MELBA Journal Machine Learning for Biomedical Imaging

Benchmarking Foundation Models for Mitotic Figure Classification

Jonas **Ammeling** [1], Jonathan Ganz [2, 1], Emely Rosbach [1], Ludwig Lausser [1], Christof A. Bertram [3], Katharina Breininger [4, 5], Marc Aubreville [6]

1 Technische Hochschule Ingolstadt, AImotion, Ingolstadt, DE
2 MIRA vision Microscopy GmbH, Göppingen, DE
3 University of Veterinary Medicine, Institute of Pathology, Vienna, AU
4 Center for AI and Data Science, Julius-Maximilians-Universität Würzburg, Würzburg, DE
5 Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, DE
6 Flensburg University of Applied Sciences, Flensburg, DE

Abstract

The performance of deep learning models is known to scale with data quantity and diversity. In pathology, as in many other medical imaging domains, the availability of labeled images for a specific task is often limited. Self-supervised learning techniques have enabled the use of vast amounts of unlabeled data to train large-scale neural networks, i.e., foundation models, that can address the limited data problem by providing semantically rich feature vectors that can generalize well to new tasks with minimal training effort increasing model performance and robustness. In this work, we investigate the use of foundation models for mitotic figure classification. The mitotic count, which can be derived from this classification task, is an independent prognostic marker for specific tumors and part of certain tumor grading systems. In particular, we investigate the data scaling laws on multiple current foundation models and evaluate their robustness to unseen tumor domains. Next to the commonly used linear probing paradigm, we also adapt the models using low-rank adaptation (LoRA) of their attention mechanisms. We compare all models against end-to-end-trained baselines, both CNNs and Vision Transformers. Our results demonstrate that LoRA-adapted foundation models provide superior performance to those adapted with standard linear probing, reaching performance levels close to 100 % data availability with only 10 % of training data. Furthermore, LoRA-adaptation of the most recent foundation models almost closes the out-of-domain performance gap when evaluated on unseen tumor domains. However, full fine-tuning of traditional architectures still yields competitive performance.

Keywords
Machine Learning, Computational Pathology, Self-Supervised Learning, Foundation Models, Mitotic Figure Classification, Low-Rank Adaptation

Article informations
https://doi.org/10.59275/j.melba.2026-a3eb

Volume 2026, Received: 2025-07-15, Published 2026-01-30

Corresponding author: jonas.ammeling@thi.de

Special issue: MELBA–BVM 2025 Special Issue
Guest editors: Andreas Maier, Thomas Deserno, Heinz Handels, Klaus Maier-Hein, Christian Palm, Thomas Tolxdorff, Katharina Breininger

1. Introduction

Self-supervised learning (SSL) is transforming the landscape of publicly available methods for computational pathology (Khan et al., 2024). By leveraging vast amounts of unlabeled data, SSL overcomes the limitations of traditional supervised approaches, which are constrained by the availability of expert-annotated datasets. This is particularly advantageous in computational pathology, where annotation is a labor-intensive and time-consuming process requiring highly trained pathologists to examine large, high-resolution whole slide images (WSIs) with diverse morphological structures across tissue types. The annotation process is further challenged by fatigue (Stec et al., 2018) and cognitive biases (Aeffner et al., 2017; Viray et al., 2013; Leiser et al., 2023), leading to variability in label quality and high inter-rater variability (Smits et al., 2014). Additionally, differences in staining protocols and scanning devices across institutions can alter tissue appearance, complicating the development and deployment of robust methods (Aubreville et al., 2024). The scarcity of high-quality and large-scale datasets ultimately limits the advancement of supervised models and hinders their generalizability across diverse clinical settings. With the advent of self-supervised learning (SSL) techniques such as SimCLR (Chen et al., 2020a,b), MoCo (He et al., 2019; Chen et al., 2020c) and DINO (Caron et al., 2021; Oquab et al., 2023), large-scale neural networks were able to train on datasets with sizes beyond 1 billion images (Goyal et al., 2021) surpassing the performance of models trained on labeled data on competitive benchmarks like ImageNet (Tomasev et al., 2022; He et al., 2019; Deng et al., 2009). These models, often large Vision Transformers (ViTs) (Dosovitskiy et al., 2020), are commonly referred to as foundation models due to their ability to adapt to a wide range of downstream tasks with little to no fine-tuning. Because of their comprehensive pretraining, they are capable of generating semantically rich embeddings, which enables them to perform well in tasks such as few-shot learning and to serve as a robust foundation for a multitude of downstream applications (Zhang et al., 2024; Zhang and Metaxas, 2024).

Computational pathology is especially well-suited for SSL, as it routinely generates large volumes of image data. Public resources such as The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015) provide access to tens of thousands of WSIs from multiple institutions, offering a diverse and abundant source of training data. Leveraging these resources, several pathology-specific foundation models have recently been developed using TCGA (Chen and Krishnan, 2022; Wang et al., 2022a; Kang et al., 2022; Filiot et al., 2023) and other public or proprietary datasets (Chen et al., 2024a; Vorontsov et al., 2023; Zimmermann et al., 2024; Filiot et al., 2024; Charlie et al., 2024; Xu et al., 2024). These models are typically evaluated on downstream tasks such as tumor subtyping, tissue classification and mutation prediction.

A recent work by Vorontsov et al. (2023) has shown that the image embeddings of foundation models can also be utilized effectively for tasks that require fine-grained morphological features, like mitotic figure classification. This task is of particular importance in computational pathology, as accurate identification and classification of mitotic figures are crucial for assessing the aggressiveness of tumors and estimating the outcome of tumor patients (prognostication) (Elston and Ellis, 1991). However, the detection of mitotic figures remains challenging due to their morphological similarity to other cellular structures (Donovan et al., 2021) and their sparse occurrence within tissue sections (Aubreville et al., 2020b). In modern detection pipelines mitotic figure classification is often employed as a second-stage process, following the initial identification of candidate objects (Li et al., 2018; Aubreville et al., 2024). The integration of foundation models into these pipelines holds promise for improving classification performance or to reduce the need for large amounts of annotated data. Foundation models are typically adapted to new downstream tasks, such as mitotic figure classification, using techniques such as linear probing or model adaptation techniques like Low-Rank Adaptation (LoRA) (Hu et al., 2022). Linear probing involves training a simple linear classifier on top of the representations produced by the frozen foundation model, providing a fast and computationally efficient way to assess the quality of learned features. In contrast, adaptation techniques employ methods that selectively fine-tune parts of the foundation model to better tailor its representations to the downstream task. For example, LoRA introduces trainable low-rank matrices into selected layers of the model, enabling more flexible fine-tuning with minimal changes to the original model parameters. Both approaches can reduce the need for extensive fine-tuning and large annotated datasets, and the limited number of trainable parameters provides a regularizing effect that helps prevent overfitting.

However, there has been no in-depth analysis of how the performance of such classifiers depends on the size of the training set or how robust they are to domains shifts arising from differences between source and target image characteristics. This work, which extends previous work published as a conference paper (Ganz et al., 2025), aims to address these gaps by systematically investigating the scaling laws of several state-of-the-art pathology-specific foundation models for mitotic figure classification. To provide a comprehensive evaluation, we benchmark foundation model-based classifiers against several baseline methods across two publicly available mitotic figure datasets and evaluate the impact of the training set size using both linear probing and LoRA.

2. Related Work

The following section outlines the foundational principles and recent developments in self-supervised learning that underpin many state-of-the-art models in computational pathology.

2.1 Self-Supervised Learning

The development of SSL techniques marked a paradigm shift by enabling the training of large-scale neural networks on massive unlabeled datasets. In comparison to supervised learning techniques, where a model is trained on a specific task based on the available labeled data, SSL learns generic representations useful across many tasks without any labels by utilizing the intrinsic structure of the data. A major branch of SSL methods in computer vision is built based upon contrastive learning, which aims to create an embedding space where data points are organized based on their assumed similarity. SimCLR (Chen et al., 2020a,b) and MoCo (He et al., 2019; Chen et al., 2020c) are the most prominent methods in this paradigm, where two views of the same image, slightly altered by standard augmentation techniques, such as changing the color or cropping, are to be mapped to similar representations, whereas representations from views of different images are pushed further apart in latent space. Another branch of SSL methods is based on self-distillation, such as DINO (Caron et al., 2021), where a teacher-student framework is employed. In this setup, both the teacher and student networks share the same architecture but receive differently augmented views of the same image. The student, who only sees a smaller image crop, is trained to match the output distribution of the teacher, who sees a larger image crop. The distillation mechanism in this setup is given by the teacher being updated as an exponential moving average of the student's parameters. Another self-distillation method that builds on DINO is iBOT (Zhou et al., 2021), where a masked image modeling objective is added that is applied in the latent space directly, such that the target reconstruction is not the original image pixels but the same patches embedded through the teacher network. DINOv2 (Oquab et al., 2023) further builds on iBOT and is used by the majority of the recently published foundation models Chen et al. (2023a); Vorontsov et al. (2023). They improved the performance by modifying the training recipe and the architecture with better hyperparameter and regularizer such as KoLeo (Sablayrolles et al., 2018) to be more effective and stable at larger model and data sizes.

Due to the ability to leverage large-scale, unlabeled datasets, SSL techniques have gained a lot of attention across many healthcare applications (Khan et al., 2024), where labels are costly and time-consuming to acquire, such as clinical language models (Lee et al., 2019; Chen et al., 2023b; Yang et al., 2022a; Singhal et al., 2023), medical image analysis (Ma et al., 2024; Chen et al., 2024a; Wang et al., 2024; Xu et al., 2024; Alber et al., 2025; Filiot et al., 2024; Dippel et al., 2024), vision and language applications (Zhang et al., 2020; Wang et al., 2022b; Huang et al., 2023; Lu et al., 2024; Ahmed et al., 2024; Chen et al., 2024b), and omics research (Yang et al., 2022b; Celaj et al., 2023; Zhou et al., 2023).

2.2 Foundation Models in Pathology

Several pathology specific foundation models were published recently with promising performance across a multitude of downstream applications (Ciga et al., 2022; Wang et al., 2022a; Xu et al., 2024; Alber et al., 2025; Filiot et al., 2025; Charlie et al., 2024; Filiot et al., 2024; Dippel et al., 2024; Zimmermann et al., 2024), with mitotic figure classification being included by some of these works (Wang et al., 2022a; Vorontsov et al., 2023; Zimmermann et al., 2024; Shen et al., 2024). In particular, Wang et al. (2022a) proposed SRCL, an SSL method based on MoCov3 (Chen et al., 2021) along with CTransPath, a model architecture that combines convolutional layers with the Swin Transfomer model (Liu et al., 2021). Besides typical downstream tasks such as tile-level and slide-level classification, they also evaluated mitotic figure detection on the MIDOG 2021 (Aubreville et al., 2023a) dataset reporting an F1 score of 0.7332 on a custom test split, showing superior performance of SRCL to other SSL frameworks such as SimCLR and DINO (Wang et al., 2022a). They used the pre-trained CTransPath encoder as the backbone for the Faster R-CNN framework and performed full fine-tuning to adapt to the downstream task.

Vorontsov et al. (2023) introduced Virchow, a ViT-H model trained with DINOv2 (Oquab et al., 2023) on a massive proprietary dataset consisting of 2 billion tiles from almost 1.5 million slides across 17 tissue types. One of their downstream evaluation included mitotic figure classification on the MIDOG++ dataset (Aubreville et al., 2023b). They extracted patches of size 224×224 for each annotation from the original regions of interest (ROIs) and performed linear probing to train a classifier to distinguish between patches of mitotic figures and non-mitotic figures. They report an F1 score of 0.787 on a custom test split, outperforming all other tested foundation models.

Building on Virchow, Zimmermann et al. (2024) released Virchow2 based on ViT-H and Virchow2G based on ViT-G, both trained with DINOv2 (Oquab et al., 2023) on 1.7 billion and 1.9 billion tiles, respectively, from 3.1 million proprietary slides. Compared to the original Virchow model, they refined the training recipe to better suit pathology applications, incorporating mixed magnification training and exploring the effects of increased model and data scale as well as greater data diversity. They evaluated mitotic figure classification on the MIDOG++ dataset as well using the same linear probing protocol and test split as in the original Virchow work and report improved F1 scores of 0.804 for Virchow2 and 0.836 for Virchow2G.

