Pathryoshka: Compressing Pathology Foundation Models via Multi-Teacher Knowledge Distillation with Nested Embeddings

Christian Grashei[1,2,3,⋆], Christian Brechenmacher[4,⋆], Rao Muhammad Umer[4], Jingsong Liu[1,3], Carsten Marr[4,6,7,8,⋆⋆], Peter Schüffler[1,2,3,8,⋆⋆], and Ewa Szczurek[4,5,⋆⋆]

[1] Institute of Pathology, Technical University of Munich, Munich, Germany
[2] Munich Data Science Institute, Munich, Germany
[3] Munich Center for Machine Learning, Munich, Germany
[4] Institute of AI for Health, Helmholtz Munich, Munich, Germany
[5] Institute of Informatics, University of Warsaw, Warsaw, Poland
[6] Department of Medicine III, Ludwig-Maximilian-University Hospital, Munich, Germany
[7] Department of Physics, Ludwig-Maximilian-University, Munich, Germany
[8] German Cancer Consortium (DKTK), Munich, Germany

Abstract. Foundation models (FMs) have driven significant progress in computational pathology. These models can easily exceed a billion parameters and produce high-dimensional embeddings, thus limiting their applicability for research or clinical use when computing resources are tight. Here we introduce Pathryoshka, a novel multi-teacher distillation framework inspired by agglomerative models and Matryoshka representation learning to reduce pathology FM sizes while allowing for adaptable embedding dimensions. We evaluate our framework with a distilled model on ten public pathology benchmarks with varying downstream tasks. Compared to its much larger teachers, Pathryoshka reduces the model size by 86-92% at on-par performance. It outperforms state-of-the-art single-teacher distillation models of comparable size by a median margin of 7.0 and other pathology multi-teacher distillation by 5.3 percentage points in accuracy. By enabling efficient deployment without sacrificing accuracy or representational richness, Pathryoshka democratizes access to state-of-the-art pathology FMs for the broader research and clinical community.

Keywords: Foundation Model · Knowledge Distillation · Computational Pathology

1 Introduction

Empirical scaling laws demonstrate consistent performance gains from increasing foundation model (FM) size. Larger models capture more complex structures

⋆ C. Grashei and C. Brechenmacher contributed equally.
⋆⋆ C. Marr, P. Schüffler and E. Szczurek supervised equally.

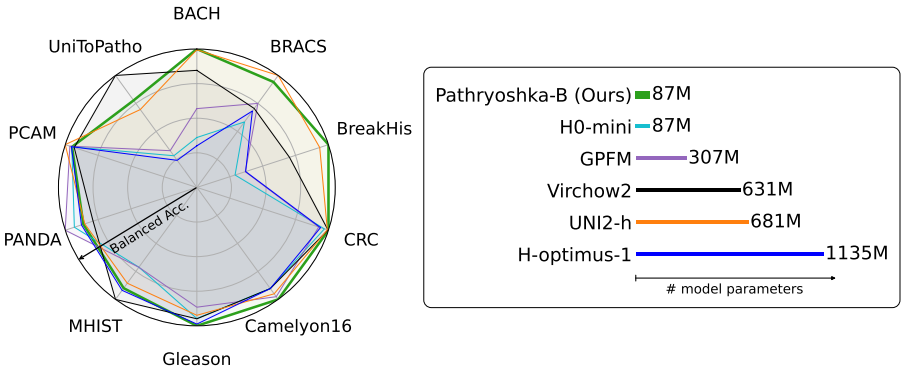

Fig. 1: Although smaller in parameter size, Pathryoshka outperforms or is on par with its state-of-the-art teacher pathology FMs on various benchmarks and outperforms current single-teacher (H0-mini) and multi-teacher-distillation (GPFM) models.

and show stronger performance across tasks [20, 27, 50]. Yet, model scaling also brings rapid cost escalation, resulting in inaccessible state-of-the-art development and deployment. This challenge is acute in pathology: Although pathology FMs promise major advances in cancer diagnosis and patient care [5, 40], their hardware and operational costs hinder adoption in resource-limited clinics and research centers. Pathology FMs contain hundreds of millions to billions of parameters, and with whole-slide images requiring thousands of tiles to be encoded, inference and storage costs become prohibitive. Thus, there is a clear need for pathology FMs with reduced model and embedding sizes that maintain performance while enabling compute- and memory-efficient deployment.

While several promising approaches exist in this direction, they still face notable limitations. Naively training smaller models on equally rich datasets does not match the performance of larger models [30]. A prominent approach to reducing model size while preserving accuracy is *knowledge distillation*, which transfers knowledge from large teacher models to a smaller student by aligning their embeddings or logits [22, 23, 36, 43]. Although highly effective compared to naive approaches such as model ensembling, and able to substantially reduce model size [13, 39], distillation typically retains large embedding sizes, leading to high inference times and memory usage in downstream tasks. To obtain compact embeddings, *Matryoshka representation learning* (MRL) [32, 48] trains a single high-dimensional embedding whose prefix sub-vectors remain semantically meaningful and effective. This nested structure enables dynamic truncation to lower dimensions with minimal loss. However, while MRL compresses embeddings, it does not guarantee overall model size reduction. To our knowledge, no prior work has combined distillation with MRL to jointly realize the benefits of both approaches.

In the context of pathology FMs, distillation is often performed using a single teacher model, such as H0-mini [16] trained by distillation from a large vision

transformer model called H-optimus-0 [44]. However, relying on a single teacher risks inheriting its biases and limits the represented diversity of clinical variation. We illustrate this behavior in Fig. 1, where the distilled H0-mini is only able to perform as good or only marginally better than H-optimus-1 [6], a model trained as successor to H-optimus-0. While multi-teacher distillation strategies such as AM-RADIO [43] have been successfully applied to general image foundation models, they are underexplored in pathology. Combining multiple pathology FMs has been done with GPFM [36], but without reduction in model capacity with respect to the teacher models. In summary, pathology FMs still lack a method that simultaneously compresses both model and embedding size while leveraging diverse multi-teacher signals.

To address these challenges, we propose Pathryoshka, a pathology FM that leverages multi-teacher knowledge distillation in combination with MRL to fundamentally reduce model size and improve computational efficiency while achieving similar performances to large teacher models [6, 13, 33, 58] and outperforming previous distilled small models [16] (Fig. 1). The hierarchical embedding structure of Pathryoshka fuses complementary signals from multiple teachers into compact and robust subembeddings. To avoid memorization of embeddings of teacher models from public training data while ensuring enough data diversity, distillation training of Pathryoshka is performed on an internal dataset of 243 million image tiles extracted from 158,233 H&E-stained images across all organs, which is augmented using a novel cropping strategy for distillation. Evaluations on multiple benchmarks, including patch-level and slide-level, demonstrate that Pathryoshka maintains parity or outperforms individual teacher models in terms of efficiency and performance. Code and model weights are available at `https://huggingface.co/SchuefflerLab/Pathryoshka-B`.

