RudolfV 2: A Family of Robust and Efficient Open-Weights Pathology Foundation Models

Timo Milbich[*1], Elias Eulig[*1], Alexandra Carpen-Amarie[*1],
Jonas Dippel[*1], Lukas Muttenthaler[*178], Stephan Tietz[*1],
Beatriz Perez Cancer[1], Jérôme Lüscher[1], Alessandro Benetti[1],
Sayed Abid Hashimi[1], Neelay Shah[1], Moritz Krügener[1], Philipp Jurmeister[29], David Horst[39],
Andrew Norgan[4], Simon Schallenberg[3], Lukas Ruff[1],
Klaus-Robert Müller[†5610 11], Frederick Klauschen[†2369 12], Maximilian Alber[†1]

[1] Aignostics, Germany
[2] Institute of Pathology, Ludwig-Maximilians-Universität München, Germany
[3] Institute of Pathology, Charité – Universitätsmedizin Berlin, Germany
[4] Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, US
[5] Machine Learning Group, Technische Universität Berlin, Germany
[6] BIFOLD – Berlin Institute for the Foundations of Learning and Data, Germany
[7] Helmholtz Munich, Germany
[8] Technical University Munich, Germany
[9] German Cancer Research Center (DKFZ) & German Cancer Consortium (DKTK),
Berlin & Munich Partner Sites, Germany
[10] Department of Artificial Intelligence, Korea University, Republic of Korea
[11] Max-Planck Institute for Informatics, Germany
[12] Bavarian Cancer Research Center (BZKF), Germany
∗ Equal contribution
† Corresponding author

Abstract

Pathology foundation models have markedly advanced computational pathology. Yet the most capable models are large, expensive to run, or released with closed weights, while openly available alternatives have trailed them in both accuracy and robustness. Here we introduce RudolfV 2, RudolfV 2-B, and RudolfV 2-S, a family of three open-weights pathology vision foundation models that narrow this gap. In a comprehensive evaluation across representative public benchmarks, RudolfV 2 is the best-performing and most robust openly available model, approaching the performance of our closed flagship Atlas 2. Its distilled ViT-B and ViT-S variants match far larger models at a fraction of the inference cost while remaining notably robust for their size. We release all three models openly for academic research.[1]

1 Introduction

The large-scale digitization of pathology, alongside virtual microscopy and artificial intelligence (AI)-driven analysis of histological images, has begun to reshape anatomic pathology, producing a growing body of encouraging proof-of-concept studies and early clinical applications [e.g., 8, 17, 45, 68, 46]. Progress, however, has been held back by models that struggle to generalize and stay robust across settings. This is compounded by the limited availability of training data, especially for rare entities [20], and by how costly and impractical it is to assemble annotations that span the full

[1] RudolfV 2 is available at: `https://huggingface.co/collections/Aignostics/rudolfv-2`

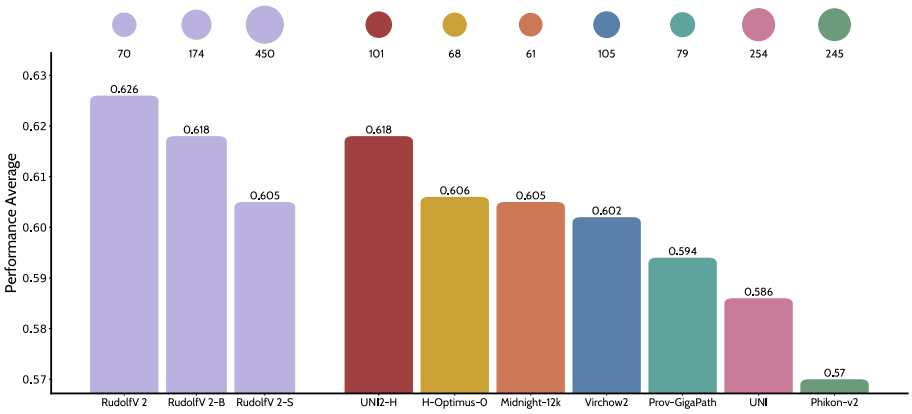

Figure 1: RudolfV 2 is the best-performing open-weights model. RudolfV 2-B and 2-S achieve prediction performance similar to competing models but are up to an order of magnitude more efficient. The performance average is based on morphology prediction tasks from *eva* [29] and molecular prediction tasks from HEST [35] in Table 1. The dots show the processing speed of each model for a 224×224 pixel image on an L4 GPU.

range of disease manifestations along with the biological and technical variation introduced by tissue handling, staining, and scanner hardware. Pathology foundation models have arisen as a response to these obstacles, offering the prospect of transferable and resilient representations that absorb such heterogeneity through self-supervised pretraining at scale [e.g., 14, 89, 86, 26, 19, 70, 3, 2]. For instance, foundation models underpin applications such as Atlas H&E-TME [75], which profiles the tumor microenvironment (TME) from routine H&E at expert pathologist-level accuracy, or OpenTME [27], a derived open dataset of cell-level profiles across thousands of TCGA slides—use cases where robustness to technical variation and efficient inference are decisive for real-world impact.

Many of the strongest models, however, are distributed with closed weights, or exposed through proprietary interfaces [e.g., 3, 9, 64, 2]—a natural choice for commercial and clinical use, but one that gives the academic community limited room to access, use, and build on them independently. Openly released models, in turn, have trailed their closed counterparts in both accuracy and robustness. A second and largely separate hurdle is computational: the leading models are expensive to run, a concern that is especially acute in pathology, where each whole slide image (WSI) decomposes into thousands of tiles and inference cost accumulates rapidly at scale. Here, we address these limitations and introduce an open-weights successor to our earlier pathology foundation model RudolfV [19].

Building on a refined training pipeline and an enlarged training corpus, we developed "RudolfV 2", an open-weights model that attains the strongest downstream performance of any openly available pathology foundation model and approaches our proprietary flagship Atlas 2 [2]. RudolfV 2 is at the same time the most robust open model to date, matching Atlas 2 in its resilience to the staining, scanner, and medical-center variation that routinely degrades foundation model representations—thereby narrowing the accuracy–robustness gap that has separated open models from their closed counterparts.

We distill [33, 7, 25] this flagship into two compact variants, "RudolfV 2-B" and "RudolfV 2-S", built on the leaner ViT-B and ViT-S backbones. These variants process a standard 224×224 image 2.5 and 6.5 times faster, respectively, and rank as the most capable openly available small models released so far, rivaling far larger networks at a small fraction of their inference cost. Notably, this efficiency comes without sacrificing reliability: both lightweight models are markedly more robust than other open models of comparable size, extending RudolfV 2's robustness advantage into the low-compute regime. Each model is trained across a wide spectrum of diseases, stain types, and scanners, and draws on several image magnifications. We release the weights of all three models openly for academic research, and evaluate them against other leading models on a broad range of downstream pathology applications, including the prediction of morphologic and molecular features. Figure 1 and Figure 2 summarize the model performance characteristics and key results.