Shen et al. (2024) introduced the Optimised Mitoses Generator Network (OMG-Net) to perform mitotic figure detection. Their 2-stage framework utilized SAM (Kirillov et al., 2023), a promptable foundation model with zero-shot capabilities to transfer to new image distributions and tasks, as first stage to outline candidate cells, followed by an adapted ResNet18 (He et al., 2015) that distinguishes mitotic figures. They combined publicly available mitotic figure datasets such as ICPR (Ludovic et al., 2013), TUPAC

(Veta et al., 2019), MIDOG++ (Aubreville et al., 2023b), two fully annotated WSI datasets for canine cutaneous mast cell tumor (CCMCT) (Bertram et al., 2019) and for canine mammary carcinoma (CMC) (Aubreville et al., 2020a), together with an in-house dataset of human soft tissue tumor (STT) to create a large database with 74620 mitotic figures to train their pipeline. They report F1 scores on a MIDOG++ test split ranging between 0.64 on neuroendocrine tumors to 0.86 for cutaneous mast cell tumors.

Xu et al. (2024) introduced Prov-GigaPath, a new foundation model for slide-level pretraining. They first trained a ViT-G architecture on tile-level using DINOv2, followed by slide-level pretraining using a masked autoencoder (He et al., 2021) and LongNet (Ding et al., 2023) to scale to thousands of image-tiles. They trained their model on 1.3 billion tiles from $171,189$ proprietary slides from Providence Health and Services. Prov-GigaPath was evaluated on 17 genomic prediction tasks and 9 cancer subtyping tasks using both Providence and TCGA data.

Chen et al. (2024a) introduced UNI, a tile-level foundation model based on ViT-L, trained with DINOv2 on more than 100 million tiles from 100K proprietary slides. They evaluated UNI on 34 clinical tasks with varying diagnostic difficulty, such as nuclear segmentation, primary and metastatic cancer detection, cancer grading and subtyping, molecular subtyping and several pan-cancer classification tasks.

Filiot et al. (2023) introduced Phikon, a ViT-B model trained with iBOT (Zhou et al., 2021), combining masked image modeling and contrastive learning. They trained their model on 43.3 million tiles from 6093 TCGA slides. They evaluated their performance across 17 downstream tasks including tile-level and slide-level tasks such as subtype classification, genomic alterations, and survival prediction.

Charlie et al. (2024) introduced H-optimus-0, a ViT-G model trained with DINOv2 on more than 500K proprietary slides. There is no exact number of tiles on which they trained their model but they mentioned more than several 100 million tiles. They evaluated across a wide range of downstream tasks as well covering tasks such as tissue classification, mutation prediction, and survival analysis.

2.3 Benchmarking Foundation Models

Due to the increasing availability of large-scale pathology foundation models of varying size, there is an increasing demand for unified and objective benchmarks of such models. Benchmarking these models is essential for providing fair and transparent comparisons across different architectures, training strategies, and datasets. Recent efforts have focused on establishing standardized evaluation protocols and curated test sets that encompass a diverse range of clinically relevant tasks.

Ma et al. (2025) introduced PathBench as a comprehensive benchmarking framework, featuring multi-center datsets, rigorous leakage prevention, and a standardized evaluation protocol across 64 diagnosis and prognosis tasks. They collected 15,888 slides from 8,549 patients and 10 hospitals and evaluated 19 recently published foundation models using a standardized preprocessing and linear probing protocol and showed that Virchow2 and H-optimus-1 are the most effective models overall.

Similarly, Campanella et al. (2025) provide a clinical benchmark dataset collected during standard hospital operation from three health systems including tasks such as disease detection and biomarker prediction. They evaluated 11 foundation models concluding that all DINO and DINOv2 trained models perform comparably, where H-optimus-0 and Prov-GigaPath performed significantly better in a few tasks.

Lee et al. (2025) performed a benchmark evaluation of four foundation models across 20 datasets in different scenarios where they address the influence of different adaptation strategies. In one scenario they tested linear probing, full fine-tuning, partial fine-tuning, parameter-efficient fine-tuning (PEFT) such as LoRA (Hu et al., 2022) and fully supervised learning and concluded that LoRA was both most efficient and effective when adapting to diverse datasets within the same classification tasks.

Breen et al. (2025) performed a task-specific benchmark study for ovarian cancer subtype classification. They compared three ImageNet-pretrained encoders and fourteen foundation models, each trained with 1,864 slides collected at Leeds Teaching Hospitals NHS Trust and validated on two external datasets. Their best performing classifier used the H-optimus-0, although UNI achieved similar results with only a quarter of the computational cost.

Neidlinger et al. (2024) benchmarked 19 foundation models on 9,528 slides from lung, colorectal, gastric, and breast cancers. They evaluated 31 weakly-supervised tasks related to morphology, biomarkers and prognostication. They report that the vision-language model CONCH (Lu et al., 2024) yielded the highest performance, when compared to vision-only models, where Virchow2 is the second best model. They also evaluated the downstream performance under different data scarcity settings. Their results indicate that while larger and more diverse pretraining datasets – slide count, patient count, and tissue site diversity – are generally associated with improved downstream performance, other factors like architecture and dataset quality also play critical roles. In scarce cohorts with only 75 to 300 patients for a specific downstream task or when rare biomarkers are involved, performance differences between models become more pronounced, with all models showing a decline as fine-tuning data decreases.

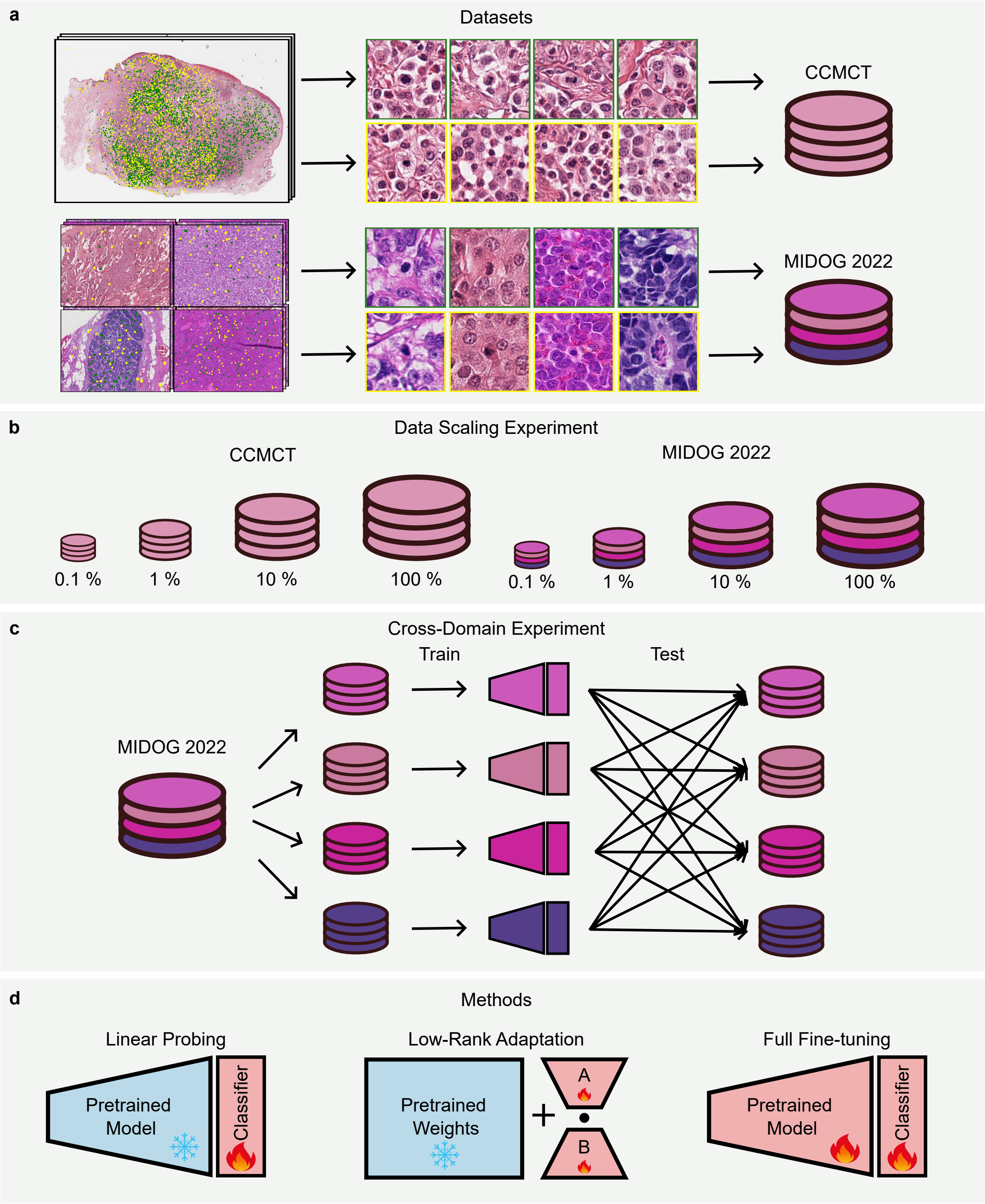

Figure 1: Benchmark study overview. **a)** Exemplary overview of datasets. Green shows mitotic figures and yellow shows hard negatives. During inference we extract patches of size 224×224 around these annotations for evaluation. **b)** Overview of dataset scaling experiments. **c)** Schematic overview of the cross-domain experiment. We train a model on each domain separately and evaluate across all domains. **d)** Overview of evaluated methods.

2.4 Benchmarking Adaptation Strategies

Adapting large-scale foundation models with billions of parameters to specific downstream tasks remains a significant challenge, particularly in resource-constrained settings. Parameter-efficient fine-tuning techniques, such as LoRA (Hu et al., 2022), have emerged as promising solutions to address these challenges by enabling effective adaptation with minimal additional parameters. Yang et al. (2024) provide a comprehensive review of LoRA, discussing its applications and associated challenges. Despite encouraging results reported in studies such as Lee et al. (2025), the use of LoRA in medical applications remains underexplored. For instance, Cui et al. (2024) adapted a ViT-B model pretrained with DINOv2 using LoRA for surgical depth estimation, demonstrating significant improvements over state-of-the-art models on the SCARED dataset which collected from da Vinci Xi endoscope surgery. Similarly, Dausort et al. (2024) investigated the application of LoRA for cytological classification, evaluating five foundation models fine-tuned with LoRA across four datasets. Their results show that LoRA fine-tuning consistently outperforms linear probing, with particularly strong gains in few-shot scenarios where labeled data is scarce. Furthermore, in a dataset scaling experiment, they demonstrated that a CLIP model fine-tuned with LoRA surpassed the state-of-the-art Hier-Swin (Cai et al., 2024) model when trained on just 70% of the data, highlighting the advantages of parameter-efficient adaptation methods like LoRA when labeled data is limited or costly to acquire.

Despite these advances, there remains a critical need to systematically evaluate how foundation models adapt to clinically relevant tasks such as mitotic figure classification, particularly when labeled data is limited. The comparative effectiveness of linear probing and LoRA for mitotic figure classification under varying data regimes remains an open question.

3. Methodology

In this work, we address this gap by benchmarking recently published pathology foundation models on the task of mitotic figure classification. We focus on comparing linear probing and LoRA-based fine-tuning in a data scaling experiment, highlighting their respective strengths and limitations in scarce data scenarios. Additionally, we investigate the robustness of these foundation models and adaptation strategies in a cross-domain setting, assessing their generalizability across different tumor types.

3.1 Datasets

Recent publicly available mitotic figure datasets (e.g., CMC (Aubreville et al., 2020a), CCMCT (Bertram et al., 2019), MIDOG 2022 (Aubreville et al., 2024), MIDOG++ (Aubreville et al., 2023b)) are larger and more diverse than before, making them ideal for benchmarking adaptation strategies. We conduct our analysis on two of these datasets, CCMCT (Bertram et al., 2019) and MIDOG 2022 (Aubreville et al., 2024) (Table 1), each chosen for their complementary properties. The CCMCT dataset is a large-scale resource focused on a single tumor domain, making it particularly well-suited for scaling experiments and in-depth analysis within a consistent biological context. In contrast, MIDOG 2022 is a highly diverse dataset spanning multiple tumor types, species, and laboratories, providing an ideal benchmark for evaluating model robustness and generalization in cross-domain settings.

3.1.1 CCMCT

The CCMCT dataset consists out of 32 fully annotated CCMCT WSIs with $44,880$ annotations for mitotic figures and $27,965$ hard negatives, including both low grade cases as well as high grade cases. The images were scanned with an Aperio ScanScope S2 WSI scanner at a resolution of 0.25 microns per pixel. For some examples see Figure 1.

3.1.2 MIDOG 2022

The MIDOG 2022 dataset is a comprehensive multi-tumor, multi-species, and multi-laboratory collection comprising 354 ROIs with a total of $11,051$ mitotic figures and $9,501$ challenging negative samples. It includes five distinct tumor types: canine cutaneous mast cell tumor (domain A), scanned at 40x magnification (0.25 microns per pixel) using the Aperio ScanScope CS2; canine lymphoma (domain B) scanned with the 3DHistech Panoramic Scan II scanner at 0.25 microns per pixel; human breast cancer (domain C), with 50 slides each scanned using Hamamatsu XR, Hamamatsu S360, and Aperio ScanScope CS2 scanners at resolutions ranging from 0.23 to 0.25 microns per pixel; human neuroendocrine tumor (domain D), scanned at 40x (0.23 microns per pixel) with the Hamamatsu XR; and canine lung cancer (domain E) digitized with the 3DHistech Panoramic Scan II scanner at 0.25 microns per pixel. Some examples are shown in Figure 1.