Our main contributions are the following:

- We introduce a multi-teacher distillation framework for unsupervised distillation of pathology FMs that relies solely on CLS and patch tokens, enabling the use of external public models.
- We propose nested embeddings for multi-teacher distillation, yielding a pathology FM with adaptable embeddings.
- We demonstrate that multiple large-scale pathology FM teacher models can be effectively compressed into a single compact model with competitive performance, substantially reducing model size and inference cost.
- We introduce a novel cropping-based augmentation strategy for multi-teacher distillation improving representation alignment across teachers.

2 Related Work

Current Pathology Foundation Models. Recent years have witnessed a surge of pathology FMs leveraging vision transformer architecture [15] and self-supervised learning on large unlabeled datasets to learn generalizable representations that can be efficiently adapted to many downstream pathology tasks

with minimal labeled data [13, 17, 18, 34, 44, 52, 53, 58]. Among most prominent pathology FMs, three large models have shown excellent performance: Virchow2, UNI2-h, and H-optimus-1. Virchow2 [58] is a 632-million-parameter vision transformer model (ViT-H/14) for pathology, extending the earlier Virchow model [52] and pretrained on an exceptionally large dataset of 3.1 million H&E-stained whole-slide images (WSIs) from 225,000 patients, totaling 2 billion image tiles. A bigger variant, Virchow2G (1.9 billion parameters, ViT-G), was introduced to examine pure model scaling effects. UNI2-h [13] is a 681-million-parameter model (ViT-H/14) pretrained on over 200 million tiles from more than 350,000 H&E WSIs across 20 major tissue types. H-optimus-1 [6] is a 1.1-billion-parameter ViT-G/14 model pretrained on over one million H&E WSIs, comprising billions of tiles. Despite these advances, adoption of current resource-hungry pathology FMs in the clinic remains limited, and no single model provides both optimal efficiency and performance across all pathology tasks.

Distillation for Pathology Foundation Models. In contrast to the extensive literature on supervised distillation, unsupervised knowledge distillation represents an underexplored frontier. This research gap extends to the pathology domain, which has seen a surge of FMs trained in an unsupervised manner. Given the scarcity of annotated data, the importance of label-free distillation methods is amplified, as they provide a vital mechanism for compressing foundational knowledge without the need for supervision labels.

So far, only *single*-teacher distillation has been studied for reducing pathology FM size. One example is H0-mini [16], a model distilled from H-optimus-0 (ViT-G) into a smaller (ViT-B) architecture with 86M parameters. Virchow2G-Mini (ViT-S), using 22M parameters, is distilled from Virchow2G (ViT-G). Notably, both models rely on the DINOv2 distillation framework [41], which necessitates access to teacher heads and prototype scores that are frequently omitted from public model releases, limiting the reproducibility and flexibility of the distillation process. Single-teacher distillation additionally inherits the weaknesses of its teacher: Figure 1 demonstrates that H0-mini performs similarly to H-optimus-1 from the same model family. GPFM [36] introduces a multi-teacher objective while also integrating DINO [11] and Masked Image Modeling [21] losses. Its student architecture is of same size or larger in scale to its teachers, thus not addressing model efficiency or significant parameter compression.

Consequently, the distillation of knowledge from multiple, substantially larger teachers into a compact student remains an open research question. Furthermore, we posit that unsupervised distillation can serve as more than a compression tool. It can act as a knowledge agglomeration to create pathology FMs that surpass the performance of individual teachers. Finally, the integration of nested embeddings within a distillation framework remains unexplored in the pathology domain, representing a missed opportunity for creating flexible representations. Flexible embeddings offer a trade-off between accuracy and efficiency where resources are scarce. Working with gigapixel whole-slide images often involves storing large amounts of embeddings in memory or disk [12, 45]. Reducing the embedding size

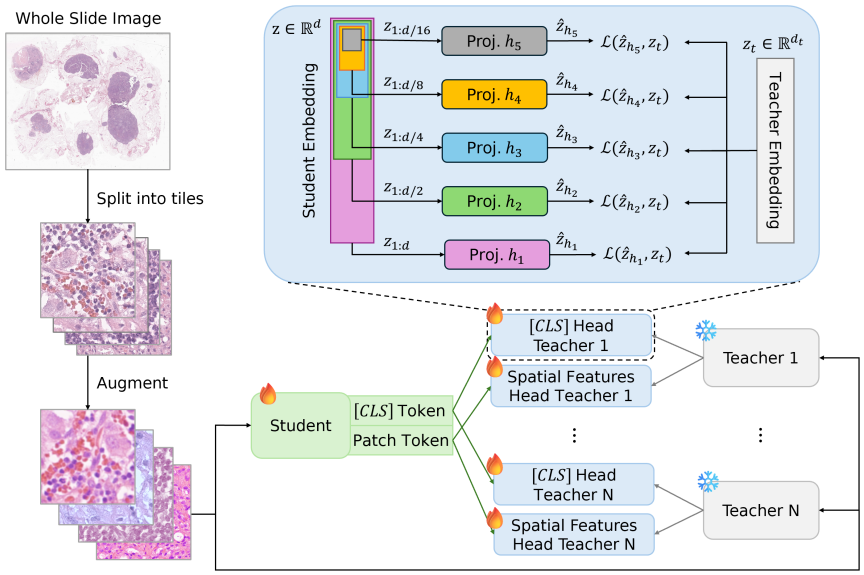

Fig. 2: In Pathryoshka, a multi-teacher distillation framework, the student model and distillation heads are jointly trained while teacher models remain frozen. Input images are augmented and processed by both the student and all teachers. The CLS and spatial feature heads align student representations with those of the teachers. Each head includes multiple MLP projection layers that form nested embeddings by minimizing a similarity loss between student and teacher projections.

alleviates the storage burden while accelerating downstream tasks where compute scales with dimensionality, such as clustering for large-scale data curation [51].