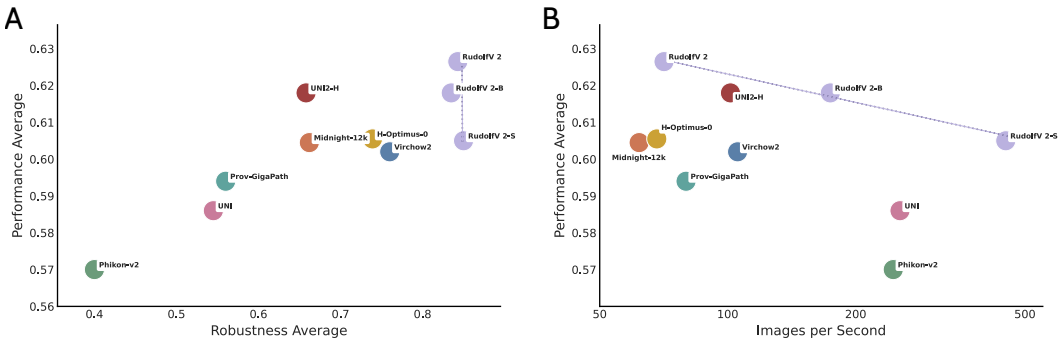

Figure 2: **(A)** RudolfV 2, RudolfV 2-B, and RudolfV 2-S are the most robust open-weights models and define the state-of-the-art performance–robustness Pareto front. The plot shows the performance average and the robustness average from Table 1. **(B)** The RudolfV 2 models define the performance–efficiency Pareto front.

2 Related Work

Pathology foundation models. In recent years, pathology foundation models [19, 3, 2, 89, 26, 86, 70, 14, 18, 64, 51, 36, 55, 67] established themselves as the basis for developing clinical-grade applications [84] in digital pathology. These models fall into two categories: tile-based and slide-based. Tile-based models [19, 3, 2, 89, 14, 9, 64, 40, 39] are trained via self-supervised learning frameworks such as DINOv2 [62] or DINOv3 [73] to encode small, fixed-sized image tiles into compact embeddings. These embeddings then typically serve as a starting point to train slide-based pathology foundation models [55, 72, 84], which learn to aggregate the per-tile information into a single representation for an entire WSI using slide-level supervision such as patient report- [55, 83, 72, 78] or molecular [80, 87] data. Key drivers behind the progress of pathology foundation models are increasing both the number of WSIs used for training [9, 89, 3] and model capacity, primarily through larger encoders [89, 9, 39]. In contrast to this prevailing focus on scale, the model presented here, RudolfV 2, attains the strongest downstream performance among openly available pathology foundation models—trained on 300k WSIs using a 1.1 billion-parameter Vision Transformer [21]—and is released together with two distilled, resource-efficient variants, RudolfV 2-B and RudolfV 2-S, based on ViT-B and ViT-S backbones.

Robustness of pathology foundation models. While pathology foundation models are trained on increasingly large and diverse pretraining datasets, their lack of robustness to scanner differences, laboratory processing artifacts, and staining variations remains problematic [47, 24, 16, 48, 31, 53, 12, 13, 58]. This can lead to a substantial decrease in performance when prediction targets are correlated with such confounding information in the representations [50, 44, 47]. Multiple investigations have analyzed these shortcomings and proposed benchmarks to quantify this robustness [24, 47]. So far, only limited work has focused on improving the robustness of pathology foundation models. Examples include stain normalization [56, 69], data augmentation [22], projecting out confounder information [60, 47], distillation [24], or applying methods during downstream training (e.g., DANN [28, 47]).

Efficient pathology foundation models. To provide efficient alternatives to large-scale frontier pathology foundation models, several groups [89, 64, 25, 2] have released lightweight counterparts, either by training small encoders from scratch [64, 38, 59, 1] or by performing knowledge distillation [33, 7] from larger foundation models [25, 89, 2].

3 Data and Methods

3.1 Dataset and Preprocessing

We curate a dataset of more more than 300k de-identified WSIs, derived from the digital archives of Charité - Universitätsmedizin Berlin and LMU Munich, from which we extract tiles at multiple resolutions, namely 0.25, 0.5, 1.0, and 2.0 microns per pixel, corresponding to objective microscopic magnifications of $40\times$, $20\times$, $10\times$, and $5\times$, respectively.

3.2 Model Framework and Compute Environment

We perform model training in several phases, including distillation.

Pretraining and model distillation. All of our model architectures are based on the Vision Transformer (ViT) [21], using a patch-token size of 8. Our main model, RudolfV 2, is based on the ViT-g architecture with 1.1 billion parameters. For the distilled versions, RudolfV 2-B and RudolfV 2-S, we use the model sizes "base" (ViT-B; 86 million parameters) and "small" (ViT-S; 22 million parameters), respectively. Our RudolfV 2 training framework builds on an extended RudolfV [19] and Atlas [3] approach, with parts based on DINOv2 [62] and selective improvements from DINOv3 [73]. For model training, we use NVIDIA A100 GPUs.

Posttraining. A novel component of our framework is a dedicated posttraining stage, which we adapt from paradigms established in other domains. Broadly, posttraining refers to a refinement stage applied after pretraining and prior to downstream task adaptation; its concrete objective, however, varies by domain, ranging from behavioral and preference alignment in large language models [e.g., 15, 71, 76, 63] to representational alignment in computer vision [e.g., 52, 54, 88, 77, 57]. To our knowledge, this represents the first introduction of a dedicated posttraining stage to improve pathology foundation models.

3.3 Evaluation Protocols

For reproducibility and comparability, we use publicly available frameworks for evaluation: HEST [35], *eva* [29], PathoROB [47], and Plismbench [24], covering tile-level as well as slide-level tasks. To further extend the pool of tasks, we additionally report a number of results for public benchmarks based on an in-house evaluation framework. To evaluate the quality of the foundation model embeddings, we keep the model encoders frozen and solely apply task-specific prediction heads. Unless stated otherwise, we report results based on a concatenation of the CLS token and the mean over patch tokens ("CLS+MEAN"). Input images are normalized using each model's specified statistics. We do not apply normalization to the output tokens unless the evaluation framework specifically does so.

We evaluate five foundation models using their publicly available weights (UNI2-H [14], H-Optimus-0 [70], Midnight-12k [39], Virchow2 [89] and Phikon-v2 [26]) and three additional foundation models (Atlas 2 [2], Pluto-4G [64] and H0-mini [24]) whose weights are either not publicly released or not accessible to us. If not accessible, we include comparisons based on published results.

HEST The HEST-Benchmark [35] is composed of tasks for gene expression prediction and designed as multivariate regression. We use the default evaluation protocol and implementation [32]. The protocol leverages a Ridge Regression with Principal Component Analysis (PCA) to solve the multivariate regression on the extracted foundation model embeddings. The results are based on the Pearson correlation coefficient as defined in [35]. Dataset details can be found in Appendix A.1.

eva The *eva* framework [29] comprises tasks for clinical prediction across various cancer types and pathology applications. Again, we use the default evaluation protocol and implementation [23]. For patch-level classification tasks, a linear classification head is trained on the CLS or CLS+MEAN tokens. For patch-level segmentation tasks, segmentation is performed on the extracted patch token embeddings. For slide-level classification tasks, *eva* applies the default Attention-based Multiple Instance Learning (ABMIL) [34] protocol. We use balanced accuracy as a performance metric for all patch-level and slide-level classification tasks. For patch-level segmentation tasks, Dice score without background is reported. Evaluation results are presented using the test split when provided, and the validation split when a test split is unavailable. Datasets details can be found in Appendix A.1.