Table 1: Summary of datasets.

Dataset	Images	Mitotic figures	Hard negatives	Tumor types	Scanner	Magnification
CCMCT	32 WSIs	44,880	27,965	1	1	40x
MIDOG 2022	354 ROIs	11,051	9,501	5	4	40x

3.2 Foundation Models

We selected six state-of-the-art pathology foundation models (Table 2) that represent a diverse range of architectures, pretraining strategies, and dataset scales. The models include Phikon (Filiot et al., 2023), UNI (Chen et al., 2024a), Virchow (Vorontsov et al., 2023), Virchow2 (Zimmermann et al., 2024), H-optimus-0 (Charlie et al., 2024), and Prov-GigaPath (Xu et al., 2024), spanning backbones from ViT-B to ViT-G and pretraining algorithms such as iBOT and DINOv2. These models were pretrained on datasets ranging from public sources like TCGA to large proprietary collections, with slide counts varying from 6,093 to over 3 million and tile counts from 43 million to 2 billion. The selection covers a broad spectrum of model sizes (86M to 1B parameters), pretraining magnifications, and feature dimensionalities, providing a comprehensive benchmark for mitotic figure classification across different data and model scales. Additionally, we compare the pathology foundation models with ViT-S DINOv3 (Siméoni et al., 2025), a compact self-supervised Vision Transformer from the DINOv3 family. Trained on large-scale web datasets, it serves as a general-purpose visual encoder that produces robust, high-quality dense representations transferable across diverse vision tasks without task-specific fine-tuning.

3.3 Linear Probing

Linear probing is a widely adopted and straightforward approach for evaluating the quality of representations learned by pre-trained models. In this method, the parameters of a pre-trained model θ are frozen, and only a linear classifier is trained on top of the extracted features to adapt the model to a specific downstream task. Given an input x, the model computes feature representations $z = f_\theta(x)$, which are then passed to a linear layer. The output of the classifier is given by:

$$y = \mathbf{W}z \tag{1}$$

where $z \in \mathbb{R}^n$ is the feature vector, and $\mathbf{W} \in \mathbb{R}^{c \times n}$ the weight matrix of the linear classifier, with n the number of extracted features and c the number of classes. Linear probing serves as a standard benchmark to assess how well the pre-trained features can be separated by a simple linear decision boundary, providing insight into the generalizability and utility of the learned representations for new tasks without updating the backbone model. This approach is particularly valuable in scenarios with limited labeled data, as it requires training only a small number of parameters.

3.4 Low-Rank Adaptation

To explore the benefits of fine-tuning foundation models beyond the classification head, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022), a parameter-efficient fine-tuning technique. LoRA updates pre-trained model weights using low-rank decomposition. Instead of directly modifying the full weight matrix $\mathbf{W} \in \mathbb{R}^{m \times n}$, LoRA introduces a weight correction $\Delta\mathbf{W} \in \mathbb{R}^{m \times n}$, which is expressed as the product of two smaller matrices: $\mathbf{A} \in \mathbb{R}^{r \times n}$ and $\mathbf{B} \in \mathbb{R}^{m \times r}$, where r is the chosen rank and typically much smaller than m or n. The forward pass is then modified as follows:

$$h = \mathbf{W}x + \gamma\Delta\mathbf{W}x = \mathbf{W}x + \gamma\mathbf{BA}x$$

Here, h denotes the hidden state at a given layer, x is the input to that layer, and γ is a scaling factor. During initialization, $\mathbf{A}$ is randomly initialized, while $\mathbf{B}$ is set to zero. Only the entries of these low-rank matrices are updated during training, while the original model weights remain frozen, significantly reducing the number of trainable parameters. At inference time, the effective weight matrix is simply the sum of the original and the low-rank update, so the computational cost remains unchanged.

3.5 Baselines

We compare the performance of the foundation models to multiple baselines. First, we use the two widely adopted feature extractors ResNet50 and ViT-H, pretrained on ImageNet, to generate embeddings for linear probing in the same manner as with the foundation models. Additionally, we train four models starting from ImageNet pretraining in a fully supervised setting. We select a ResNet50 and three Vision Transformer (ViT-S, ViT-B, ViT-H) covering a range of model scales to directly compare with their foundation model counterparts. For these supervised baselines, standard image augmentations are applied during training, including random color jitter, Gaussian blur, flipping, and random rotations. This setup allows for a direct comparison between foundation model adaptation strategies and traditional supervised learning approaches.

3.6 Dataset Scaling Experiment

To systematically evaluate how foundation models and adaptation strategies perform under varying data availability, we conducted a dataset scaling experiment using both the large-scale, single-domain CCMCT dataset and the diverse, multi-domain MIDOG 2022 dataset. For each model, we trained on four different fractions of the available data (0.1%, 1%, 10%, and 100%), enabling us to assess model robustness and adaptation in both data-rich and data-scarce scenarios. To enhance the statistical reliability of our findings, we employed five-fold Monte Carlo cross-validation for each combination of model and training set size, resulting in 20 independent training runs per model. In the beginning, 20% of all annotations from the respective dataset were randomly selected as the test set to ensure a fair comparison between the different fractions of the data. Then, for each run, we set aside 20% of the remaining training annotations for validation. We used a case-level split to avoid data leakage between the splits. All models were trained and evaluated on identical splits to ensure fair comparison between the models. It is important to note that, due to the smaller overall size of MIDOG 2022, the absolute number of training samples at each percentage level was substantially lower than in CCMCT, providing a stringent test of model performance in low-data regimes.

3.7 Cross-Domain Experiment

To further investigate the generalization capabilities of the models, we performed a cross-domain experiment using the MIDOG 2022 dataset, which is especially suitable for multi-domain evaluation. In this setup, we used each of the five tumor domains as the training domain, while using the remaining domains exclusively for testing, thereby simulating real-world scenarios where models are deployed on previ-

Table 2: Summary of pathology foundation models.

Model	Backbone	Pretraining algorithm	Parameters	Data source	Pretraining magnifications	Slide count	Tile count	Patch features
Phikon	ViT-B	iBOT	86M	TCGA	20x	6093	43,3M	768
UNI	ViT-L	DINOv2	304M	Proprietary	20x	100K	100M	1024
Virchow	ViT-H	DINOv2	632M	Proprietary	20x	1.5M	2B	1280
Virchow2	ViT-H	DINOv2	632M	Proprietary	5x, 10x, 20x, 40x	3.1M	1.7B	1280
H-optimus-0	ViT-G	DINOv2	1B	Proprietary	20x	500K	NA	1536
Prov-GigaPath	ViT-G	DINOv2	1B	Proprietary	20x	170K	1.3B	1536

ously unseen data distributions. For in-domain evaluation, 20% of the data from the training domain was withheld and included as further test set for in-domain evaluation. As with the scaling experiment, we conducted five independent training runs per domain and per model, resulting in a total of 25 training sessions per model. This experimental design allows for a comprehensive assessment of both in-domain and cross-domain robustness, leveraging the diversity of MIDOG 2022.

3.8 Implementation Details

For the mitotic figure classification task we define $c = 2$ to distinguish between mitotic figures and mitotic figure look-alikes (hard negatives), which both datasets provide. We extracted image patches of size 224×224 pixels centered around the mitotic figure and hard negative annotations (Vorontsov et al., 2023). The patch embeddings for the linear probing experiments were created by passing the patches through the frozen encoder of each model. The patches were normalized according to the means and standard deviations provided in the respective works. No additional data augmentation was applied. For Virchow and Virchow2, these patch embeddings were created by concatenating the class token and the mean across all other 256 predicted patch tokens, as described in Vorontsov et al. (2023). For all other models, only the class token was used (Vorontsov et al., 2023). We use the linear probing implementation by Chen et al. (2024a).

We apply LoRA to the query, key, and value projection layer and the output projection layer of the attention blocks and the first and second fully connected layers in the multi-layer perceptron (MLP) sections. The rank is set to 16, the scaling factor is set to 16, the dropout is set to 0.1.

The fully supervised learning of the baselines and the LoRA-adaptation of the foundation models is performed using the Adam optimizer with default parameter values ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e - 8$), batch size of 16, patch size of 224×224, and standard image augmentations. Binary cross-entropy loss is adopted as loss function. In each pseudo epoch, patches are sampled randomly with 50% of the patches containing mitotic figures, and the other 50% containing either a hard negative patch or a completely random patch, each with 25% probability. The pseudo-epoch length is set to 1280. We train the models for 100 pseudo epochs using a one-cycle learning rate policy (Smith and Topin, 2017), including a linear warum-up phase followed by cosine annealing, with a maximum learning rate of 10^{-4}, and select the best model retrospectively based on the validation loss per epoch. During inference, we evaluate only patches that contain annotations for mitotic figures or hard negatives. The classification experiments are evaluated with balanced accuracy and weighted F1 scores due to the class imbalance between mitotic figures and hard negatives, and additionally with the AUROC score. All experiments were executed on a workstation equipped with a single NVIDIA RTX 3090 GPU.

4. Results

4.1 Data Scaling Experiment

The results of the data scaling experiment are shown in Figure 2 and Figure 3. The experiments on both the single-domain CCMCT and the multi-domain MIDOG 2022 datasets consistently demonstrate the superior adaptability and data efficiency of pathology foundation models, particularly when adapted with parameter efficient fine-tuning methods such as LoRA. Across both datasets, foundation models outperform traditional feature extractors such as ResNet50 and ViT-H pretrained on ImageNet at every data regime, with the performance gap being most pronounced in low-data settings.

On the CCMCT dataset, which represents a large-scale, single-domain scenario, foundation models fine-tuned with LoRA achieve substantial gains in AUROC as the training set size increases. Already at the smallest data fraction of 0.1% the LoRA adapted foundation models surpass the baselines, and this advantage becomes more pronounced as more data is made available. At full data scale, these models approach AUROC values of 0.9, while baselines plateau at lower levels. Interestingly, the performance of LoRA fine-tuned foundation models already reaches near full-data scale performance already at 10% of the available data, especially with the H-optimus-0 model. The ResNet50 End-to-End model consistently outperforms all foundation models adapted with standard linear probing and is only outperformed by Virchow2 and H-optimus-0 adapted with

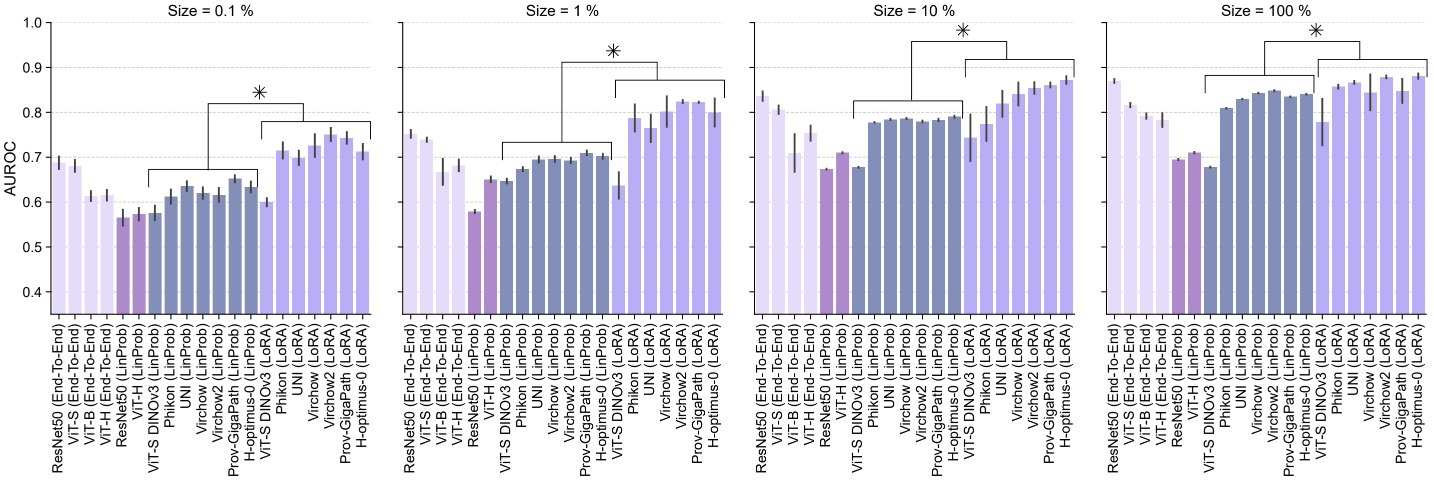

Figure 2: Results of the data scaling experiment on the CCMCT dataset. (*) indicates statistical significance ($\alpha < 0.05$) between the pooled scores of LoRA and LinProb models.