3 Methodology

We introduce Pathryoshka, a framework for distilling multiple teacher FMs into a single, compact vision transformer (Fig. 2). The only prerequisite for the teachers is the provision of token-based representations, specifically a global class (CLS) token and spatial patch tokens, to align with the student's output. Pathryoshka draws inspiration from the AM-RADIO [43] approach for agglomerative distillation using multiple heterogeneous teachers, as well as from Matryoshka representation learning [32] for flexible embeddings of varying dimensionalities. While demonstrated for histopathology, our proposed setup is domain-agnostic, enabling application in other fields where multiple homogeneous models are available. It operates in a fully unsupervised manner, requiring only a large-scale unlabeled dataset.

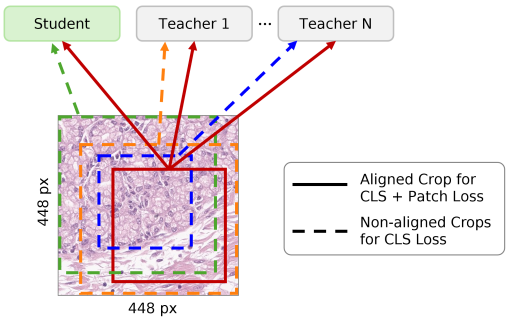

Fig. 3: Robustness-Enhancing Cropping Strategy. A first random crop (red) is obtained from the input image. This *aligned* crop is individually color augmented for the student and all teachers. Both CLS and patch loss are computed for this aligned crop. Additionally, each model receives a second random individual crop (dashed lines) which aims to improve magnification invariance. We only compute CLS loss on this *non-aligned* crop.

3.1 Dataset

To train our model, we use a large-scale in-house dataset of 158,233 H&E stained whole-slide images (WSIs). We identify tissue regions using CLAM [35]. We randomly sample a maximum of 2,250 tiles of size 448x448 pixels per WSI at 10x, 20x and 40x magnifications in proportions of 20%, 40% and 40%, respectively, to cover multiple morphological scales. If there was not enough tissue on a WSI, sampling is stopped earlier, resulting in a total of 243 million tiles. While substantial in absolute terms, large-scale foundation models often exceed this training dataset size [6, 58]. We intentionally exclude public datasets from training to ensure a fair and unbiased evaluation. In the context of distillation, separating training and evaluation data is particularly important to avoid the student model to memorize teacher embeddings for seen data, and to be able to demonstrate generalization to unseen inputs.

Data Augmentation. Since the dataset originates from a single institution and was digitized using a single scanner, we apply extensive augmentations to enhance its diversity, including color augmentation, random flipping and Gaussian blur. For color augmentation, we employ H&E-specific stain perturbations from the Albumentations library [10] that decomposes an image into the hematoxylin and eosin channels before perturbing their intensities to simulate more realistic stain variations. Additionally, we incorporate a novel cropping-based augmentation dedicated to our multi-teacher framework as explained in Sec. 3.2 and Fig. 3.

3.2 Pathryoshka Distillation Framework

Pathryoshka aims to obtain a compressed student model by distillation from multiple teachers and training structured data representations. To this end,

we introduce a novel cropping strategy to enhance robustness, employ multi-teacher nested embeddings, and design adjusted distillation heads suited to this architecture. The training objective is further supported by a dedicated loss formulation that aligns the student with the structured teacher signals. These components of our framework are detailed in the following sections.

Robustness-Enhancing Cropping Strategy. We generate multiple crops per tile to train the model (Fig. 3). Each crop covers between 25% and 100% of the tile area with an aspect ratio sampled from $[0.9, 1.1]$, followed by random horizontal and vertical flips. The resulting region is resized to 224×224 px. Formally, let x denote an image tile from the dataset, a_{spatial} the random cropping and flipping augmentation, and a_{visual} a random H&E stain and blur augmentation. The *aligned crop* x_c is defined as:

$$x_c = a_{\text{spatial}}(x) \tag{1}$$

The student model and all teacher models receive a color augmented version of that aligned crop. Additionally, the student model and all teacher models receive a individual random crop called *non-aligned crop* that is color augmented as well. Those crops cover different areas of the original tile, effectively simulating variations in magnification and thereby improving robustness. Formally, let $\mathcal{M} = f_s, f_{t_1}, \ldots, f_{t_N}$ denote the set of student and teachers. Each model $f_m \in \mathcal{M}$ produces two outputs:

$$z_m^{\text{a}} = f_m(a_{\text{visual}}(x_c)),\ z_m^{\text{g}} = f_m(a_{\text{visual}}(a_{\text{spatial}}(x))). \tag{2}$$

Here, z_m^{a} is the output for the aligned crop and z_m^{g} the output for the non-aligned crop. Each output z_m comprises the CLS token and patch embeddings of the respective model. For the aligned crop we compute a loss for both the CLS token and the patch tokens. For the non-aligned crop, we only compute a loss for the CLS token that doesn't require strict spatial correspondence.

Note, that *DINOv2 distillation* [41] would require the teachers to forward the CLS token and patch tokens through the DINO and iBOT head, respectively, to obtain the high-dimensional teacher *prototype scores*. However, the weights for the DINO and iBOT head are often not publicly available making DINOv2 distillation unfeasible.

Nested Embeddings. We incorporate nested embeddings [32], which allow the distilled model to generate representations that adapt to varying computational budgets and downstream requirements.

Given an input x, the distilled model produces an embedding $z \in \mathbb{R}^d$. We aim to learn embeddings where any prefix $z_{1:m}$, with $m \leq d$, forms a semantically meaningful representation. Let M denote the set of target embedding sizes. For each $m \in M$, we enforce representational consistency through a loss-based constraint. Following [48], we adopt a geometrically decreasing series of nested

embedding sizes $M = \{768, 384, 192, 96, 48\}$, obtained by halving the dimension for each level. At training time, we compute distillation losses at all nesting levels, ensuring that truncated representations $z_{1:m}$ remain semantically aligned with teacher features.

Distillation Heads. We employ three-layer MLPs to project the outputs of the distilled student to the embedding dimension of the respective teacher. The size of the first MLP layer matches the student embedding and the last layer the embedding size of the teacher model. We set the intermediate layer to the same dimensionality as the teacher embeddings to avoid creating an additional bottleneck. This design ensures that the only compression in the network occurs between the teacher and student embedding spaces.

Two sets of projection heads are used per teacher, one being for the CLS token, the other for the patch tokens encoding spatial information. Each set of projection heads consists of $|M|$ MLPs, where $|M|$ denotes the number of nesting levels. Each MLP is responsible for one nesting level of dimension m and receives the first m features of the embedding.

3.3 Loss Formulation

The training targets of our student model are the outputs of multiple teacher models. Since all teachers are vision transformers, each produces a global CLS token, which captures image-level information, and a set of patch tokens that encode spatial features.