PathoROB The PathoROB Robustness Index benchmark [47] is designed to evaluate model robustness with respect to non-biological, confounding features. We use their default evaluation protocol and implementation [65]. Our average values for Virchow2 differ slightly from official values due to using bicubic image resizing instead of bilinear. This does not affect the ranking. We provide dataset details in Appendix A.1.

	RudolfV 2	RudolfV 2-B	RudolfV 2-S	UNI2-H	H-Optimus-0	Midnight-12k	Virchow2	H0-mini*	Phikon-v2	Atlas 2	Pluto-4G*
HEST-ccRCC	28.6	27.6	26.1	28.1	**29.2**	20.9	27.2	26.4	27.4	27.9	-
HEST-COAD	**33.6**	30.1	31.3	33.2	30.9	31.8	25.8	27.0	25.6	34.9	-
HEST-IDC	**61.9**	60.6	59.7	60.5	61.1	59.9	59.7	59.1	56.8	62.7	-
HEST-LUNG	57.5	55.7	55.2	57.4	57.5	**58.3**	56.8	56.3	55.0	59.4	-
HEST-LYMPH-IDC	26.8	26.2	26.0	27.4	26.6	**27.5**	25.7	26.4	24.8	28.6	-
HEST-PAAD	50.6	50.8	50.6	**52.3**	51.1	50.2	47.8	50.7	47.7	54.2	-
HEST-PRAD	37.0	36.6	37.1	37.5	36.2	37.1	35.3	36.3	**37.9**	39.7	-
HEST-RECTUM	23.2	22.1	19.8	22.7	**24.0**	20.3	20.7	20.5	18.7	25.4	-
HEST-SKCM	65.7	61.5	57.8	**68.3**	66.1	64.8	64.0	61.2	58.4	70.0	-
MSI CRC (patch)	**73.6**	71.8	72.3	72.0	70.0	70.8	71.9	-	68.3	78.0	-
MSI STAD (patch)	**75.6**	74.6	75.4	73.1	72.2	73.7	72.5	-	69.2	76.1	-
HEST-Average	42.8	41.2	40.4	**43.0**	42.5	41.2	40.3	40.4	39.1	44.8	43.2
Molecular-Average	**48.6**	47.1	46.5	48.4	47.7	46.8	46.1	-	44.5	50.6	-
BACH	89.5	90.9	89.4	**92.2**	74.7	90.6	88.6	-	73.6	91.7	93.2
BreakHis	90.2	**90.3**	82.4	86.3	81.0	83.4	81.1	-	70.9	88.3	81.8
CoNSeP	**66.9**	66.7	66.1	63.6	64.5	62.4	64.5	62.9	62.9	66.1	65.0
CRC-100k	**97.1**	**97.1**	96.9	96.9	96.4	96.7	96.6	-	95.3	97.2	96.8
Gleason	78.2	78.0	77.2	77.6	77.3	**79.0**	78.5	-	75.3	80.7	78.5
MHIST	86.5	**87.2**	84.5	82.6	85.1	81.2	86.5	-	79.6	88.5	87.9
MoNuSAC	**71.6**	69.6	68.6	64.5	68.1	65.8	66.8	64.3	64.3	69.4	70.4
PANDA	67.0	**67.1**	66.2	66.2	66.2	65.4	64.7	-	62.1	68.3	66.6
PCAM	**95.4**	95.1	94.4	95.1	94.4	93.0	93.9	-	90.0	96.1	95.2
CAMELYON16 (0.25 MPP)	85.9	**87.4**	86.4	85.9	84.7	86.1	85.7	-	80.7	86.8	-
TCGA Uniform (1.0 MPP)	79.0	78.0	75.4	79.4	76.0	**83.7**	78.0	-	70.1	84.1	-
TCGA Uniform (0.5 MPP)	77.2	76.0	73.3	78.8	78.8	**83.2**	77.5	-	76.7	82.7	-
Morphology-Average-Pluto-4G	**82.5**	82.4	80.6	80.6	78.6	79.7	80.1	-	74.9	82.9	81.7
Morphology-Average	**82.0**	**82.0**	80.1	80.8	78.9	80.9	80.2	-	75.1	83.3	-
Prediction-Average-Pluto-4G	**62.6**	61.8	60.5	61.8	60.6	60.5	60.2	-	57.0	63.8	62.5
Prediction-Average	**66.0**	65.3	64.0	65.3	64.0	64.6	63.9	-	60.5	67.7	-
Robustness Index (Camelyon)	89.7	87.7	**93.0**	54.4	70.5	47.8	79.9	71.8	1.9	94.0	-
Robustness Index (TCGA)	**86.1**	85.3	85.7	80.3	81.2	85.8	82.2	79.4	61.9	87.9	-
Robustness Index (Tolkach)	96.4	95.9	**96.5**	92.3	91.8	94.1	95.4	93.2	76.8	96.4	-
Plismbench (Leaderboard metric)	65.0	**65.1**	64.7	36.0	52.2	37.3	46.4	-	19.6	64.5	-
Robustness-Average	84.3	83.5	**85.0**	65.8	73.9	66.2	76.0	-	40.0	85.7	-

*Results for H0-mini taken from [24] (HEST, *eva*) and [65] (PathoROB). Results for Pluto-4G taken from [64].

Table 1: Comparison of RudolfV 2 models to state-of-the-art open-weights and closed-weights models (Atlas 2 [2], Pluto-4G [64]). The three sections of the table show prediction results for tasks with morphology and molecular targets as well as for robustness benchmarks. The **best** open-weights model result per row is in bold, the second-best open-weights model result is underlined. The presented RudolfV 2 models exhibit the best average performance in each of the three sections.

Plismbench The Plismbench benchmark [24] measures the representation consistency across different scanning and staining conditions. We use the default evaluation protocol and implementation [66]. We report the leaderboard metric, defined as the average of four metrics: cosine similarity for all pairs, top-10 accuracy for cross-scanner pairs, top-10 accuracy for cross-staining pairs, and top-10 accuracy for cross-scanner and cross-staining pairs and use $n = 8{,}139$ tiles, the reference value reported in the repository. The official Plismbench leaderboard [66] reports results for either the CLS+MEAN or the CLS token per model. Here, we report results for both the CLS+MEAN (in Table 1) and CLS tokens (in Table 2). Further details can be found in Appendix A.1.

Additional benchmarks The public benchmarks MSI CRC [43, 37], MSI STAD [43, 37], and TCGA Uniform [49] are not part of any public evaluation framework. For these benchmarks, we report an evaluation based on an internal framework. For each benchmark, we apply linear probing using logistic regression. Results are computed for 5 seeds over data split, shuffling, and initialization

of the prediction head per foundation model and task. The mean across seeds is reported. Dataset splits are described in detail in Appendix A.1.