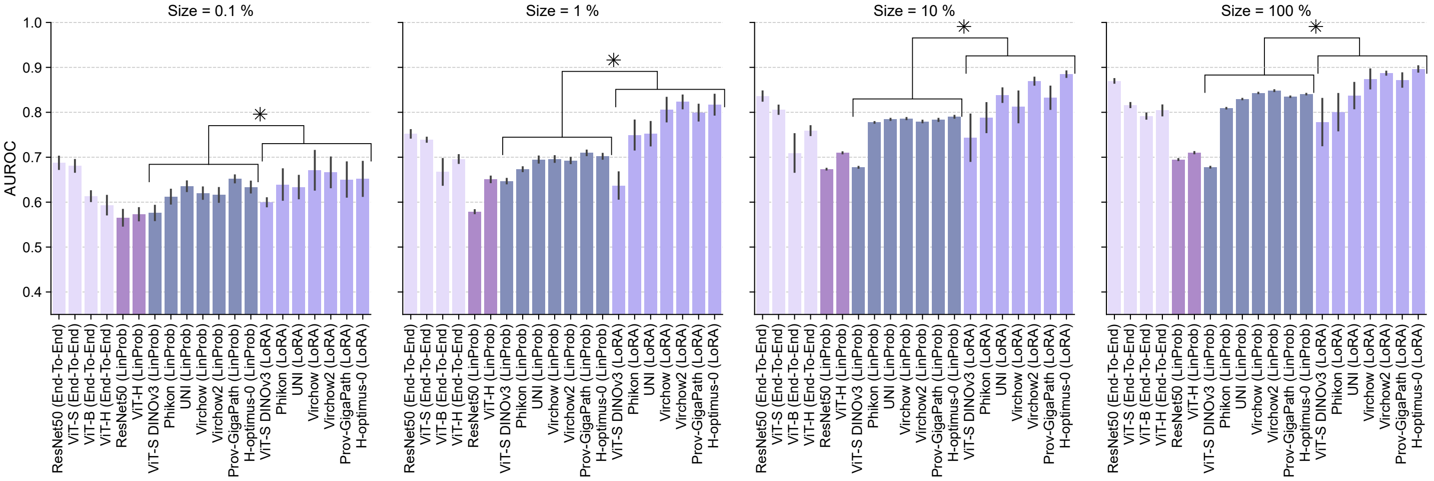

Figure 3: Results of the data scaling experiment on the MIDOG 2022 dataset. (*) indicates statistical significance ($\alpha < 0.05$) between the pooled scores of LoRA and LinProb models.

LoRA at full-data scale (Table 3). The ViT End-to-End variants show similar performance across the different data sizes with a clear advantage of the more compact ViT-S model at all data scales. However, their AUROC scores are considerably lower compared to the ResNet50 End-to-End model and moderately lower compared to their foundation model counterparts, especially for ViT-B and ViT-H. The performance of adapted ViT-S DINOv3 model lacks behind the pathology foundation models in both linear probing and LoRA settings. Even in the LoRA setting, the performance cannot match its ViT-S End-to-End counterpart, indicating that the domain shift from general purpose weights to histopathology data requires more extensive adaptation than LoRA fine-tuning.

A similar trend is observed on the MIDOG 2022 dataset, which is characterized by greater diversity in tumor types, species, and imaging conditions. Compared to CCMCT, the more challenging conditions and reduced absolute training sample sizes in MIDOG 2022 are reflected in generally lower AUROC scores, particularly for linear probing baselines and foundation models at the lowest data regimes (0.1% to 1%). However, the benefits of foundation models with LoRA adaptation become more clear at higher data fractions (1% to 100%), where these models substantially outperform their linear probing counterparts. The only exceptions are UNI and Phikon, which lag slightly behind at full data scale. Again, the ResNet50 End-to-End baseline outperforms linear probing of all foundation models and is only outperformed by Virchow2 and H-optimus-0 adapted with LoRA at full data scale (Table 4).

For each data scale, we pooled the performance scores from all foundation models fine-tuned using linear probing or LoRA, yielding 35 paired observations per adaptation strategy (7 models × 5 repetitions). To assess whether performance differed between the two adaptation strategies, we conducted Wilcoxon signed-rank tests Wilcoxon (1945).

Across all data scales, the results indicated that linear probing achieved significantly lower scores than LoRA ($\alpha = 0.05$).

Table 3: Results at 100 % dataset size of CCMCT dataset.

Model	AUROC	Balanced ACC	Weighted F1
ResNet50 (End-To-End)	0.87±0.01	0.79±0.01	0.78±0.01
ViT-S (End-To-End)	0.82±0.01	0.73±0.01	0.72±0.03
ViT-B (End-To-End)	0.79±0.01	0.70±0.02	0.71±0.03
ViT-H (End-To-End)	0.78±0.03	0.67±0.04	0.67±0.11
ResNet50 (LinProb)	0.69±0.00	0.61±0.00	0.65±0.00
ViT-H (LinProb)	0.71±0.00	0.61±0.00	0.65±0.00
ViT-S DINOv3 (LinProb)	0.68±0.00	0.57±0.00	0.60±0.00
Phikon (LinProb)	0.81±0.00	0.71±0.00	0.74±0.00
UNI (LinProb)	0.83±0.00	0.73±0.00	0.76±0.00
Virchow (LinProb)	0.84±0.00	0.75±0.00	0.78±0.00
Virchow2 (LinProb)	0.85±0.00	0.76±0.00	0.78±0.00
Prov-GigaPath (LinProb)	0.83±0.00	0.74±0.00	0.77±0.00
H-optimus-0 (LinProb)	0.84±0.00	0.75±0.00	0.78±0.00
ViT-S DINOv3 (LoRA)	0.78±0.12	0.70±0.09	0.71±0.09
Phikon (LoRA)	0.86±0.01	0.77±0.01	0.78±0.01
UNI (LoRA)	0.87±0.01	0.78±0.01	0.78±0.01
Virchow (LoRA)	0.84±0.09	0.75±0.07	0.76±0.06
Virchow2 (LoRA)	0.88±0.01	0.78±0.01	0.80±0.01
Prov-GigaPath (LoRA)	0.85±0.06	0.76±0.05	0.77±0.04
H-optimus-0 (LoRA)	**0.88±0.01**	**0.79±0.02**	**0.80±0.01**

Table 4: Results at 100% dataset size of MIDOG dataset.

Model	AUROC	Balanced ACC	Weighted F1
ResNet50 (End-To-End)	0.87±0.009	0.79±0.009	0.78±0.010
ViT-S (End-To-End)	0.82±0.010	0.73±0.012	0.72±0.027
ViT-B (End-To-End)	0.79±0.013	0.70±0.021	0.71±0.025
ViT-H (End-To-End)	0.80±0.035	0.70±0.051	0.70±0.087
ResNet50 (LinProb)	0.69±0.003	0.61±0.002	0.65±0.002
ViT-H (LinProb)	0.71±0.003	0.61±0.002	0.65±0.002
ViT-S DINOv3 (LinProb)	0.68±0.002	0.57±0.003	0.60±0.005
Phikon (LinProb)	0.81±0.001	0.71±0.002	0.74±0.002
UNI (LinProb)	0.83±0.001	0.73±0.003	0.76±0.002
Virchow (LinProb)	0.84±0.001	0.75±0.002	0.78±0.002
Virchow2 (LinProb)	0.85±0.002	0.76±0.001	0.78±0.001
Prov-GigaPath (LinProb)	0.83±0.001	0.74±0.002	0.77±0.002
H-optimus-0 (LinProb)	0.84±0.001	0.75±0.002	0.78±0.002
ViT-S DINOv3 (LoRA)	0.78±0.115	0.70±0.092	0.71±0.095
Phikon (LoRA)	0.80±0.128	0.73±0.102	0.73±0.105
UNI (LoRA)	0.84±0.090	0.76±0.072	0.76±0.072
Virchow (LoRA)	0.87±0.067	0.78±0.060	0.79±0.051
Virchow2 (LoRA)	0.89±0.011	0.80±0.022	0.81±0.014
Prov-GigaPath (LoRA)	0.87±0.047	0.79±0.044	0.79±0.037
H-optimus-0 (LoRA)	**0.90±0.019**	**0.81±0.022**	**0.81±0.015**

4.2 Cross-Domain Experiment

The results of the cross-domain experiment are summarized in Table 5 and Figure 4. As expected, all models perform better in-domain than out-of-domain, reflecting the inherent challenge of domain shifts in histopathology. Traditional feature extractors such as ResNet50 and ViT-H pretrained on ImageNet show limited generalization, with out-of-domain AUROC scores dropping to 0.53 and 0.56, respectively.

Foundation models adapted with standard linear probing demonstrate moderate improvements, with out-of-domain AUROC values in the range of 0.61-0.66 (Table 5). However, the most substantial gains are observed when foundation models are fine-tuned with LoRA. In this setting, models such as H-optimus-0, Virchow2, and Prov-Gigapath achieve out-of-domain AUROC scores of 0.88, 0.87, and 0.85 respectively, with only a minimal drop compared to their in-domain performance. This trend is consistent across all evaluation metrics, including balanced accuracy and weighted F1, highlighting the effectiveness of LoRA in enhancing cross-domain robustness.

We performed planned pairwise comparisons to assess whether the performance differences between the linear probing and LoRA models were statistically significant, and whether LoRA performance differed from the ResNet50 End-to-End baseline, yielding 14 model comparisons. For each in-domain comparison, we analyzed 25 paired observations (5 domains × 5 repetitions), and for each out-of-domain comparison, 100 paired observations (5 domains × 4 out-of-domain evaluations × 5 repetitions). All planned comparisons were conducted separately for each evaluation metric and domain setting (in-domain and out-of-domain) using the Wilcoxon signed-rank test. Resulting p-values were adjusted for multiple comparisons using the Holm procedure (Holm, 1979) for each metric and domain setting. Nearly all LoRA models significantly outperformed their linear-probing counterparts for every evaluation metric, except for Phikon, UNI, and Virchow, where not all differences were statistically significant. The general-purpose ViT-S DINOv3 model was also significantly outperformed by the ResNet50 End-to-End baseline. The strongest improvements over the ResNet50 End-to-End baseline were observed for Virchow2 and H-optimus-0, where both F1-scores and AUROC metrics differed significantly.

Looking at individual scenarios in Figure 4 we can clearly observe the strong gains of LoRA adaptation for the foundation models. Especially when looking at models such as Prov-Gigapath, H-opimus-0, and Virchow2 where the LoRA-adapted models nearly closed the out-of-domain performance gap between any scenario. Despite strong gains with LoRA some scenarios were still particularly challenging. Domain D (human neuroendocrine tumor) and E (canine lung cancer) were most challenging for all models. Training on either of these domains led to the worst performances across all models, with the biggest differences observed in Phikon, UNI, and Virchow.

Table 5: Results of the cross-domain experiment. The results are averaged across all in-domain and out-domain scenarios. Displayed are the mean score and the standard deviation. (*) indicates statistically significant differences compared to the LinProb counterpart ($\alpha = 0.05$). (†) indicates statistically significant differences compared to the ResNet50 End-to-End baseline ($\alpha = 0.05$).