Let x denote the augmented input image. The student network, parameterized by Θ, outputs a sequence of tokens:

$$\mathbf{z} = f(x|\Theta) = [z_s, z_p^{(1)}, z_p^{(2)}, \ldots, z_p^{(N)}], \tag{3}$$

where $z_s \in \mathbb{R}^d$ is the global CLS token and $z_p^{(i)} \in \mathbb{R}^d$ are the N patch tokens, each representing a spatial region of the input.

Similarly, the output for teacher t with parameters Θ_t is denoted as

$$\mathbf{z^t} = f_t(x|\Theta_t) = [z_s^{(t)}, z_p^{(t,1)}, z_p^{(t,2)}, \ldots, z_p^{(t,N)}], \tag{4}$$

where $z_s^{(t)}$, $z_p^{(t,i)} \in \mathbb{R}^{d_t}$.

For every teacher t, we attach two families of distillation heads:

$$g_s^{(t,m)} : \mathbb{R}^m \to \mathbb{R}^{d_t}, \quad g_p^{(t,m)} : \mathbb{R}^m \to \mathbb{R}^{d_t}, \tag{5}$$

where $m \in M$ denotes the nesting dimension (i.e., the subset of the first m features) and M is the set of considered nesting levels. Each $g_s^{(t,m)}$ projects the first m dimensions of the student's CLS token to the teacher's embedding space, while $g_p^{(t,m)}$ performs an analogous projection for each patch token.

The projected features are defined as:

$$\hat{z}_s^{(t,m)} = g_s^{(t,m)}(z_{s,1:m}), \tag{6}$$

$$\hat{z}_p^{(t,i,m)} = g_p^{(t,m)}(z_{p,1:m}^{(i)}), \tag{7}$$

where $z_{s,1:m}$ and $z_{p,1:m}^{(i)}$ denote the first m dimensions of the corresponding embeddings.

The loss for the CLS token is computed as:

$$\mathcal{L}_{\text{cls}} = \sum_t \sum_m \mathcal{L}_{\cos}(\hat{z}_s^{(t,m)}, z_s^{(t)}), \tag{8}$$

where $\mathcal{L}_{\cos}$ is the cosine similarity loss and $z_s^{(t)}$ denotes the teacher's CLS token.

Before computing the patch-level loss, we standardize each teacher's patch embeddings to account for differences in embedding magnitude, following [42]:

$$\tilde{z}_p^{(t,i)} = \frac{z_p^{(t,i)} - \mu_c^t}{\sigma_c^t}, \tag{9}$$

where μ_c^t and σ_c^t are the mean and standard deviation computed over channel c of teacher t for the current batch.

Finally, the patch-level distillation loss is defined as:

$$\mathcal{L}_{\text{patch}} = \sum_t \sum_i \sum_m \mathcal{L}_{\text{MSE}}(\hat{z}_p^{(t,i,m)}, \tilde{z}_p^{(t,i)}), \tag{10}$$

where $\mathcal{L}_{\text{MSE}}$ denotes the mean squared error loss, utilized here in accordance with established distillation paradigms [43].

The loss formulation would allow assigning a weight λ_t to each teacher. Considering the large exploration space and the significant training cost, we chose to give each teacher equal weight. This ensures that the student model learns a balanced representation derived from the collective expertise of the ensemble.

3.4 Teacher Selection

Today, a plethora of pathology FMs exist [13, 17, 18, 24, 28, 34, 37, 44, 52, 54, 56–58]. We selected a representative ensemble of the most recent versions of high-capacity teachers: Virchow2 [58], UNI2-h [13] and H-optimus-1 [6]. These models were chosen for their state-of-the-art performance across diverse pathological benchmarks [38], with each exceeding 600 million parameters.

All of them are trained by distinct institutions with proprietary data increasing their heterogeneity compared to models trained on overlapping public datasets. Their complementarity is confirmed in the evaluation in Tab. 1. There is no single teacher dominating in all benchmark datasets.

To quantify the representational overlap, we computed the centered linear kernel alignment (CKA) [31] between the teacher models on a subset of 10,000 images of our training dataset. CKA measures the similarity between representations while being invariant to linear transformations and ranges from 0 (orthogonal) to 1 (identical). We observed pairwise CKA scores between 0.64 and 0.71 for

Table 1: Benchmark on public datasets for our models in comparison to the baselines H0-mini and GPFM, and the teacher models Virchow2, UNI2-h and H-optimus-1. We report average multiclass accuracy for multi-class classification, and binary balanced accuracy for binary classification tasks over five runs as reported by the *eva* framework [26]. † indicates that the dataset was part of the training data of this model, therefore generalization may be overestimated. Highlighted are **best** and second best models.

	Teacher					Ours	
Dataset	Virchow2	UNI2-h	H-optimus-1	H0-mini	GPFM	**Pathryoshka-S**	**Pathryoshka-B**
BACH	88.0 ±0.5	90.7 ±1.1	78.1 ±0.8	79.2 ±0.5	83.0† ±0.4	85.8 ±0.5	**90.8** ±0.3
BRACS	62.2 ±0.7	**66.1** ±0.9	61.9 ±0.5	60.6 ±0.6	62.8† ±0.5	61.7 ±1.0	65.3 ±0.2
BreakHis	81.9 ±0.6	85.9 ±0.3	76.0 ±1.2	74.7 ±0.9	76.0† ±0.5	76.2 ±0.4	**87.1** ±0.7
CRC	96.6 ±0.2	96.5 ±0.2	95.5 ±0.0	96.1 ±0.2	95.2† ±0.1	96.0 ±0.0	**96.7** ±0.1
Gleason	77.9 ±0.6	77.4 ±0.3	78.5 ±0.5	77.9 ±0.5	76.6 ±0.6	**78.7** ±0.8	**78.7** ±0.6
MHIST	**86.3** ±0.1	82.6 ±0.3	83.7 ±0.3	79.0 ±0.2	81.4 ±0.1	82.2 ±0.4	84.0 ±0.2
Patch Cam.	93.8 ±0.1	**95.1** ±0.1	94.2 ±0.1	94.2 ±0.1	94.4 ±0.0	92.1 ±0.1	94.1 ±0.1
CAM. 16	93.1 ±1.4	93.8 ±1.3	92.1 ±1.3	93.1 ±1.7	94.9† ±2.6	93.5 ±1.3	**95.0** ±1.0
PANDA	75.8 ±1.0	76.8 ±1.1	77.2 ±1.4	78.1 ±0.4	**79.1**† ±0.9	76.2 ±0.9	77.0 ±0.8
UniToPatho	**57.5** ±0.3	54.0 ±0.3	48.7 ±0.6	49.2 ±0.3	49.7† ±0.1	52.7 ±0.4	54.9 ±0.1
Median	84.1	84.3	78.3	78.6	80.3	80.5	**85.6**

our pathology-specific teacher models suggesting that their representations are domain-aligned but can contribute distinctive views. In comparison, a ViT-L model pre-trained with DINOv2 on the LVD142M natural image dataset [41] and our selected teachers have pairwise CKA scores between 0.35 and 0.40. To obtain a baseline for model-to-model consistency, we measured CKA of two models with different sizes (ViT-L and ViT-g) pre-trained on the same dataset (LVD142M) yielding a high similarity score of 0.88 for our subset.