4 Results

The following analysis is based on public benchmark datasets from four foundation model evaluation frameworks: HEST [35], *eva* [29], PathoROB [47], and Plismbench [24], as well as additional (TME) and MSI prediction benchmark datasets. The tasks span TME tissue and cell typing, morphological pattern recognition, cancer subtyping, and gene mutation prediction from histology. The evaluation protocols and task descriptions are detailed in Section 3.3 and Appendix A.1, respectively.

	RudolfV 2	RudolfV 2-B	RudolfV 2-S	UNI2-H	H-Optimus-0	Midnight-12k	Virchow2	H0-mini*	Phikon-v2	Atlas 2	Pluto-4G*
HEST-ccRCC	28.1	**28.3**	26.3	26.4	26.8	21.3	27.4	26.7	26.6	26.0	28.9
HEST-COAD	**31.0**	28.0	27.7	30.1	30.9	29.1	25.9	24.9	25.0	32.3	31.6
HEST-IDC	**61.3**	59.7	58.7	59.0	59.8	58.2	59.2	58.6	54.1	62.1	60.6
HEST-LUNG	**56.6**	53.9	51.0	55.8	55.9	55.8	55.3	54.8	54.2	57.2	56.9
HEST-LYMPH-IDC	25.8	25.3	25.4	**27.2**	25.9	26.4	25.6	26.3	24.4	28.2	27.3
HEST-PAAD	49.6	48.0	46.8	**50.0**	49.1	49.0	47.2	49.2	44.5	52.3	51.1
HEST-PRAD	35.9	35.3	35.4	35.7	**38.5**	33.7	34.8	36.8	35.5	38.5	37.4
HEST-RECTUM	22.0	20.5	17.7	**22.3**	22.2	18.5	20.9	18.6	17.5	24.0	23.3
HEST-SKCM	63.8	59.6	51.6	**65.9**	64.5	63.6	61.9	60.1	55.5	67.8	67.0
MSI CRC (patch)	**74.0**	72.0	71.9	71.8	69.7	69.9	71.6	-	67.6	78.2	-
MSI STAD (patch)	**76.3**	75.9	75.6	72.8	72.6	72.6	73.0	-	68.6	75.8	-
HEST-Average	**41.6**	39.8	37.8	41.4	41.5	39.5	39.8	39.6	37.5	43.2	42.7
Molecular-Average	**47.7**	46.0	44.4	47.0	46.9	45.3	45.7	-	43.0	49.3	-
BACH	89.8	**91.5**	90.0	**91.5**	75.9	90.5	88.3	77.4	73.3	91.1	93.8
BreakHis	**90.0**	89.5	81.4	85.5	80.1	81.2	82.1	-	71.3	85.6	81.5
CoNSeP	**66.9**	66.7	66.1	63.6	64.5	62.4	64.5	62.9	62.9	66.1	65.0
CRC-100k	97.0	**97.1**	96.9	96.5	95.5	96.6	96.7	96.1	93.9	97.1	96.4
Gleason	78.9	79.1	77.5	77.5	77.0	**80.0**	78.3	-	75.7	79.5	79.3
MHIST	86.2	**87.3**	84.5	82.6	84.4	80.4	86.1	79.0	77.7	88.0	87.5
MoNuSAC	**71.6**	69.6	68.6	64.5	68.1	65.8	66.8	64.3	64.3	69.4	70.4
PANDA	67.0	65.9	66.5	64.6	**67.2**	65.2	64.9	66.7	62.4	68.3	66.8
PCAM	**95.3**	94.9	94.4	95.0	94.3	92.9	93.8	94.2	89.4	96.1	95.1
CAMELYON16 (0.25 MPP)	85.8	**87.1**	86.8	85.2	83.1	84.0	85.9	84.2	80.2	86.5	-
TCGA Uniform (1.0 MPP)	78.9	78.1	75.3	79.3	75.6	**82.7**	78.2	-	69.8	84.2	-
TCGA Uniform (0.5 MPP)	77.2	75.9	73.4	78.8	78.7	**82.5**	77.8	-	76.7	82.6	-
Morphology-Average-Pluto-4G	**82.5**	82.4	80.7	80.1	78.6	79.4	80.2	-	74.5	82.4	81.8
Morphology-Average	**82.0**	81.9	80.1	80.4	78.7	80.4	80.3	-	74.8	82.9	-
Prediction-Average-Pluto-4G	**62.0**	61.1	59.2	60.8	60.0	59.5	60.0	-	56.0	62.8	62.2
Prediction-Average	**65.6**	64.7	63.0	64.4	63.5	63.6	63.7	-	59.6	66.8	-
Robustness Index (Camelyon)	89.5	87.8	**93.0**	51.5	69.0	46.7	74.1	-	1.6	93.6	-
Robustness Index (TCGA)	**86.1**	85.1	85.5	79.6	80.4	85.4	82.7	-	58.8	87.8	-
Robustness Index (Tolkach)	**96.2**	95.5	96.0	91.4	91.0	93.9	95.2	-	73.3	95.9	-
Plismbench (Leaderboard metric)	**61.5**	58.1	56.6	33.2	48.0	33.8	44.8	54.1	16.4	60.5	-
Robustness-Average	**83.3**	81.6	82.8	63.9	72.1	65.0	74.2	-	37.5	84.4	-

*Results for H0-mini taken from [24] (*eva*), [66] (Plismbench) and [32] (HEST). Results for Pluto-4G taken from [64].

Table 2: Same table content and formatting as in Table 1, but for the CLS token instead of CLS+MEAN.

Table 1 demonstrates that RudolfV 2 achieves the best performance among open-weights foundation models on morphology prediction tasks (up to 1.1 p.p. improved average performance over the runner-up Midnight-12k [39]), molecular prediction tasks (up to 0.2 p.p. improved average performance over the runner-up UNI2-H [14]) and model robustness evaluation (up to 8.3 p.p. improved average performance over the runner-up Virchow2 [89]). Notably, RudolfV 2 also performs on par with

the closed-weights model Pluto-4G [64] and is only surpassed by the state-of-the-art pathology foundation model Atlas 2 [2]. On model robustness evaluation, RudolfV 2 comes close to Atlas 2 [2], narrowing the gap to 1.4 p.p. (84.3 vs. 85.7).

Our distilled models RudolfV 2-B and RudolfV 2-S are the leading efficient open-weights models for morphology prediction, molecular prediction, and model robustness in comparison to H0-mini [24] and Phikon-v2 [26]. For morphology prediction tasks, RudolfV 2-B even matches the average performance of our open-weights flagship RudolfV 2 and outperforms larger open-weights models.

Table 2 provides additional evaluation results when using the CLS token instead of CLS+MEAN tokens, with conclusions similar to those in Table 1.

5 Conclusion

With RudolfV 2, we set out to narrow the gap that has separated openly available pathology foundation models from their closed, proprietary counterparts. Across representative public benchmarks, RudolfV 2 is the best-performing and most robust open-weights model to date, approaching the performance and matching the robustness of our closed flagship Atlas 2. Moreover, its distilled variants, RudolfV 2-B and RudolfV 2-S, retain much of this capability and robustness at a fraction of the inference cost, making them well suited to research settings with limited computational resources. By releasing the weights of all three models for research use, we hope to lower the barrier to reproducible work in computational pathology and to establish a stronger foundation for the community to build upon.