Model	AUROC		Balanced ACC		Weighted F1	
	In-domain	Out-of-domain	In-domain	Out-of-domain	In-domain	Out-of-domain
ResNet50 (End-to-End)	0.87±0.04	0.74 ±0.07	0.79±0.03	0.67±0.06	0.81±0.03	0.67±0.09
ViT-S (End-to-End)	0.84±0.03	0.76±0.04	0.74±0.03	0.68±0.04	0.75±0.02	0.67±0.06
ViT-B (End-to-End)	0.83±0.03	0.75±0.05	0.75±0.03	0.67±0.04	0.76±0.04	0.65±0.07
ViT-H (End-to-End)	0.83±0.03	0.75±0.04	0.75±0.02	0.66±0.04	0.76±0.03	0.65±0.08
ResNet50 (LinProb)	0.63±0.04	0.53±0.03	0.58±0.02	0.51±0.01	0.62±0.05	0.48±0.08
ViT-H (LinProb)	0.68±0.05	0.56±0.04	0.59±0.03	0.53±0.02	0.63±0.05	0.49±0.10
ViT-S DINOv3 (LinProb)	0.65±0.06	0.58±0.04	0.56±0.02	0.53±0.03	0.58±0.07	0.47±0.12
Phikon (LinProb)	0.76±0.03	0.61±0.05	0.68±0.03	0.56±0.04	0.71±0.03	0.55±0.09
UNI (LinProb)	0.76±0.03	0.64±0.05	0.69±0.03	0.59±0.03	0.71±0.02	0.59±0.06
Virchow (LinProb)	0.79±0.04	0.66±0.06	0.72±0.04	0.60±0.03	0.74±0.02	0.60±0.06
Virchow2 (LinProb)	0.79±0.04	0.66±0.05	0.71±0.04	0.60±0.04	0.74±0.02	0.59±0.07
Prov-GigaPath (LinProb)	0.79±0.03	0.66±0.07	0.71±0.03	0.61±0.04	0.74±0.02	0.61±0.05
H-optimus-0 (LinProb)	0.79±0.04	0.66±0.07	0.72±0.03	0.60±0.05	0.74±0.02	0.61±0.07
ViT-S DINOv3 (LoRA)	0.82±0.05*	0.75±0.05*	0.74±0.04*†	0.68±0.04*†	0.74±0.04*†	0.67±0.06*†
Phikon (LoRA)	0.81±0.05*	0.76±0.09*	0.74±0.07	0.69±0.07	0.76±0.07	0.68±0.09
UNI (LoRA)	0.86±0.05*†	0.80±0.06*†	0.78±0.05	0.72±0.06	0.79±0.04*	0.72±0.08*
Virchow (LoRA)	0.86±0.08*†	0.82±0.08*†	0.78±0.06	0.75±0.07	0.79±0.07	0.75±0.09
Virchow2 (LoRA)	0.89±0.02*†	0.87±0.02*†	0.82±0.02*	0.79±0.03*	0.83±0.01*†	0.80±0.05*†
Prov-GigaPath (LoRA)	0.89±0.02*†	0.85±0.04*†	0.80±0.03*	0.76±0.05*	0.81±0.02*	0.77±0.06*
H-optimus-0 (LoRA)	**0.90±0.02***†	**0.88±0.02***†	**0.82±0.02***	**0.80±0.03***	**0.83±0.01***†	**0.81±0.04***†

5. Discussion

Our systematic benchmarking of pathology foundation models for mitotic figure classification across both single-domain and multi-domain datasets provides several important insights for the field of computational pathology. Most notably, our results demonstrate that foundation models, particularly when adapted with parameter-efficient fine-tuning methods such as LoRA, offer substantial advantages in both data-scarce and cross-domain scenarios.

The data scaling experiments reveal that foundation models consistently outperform traditional feature extractors, with the performance gap being most pronounced in low-data regimes. This suggests that rich, transferable representations learned during large-scale pretraining can be effectively leveraged for a new task such as mitotic figure classification when only a small amount of labeled data is available. The ability of LoRA-adapted models to reach near full-data scale performance with as little as 10% of the available data highlights the practical value of parameter-efficient adaptation strategies for real-world applications where annotation is often costly and time-consuming. Furthermore, the comparison between fully fine-tuning large vision transformers such as ViT-B and ViT-H and their LoRA-adapted counterparts highlights how efficient LoRA acts as a regularization when fine-tuning such large-scale models to a new task. On the other hand, the more commonly used linear probing does not utilize the full potential of foundation models, demonstrating inferior performance to LoRA-adapted models and fully fine-tuned traditional architectures such as ResNet50.

In the more challenging, heterogeneous setting of the MIDOG 2022 dataset, foundation models again demonstrate superior robustness, with LoRA adaptation further narrowing the gap between in-domain and out-of-domain performance. They maintain high evaluation scores even when tested on previously unseen tumor types, especially models such as Virchow2 and H-optimus-0 where cross-domain results are very similar in every scenario. This robustness is critical for clinical deployment, where models should be able to generalize to new data distributions and avoid overfitting to specific domains.

Adapting foundation models with parameter-efficient methods like LoRA offers a strong compute–performance trade-off. LoRA updates a small fraction of weights (typically ¡5%), significantly reducing memory and training time compared to full fine-tuning. While linear probing is most

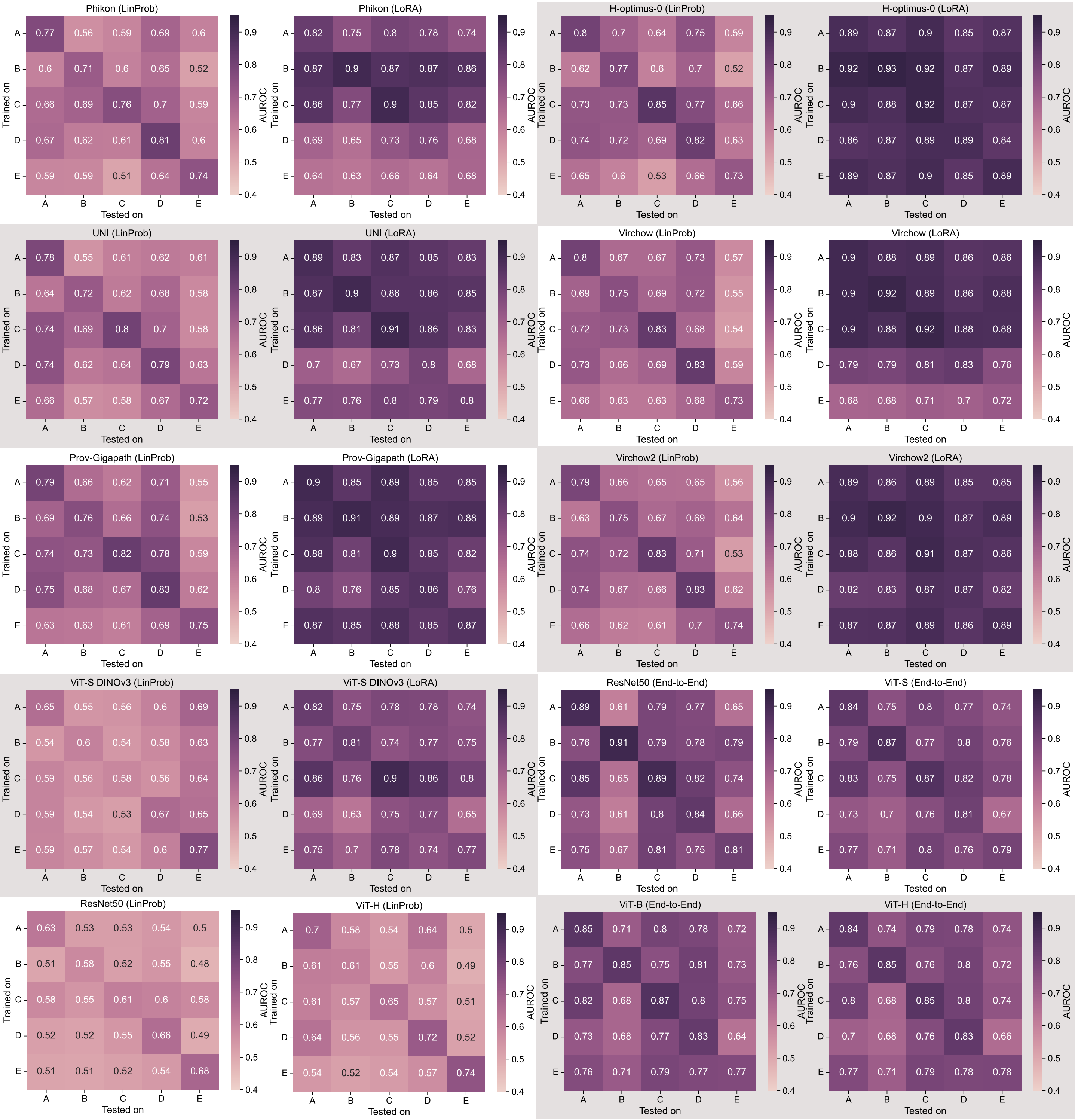

Figure 4: Results of the cross-domain experiment. We show each individual scenario with its averaged AUROC score over all training sessions. A: canine mast cell tumor. B: canine lymphoma. C: human breast cancer. D: human neuroendocrine tumor. E: canine lung cancer.

efficient, LoRA provides substantial performance gains with only modest additional compute. This makes adapted foundation models a pragmatic choice over training large Vision Transformers end-to-end, especially when compute, time, or data is limited. They offer strong accuracy with reduced training costs and deployment complexity. Rigorous quantification of these efficiencies is a key area for future research.

While we freeze or adapt transformer layers uniformly in our experiments, the need for adaptation is likely heterogeneous across depth. Early blocks tend to encode low-level features (e.g., edges, textures), whereas deeper blocks capture higher-level, task- and domain-specific concepts. Under domain shift, both strata can degrade. Early blocks may require modest recalibration when low-level statistics shift (e.g., noise, color, resolution), while later blocks may demand stronger adaptation to realign semantic representations. This motivates a more in-depth investigation into feature discrimination across model depth for future work (Ammeling et al., 2025).

Despite the promising results, several limitations should be acknowledged. First, while our experiments cover a range of diverse foundation models and common adaptation strategies, the scope of this benchmark study is limited to mitotic figure classification. The generalizability to other specific histopathological tasks, remains to be established. Second, mitotic figure classification is often performed as a secondary stage after detecting initial candidate objects through a detection or segmentation pipeline, hence the integration into a full mitotic figure detection pipeline and its evaluation needs further work for full integration into clinical practice. Third, while LoRA proved highly effective, we did not exhaustively explore the effects of the hyperparameter or other parameter-efficient fine-tuning methods or combinations thereof, which could yield further improvements.

The ResNet50 end-to-end baseline performed strongly, sometimes outperforming foundation models with linear probing or LoRA adaptation. This suggests that, in some cases, traditional architectures with full fine-tuning can still be competitive, especially when sufficient labeled data is available. Future work should further investigate the conditions under which foundation models provide the greatest benefits over conventional approaches.

6. Conclusion

Our findings support the growing consensus that foundation models, when paired with efficient adaptation strategies, are poised to transform computational pathology by enabling robust, scalable solutions that generalize across tasks and domains. The demonstrated data efficiency and cross-domain robustness of LoRA-adapted models are particularly relevant for clinical translation, where data heterogeneity and annotation scarcity are persistent challenges. Future research should extend these benchmarks to additional tasks, datasets, and adaptation methods, and explore strategies for further improving out-of-domain generalization.

Acknowledgments

This work was supported by the Bavarian State Ministry of Science and the Arts (project Fokus-TML) and by the German Research Foundation (DFG) (projects 520330054, 460333672 CRC1540 EBM).

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we do not have conflicts of interest.

Data availability

The CCMCT dataset can be found at `https://doi.org/10.6084/m9.figshare.c.4552445.v1`. The MIDOG 2022 dataset can be found at `https://zenodo.org/records/6547151`. The code can be found at `https://github.com/DeepMicroscopy/FoundationModelComparison`.

References

Famke Aeffner, Kristin Wilson, Nathan T Martin, Joshua C Black, Cris L Luengo Hendriks, Brad Bolon, Daniel G Rudmann, Roberto Gianani, Sally R Koegler, Joseph Krueger, et al. The gold standard paradox in digital image analysis: manual versus automated scoring as ground truth. *Archives of pathology & laboratory medicine*, 141 (9):1267–1275, 2017.

Faruk Ahmed, Andrew Sellergren, Lin Yang, Shawn Xu, Boris Babenko, Abbi Ward, Niels Olson, Arash Mohtashamian, Yossi Matias, Greg S. Corrado, Quang Duong, Dale R. Webster, Shravya Shetty, Daniel Golden, Yun Liu, David F. Steiner, and Ellery Wulczyn. Pathalign: A vision-language model for whole slide images in histopathology. 6 2024. ISSN 26403498. URL `http://arxiv.org/abs/2406.19578`.

Maximilian Alber, Stephan Tietz, Jonas Dippel, Timo Milbich, Timothée Lesort, Panos Korfiatis, Moritz Krügener, Beatriz Perez Cancer, Neelay Shah, Alexander Möllers, Philipp Seegerer, Alexandra Carpen-Amarie, Kai Standvoss, Gabriel Dernbach, Edwin de Jong, Simon Schallenberg, Andreas Kunft, Helmut Hoffer von Ankershoffen, Gavin Schaeferle, Patrick Duffy, Matt Redlon, Philipp Jurmeister, David Horst, Lukas Ruff, Klaus-Robert Müller, Frederick Klauschen, and Andrew Norgan. Atlas: A novel pathology foundation model by mayo clinic, charité, and aignostics. 1 2025. URL `http://arxiv.org/abs/2501.05409`.

Jonas Ammeling, Jonathan Ganz, Frauke Wilm, Katharina Breininger, and Marc Aubreville. Investigation of class separability within object detection models in histopathology. *IEEE transactions on medical imaging*, 44:3162–3174, 2025. ISSN 1558-254X. . URL https://pubmed.ncbi.nlm.nih.gov/40232918/.