4 Experiments and Results

We evaluate Pathryoshka in two sizes: Pathryoshka-B (86M parameters) and Pathryoshka-S (22M), and compare them to their teacher models, to the state-of-the-art single-teacher distilled pathology FM, H0-mini, and to the multi-teacher model GPFM. Because H0-mini and GPFM are distilled models, they are serving as baselines for single- and multi-teacher distillation. Furthermore, we evaluate how the different nesting levels behave compared to FMs without such an ordering in their embeddings.

4.1 Classification Benchmark Results

We evaluate all models on ten commonly used public pathology benchmarks. Eight of those are patch-level tasks (BACH [1], BRACS [7], BreakHis [46], CRC 100k [29], Gleason Grading [2], MHIST [55], Patch Camelyon [47], and UniToPatho [3]), and two are slide-level tasks (CAMELYON16 [4] and PANDA [9]). The selected benchmarks cover a wide range of magnifications levels from 0.25

Table 2: Comparison of computational efficiency across our models (Path.-S/B), teacher models, and baselines including parameter count, FLOPs, and throughput measured on an RTX 3090 GPU.

	Virchow2	UNI2-h	H-optimus-1	H0-mini	GPFM	**Path.-S**	**Path.-B**
Params. [M] (↓)	631	681	1,135	86	303	**22**	86
FLOPs [G] (↓)	329.1	360.7	591.8	44.6	155.6	**11.1**	44.6
Throughput [img/s] (↑)	139 ±2	136 ±1	82 ±1	842 ±4	278 ±1	**2338** ±14	842 ±4

microns per pixel (mpp) to 3.5 mpp. For all public benchmarks we used the *eva* [26] benchmark framework. This framework uses a single linear layer for patch classification tasks and an ABMIL [25] head for slide-level classification tasks to classify the CLS embedding of the respective model. To ensure a fair comparison of our models with the baseline and the teachers, we run the benchmarks for all models using *eva*'s default configuration. We report the average of five runs of the multiclass classification accuracy for multiclass tasks and binary balanced accuracy for binary classification tasks. The results are shown in Table 1.

Compared to its much larger teacher models, Pathryoshka-B delivers competitive performance across the benchmark datasets. It exceeds the median accuracy of UNI2-h by 1.3 points, Virchow2 by 1.5 points, and H-optimus-1 by 7.3 points, despite requiring 86%, 87%, and 92% fewer parameters, respectively. Furthermore, it surpasses the distilled models by 7.0 (H0-mini) and by 5.3 (GPFM) points in median accuracy, while additionally providing nested embeddings. With a four-fold reduction in parameters compared to Pathryoshka-B, Pathryoshka-S shows a 5.1-point drop in median accuracy. However, its median balanced accuracy still exceeds that of H0-mini, H-optimus-1, and GPFM. The advantage of Pathryoshka over the top-performing teacher model UNI2-h, as well as H0-mini and GPFM baselines is statistically significant as determined using a one-sided Wilcoxon Signed-Rank Test [14] across all datasets with p-values of $p = 0.042$, $p = 0.010$, and $p = 0.019$ respectively. Tests for teacher superiority over Pathryoshka-B yield consistently high p-values (UNI2-h: 0.967, Virchow2: 0.935, H-Optimus-1: 0.995), confirming the teachers fail to significantly outperform our student.

An ablation study of the random-cropping strategy introduced in Section 3 demonstrates its effectiveness: incorporating random cropping improves the average accuracy of Pathryoshka-B by 0.8 points and median by 0.2 across the public benchmarks as shown in Table 3. Another ablation of the number of teachers demonstrated that a single teacher variant inherited its teacher's weaknesses and performed worse in median balanced accuracy than our multi-teacher distilled model (Supplementary Table S6).

Table 2 compares parameter counts and inference efficiency of the evaluated models, reporting FLOPs and throughput (images per second) measured on a consumer-grade RTX 3090 GPU. Throughput is computed in FP16 precision with a batch size of 32, and we report the mean and standard deviation over 500 batches. Our best-performing model, Pathryoshka-B, uses 92% fewer parameters and achieves roughly 11× higher throughput compared to its largest teacher,

Table 3: Random cropping augmentation matches or exceeds performance framework on 9 out of 10 benchmarks, showing distinct improvements in 7 instances.

Dataset	Pathryoshka-B **crop**	Pathryoshka-B **no crop**
BACH	**90.8** ±0.3	90.0 ±0.4
BRACS	**65.3** ±0.2	64.6 ±0.7
BreakHis	**87.1** ±0.7	85.5 ±0.6
CRC	**96.7** ±0.1	96.0 ±0.1
Gleason	**78.7** ±0.6	77.7 ±0.3
MHIST	84.0 ±0.2	**85.1** ±0.2
Patch Cam.	**94.1** ±0.1	93.6 ±0.0
CAM. 16	**95.0** ±1.0	91.6 ±1.5
PANDA	**77.0** ±0.8	**77.0** ±1.2
UniToPatho	**54.9** ±0.1	**54.9** ±0.2
Average	**82.4**	81.6
Median	**85.6**	85.3

H-optimus-1. In whole slide image analysis, where thousands of image tiles must be encoded for a single prediction, this efficiency translates into substantial reductions in total processing time.

4.2 Evaluation of Nested Embeddings

k-NN-Classification. In Fig. 4, we benchmark the quality of the learned representations by progressively reducing the embedding size. We evaluate k-NN classification ($k = 10$, cosine similarity) on six multiclass patch classification datasets. Pathryoshka-B is compared against the best-performing teacher model, UNI2-h, and the two baselines H0-mini and GPFM. Note that the full embedding dimension varies across models: UNI2-h uses 1536, GPFM uses 1024, while both Pathryoshka-B and H0-mini use 768.

