.6 remains competitive with LiteFM at 2.2M parameters, and sub-million MobileOne variants retain 66.5–66.9% external AUROC with 327–605× Raspberry-Pi

speedups over Virchow2. From a clinical perspective, encoders that run directly on departmental workstations or microscope-attached edge devices remove the data-transfer and compute barriers that currently keep PFM-based decision support out of routine diagnostic workflows.

Acknowledgments. The authors gratefully acknowledge computing-time support from the GWK through ZIH at TU Dresden and from the Gauss Centre for Supercomputing e.V. through NIC on the JUWELS and JUPITER Booster modules at Jülich Supercomputing Centre. We also acknowledge the TCGA Research Network and the Clinical Proteomic Tumor Analysis Consortium (CPTAC), which generated the data on which the results shown in this study are based.

Disclosure of Interests. MG received honoraria for lectures sponsored by Sartorius AG and AstraZeneca, UK. NGR received compensation for travel expenses from nanoString, a Bruker company. JNK holds shares in StratifAI, Synagen, Spira Labs, Tremont AI, and Saterra AI; is Co-PI on institutional research grants from GSK and AstraZeneca, and declares honoraria or consulting fees from AstraZeneca, Bayer, Bioptimus, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, Novartis, BMS, Roche, and Pfizer.

References

1. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., Jemal, A.: Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **74**(3), 229–263 (2024). https://doi.org/10.3322/caac.21834
2. Cai, Y., Jin, C., Ma, J., et al.: A deployment-friendly foundational framework for efficient computational pathology. arXiv preprint arXiv:2602.14010 (2026). https://doi.org/10.48550/arXiv.2602.14010
3. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., 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 **30**, 850–862 (2024). https://doi.org/10.1038/s41591-024-02857-3
4. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
5. Choudhury, D., Dolezal, J.M., Dyer, E., Kochanny, S., Ramesh, S., Howard, F.M., Margalus, J.R., Schroeder, A., Schulte, J., Garassino, M.C., Kather, J.N., Pearson, A.T.: Developing a low-cost, open-source, locally manufactured workstation and computational pipeline for automated histopathology evaluation using deep learning. EBioMedicine **107**, 105276 (2024). https://doi.org/10.1016/j.ebiom.2024.105276
6. El Nahhas, O.S.M., van Treeck, M., Wölflein, G., Unger, M., Ligero, M., Lenz, T., Wagner, S.J., Hewitt, K.J., et al.: From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology. Nature Protocols **20**, 293–316 (2025). https://doi.org/10.1038/s41596-024-01047-2
7. Ellis, M.J., Gillette, M., Carr, S.A., Paulovich, A.G., Smith, R.D., Rodland, K.K., Townsend, R.R., Kinsinger, C., Mesri, M., Rodriguez, H., Liebler, D.C.: Connecting genomic alterations to cancer biology with proteomics: The NCI clinical proteomic tumor analysis consortium. Cancer Discovery **3**(10), 1108–1112 (2013). https://doi.org/10.1158/2159-8290.CD-13-0219

8. Filiot, A., Dop, N., Tchita, O., Riou, A., Dubois, R., Peeters, T., Valter, D., Scalbert, M., Saillard, C., Robin, G., Olivier, A.: Distilling foundation models for robust and efficient models in digital pathology. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. Lecture Notes in Computer Science, vol. 15966, pp. 162–172. Springer (2026). https://doi.org/10.1007/978-3-032-04981-0_16
9. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop (2015). https://doi.org/10.48550/arXiv.1503.02531
10. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H.: Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1314–1324 (2019)
11. Lenz, T., Neidlinger, P., Ligero, M., Wölflein, G., van Treeck, M., Kather, J.N.: Unsupervised foundation model-agnostic slide-level representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025). https://doi.org/10.48550/arXiv.2411.13623
12. Li, Y., Yuan, G., Wen, Y., Hu, E., Evangelidis, G., Tulyakov, S., Wang, Y., Ren, J.: Efficientformer: Vision transformers at MobileNet speed. In: Advances in Neural Information Processing Systems. vol. 35, pp. 12934–12949 (2022)
13. Ma, J., Guo, Z., Zhou, F., Wang, Y., Xu, Y., Li, J., Yan, F., Cai, Y., Zhu, Z., Jin, C., et al.: A generalizable pathology foundation model using a unified knowledge distillation pretraining framework. Nature Biomedical Engineering (2025). https://doi.org/10.1038/s41551-025-01488-4
14. Mahmood Lab: UNI2-h: Pathology foundation model. Hugging Face model card (2025), https://huggingface.co/MahmoodLab/UNI2-h
15. Mallya, M., Mirabadi, A.K., Farahani, H., Bashashati, A.: HistoMILKD: A multiple instance learning based multi-teacher knowledge distillation framework for whole slide image classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3390–3400 (2026)
16. Neidlinger, P., El Nahhas, O.S.M., Muti, H.S., Lenz, T., Hoffmeister, M., Brenner, H., et al.: Benchmarking foundation models as feature extractors for weakly supervised computational pathology. Nature Biomedical Engineering **10**, 1113–1123 (2026). https://doi.org/10.1038/s41551-025-01516-3
17. Neidlinger, P., Lenz, T., Foersch, S., Loeffler, C.M.L., Clusmann, J., Gustav, M., Shaktah, L.A., et al.: A deep learning framework for efficient pathology image analysis (2025). https://doi.org/10.48550/arXiv.2502.13027
18. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2018)
19. Saillard, C., Jenatton, R., Llinares-Lopez, F., Mariet, Z., Cahane, D., Durand, E., Vert, J.P.: H-optimus-0. Model release (2024), https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0
20. Scalbert, M., Saillard, C., Peeters, T., Gonzalez, L., Valter, D., Llinares-Lopez, F., et al.: H-optimus-1: A foundation model for computational histopathology. In: Proceedings of the American Association for Cancer Research Annual Meeting 2026; Part 2 (Late-Breaking, Clinical Trial, and Invited Abstracts). vol. 86, p. LB174. AACR (2026). https://doi.org/10.1158/1538-7445.AM2026-LB174
21. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=SkgpBJrtvS

22. Vasu, P.K.A., Gabriel, J., Zhu, J., Tuzel, O., Ranjan, A.: MobileOne: An improved one millisecond mobile backbone. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7907–7917 (2023)
23. Vorontsov, E., Bozkurt, A., Casson, A., et al.: A foundation model for clinical-grade computational pathology and rare cancers detection. Nature Medicine **30**, 2924–2935 (2024). https://doi.org/10.1038/s41591-024-03141-0
24. Walsh, E., Orsi, N.M.: The current troubled state of the global pathology workforce: A concise review. Diagnostic Pathology **19**, 163 (2024). https://doi.org/10.1186/s13000-024-01590-2
25. Wang, A., Chen, H., Lin, Z., Han, J., Ding, G.: RepViT: Revisiting mobile CNN from ViT perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15909–15920 (2024)
26. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., et al.: The cancer genome atlas pan-cancer analysis project. Nature Genetics **45**, 1113–1120 (2013). https://doi.org/10.1038/ng.2764
27. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**, 181–188 (2024). https://doi.org/10.1038/s41586-024-07441-w
28. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., et al.: Virchow2: Scaling self-supervised mixed magnification models in pathology (2024). https://doi.org/10.48550/arXiv.2408.00738