Marc Aubreville, Christof A. Bertram, Taryn A. Donovan, Christian Marzahl, Andreas Maier, and Robert Klopfleisch. A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research. *Scientific Data 2020 7:1*, 7:1–10, 11 2020a. ISSN 2052-4463. . URL https://www.nature.com/articles/s41597-020-00756-z.

Marc Aubreville, Christof A. Bertram, Christian Marzahl, Corinne Gurtner, Martina Dettwiler, Anja Schmidt, Florian Bartenschlager, Sophie Merz, Marco Fragoso, Olivia Kershaw, Robert Klopfleisch, and Andreas Maier. Deep learning algorithms out-perform veterinary pathologists in detecting the mitotically most active tumor region. *Scientific Reports*, 10, 12 2020b. ISSN 20452322. .

Marc Aubreville, Nikolas Stathonikos, Christof A. Bertram, Robert Klopfleisch, Natalie ter Hoeve, Francesco Ciompi, Frauke Wilm, Christian Marzahl, Taryn A. Donovan, Andreas Maier, Jack Breen, Nishant Ravikumar, Youjin Chung, Jinah Park, Ramin Nateghi, Fattaneh Pourakpour, Rutger H.J. Fick, Saima Ben Hadj, Mostafa Jahanifar, Adam Shephard, Jakob Dexl, Thomas Wittenberg, Satoshi Kondo, Maxime W. Lafarge, Viktor H. Koelzer, Jingtang Liang, Yubo Wang, Xi Long, Jingxin Liu, Salar Razavi, April Khademi, Sen Yang, Xiyue Wang, Ramona Erber, Andrea Klang, Karoline Lipnik, Pompei Bolfa, Michael J. Dark, Gabriel Wasinger, Mitko Veta, and Katharina Breininger. Mitosis domain generalization in histopathology images — the midog challenge. *Medical Image Analysis*, 84:102699, 2 2023a. ISSN 1361-8415. .

Marc Aubreville, Frauke Wilm, Nikolas Stathonikos, Katharina Breininger, Taryn A. Donovan, Samir Jabari, Mitko Veta, Jonathan Ganz, Jonas Ammeling, Paul J. van Diest, Robert Klopfleisch, and Christof A. Bertram. A comprehensive multi-domain dataset for mitotic figure detection. *Scientific Data*, 10, 12 2023b. ISSN 20524463. .

Marc Aubreville, Nikolas Stathonikos, Taryn A. Donovan, Robert Klopfleisch, Jonas Ammeling, Jonathan Ganz, Frauke Wilm, Mitko Veta, Samir Jabari, Markus Eckstein, Jonas Annuscheit, Christian Krumnow, Engin Bozaba, Sercan Çayır, Hongyan Gu, Xiang 'Anthony' Chen, Mostafa Jahanifar, Adam Shephard, Satoshi Kondo, Satoshi Kasai, Sujatha Kotte, V. G. Saipradeep, Maxime W. Lafarge, Viktor H. Koelzer, Ziyue Wang, Yongbing Zhang, Sen Yang, Xiyue Wang, Katharina Breininger, and Christof A. Bertram. Domain generalization across tumor types, laboratories, and species — insights from the 2022 edition of the mitosis domain generalization challenge. *Medical Image Analysis*, 94:103155, 5 2024. ISSN 1361-8415. . URL https://www.sciencedirect.com/science/article/pii/S136184152400080X.

Christof A. Bertram, Marc Aubreville, Christian Marzahl, Andreas Maier, and Robert Klopfleisch. A large-scale dataset for mitotic figure assessment on whole slide images of canine cutaneous mast cell tumor. *Sci. Data*, 6, 12 2019. ISSN 20524463. .

Jack Breen, Katie Allen, Kieran Zucker, Lucy Godson, Nicolas M. Orsi, and Nishant Ravikumar. A comprehensive evaluation of histopathology foundation models for ovarian cancer subtype classification. *npj Precision Oncology 2025 9:1*, 9:1–12, 1 2025. ISSN 2397-768X. . URL https://www.nature.com/articles/s41698-025-00799-8.

De Cai, Jie Chen, Junhan Zhao, Yuan Xue, Sen Yang, Wei Yuan, Min Feng, Haiyan Weng, Shuguang Liu, Yulong Peng, Junyou Zhu, Kanran Wang, Christopher Jackson, Hongping Tang, Junzhou Huang, and Xiyue Wang. Hicervix: An extensive hierarchical dataset and benchmark for cervical cytology classification. *IEEE Transactions on Medical Imaging*, 43:4344–4355, 2024. ISSN 1558254X. . URL https://pubmed.ncbi.nlm.nih.gov/38923481/.

Gabriele Campanella, Shengjia Chen, Manbir Singh, Ruchika Verma, Silke Muehlstedt, Jennifer Zeng, Aryeh Stock, Matt Croken, Brandon Veremis, Abdulkadir Elmas, Ivan Shujski, Noora Neittaanmäki, Kuan Lin Huang, Ricky Kwan, Jane Houldsworth, Adam J. Schoenfeld, and Chad Vanderbilt. A clinical benchmark of public self-supervised pathology foundation models. *Nature Communications*, 16:1–12, 12 2025. ISSN 20411723. . URL https://www.nature.com/articles/s41467-025-58796-1.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE International Conference on Computer Vision*, pages 9630–9640, 4 2021. ISSN 15505499. . URL https://arxiv.org/pdf/2104.14294.

Albi Celaj, Alice Jiexin Gao, Tammy T.Y. Lau, Erle M. Holgersen, Alston Lo, Varun Lodaya, Christopher B. Cole, Robert E. Denroche, Carl Spickett, Omar Wagih, Pedro O.

Pinheiro, Parth Vora, Pedrum Mohammadi-Shemirani, Steve Chan, Zach Nussbaum, Xi Zhang, Helen Zhu, Easwaran Ramamurthy, Bhargav Kanuparthi, Michael Iacocca, Diane Ly, Ken Kron, Marta Verby, Kahlin Cheung-Ong, Zvi Shalev, Brandon Vaz, Sakshi Bhargava, Farhan Yusuf, Sharon Samuel, Sabriyeh Alibai, Zahra Baghestani, Xinwen He, Kirsten Krastel, Oladipo Oladapo, Amrudha Mohan, Arathi Shanavas, Magdalena Bugno, Jovanka Bogojeski, Frank Schmitges, Carolyn Kim, Solomon Grant, Rachana Jayaraman, Tehmina Masud, Amit Deshwar, Shreshth Gandhi, and Brendan J. Frey. An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics. *bioRxiv*, page 2023.09.20.558508, 9 2023. . URL https://www.biorxiv.org/content/10.1101/2023.09.20.558508v1.

Charlie, Rodolphe Jenatton, Felipe Llinares-López, Zelda Mariet, David Cahané, Eric Durand, and Jean-Philippe Saillard Vert. H-optimus-0, 2024. URL https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0.

Richard J. Chen and Rahul G. Krishnan. Self-supervised vision transformers learn visual concepts in histopathology. 3 2022. URL https://arxiv.org/pdf/2203.00585.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H. Song, Muhammad Shaban, Mane Williams, Anurag Vaidya, Sharifa Sahai, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Walt Williams, Long Phi Le, Georg Gerber, and Faisal Mahmood. A general-purpose self-supervised model for computational pathology. 8 2023a. URL http://arxiv.org/abs/2308.15474.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F.K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30:850–862, 3 2024a. ISSN 1546170X. .

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *37th International Conference on Machine Learning, ICML 2020*, PartF168147-3:1575–1585, 2 2020a. URL https://arxiv.org/pdf/2002.05709.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, 2020-December, 6 2020b. ISSN 10495258. URL https://arxiv.org/pdf/2006.10029.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. 3 2020c. URL https://arxiv.org/pdf/2003.04297.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *Proceedings of the IEEE International Conference on Computer Vision*, pages 9620–9629, 4 2021. ISSN 15505499. . URL https://arxiv.org/pdf/2104.02057.

Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Bin Zhang, Nana Pei, Rongshan Yu, Yu Qiao, and Junjun He. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. 10 2024b. URL http://arxiv.org/abs/2410.11761.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, ‡ Antoine Bonnet, ‡ Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and † Antoine Bosselut. Meditron-70b: Scaling medical pretraining for large language models. 11 2023b. URL https://arxiv.org/pdf/2311.16079.

Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 3 2022. ISSN 2666-8270. . URL https://www.sciencedirect.com/science/article/pii/S2666827021000992.

Beilei Cui, Mobarakol Islam, Long Bai, and Hongliang Ren. Surgical-dino: adapter learning of foundation models for depth estimation in endoscopic surgery. *International Journal of Computer Assisted Radiology and Surgery*, 19:1013–1020, 6 2024. ISSN 18616429. .

Manon Dausort, Tiffanie Godelaine, Maxime Zanella, Karim El Khoury, Isabelle Salmon, and Benoît Macq. Exploring foundation models fine-tuning for cytology classification. 11 2024. URL https://arxiv.org/pdf/2411.14975.

Jia Deng, Wei Dong, Richard Socher, Li Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pages 248–255, 2009. .

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei.

Longnet: Scaling transformers to 1,000,000,000 tokens. 7 2023. URL https://arxiv.org/pdf/2307.02486.

Jonas Dippel, Barbara Feulner, Tobias Winterhoff, Timo Milbich, Stephan Tietz, Simon Schallenberg, Gabriel Dernbach, Andreas Kunft, Simon Heinke, Marie-Lisa Eich, Julika Ribbat-Idel, Rosemarie Krupar, Philipp Anders, Niklas Prenißl, Philipp Jurmeister, David Horst, Lukas Ruff, Klaus-Robert Müller, Frederick Klauschen, and Maximilian Alber. Rudolfv: A foundation model by pathologists for pathologists. 1 2024. URL http://arxiv.org/abs/2401.04079.

Taryn A. Donovan, Frances M. Moore, Christof A. Bertram, Richard Luong, Pompei Bolfa, Robert Klopfleisch, Harold Tvedten, Elisa N. Salas, Derick B. Whitley, Marc Aubreville, and Donald J. Meuten. Mitotic figures—normal, atypical, and imposters: A guide to identification. *Veterinary Pathology*, 58:243–257, 3 2021. ISSN 15442217. . URL https://journals.sagepub.com/doi/pdf/10.1177/0300985820980049.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR 2021 - 9th International Conference on Learning Representations*, 10 2020. URL https://arxiv.org/pdf/2010.11929.

C. W. Elston and I. O. Ellis. Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19:403–410, 1991. ISSN 13652559. . URL https://pubmed.ncbi.nlm.nih.gov/1757079/.

Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Axel Camara, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. 7 2023. . URL http://medrxiv.org/lookup/doi/10.1101/2023.07.21.23292757.

Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard Owkin. Phikon-v2, a large and public feature extractor for biomarker prediction. 9 2024. URL https://arxiv.org/pdf/2409.09173.

Alexandre Filiot, Nicolas Dop, Oussama Tchita, Auriane Riou, Rémy Dubois, Thomas Peeters, Daria Valter, Marin Scalbert, Charlie Saillard, Geneviève Robin, and Antoine Olivier. Distilling foundation models for robust and efficient models in digital pathology. 2 2025. URL https://arxiv.org/pdf/2501.16239v1.

Jonathan Ganz, Jonas Ammeling, Emely Rosbach, Ludwig Lausser, Christof A. Bertram, Katharina Breininger, and Marc Aubreville. Is self-supervision enough? *Informatik aktuell*, pages 63–68, 2025. ISSN 2628-8958. . URL https://link.springer.com/chapter/10.1007/978-3-658-47422-5_15.

Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, and Piotr Bojanowski. Self-supervised pretraining of visual features in the wild. 3 2021. URL https://arxiv.org/pdf/2103.01988.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December:770–778, 12 2015. ISSN 10636919. . URL https://arxiv.org/abs/1512.03385v1.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 9726–9735, 11 2019. ISSN 10636919. . URL https://arxiv.org/pdf/1911.05722.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June:15979–15988, 11 2021. ISSN 10636919. . URL https://arxiv.org/pdf/2111.06377.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 1979. .

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ICLR 2022 - 10th International Conference on Learning Representations*, 2022.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nat. Med.*, 29:2307–2316, 9 2023. ISSN 1546170X. .

Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2023-June:3344–3354, 12 2022.

ISSN 10636919. . URL https://arxiv.org/pdf/2212.04690.

Wasif Khan, Seowung Leem, Kyle B. See, Joshua K. Wong, Shaoting Zhang, and Ruogu Fang. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering*, 2024. .

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *Proceedings of the IEEE International Conference on Computer Vision*, pages 3992–4003, 4 2023. ISSN 15505499. . URL https://arxiv.org/pdf/2304.02643.