Due to its nested embedding design, the features of our model are inherently ordered by relevance. To assess this property, we evaluate performance using only the first $\frac{\text{dim}}{n}$ features of our model for $n \in \{1, 2, 4, 8, 16, 32, 64\}$. To ensure a fair comparison, we match these dimensionalities for the other models by randomly sampling $\frac{dim}{n}$ features, reporting the average over five runs. For UNI2-h, we additionally evaluate $n = 128$ so that its lowest feature count matches ours. Detailed results are provided in the Supplementary Material in Table S4.

Interestingly, performance remains relatively stable for all models up to $\frac{\text{dim}}{4}$, suggesting a degree of redundancy in the embeddings of UNI2-h, H0-mini and GPFM. Beyond this threshold, UNI2-h, H0-mini and GPFM exhibit a significant drop in accuracy. In contrast, Pathryoshka-B degrades much more gracefully. The average relative performance decrease from the full embedding down to $\frac{dim}{64}$ across the six datasets is only 9.4% for our model, compared to 17.7% (UNI2-h), 28.4% (H0-mini) and 22.64% (GPFM). At an embedding size of only 12 it is outperforming the other models by a large margin. Notably, even though

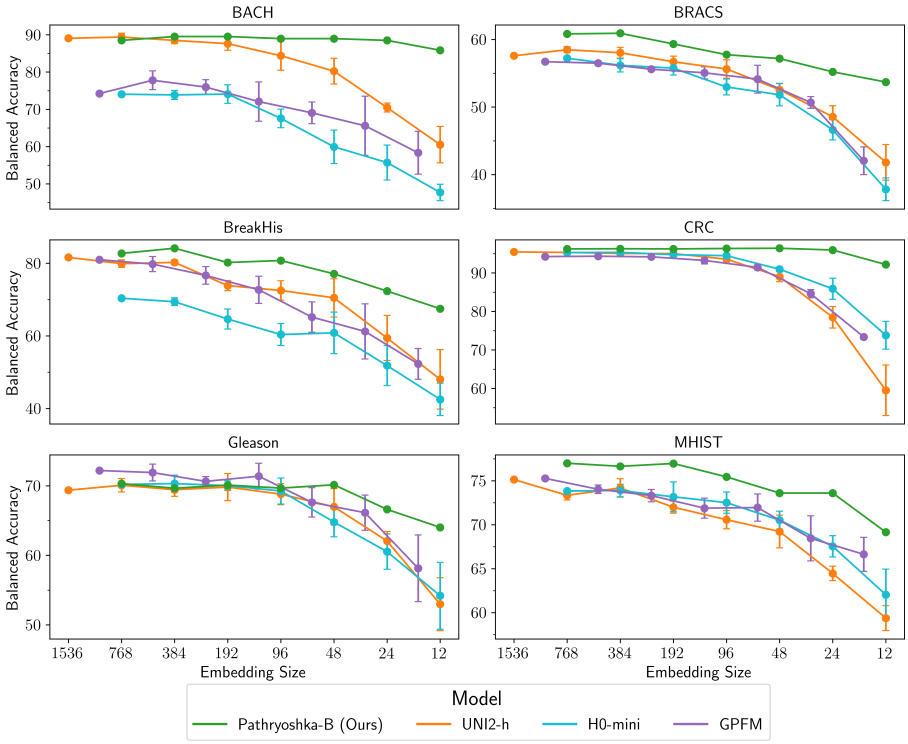

Fig. 4: Balanced k-NN classification accuracy between Pathryoshka-B, the best-performing teacher model UNI2-h, and the baselines H0-mini and GPFM, across six patch-based multiclass classification benchmarks. Each model is evaluated using successively halved subsets of its original embedding, down to a minimum embedding size of 12. Error bars indicate the standard deviation across five random feature sampling runs for the baseline models. Benefiting from its nested embedding structure, accuracy for Pathryoshka-B degrades gracefully even at severely reduced dimensionalities.

Pathryoshka-B is explicitly trained only for nested sizes down to $\frac{\dim}{16}$, further truncated embeddings still maintain competitive accuracy.

This behavior is further validated by the visualization of patch embeddings in Figure 5. Despite significant truncation, Pathryoshka-B preserves semantic consistency in its patch tokens. To demonstrate this, we projected the first $\frac{dim}{n}$ dimensions ($n \in \{1, 4, 16, 64, 256\}$) into RGB using PCA. When only a fraction of the original dimensions remain, UNI2-h exhibits inconsistent representations for similar histological structures while Pathryoshka-B maintains morphological consistency.

Patch Retrieval. We further evaluate the quality of nested embeddings via a patch retrieval task. For each query patch, the top-k nearest patches are retrieved

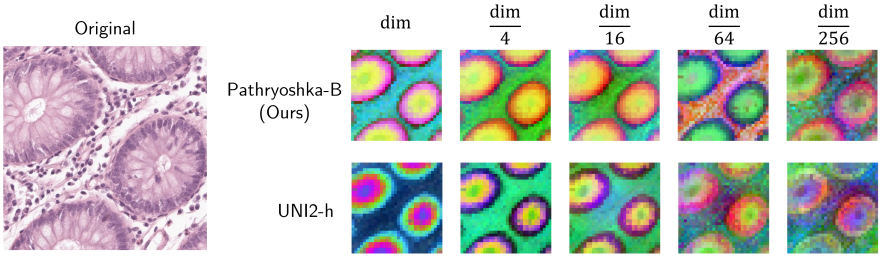

Fig. 5: Visualization of the semantic degradation of patch features when truncating the embeddings of Pathryoshka-B and UNI2-h, by mapping the first three PCA components to the RGB color space. For the given image of intestinal glands from our training dataset, our model is able to sharpely delineate glands from connective tissue even with a dimensionality of three. For UNI2-h, which does not have a nested embedding structure, the glands start to blend in with surrounding tissue below $\frac{dim}{64}$.

Table 4: Patch-retrieval experiment results on the CCRCC dataset. We report the average Recall@5 over all queries. Results for $K = 1$ and $K = 10$ are provided in the Supplementary Material Tab. S1. † indicates that CCRCC is contained in the training data. Pathryoshka-B exhibits minimal degradation when the feature dimensionality is reduced compared to the larger teachers and baseline models.