Acknowledgments

We would like to thank the teams at Aignostics, Charité - Universitätsmedizin Berlin - Pathology Department, and LMU Munich - Pathology Department for supporting this work.

The benchmark results shown here are in part based upon data generated by the TCGA Research Network: `https://www.cancer.gov/tcga`.

This work was in part supported by the German Ministry of Research, Technology and Space (BMFTR) under Grants 13GW0744A, 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18025A, 031L0207D, and 01IS18037A. K.R.M. was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University and No. 2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation).

References

[1] Nanne Aben, Edwin D. de Jong, Ioannis Gatopoulos, Nicolas Känzig, Mikhail Karasikov, Axel Lagré, Roman Moser, Joost van Doorn, and Fei Tang. Towards Large-Scale Training of Pathology Foundation Models, March 2024. arXiv:2404.15217.

[2] Maximilian Alber, Timo Milbich, Alexandra Carpen-Amarie, Stephan Tietz, Jonas Dippel, Lukas Muttenthaler, Beatriz Perez Cancer, Alessandro Benetti, Panos Korfiatis, Elias Eulig, et al. Atlas 2-foundation models for clinical deployment. *arXiv preprint arXiv:2601.05148*, 2026.

[3] 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, 2025.

[4] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al.

BACH: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019.

[5] Eirini Arvaniti, Kim Fricker, Michael Moret, Niels Rupp, Thomas Hermanns, Christian Fankhauser, Norbert Wey, Peter Wild, Jan Rueschoff, and Manfred Claassen. Automated gleason grading of prostate cancer tissue microarrays via deep learning, 03 2018.

[6] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017.

[7] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[8] Alexander Binder, Michael Bockmayr, Miriam Hägele, Stephan Wienert, Daniel Heim, Katharina Hellweg, Masaru Ishii, Albrecht Stenzinger, Andreas Hocke, Carsten Denkert, et al. Morphological and molecular breast cancer profiling through explainable machine learning. *Nature Machine Intelligence*, 3(4):355–366, 2021.

[9] Bioptimus. H-optimus-1. `https://huggingface.co/bioptimus/H-optimus-1`, 2025.

[10] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester Van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the PANDA challenge. *Nature medicine*, 28(1):154–163, 2022.

[11] Péter Bándi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halıcı, Hunter Jackson, Richard Chen, Fabian Both, Jörg Franke, Heidi Küsters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019.

[12] Gianluca Carloni, Biagio Brattoli, Seongho Keum, Jongchan Park, Taebum Lee, Chang Ho Ahn, and Sergio Pereira. Pathology foundation models are scanner sensitive: Benchmark and mitigation with contrastive scangen loss. In *International Workshop on Foundation Models for General Medical AI*, pages 44–53. Springer, 2025.

[13] Binghao Chai, Jianan Chen, Paul Cool, Fatine Oumlil, Anna Tollitt, David F Steiner, Tapabrata Chakraborti, and Adrienne M Flanagan. Impact of tissue staining and scanner variation on the performance of pathology foundation models: a study of sarcomas and their mimics. *bioRxiv*, pages 2025–08, 2025.

[14] 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(3):850–862, March 2024.

[15] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[16] Edwin D de Jong, Eric Marcus, and Jonas Teuwen. Current pathology foundation models are unrobust to medical center differences. *arXiv preprint arXiv:2501.18055*, 2025.

[17] James A Diao, Jason K Wang, Wan Fung Chui, Victoria Mountain, Sai Chowdary Gullapally, Ramprakash Srinivasan, Richard N Mitchell, Benjamin Glass, Sara Hoffman, Sudha K Rao, et al. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nature communications*, 12(1):1613, 2021.

[18] Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, pages 1–13, 2025.

[19] 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, June 2024. arXiv:2401.04079.

[20] Jonas Dippel, Niklas Prenißl, Julius Hense, Philipp Liznerski, Tobias Winterhoff, Simon Schallenberg, Marius Kloft, Oliver Buchstab, David Horst, Maximilian Alber, Lukas Ruff, Klaus-Robert Müller, and Frederick Klauschen. Ai-based anomaly detection for clinical-grade histopathological diagnostics. *NEJM AI*, 1(11):AIoa2400468, 2024.

[21] 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.

[22] David Jacob Drexlin, Jonas Dippel, Julius Hense, Niklas Prenißl, Grégoire Montavon, Frederick Klauschen, and Klaus-Robert Müller. Medi: Metadata-guided diffusion models for mitigating biases in tumor classification. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 379–388, 2025.

[23] eva: Evaluation framework for oncology foundation models (FMs). `https://github.com/kaiko-ai/eva/tree/0.4.2`, 2025. v0.4.2.

[24] 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. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pages 162–172, Cham, 2026. Springer Nature Switzerland.

[25] Alexandre Filiot, Nicolas Dop, Oussama Tchita, Auriane Riou, Thomas Peeters, Daria Valter, Marin Scalbert, Charlie Saillard, Geneviève Robin, and Antoine Olivier. Distilling foundation models for robust and efficient models in digital pathology, 2025.

[26] Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, A large and public feature extractor for biomarker prediction, September 2024. arXiv:2409.09173.

[27] Maaike Galama, Nina Kozar-Gillan, Christina Embacher, Todd Dembo, Cornelius Böhm, Evelyn Ramberger, Julika Ribbat-Idel, Rosemarie Krupar, Verena Aumiller, Miriam Hägele, Kai Standvoss, Gerrit Erdmann, Blanca Pablos, Ari Angelo, Simon Schallenberg, Andrew Norgan, Viktor Matyas, Klaus-Robert Müller, Maximilian Alber, Lukas Ruff, and Frederick Klauschen. OpenTME: An open dataset of ai-powered H&E tumor microenvironment profiles from TCGA. *arXiv preprint arXiv:2604.12075*, 2026.

[28] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35, 2016.

[29] Ioannis Gatopoulos, Nicolas Känzig, Roman Moser, and Sebastian Otálora. eva: Evaluation framework for pathology foundation models. In *Medical Imaging with Deep Learning*, 2024.

[30] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Jin Tae Kwak, and Nasir M. Rajpoot. XY network for nuclear segmentation in multi-tissue histology images. *CoRR*, abs/1812.06499, 2018.

[31] Fredrik K Gustafsson and Mattias Rantalainen. Evaluating computational pathology foundation models for prostate cancer grading under distribution shifts. *arXiv preprint arXiv:2410.06723*, 2024.

[32] Hest-library: Bringing spatial transcriptomics and histopathology together. `https://github.com/mahmoodlab/HEST/tree/v1.2.0`, 2024. v1.2.0.

[33] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS 2014 - Deep Learning Workshop*, 2015.

[34] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, pages 2127–2136, 2018.