Jeaung Lee, Jeewoo Lim, Keunho Byeon, and Jin Tae Kwak. Benchmarking pathology foundation models: Adaptation strategies and scenarios. *Computers in Biology and Medicine*, 190:110031, 5 2025. ISSN 0010-4825. . URL http://arxiv.org/abs/2410.16038.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234–1240, 10 2019. . URL http://arxiv.org/abs/1901.08746http://dx.doi.org/10.1093/bioinformatics/btz682.

Florian Leiser, Simon Warsinsky, Marie Daum, Manuel Schmidt-Kraepelin, Scott Thiebes, Martin Wagner, and Ali Sunyaev. Understanding the role of expert intuition in medical image annotation: A cognitive task analysis approach. In *Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS 2023). Ed.: T.X. Bui*, pages 2850–2859, 2023. ISBN 978-0-9981331-6-4.

Chao Li, Xinggang Wang, Wenyu Liu, and Longin Jan Latecki. Deepmitosis: Mitosis detection via deep detection, verification and segmentation networks. *Medical image analysis*, 45:121–133, 2018.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE International Conference on Computer Vision*, pages 9992–10002, 3 2021. ISSN 15505499. . URL https://arxiv.org/pdf/2103.14030.

Ming Y. Lu, Bowen Chen, Drew F.K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 3 2024. ISSN 1546170X. .

Roux Ludovic, Racoceanu Daniel, Loménie Nicolas, Kulikova Maria, Irshad Humayun, Klossa Jacques, Capron Frédérique, Genestie Catherine, Le Naour Gilles, and Gurcan Metin N. Mitosis detection in breast cancer histological images an icpr 2012 contest. *Journal of Pathology Informatics*, 4:8, 1 2013. ISSN 21533539. . URL https://pubmed.ncbi.nlm.nih.gov/23858383/.

Jiabo Ma, Yingxue Xu, Fengtao Zhou, Yihui Wang, Jin Cheng, Zhengrui Guo, Jianfeng Wu, On Ki Tang, Huajun Zhou, Xi Wang, Luyang Luo, Zhengyu Zhang, Du Cai, Zizhao Gao, Wei Wang, Yueping Liu, Jiankun He, Jing Cui, Zhenhui Li, Jing Zhang, Feng Gao, Xiuming Zhang, Li Liang, Ronald Cheong, Kin Chan, Zhe Wang, and Hao Chen. Pathbench: A comprehensive comparison benchmark for pathology foundation models towards precision oncology. 5 2025. URL https://arxiv.org/pdf/2505.20202.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications 2024 15:1*, 15:1–9, 1 2024. ISSN 2041-1723. . URL https://www.nature.com/articles/s41467-024-44824-z.

Peter Neidlinger, Omar S. M. El Nahhas, Hannah Sophie Muti, Tim Lenz, Michael Hoffmeister, Hermann Brenner, Marko van Treeck, Rupert Langer, Bastian Dislich, Hans Michael Behrens, Christoph Röcken, Sebastian Foersch, Daniel Truhn, Antonio Marra, Oliver Lester Saldanha, and Jakob Nikolas Kather. Benchmarking foundation models as feature extractors for weakly-supervised computational pathology. *Christoph Röcken*, 6:12, 8 2024. URL https://arxiv.org/pdf/2408.15823.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. 4 2023. ISSN 28358856. URL https://arxiv.org/pdf/2304.07193.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. *7th International Conference on Learning Representations, ICLR 2019*, 6 2018. URL https://arxiv.org/pdf/1806.03198.

Zhuoyan Shen, Mikaël Simard, Douglas Brand, Vanghelita Andrei, Ali Al-Khader, Fatine Oumlil, Katherine Trevers, Thomas Butters, Simon Haefliger, Eleanna Kara, Fernanda Amary, Roberto Tirabosco, Paul Cool, Gary Royle, Maria A. Hawkins, Adrienne M. Flanagan, and Charles Antoine Collins-Fekete. A deep learning framework deploying segment anything to detect pan-cancer mitotic figures from haematoxylin and eosin-stained slides. *Communications Biology 2024 7:1*, 7:1–11, 12 2024. ISSN 2399-3642. . URL https://www.nature.com/articles/s42003-024-07398-6.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620:172–180, 8 2023. ISSN 14764687. . URL https://www.nature.com/articles/s41586-023-06291-2.

Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. arxiv. *arXiv preprint arXiv:1708.07120*, 6, 2017.

Alexander Smits, Ja Kummer, Pc de Bruin, Mijke Bol, jg van den tweel, Kees Seldenrijk, Stefan Willems, George Offerhaus, Roel Weger, Paul Diest, and Aryan Vink. The estimation of tumor cell percentage for molecular testing by pathologists is not accurate. *Modern pathology*, 27 (2):168–174, 07 2014. .

Nadia Stec, Danielle Arje, Alan R Moody, Elizabeth A Krupinski, and Pascal N Tyrrell. A systematic review of fatigue in radiology: is it a problem? *American Journal of Roentgenology*, 210(4):799–806, 2018.

Nenad Tomasev, Ioana Bica, Brian McWilliams, Lars Buesing, Razvan Pascanu, Charles Blundell, and Jovana Mitrovic. Pushing the limits of self-supervised resnets: Can we outperform supervised learning without labels on imagenet? 1 2022. URL https://arxiv.org/pdf/2201.05119.

Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015:68–77, 2015. ISSN 1428-2526. . URL https://tcga-data.nci.nih.gov/datareports/codeTablesReport.

Mitko Veta, Yujing J. Heng, Nikolas Stathonikos, Babak Ehteshami Bejnordi, Francisco Beca, Thomas Wollmann, Karl Rohr, Manan A. Shah, Dayong Wang, Mikael Rousson, Martin Hedlund, David Tellez, Francesco Ciompi, Erwan Zerhouni, David Lanyi, Matheus Viana, Vassili Kovalev, Vitali Liauchuk, Hady Ahmady Phoulady, Talha Qaiser, Simon Graham, Nasir Rajpoot, Erik Sjöblom, Jesper Molin, Kyunghyun Paeng, Sangheum Hwang, Sunggyun Park, Zhipeng Jia, Eric I.Chao Chang, Yan Xu, Andrew H. Beck, Paul J. van Diest, and Josien P.W. Pluim. Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. *Medical Image Analysis*, 54:111–121, 5 2019. ISSN 1361-8415. . URL https://www.sciencedirect.com/science/article/pii/S1361841518305231.

Hollis Viray, Kevin Li, Thomas A Long, Patricia Vasalos, Julia A Bridge, Lawrence J Jennings, Kevin C Halling, Meera Hameed, and David L Rimm. A prospective, multi-institutional diagnostic trial to determine pathologist accuracy in estimation of percentage of malignant cells. *Archives of Pathology and Laboratory Medicine*, 137(11): 1545–1549, 2013.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Juan Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David Klimstra, Brandon Rothrock, and Thomas J. Fuchs. Virchow: A million-slide digital pathology foundation model. 9 2023. URL http://arxiv.org/abs/2309.07778.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 10 2022a. ISSN 1361-8415.

. URL https://www.sciencedirect.com/science/article/pii/S1361841522002043.

Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, Fang Wang, Yulong Peng, Junyou Zhu, Jing Zhang, Christopher R. Jackson, Jun Zhang, Deborah Dillon, Nancy U. Lin, Lynette Sholl, Thomas Denize, David Meredith, Keith L. Ligon, Sabina Signoretti, Shuji Ogino, Jeffrey A. Golden, MacLean P. Nasrallah, Xiao Han, Sen Yang, and Kun Hsing Yu. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature 2024 634:8035*, 634:970–978, 9 2024. ISSN 1476-4687. . URL https://www.nature.com/articles/s41586-024-07894-z.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 3876–3887, 10 2022b. . URL https://arxiv.org/pdf/2210.10163.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80, 12 1945. ISSN 00994987. .

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630:181–188, 6 2024. ISSN 14764687. .

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4:852–866, 10 2022a. ISSN 25225839. . URL https://www.nature.com/articles/s42256-022-00534-z.

Menglin Yang, Jialin Chen, Yifei Zhang, Jiahong Liu, Jiasheng Zhang, Qiyao Ma, Harshit Verma, Qianru Zhang, Min Zhou, Irwin King, and Rex Ying. Low-rank adaptation for foundation models: A comprehensive review. 12 2024. URL https://arxiv.org/pdf/2501.00365.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. A large language model for electronic health records. *npj Digital Medicine*, 5:1–9, 12 2022b. ISSN 23986352. . URL https://www.nature.com/articles/s41746-022-00742-2.

Bowen Zhang, Ying Chen, Long Bai, Yan Zhao, Yuxiang Sun, Yixuan Yuan, Jian hua Zhang, and Hongliang Ren. Learning to adapt foundation model dinov2 for capsule endoscopy diagnosis. *Procedia Computer Science*, 2024. .

Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *Medical Image Analysis*, 91:102996, 1 2024. ISSN 1361-8415. . URL https://www.sciencedirect.com/science/article/pii/S1361841523002566.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, Curtis P Langlotz, Y Zhang, H Jiang, Y Miura, C D Manning, and C P Langlotz. Contrastive learning of medical visual representations from paired images and text. *Proceedings of Machine Learning Research*, 182:2–25, 10 2020. ISSN 26403498. URL https://arxiv.org/pdf/2010.00747.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *ICLR 2022 - 10th International Conference on Learning Representations*, 11 2021. URL https://arxiv.org/pdf/2111.07832.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V. Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *12th International Conference on Learning Representations, ICLR 2024*, 6 2023. URL https://arxiv.org/pdf/2306.15006.

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, Thomas Fuchs, Nicoì O Fusi, Siqi Liu, and Kristen Severson. Virchow2: Scaling self-supervised mixed magnification models in pathology. 8 2024. URL https://arxiv.org/pdf/2408.00738.

Appendix A. Additional Results from Dataset Scaling Experiments

Table 6: Results at 0.1% dataset size of CCMCT dataset.

Model	AUROC	Balanced ACC	Weighted F1
ResNet50 (End-To-End)	0.69±0.03	0.63±0.03	0.59±0.07
ViT-S (End-To-End)	0.68±0.03	0.62±0.03	0.61±0.07
ViT-B (End-To-End)	0.61±0.02	0.58±0.02	0.56±0.07
ViT-H (End-To-End)	0.62±0.03	0.58±0.02	0.57±0.02
ResNet50 (LinProb)	0.57±0.04	0.54±0.02	0.58±0.02
ViT-H (LinProb)	0.57±0.03	0.55±0.02	0.58±0.03
ViT-S DINOv3 (LinProb)	0.58±0.04	0.55±0.03	0.58±0.03
Phikon (LinProb)	0.61±0.03	0.58±0.03	0.61±0.03
UNI (LinProb)	0.64±0.02	0.60±0.02	0.62±0.02
Virchow (LinProb)	0.62±0.03	0.58±0.02	0.61±0.02
Virchow2 (LinProb)	0.62±0.03	0.58±0.02	0.61±0.03
Prov-GigaPath (LinProb)	0.65±0.02	0.61±0.02	0.63±0.01
H-optimus-0 (LinProb)	0.63±0.03	0.60±0.02	0.62±0.03
ViT-S DINOv3 (LoRA)	0.60±0.02	0.57±0.02	0.54±0.05
Phikon (LoRA)	0.72±0.04	0.65±0.04	0.65±0.07
UNI (LoRA)	0.70±0.04	0.64±0.03	0.60±0.08
Virchow (LoRA)	0.73±0.06	0.66±0.05	0.66±0.09
Virchow2 (LoRA)	0.75±0.03	0.69±0.03	0.69±0.03
Prov-GigaPath (LoRA)	0.74±0.03	0.68±0.03	0.66±0.05
H-optimus-0 (LoRA)	0.71±0.04	0.65±0.03	0.65±0.05

Table 7: Results at 1% dataset size of CCMCT dataset.