Model	dim	$\frac{\text{dim}}{2}$	$\frac{\text{dim}}{4}$	$\frac{\text{dim}}{8}$	$\frac{\text{dim}}{16}$	$\frac{\text{dim}}{32}$	$\frac{\text{dim}}{64}$
Virchow2	97.2	97.2	97.1	97.1	**97.0**	**96.6**	94.8
UNI2-h	97.2	97.1	97.0	97.0	96.7	95.6	92.4
H-optimus-1	97.1	97.1	97.0	97.0	96.6	95.7	92.7
H0-mini†	**97.4**	**97.3**	**97.2**	97.0	96.6	95.5	92.6
GPFM†	97.3	**97.3**	**97.2**	**97.2**	96.9	95.7	92.0
Pathryoshka-B	97.1	97.1	97.0	97.0	96.9	**96.6**	**95.2**

based on cosine similarity, and accuracy is measured by label agreement with the query. The experiment is conducted on the CCRCC dataset [8]. Table 4 shows Pathryoshka-B matches the retrieval performance of large-scale teacher models while surpassing all baselines at compressed embedding sizes of $\frac{dim}{32}$ and $\frac{dim}{64}$.

Segmentation. Finally, we performed quantitative verification of Pathryoshka's nested patch tokens using the *eva* evaluation framework and reporting the average Dice score of five independent runs for the ConSep [19] and MoNuSAC [49] datasets. In this task, we compared to the two distilled pathology FM baselines, H0-mini and GPFM. The segmentation masks are produced by forwarding the patch tokens and the raw input image to a task-specific head featuring a lightweight convolutional decoder. The resulting performance is detailed in Tab. 5. At a heavily reduced embedding size of $\frac{\text{dim}}{16}$, our nested model maintains robust accuracy with Dice score drops restricted to 1.6% (ConSep) and 3.9% (MoNuSAC), while significantly outperforming both distilled baselines.

Table 5: Evaluation of nested embeddings of the patch tokens on segmentation tasks using the ConSep and MoNuSAC dataset. We compare the performance of Pathryoshka-B with full and reduced embedding size of dim/16 against GPFM and H0-mini using the average Dice score of five runs.

	ConSep		**MoNuSAC**	
Model	full dim	dim/16	full dim	dim/16
Pathryoshka-B	**63.6** ±0.3	**62.6** ±0.2	63.8 ±0.7	**61.3** ±0.8
GPFM	47.8 ±0.2	47.6 ±0.4	48.3 ±0.4	47.4 ±0.8
H0-mini	63.1 ±0.4	54.8 ±0.1	**64.0** ±1.0	55.4 ±1.0

5 Discussion and Conclusion

In this work, we introduced Pathryoshka, a pathology foundation model that integrates multi-teacher distillation with nested embeddings. By employing a multi-teacher strategy with cropping that fosters robustness and distillation-adapted augmentation, Pathryoshka integrates complementary knowledge from multiple experts, mitigating the individual biases inherent in single-teacher distillation to produce a more robust and generalizable student. Furthermore, the implementation of nested embeddings provides a capability to reduce embedding size based on downstream task requirements enabling a trade-off between accuracy and computational cost. It allows for severe truncation of the embedding with only minimal degradation in accuracy. While validated on digital pathology, our approach is domain-agnostic and applicable to other fields.

From a deployment perspective, the model is already well suited for clinical settings: Pathryoshka delivers a lightweight architecture and fast inference while maintaining parity with state-of-the-art large-scale FMs and consistently outperforming state-of-the-art single-teacher and multi-teacher distillation pathology FMs. While our evaluation already masters well-established classification and representation benchmarks capturing many practical needs in pathology, future work includes more extensive evaluation of clinical utility.

Our approach challenges the conventional scaling paradigm, demonstrating that performance does not strictly depend on increasing model parameters or dataset volume. These findings suggest that current large-scale FMs underutilize their capacity, with high-dimensional embeddings containing significant redundancy. These results underscore the need for more efficient pre-training strategies over raw scaling, a direction we hope will spark further research in the field.

Acknowledgements

We acknowledge support by the Munich Data Science Institute (MDSI) at Technical University of Munich (TUM) via the MDSI Doctoral Fellowship program, and the BMFTR-funded SATURN3 project (01KD2206C).

Bibliography

[1] Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S.S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., et al.: Bach: Grand challenge on breast cancer histology images. Medical image analysis **56**, 122–139 (2019)

[2] Arvaniti, E., Fricker, K.S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P.J., Rueschoff, J.H., Claassen, M.: Automated gleason grading of prostate cancer tissue microarrays via deep learning. Scientific reports **8**(1), 12054 (2018)

[3] Barbano, C.A., Perlo, D., Tartaglione, E., Fiandrotti, A., Bertero, L., Cassoni, P., Grangetto, M.: Unitopatho, a labeled histopathological dataset for colorectal polyps classification and adenoma dysplasia grading. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 76–80. IEEE (2021)

[4] Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama **318**(22), 2199–2210 (2017)

[5] Bilal, M., Aadam, Raza, M., Altherwy, Y., Alsuhaibani, A., Abduljabbar, A., Almarshad, F., Golding, P., Rajpoot, N.: Foundation models in computational pathology: A review of challenges, opportunities, and impact (2025), https://arxiv.org/abs/2502.08333

[6] Bioptimus: H-optimus-1 (2025), https://huggingface.co/bioptimus/H-optimus-1

[7] 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)

[8] Brummer, O., Pölönen, P., Mustjoki, S., Brück, O.: Computational textural mapping harmonises sampling variation and reveals multidimensional histopathological fingerprints. British Journal of Cancer **129**(4), 683–695 (2023)

[9] Bulten, W., Kartasalo, K., Chen, P.H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., Van Boven, H., Vink, R., et al.: Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. Nature medicine **28**(1), 154–163 (2022)

[10] Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: Fast and flexible image augmentations. Information **11**(2) (2020). https://doi.org/10.3390/info11020125, https://www.mdpi.com/2078-2489/11/2/125

[11] 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 (ICCV). pp. 9650–9660 (October 2021)
[12] Chen, C., Lu, M.Y., Williamson, D.F., Chen, T.Y., Schaumberg, A.J., Mahmood, F.: Fast and scalable search of whole-slide images via self-supervised deep learning. Nature biomedical engineering **6**(12), 1420–1434 (2022)
[13] 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)
[14] Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research **7**(Jan), 1–30 (2006)
[15] Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
[16] 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: Gee, J.C., Alexander, D.C., Hong, J., Iglesias, J.E., Sudre, C.H., Venkataraman, A., Golland, P., Kim, J.H., Park, J. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. pp. 162–172. Springer Nature Switzerland, Cham (2026)
[17] Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Kain, A.M., Saillard, C., Schiratti, J.B.: Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling (Sep 2023). https://doi.org/10.1101/2023.07.21.23292757
[18] Filiot, A., Jacob, P., Kain, A.M., Saillard, C.: Phikon-v2, A large and public feature extractor for biomarker prediction (Sep 2024). https://doi.org/10.48550/arXiv.2409.09173
[19] Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N.: Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical image analysis **58**, 101563 (2019)
[20] Hashimoto, T.: Model performance scaling with multiple data sources. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 4107–4116. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/hashimoto21a.html
[21] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
[22] Heinrich, G., Ranzinger, M., Yin, H., Lu, Y., Kautz, J., Tao, A., Catanzaro, B., Molchanov, P.: Radiov2. 5: Improved baselines for agglomerative vision foundation models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22487–22497 (2025)
[23] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv [stat.ML] (Mar 2015)