[35] Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro Pérez, Sophia J Wagner, Anurag Jayant Vaidya, Richard J. Chen, Drew FK Williamson, Ahrong Kim, and Faisal Mahmood. HEST-1k: A dataset for spatial transcriptomics and histology image analysis. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[36] Dinkar Juyal, Harshith Padigela, Chintan Shah, Daniel Shenker, Natalia Harguindeguy, Yi Liu, Blake Martin, Yibo Zhang, Michael Nercessian, Miles Markey, Isaac Finberg, Kelsey Luu, Daniel Borders, Syed Ashar Javed, Emma L Krause, Raymond Biju, Aashish Sood, Allen Ma, Jackson Nyman, John Shamshoian, Guillaume Chhor, Darpan Sanghavi, Marc Thibault, Limin Yu, Fedaa Najdawi, Jennifer A. Hipp, Darren Fahy, Benjamin Glass, Eric Walk, John Abel, Harsha Vardhan pokkalla, Andrew H. Beck, and Sean Grullon. PLUTO: Pathology-universal transformer. In *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*, 2024.

[37] Jakub R Kaczmarzyk, Rajarsi Gupta, Tahsin M Kurc, Shahira Abousamra, Joel H Saltz, and Peter K Koo. Champkit: A framework for rapid evaluation of deep neural networks for patch-based histopathology classification. *Computer methods and programs in biomedicine*, 239:107631, 2023.

[38] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[39] Daniel Kaplan, Ratna Sagari Grandhi, Connor Lane, Benjamin Warner, Tanishq Mathew Abraham, and Paul S. Scotti. How to train a state-of-the-art pathology foundation model with $1.6k, 2025.

[40] Mikhail Karasikov, Joost van Doorn, Nicolas Känzig, Melis Erdal Cesur, Hugo Mark Horlings, Robert Berke, Fei Tang, and Sebastian Otálora. Training state-of-the-art pathology foundation models with orders of magnitude less data. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15967, pages 573–583. Springer Nature Switzerland, October 2025.

[41] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue (v0.1) [Data set]. Zenodo. `https://doi.org/10.5281/zenodo.1214456`, May 2018.

[42] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730, 2019.

[43] Jakob Nikolas Kather, Alexander T Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature medicine*, 25(7):1054–1056, 2019.

[44] Jacob Kauffmann, Jonas Dippel, Lukas Ruff, Wojciech Samek, Klaus-Robert Müller, and Grégoire Montavon. Explainable AI reveals Clever Hans effects in unsupervised learning models. *Nature Machine Intelligence*, 7:412—-422, 2025.

[45] Philipp Keyl, Michael Bockmayr, Daniel Heim, Gabriel Dernbach, Grégoire Montavon, Klaus-Robert Müller, and Frederick Klauschen. Patient-level proteomic network prediction by explainable artificial intelligence. *NPJ Precision Oncology*, 6(1):35, 2022.

[46] Frederick Klauschen, Jonas Dippel, Philipp Keyl, Philipp Jurmeister, Michael Bockmayr, Andreas Mock, Oliver Buchstab, Maximilian Alber, Lukas Ruff, Grégoire Montavon, et al. Toward explainable artificial intelligence for precision pathology. *Annual Review of Pathology: Mechanisms of Disease*, 19(1):541–570, 2024.

[47] Jonah Kömen, Edwin D de Jong, Julius Hense, Hannah Marienwald, Jonas Dippel, Philip Naumann, Eric Marcus, Lukas Ruff, Maximilian Alber, Jonas Teuwen, et al. Towards robust foundation models for digital pathology. *Nature Communications*, 17(1):5218, 2026.

[48] Jonah Kömen, Hannah Marienwald, Jonas Dippel, and Julius Hense. Do histopathological foundation models eliminate batch effects? A comparative study. 2024. Presented at: NeurIPS 2024 Workshop on Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond.

[49] Daisuke Komura, Akihiro Kawabe, Keisuke Fukuta, Kyohei Sano, Toshikazu Umezaki, Hirotomo Koda, Ryohei Suzuki, Ken Tominaga, Mieko Ochi, Hiroki Konishi, et al. Universal encoding of pan-cancer histology by deep texture representations. *Cell Reports*, 38(9), 2022.

[50] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019.

[51] Tim Lenz, Peter Neidlinger, Marta Ligero, Georg Wölflein, Marko van Treeck, and Jakob Nikolas Kather. Unsupervised foundation model-agnostic slide-level representation learning, 2024. arXiv:2411.13623.

[52] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[53] Weiping Lin, Shen Liu, Runchen Zhu, and Liansheng Wang. Unveiling institution-specific bias in pathology foundation models: Detriments, causes, and potential solutions. *arXiv preprint arXiv:2502.16889*, 2025.

[54] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[55] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024.

[56] Marc Macenko, Marc Niethammer, J. Marron, David Borland, John Woosley, Xiaojun Guan, Charles Schmitt, and Nancy Thomas. A method for normalizing histology slides for quantitative analysis. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI).*, 2009.

[57] Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K Lampinen. Aligning machine and human visual representations across abstraction levels. *Nature*, 647(8089):349–355, 2025.

[58] Alexander Möllers, Julius Hense, Florian Schulz, Timo Milbich, Maximilian Alber, and Lukas Ruff. Mind the gap: Continuous magnification sampling for pathology foundation models, 2026.

[59] Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. Hibou: A Family of Foundational Vision Transformers for Pathology, June 2024. arXiv:2406.05074.

[60] Hai Cao Truong Nguyen and David Joon Ho. fmMAP: A framework reducing site-bias batch effect from foundation models in pathology. In *MICCAI Workshop on Computational Pathology with Multimodal Data (COMPAYL)*, 2025.

[61] Masaki Ochi, Daisuke Komura, Takumi Onoyama, et al. Registered multi-device/staining histology image dataset for domain-agnostic machine learning models. *Scientific Data*, 11:330, 2024.

[62] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification.

[63] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 27730–27744, 2022.

[64] Harshith Padigela, Shima Nofallah, Atchuth Naveen Chilaparasetti, Ryun Han, Andrew Walker, Judy Shen, Chintan Shah, Blake Martin, Aashish Sood, Elliot Miller, Ben Glass, Andy Beck, Harsha Pokkalla, and Syed Ashar Javed. Pluto-4: Frontier pathology foundation models, 2025.

[65] Pathorob: Benchmark for pathology foundation models robustness to non-biological medical center differences. `https://github.com/bifold-pathomics/PathoROB/tree/ac1abe4df8c4d5b03aab13d6d9aabbc7205061e6`, 2025.

[66] Plismbench: A robustness benchmark of pathology foundation models. `https://github.com/owkin/plism-benchmark/tree/b4f32b7c233e0f5ba8340da89ef65803cfc2ce49`, 2025.

[67] Zhongwei Qiu, Hanqing Chao, Tiancheng Lin, Wanxing Chang, Zijiang Yang, Wenpei Jiao, Yixuan Shen, Yunshuo Zhang, Yelin Yang, Wenbin Liu, Hui Jiang, Yun Bian, Ke Yan, Dakai Jin, and Le Lu. Bridging local inductive bias and long-range dependencies with pixel-mamba for end-to-end whole slide image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22738–22747, October 2025.