Model	AUROC	Balanced ACC	Weighted F1
ResNet50 (End-To-End)	0.75±0.02	0.68±0.02	0.70±0.02
ViT-S (End-To-End)	0.74±0.01	0.67±0.01	0.67±0.03
ViT-B (End-To-End)	0.67±0.06	0.62±0.05	0.60±0.07
ViT-H (End-To-End)	0.68±0.03	0.62±0.03	0.61±0.03
ResNet50 (LinProb)	0.58±0.01	0.55±0.01	0.58±0.01
ViT-H (LinProb)	0.65±0.01	0.60±0.01	0.63±0.01
ViT-S DINOv3 (LinProb)	0.65±0.01	0.59±0.01	0.62±0.01
Phikon (LinProb)	0.67±0.01	0.62±0.01	0.64±0.01
UNI (LinProb)	0.69±0.02	0.64±0.01	0.66±0.01
Virchow (LinProb)	0.70±0.01	0.64±0.01	0.66±0.01
Virchow2 (LinProb)	0.69±0.01	0.64±0.01	0.66±0.01
Prov-GigaPath (LinProb)	0.71±0.01	0.65±0.01	0.67±0.01
H-optimus-0 (LinProb)	0.70±0.01	0.64±0.01	0.67±0.01
ViT-S DINOv3 (LoRA)	0.64±0.07	0.59±0.06	0.55±0.13
Phikon (LoRA)	0.79±0.07	0.72±0.05	0.73±0.04
UNI (LoRA)	0.76±0.07	0.69±0.06	0.70±0.05
Virchow (LoRA)	0.80±0.08	0.71±0.07	0.72±0.10
Virchow2 (LoRA)	0.82±0.01	0.75±0.01	0.76±0.00
Prov-GigaPath (LoRA)	0.82±0.00	0.74±0.02	0.75±0.02
H-optimus-0 (LoRA)	0.80±0.07	0.73±0.06	0.74±0.06

Table 8: Results at 10% dataset size of CCMCT dataset.

Model	AUROC	Balanced ACC	Weighted F1
ResNet50 (End-To-End)	0.84±0.02	0.75±0.02	0.75±0.03
ViT-S (End-To-End)	0.81±0.02	0.70±0.03	0.66±0.07
ViT-B (End-To-End)	0.71±0.09	0.63±0.07	0.61±0.04
ViT-H (End-To-End)	0.75±0.04	0.67±0.06	0.62±0.13
ResNet50 (LinProb)	0.67±0.00	0.61±0.00	0.64±0.00
ViT-H (LinProb)	0.71±0.00	0.62±0.01	0.65±0.01
ViT-S DINOv3 (LinProb)	0.68±0.00	0.57±0.01	0.60±0.01
Phikon (LinProb)	0.78±0.00	0.69±0.00	0.72±0.00
UNI (LinProb)	0.78±0.00	0.70±0.00	0.72±0.00
Virchow (LinProb)	0.79±0.00	0.71±0.00	0.73±0.00
Virchow2 (LinProb)	0.78±0.00	0.70±0.00	0.72±0.00
Prov-GigaPath (LinProb)	0.78±0.01	0.70±0.00	0.73±0.00
H-optimus-0 (LinProb)	0.79±0.00	0.71±0.01	0.74±0.01
ViT-S DINOv3 (LoRA)	0.74±0.12	0.67±0.09	0.67±0.11
Phikon (LoRA)	0.77±0.08	0.70±0.07	0.69±0.08
UNI (LoRA)	0.82±0.06	0.74±0.05	0.74±0.06
Virchow (LoRA)	0.84±0.06	0.76±0.05	0.76±0.05
Virchow2 (LoRA)	0.85±0.03	0.76±0.03	0.78±0.02
Prov-GigaPath (LoRA)	0.86±0.01	0.77±0.01	0.78±0.01
H-optimus-0 (LoRA)	0.87±0.02	0.78±0.03	0.79±0.02

Table 9: Results at 100% dataset size of CCMCT dataset.

Model	AUROC	Balanced ACC	Weighted F1
ResNet50 (End-To-End)	0.84±0.02	0.75±0.02	0.75±0.03
ViT-S (End-To-End)	0.81±0.02	0.70±0.03	0.66±0.07
ViT-B (End-To-End)	0.71±0.09	0.63±0.07	0.61±0.04
ViT-H (End-To-End)	0.75±0.04	0.67±0.06	0.62±0.13
ResNet50 (LinProb)	0.67±0.00	0.61±0.00	0.64±0.00
ViT-H (LinProb)	0.71±0.00	0.62±0.01	0.65±0.01
ViT-S DINOv3 (LinProb)	0.68±0.00	0.57±0.01	0.60±0.01
Phikon (LinProb)	0.78±0.00	0.69±0.00	0.72±0.00
UNI (LinProb)	0.78±0.00	0.70±0.00	0.72±0.00
Virchow (LinProb)	0.79±0.00	0.71±0.00	0.73±0.00
Virchow2 (LinProb)	0.78±0.00	0.70±0.00	0.72±0.00
Prov-GigaPath (LinProb)	0.78±0.01	0.70±0.00	0.73±0.00
H-optimus-0 (LinProb)	0.79±0.00	0.71±0.01	0.74±0.01
ViT-S DINOv3 (LoRA)	0.74±0.12	0.67±0.09	0.67±0.11
Phikon (LoRA)	0.77±0.08	0.70±0.07	0.69±0.08
UNI (LoRA)	0.82±0.06	0.74±0.05	0.74±0.06
Virchow (LoRA)	0.84±0.06	0.76±0.05	0.76±0.05
Virchow2 (LoRA)	0.85±0.03	0.76±0.03	0.78±0.02
Prov-GigaPath (LoRA)	0.86±0.01	0.77±0.01	0.78±0.01
H-optimus-0 (LoRA)	0.87±0.02	0.78±0.03	0.79±0.02

Table 10: Results at 0.1% dataset size of MIDOG dataset.

Model	AUROC	Balanced ACC	Weighted F1
ResNet50 (End-To-End)	0.69±0.03	0.63±0.03	0.59±0.07
ViT-S (End-To-End)	0.68±0.03	0.62±0.03	0.61±0.07
ViT-B (End-To-End)	0.61±0.02	0.58±0.02	0.56±0.07
ViT-H (End-To-End)	0.59±0.07	0.56±0.04	0.51±0.09
ResNet50 (LinProb)	0.57±0.04	0.54±0.02	0.58±0.02
ViT-H (LinProb)	0.57±0.03	0.55±0.02	0.58±0.03
ViT-S DINOv3 (LinProb)	0.58±0.04	0.55±0.03	0.58±0.03
Phikon (LinProb)	0.61±0.03	0.58±0.03	0.61±0.03
UNI (LinProb)	0.64±0.02	0.60±0.02	0.62±0.02
Virchow (LinProb)	0.62±0.03	0.58±0.02	0.61±0.02
Virchow2 (LinProb)	0.62±0.03	0.58±0.02	0.61±0.03
Prov-GigaPath (LinProb)	0.65±0.02	0.61±0.02	0.63±0.01
H-optimus-0 (LinProb)	0.63±0.03	0.60±0.02	0.62±0.03
ViT-S DINOv3 (LoRA)	0.60±0.02	0.57±0.02	0.54±0.05
Phikon (LoRA)	0.64±0.11	0.58±0.08	0.54±0.13
UNI (LoRA)	0.63±0.08	0.57±0.07	0.50±0.12
Virchow (LoRA)	0.67±0.14	0.60±0.08	0.56±0.14
Virchow2 (LoRA)	0.67±0.11	0.60±0.10	0.55±0.16
Prov-GigaPath (LoRA)	0.65±0.12	0.59±0.09	0.52±0.15
H-optimus-0 (LoRA)	0.65±0.12	0.58±0.08	0.52±0.14

Table 11: Results at 1% dataset size of MIDOG dataset.

Model	AUROC	Balanced ACC	Weighted F1
ResNet50 (End-To-End)	0.75±0.02	0.68±0.02	0.70±0.02
ViT-S (End-To-End)	0.74±0.01	0.67±0.01	0.67±0.03
ViT-B (End-To-End)	0.67±0.06	0.62±0.05	0.60±0.07
ViT-H (End-To-End)	0.70±0.03	0.63±0.03	0.61±0.04
ResNet50 (LinProb)	0.58±0.01	0.55±0.01	0.58±0.01
ViT-H (LinProb)	0.65±0.01	0.60±0.01	0.63±0.01
ViT-S DINOv3 (LinProb)	0.65±0.01	0.59±0.01	0.62±0.01
Phikon (LinProb)	0.67±0.01	0.62±0.01	0.64±0.01
UNI (LinProb)	0.69±0.02	0.64±0.01	0.66±0.01
Virchow (LinProb)	0.70±0.01	0.64±0.01	0.66±0.01
Virchow2 (LinProb)	0.69±0.01	0.64±0.01	0.66±0.01
Prov-GigaPath (LinProb)	0.71±0.01	0.65±0.01	0.67±0.01
H-optimus-0 (LinProb)	0.70±0.01	0.64±0.01	0.67±0.01
ViT-S DINOv3 (LoRA)	0.64±0.07	0.59±0.06	0.55±0.13
Phikon (LoRA)	0.75±0.10	0.67±0.10	0.66±0.13
UNI (LoRA)	0.75±0.08	0.66±0.08	0.65±0.11
Virchow (LoRA)	0.81±0.08	0.72±0.09	0.72±0.11
Virchow2 (LoRA)	0.82±0.05	0.74±0.08	0.73±0.11
Prov-GigaPath (LoRA)	0.80±0.06	0.70±0.08	0.70±0.11
H-optimus-0 (LoRA)	0.82±0.07	0.73±0.08	0.73±0.10

Table 12: Results at 10% dataset size of MIDOG dataset.

Model	AUROC	Balanced ACC	Weighted F1
ResNet50 (End-To-End)	0.84±0.02	0.75±0.02	0.75±0.03
ViT-S (End-To-End)	0.81±0.02	0.70±0.03	0.66±0.07
ViT-B (End-To-End)	0.71±0.09	0.63±0.07	0.61±0.04
ViT-H (End-To-End)	0.76±0.03	0.67±0.05	0.65±0.09
ResNet50 (LinProb)	0.67±0.00	0.61±0.00	0.64±0.00
ViT-H (LinProb)	0.71±0.00	0.62±0.01	0.65±0.01
ViT-S DINOv3 (LinProb)	0.68±0.00	0.57±0.01	0.60±0.01
Phikon (LinProb)	0.78±0.00	0.69±0.00	0.72±0.00
UNI (LinProb)	0.78±0.00	0.70±0.00	0.72±0.00
Virchow (LinProb)	0.79±0.00	0.71±0.00	0.73±0.00
Virchow2 (LinProb)	0.78±0.00	0.70±0.00	0.72±0.00
Prov-GigaPath (LinProb)	0.78±0.01	0.70±0.00	0.73±0.00
H-optimus-0 (LinProb)	0.79±0.00	0.71±0.01	0.74±0.01
ViT-S DINOv3 (LoRA)	0.74±0.12	0.67±0.09	0.67±0.11
Phikon (LoRA)	0.79±0.10	0.72±0.09	0.71±0.09
UNI (LoRA)	0.84±0.05	0.76±0.04	0.76±0.04
Virchow (LoRA)	0.81±0.11	0.73±0.10	0.74±0.11
Virchow2 (LoRA)	0.87±0.03	0.79±0.03	0.80±0.02
Prov-GigaPath (LoRA)	0.83±0.08	0.75±0.07	0.76±0.07
H-optimus-0 (LoRA)	0.88±0.02	0.80±0.03	0.81±0.02

Table 13: Results at 100% dataset size of MIDOG dataset.

Model	AUROC	Balanced ACC	Weighted F1
ResNet50 (End-To-End)	0.87±0.01	0.79±0.01	0.78±0.01
ViT-S (End-To-End)	0.82±0.01	0.73±0.01	0.72±0.03
ViT-B (End-To-End)	0.79±0.01	0.70±0.02	0.71±0.03
ViT-H (End-To-End)	0.80±0.04	0.70±0.05	0.70±0.09
ResNet50 (LinProb)	0.69±0.00	0.61±0.00	0.65±0.00
ViT-H (LinProb)	0.71±0.00	0.61±0.00	0.65±0.00
ViT-S DINOv3 (LinProb)	0.68±0.00	0.57±0.00	0.60±0.00
Phikon (LinProb)	0.81±0.00	0.71±0.00	0.74±0.00
UNI (LinProb)	0.83±0.00	0.73±0.00	0.76±0.00
Virchow (LinProb)	0.84±0.00	0.75±0.00	0.78±0.00
Virchow2 (LinProb)	0.85±0.00	0.76±0.00	0.78±0.00
Prov-GigaPath (LinProb)	0.83±0.00	0.74±0.00	0.77±0.00
H-optimus-0 (LinProb)	0.84±0.00	0.75±0.00	0.78±0.00
ViT-S DINOv3 (LoRA)	0.78±0.12	0.70±0.09	0.71±0.09
Phikon (LoRA)	0.80±0.13	0.73±0.10	0.73±0.11
UNI (LoRA)	0.84±0.09	0.76±0.07	0.76±0.07
Virchow (LoRA)	0.87±0.07	0.78±0.06	0.79±0.05
Virchow2 (LoRA)	0.89±0.01	0.80±0.02	0.81±0.01
Prov-GigaPath (LoRA)	0.87±0.05	0.79±0.04	0.79±0.04
H-optimus-0 (LoRA)	0.90±0.02	0.81±0.02	0.81±0.02