[24] Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)

[25] Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)

[26] kaiko.ai, Gatopoulos, I., Känzig, N., Moser, R., Otálora, S.: eva: Evaluation framework for pathology foundation models. In: Medical Imaging with Deep Learning (2024), https://openreview.net/forum?id=FNBQOPj18N

[27] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models (2020), https://arxiv.org/abs/2001.08361

[28] Karasikov, M., van Doorn, J., Känzig, N., Erdal Cesur, M., Horlings, H.M., Berke, R., Tang, F., Otálora, S.: Training state-of-the-art pathology foundation models with orders of magnitude less data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 573–583. Springer (2025)

[29] Kather, J.N., Halama, N., Marx, A.: 100,000 histological images of human colorectal cancer and healthy tissue (v0.1). https://doi.org/10.5281/zenodo.1214456 (2018). https://doi.org/10.5281/zenodo.1214456, zenodo, Dataset

[30] Kawai, M., Ota, N., Yamaoka, S.: Large-scale pretraining on pathological images for fine-tuning of small pathological benchmarks. In: Xue, Z., Antani, S., Zamzmi, G., Yang, F., Rajaraman, S., Huang, S.X., Linguraru, M.G., Liang, Z. (eds.) Medical Image Learning with Limited and Noisy Data. pp. 257–267. Springer Nature Switzerland, Cham (2023)

[31] Kornblith, S., , et al.: Similarity of neural network representations revisited. In: International Conference on Machine Learning. pp. 3519–3529. PMlR (2019)

[32] Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., et al.: Matryoshka representation learning. Advances in Neural Information Processing Systems **35**, 30233–30249 (2022)

[33] Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., Parwani, A.V., Zhang, A., Mahmood, F.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (Mar 2024)

[34] Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., Parwani, A.V., Zhang, A., Mahmood, F.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (mar 2024). https://doi.org/10.1038/s41591-024-02856-4

[35] Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)

[36] 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 pp. 1–20 (2025)
[37] Nechaev, D., Pchelnikov, A., Ivanova, E.: Hibou: A family of foundational vision transformers for pathology. arXiv preprint arXiv:2406.05074 (2024)
[38] Neidlinger, P., El Nahhas, O.S., Muti, H.S., Lenz, T., Hoffmeister, M., Brenner, H., van Treeck, M., Langer, R., Dislich, B., Behrens, H.M., et al.: Benchmarking foundation models as feature extractors for weakly supervised computational pathology. Nature biomedical engineering pp. 1–11 (2025)
[39] Neidlinger, P., Nahhas, O.S.M.E., Muti, H.S., Lenz, T., Hoffmeister, M., Brenner, H., van Treeck, M., Langer, R., Dislich, B., Behrens, H.M., Röcken, C., Foersch, S., Truhn, D., Marra, A., Saldanha, O.L., Kather, J.N.: Benchmarking foundation models as feature extractors for weakly-supervised computational pathology. arXiv [eess.IV] (Aug 2024)
[40] Ochi, M., Komura, D., Ishikawa, S.: Pathology foundation models (2024), https://arxiv.org/abs/2407.21317
[41] 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)
[42] Ranzinger, M., Barker, J., Heinrich, G., Molchanov, P., Catanzaro, B., Tao, A.: Phi-s: Distribution balancing for label-free multi-teacher distillation. arXiv preprint arXiv:2410.01680 (2024)
[43] Ranzinger, M., Heinrich, G., Kautz, J., Molchanov, P.: Am-radio: Agglomerative vision foundation model reduce all domains into one. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12490–12500 (2024)
[44] Saillard, C., Jenatton, R., Llinares-López, F., Mariet, Z., Cahané, D., Durand, E., Vert, J.P.: H-optimus-0 (2024), https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0
[45] Schirris, Y., Gavves, E., Nederlof, I., Horlings, H.M., Teuwen, J.: Deepsmile: Contrastive self-supervised pre-training benefits msi and hrd classification directly from h&e whole-slide images in colorectal and breast cancer. Medical image analysis **79**, 102464 (2022)
[46] Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. Ieee transactions on biomedical engineering **63**(7), 1455–1462 (2015)
[47] Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant cnns for digital pathology. In: International Conference on Medical image computing and computer-assisted intervention. pp. 210–218. Springer (2018)
[48] Venkataramanan, S., Pariza, V., Salehi, M., Knobel, L., Gidaris, S., Ramzi, E., Bursuc, A., Asano, Y.M.: Franca: Nested matryoshka clustering for scalable visual representation learning. arXiv preprint arXiv:2507.14137 (2025)

[49] Verma, R., Kumar, N., Patil, A., Kurian, N.C., Rane, S., Graham, S., Vu, Q.D., Zwager, M., Raza, S.E.A., Rajpoot, N., et al.: Monusac2020: A multi-organ nuclei segmentation and classification challenge. IEEE Transactions on Medical Imaging **40**(12), 3413–3423 (2021)

[50] Villalobos, P., Sevilla, J., Besiroglu, T., Heim, L., Ho, A., Hobbhahn, M.: Machine learning model sizes and the parameter gap (2022), https://arxiv.org/abs/2207.02852

[51] Vo, H.V., Khalidov, V., Darcet, T., Moutakanni, T., Smetanin, N., Szafraniec, M., Touvron, H., Couprie, C., Oquab, M., Joulin, A., et al.: Automatic data curation for self-supervised learning: A clustering-based approach. arXiv preprint arXiv:2405.15613 (2024)

[52] 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)

[53] Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical Image Analysis **81**, 102559 (oct 2022). https://doi.org/10.1016/j.media.2022.102559

[54] Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**, 102559 (2022)

[55] Wei, J., Suriawinata, A., Ren, B., Liu, X., Lisovsky, M., Vaickus, L., Brown, C., Baker, M., Tomita, N., Torresani, L., et al.: A petri dish for histopathology image analysis. In: International Conference on Artificial Intelligence in Medicine. pp. 11–24. Springer (2021)

[56] 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)

[57] Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. Nejm Ai **2**(1), AIoa2400640 (2025)

[58] 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)