[68] Patricia Raciti, Jillian Sue, Juan A Retamero, Rodrigo Ceballos, Ran Godrich, Jeremy D Kunz, Adam Casson, Dilip Thiagarajan, Zahra Ebrahimzadeh, Julian Viret, et al. Clinical validation of artificial intelligence–augmented pathology diagnosis demonstrates significant gains in diagnostic accuracy in prostate cancer detection. *Archives of Pathology & Laboratory Medicine*, 147(10):1178–1185, 2023.

[69] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

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

[71] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[72] George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D. Kunz, Juan A. Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, Matthew Hanna, Michal Zelechowski, Julian Viret, Neil Tenenholtz, James Hall, Nicolo Fusi,

Razik Yousfi, Peter Hamilton, William A. Moye, Eugene Vorontsov, Siqi Liu, and Thomas J. Fuchs. PRISM: A Multi-Modal Generative Foundation Model for Slide-Level Histopathology, May 2024. arXiv:2405.10254.

[73] 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.

[74] Fabio A. Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016.

[75] Kai Standvoss, Miriam Hägele, Rosemarie Krupar, Julika Ribbat-Idel, Jennifer Altschüler, Gerrit Erdmann, Hans Pinckaers, Evelyn Ramberger, Madleen Drinkwitz, Ádám Nárai, Alexander Möllers, Katja Lingelbach, Sebastian Kons, Lukas Hönig, Recepcan Adigüzel, Joana Baião, Alberto Megina Gonzalo, Marius Teodorescu, Marie-Lisa Eich, Paolo Chetta, Shakil Merchant, Verena Aumiller, Simon Schallenberg, Andrew Norgan, Klaus-Robert Müller, Lukas Ruff, Maximilian Alber, and Frederick Klauschen. Atlas H&E-TME: Scalable ai-based tissue profiling at expert pathologist-level accuracy. *arXiv preprint arXiv:2606.12346*, 2026.

[76] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[77] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine Hermann, Kerem Oktar, Klaus Greff, Martin N Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas O'Connell, Thomas Unterthiner, Andrew Kyle Lampinen, Klaus Robert Muller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *Transactions on Machine Learning Research*, 2025. Survey Certification, Expert Certification.

[78] Yuxuan Sun, Yixuan Si, Chenglu Zhu, Xuan Gong, Kai Zhang, Pingyi Chen, Ye Zhang, Zhongyi Shui, Tao Lin, and Lin Yang. Cpath-omni: A unified multimodal foundation model for patch and whole slide image analysis in computational pathology. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10360–10371, 2025.

[79] Yuri Tolkach, Lisa Marie Wolgast, Alexander Damanakis, Alexey Pryalukhin, Simon Schallenberg, Wolfgang Hulla, Marie-Lisa Eich, Wolfgang Schroeder, Anirban Mukhopadhyay, Moritz Fuchs, et al. Artificial intelligence for tumour tissue detection and histological regression grading in oesophageal adenocarcinomas: a retrospective algorithm development and validation study. *The Lancet Digital Health*, 5(5):e265–e275, 2023.

[80] Anurag Vaidya, Andrew Zhang, Guillaume Jaume, Andrew H. Song, Tong Ding, Sophia J. Wagner, Ming Y. Lu, Paul Doucet, Harry Robertson, Cristina Almagro-Perez, Richard J. Chen, Dina ElHarouni, Georges Ayoub, Connor Bossi, Keith L. Ligon, Georg Gerber, Long Phi Le, and Faisal Mahmood. Molecular-driven foundation model for oncologic pathology, 2025.

[81] Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 210–218, Cham, 2018. Springer International Publishing.

[82] Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, Simon Graham, Quoc Dang Vu, Mieke Zwager, Shan E. Ahmed Raza, Nasir Rajpoot, Xiyi Wu, Huai Chen, Yijie Huang, Lisheng Wang, Hyun Jung, G. Thomas Brown, Yanling Liu, Shuolin Liu, Seyed Alireza Fatemi Jahromi, Ali Asghar Khani, Ehsan Montahaei, Mahdieh Soleymani Baghshah, Hamid Behroozi, Pavel Semkin, Alexandr Rassadin, Prasad Dutande, Romil Lodaya,

Ujjwal Baid, Bhakti Baheti, Sanjay Talbar, Amirreza Mahbod, Rupert Ecker, Isabella Ellinger, Zhipeng Luo, Bin Dong, Zhengyu Xu, Yuehan Yao, Shuai Lv, Ming Feng, Kele Xu, Hasib Zunair, Abdessamad Ben Hamza, Steven Smiley, Tang-Kai Yin, Qi-Rui Fang, Shikhar Srivastava, Dwarikanath Mahapatra, Lubomira Trnavska, Hanyun Zhang, Priya Lakshmi Narayanan, Justin Law, Yinyin Yuan, Abhiroop Tejomay, Aditya Mitkari, Dinesh Koka, Vikas Ramachandra, Lata Kini, and Amit Sethi. Monusac2020: A multi-organ nuclei segmentation and classification challenge. *IEEE Transactions on Medical Imaging*, 40(12):3413–3423, 2021.

[83] Eugene Vorontsov, George Shaikovski, Adam Casson, Julian Viret, Eric Zimmermann, Neil Tenenholtz, Yi Kan Wang, Jan H. Bernhard, Ran A. Godrich, Juan A. Retamero, Jinru Shia, Mithat Gonen, Martin R. Weiser, David S. Klimstra, Razik Yousfi, Nicolo Fusi, Thomas J. Fuchs, Kristen Severson, and Siqi Liu. Prism2: Unlocking multi-modal general pathology ai with clinical dialogue, 2025.

[84] 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, and Christopher. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, October 2024.

[85] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Naofumi Tomita, Lorenzo Torresani, et al. A petri dish for histopathology image analysis. In *Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings*, pages 11–24. Springer, 2021.

[86] 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(8015):181–188, June 2024.

[87] Yingxue Xu, Yihui Wang, Fengtao Zhou, Jiabo Ma, Cheng Jin, Shu Yang, Jinbang Li, Zhengyu Zhang, Chenglong Zhao, Huajun Zhou, Zhenhui Li, Huangjing Lin, Xin Wang, Jiguang Wang, Anjia Han, Ronald Cheong Kin Chan, Li Liang, Xiuming Zhang, and Hao Chen. A multi-modal knowledge-enhanced whole-slide pathology foundation model. *Nature Communications*, 16:11406, 2025.

[88] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[89] Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Severson. Virchow 2: Scaling Self-Supervised Mixed Magnification Models in Pathology, August 2024. arXiv:2408.00738.

A Appendix

A.1 Task Descriptions

HEST-Benchmark The HEST-Benchmark was introduced by [35] for benchmarking foundation models for histology on the task of gene expression prediction from H&E-stained images. The benchmark includes 72 spatial transcriptomics profiles (using Xenium or Visium technology) with corresponding H&E-stained images from 47 patients grouped into 10 tasks based on organ. Each task involves predicting the expression levels of the 50 most variable genes (highest normalized variance) from 112×112 µm H&E-stained image patches (equivalent to 224×224 pixels at 20× magnification) centered on each spatial transcriptomics spot, formulated as a multivariate regression problem. We use the default Ridge Regression with PCA (256 factors) evaluation protocol to solve the multivariate regression problem for extracted embeddings [35]. Specifically, the objective of the 10 tasks is to predict gene expression levels for invasive ductal carcinoma (breast cancer, IDC), prostate adenocarcinoma (prostate cancer, PRAD), pancreatic adenocarcinoma (pancreatic cancer, PAAD), skin cutaneous melanoma (skin cancer, SKCM), colonic adenocarcinoma (colon cancer, COAD), rectal adenocarcinoma (rectum cancer, READ), clear cell renal cell carcinoma (kidney cancer, ccRCC), hepatocellular carcinoma (liver cancer, HCC), lung adenocarcinoma (lung cancer, LUAD), and axillary lymph nodes in IDC (metastatic, LYMPH-IDC). The benchmark applies patient-stratified splits to avoid any train/test patient-level data leakage, resulting in a k-fold cross-validation where k is the number of patients [35]. Performance is measured as the Pearson correlation coefficient between the predicted and measured gene expression. Results are reported as the mean across folds.

BACH The BACH dataset comprises 400 H&E microscopy images (2048×1536 pixels at 20× magnification) of breast cancer biopsies. It was introduced as part of the ICIAR 2018 Grand Challenge on BreAst Cancer Histology images (BACH) [4]. The task of the challenge is to classify each image into one of the following four classes: "normal", "benign", "in situ carcinoma", and "invasive carcinoma". The dataset is split into 268 training images (67 images per class) and 132 test images (33 images per class). The patches are resized and cropped to 224×224 pixels.

CRC-100k The CRC-100k dataset contains 107,180 H&E images (224×224 pixels at 20× magnification) extracted from colorectal cancer (CRC) tissue samples [42]. The tissue samples originate from CRC primary tumors and CRC liver metastases. The task of this benchmark is to classify each image into one of the following nine tissue classes: "adipose", "background", "debris", "lymphocytes", "mucus", "smooth muscle", "normal colon mucosa", "cancer-associated stroma", and "colorectal adenocarcinoma epithelium". The dataset is split into 100,000 training images (NCT-CRC-HE-100K-NONORM) and 7,180 test images (CRC-VAL-HE-7K). We use the original (no-norm) images without Macenko color normalization [41].

MHIST The task of MHIST is to classify images of colorectal polyps into hyperplastic polyps (HPs) vs. sessile serrated adenomas (SSAs) [85]. This distinction is clinically important as HPs are typically benign whereas SSAs are precancerous lesions that can turn into cancer if left untreated. The task is challenging for pathologists, often showing considerable inter-pathologist variability. The MHIST dataset consists of 3,152 H&E images (224×224 pixels at 8× magnification) of colorectal polyps. Labels are derived from the majority vote of seven pathologists. The dataset is split into 2,162 training images (1,545 HP and 617 SSA) and 990 test images (630 HP and 360 SSA).

PCAM PCAM (PatchCamelyon) defines the clinically-relevant task of metastasis detection as a binary image classification task [81]. The dataset consists of 327,680 H&E images (96×96 pixels at 10× magnification) extracted from scans of sentinel lymph node sections. Each image is annotated with a binary label indicating the presence of metastatic breast cancer tissue. An image is labeled as metastatic if the center (32×32 pixels) region contains at least one pixel of tumor tissue. The dataset is split by ratios of 80:10:10 into training, validation, and test sets with no overlap of WSIs/cases between the splits. Moreover, every split has a 50:50 balance of positive and negative examples. For evaluation, we resize each image to 224×224 pixels. PCAM has been derived from the CAMELYON16 Challenge [6].

CAMELYON16 The task of the CAMELYON16 challenge [6] is to classify whole slide images (WSIs) of lymph node tissue sections into metastatic breast cancer tissue being present or not. The

dataset comprises 399 H&E-stained WSIs of sentinel lymph node sections, which were acquired and scanned (40× magnification) at two centers from the Netherlands [6]. The dataset is split into 270 training slides (110 with and 160 without metastasis) and 129 test slides (49 with and 80 without metastasis). Here, we report results for the CAMELYON16 (small) setup in *eva* [29], which randomly samples max. 1,000 patches per slide.

PANDA The PANDA challenge [10] is concerned with the challenging task of tumor grading of whole slide images (WSIs) of prostate cancer biopsies, which suffers from significant inter-observer variability between pathologists. Prostate cancer grading follows the Gleason grading system (3, 4, or 5) based on architectural growth patterns of the tumor, which are then converted into an ISUP grade on a scale of 1–5 for usage as a prognostic marker. The dataset features 9,555 H&E-stained WSIs (subset with noisy labels removed) of prostate tissue biopsies from two medical centers scanned at 20× magnification. Specifically, the task is to classify each WSI into an ISUP grade of 0–5 (six classes), where 0 means that a biopsy does not contain any cancer. The dataset is split into 6,686 training slides, 1,430 validation slides, and 1,439 test slides in a class-stratified manner. Here, we report results for the PANDA (small) setup in *eva* [29], which considers a fewer number of total slides (955 train, 477 validation, 477 test) and fewer randomly sampled patches per slide (200).

BreakHis The Breast Cancer Histopathological Image Classification (BreakHis) dataset was introduced for the task of classifying breast tumor tissue into benign and malignant histological subtypes [74]. The task involves multi-class classification of microscopic images into eight categories: "adenosis" (A, benign), "fibroadenoma" (F, benign), "phyllodes tumor" (PT, benign), "tubular adenona" (TA), "carcinoma" (DC, malignant), "lobular carcinoma" (LC, malignant), "mucinous carcinoma" (MC, malignant) and "papillary carcinoma" (PC, malignant). The original dataset consists of 9,109 microscopic images collected from 82 patients at multiple magnification factors, 40×, 100×, 200×, and 400×. For this benchmark, we follow the *eva* [29] evaluation protocol and use only the 40× magnification subset, which consists of 1,995 H&E-stained images (700×460 pixels at 40×magnification, 0.49 µm/pixel) selected from classes with at least seven patients to ensure statistical robustness: TA, MC, F & DC. The dataset is split using patient-stratified partitioning, resulting in 1,132 training images and 339 validation images with approximate class balance maintained across splits.

CoNSeP The CoNSeP (Colorectal Nuclear Segmentation and Phenotypes) dataset was introduced for the task of nuclear segmentation and classification in colorectal tissue [30]. The task involves segmenting and classifying cell nuclei into four categories: "miscellaneous", "inflammatory", "epithelial" (combining normal and malignant/dysplastic epithelial cells), and "spindle-shaped" (combining fibroblast, muscle, and endothelial cells). The dataset consists of 41 H&E-stained images (1,000×1,000 pixels at 40× magnification) extracted from 16 whole-slide images of unique colorectal cancer patients, containing 24,319 annotated nuclei in total. Segmentation masks are provided as instance-level annotations indicating both the spatial extent and class label of each nucleus. The dataset is split into 27 training images and 14 test images. Following the *eva* [29] protocol, we use the test split as the validation split.

Gleason Arvaniti The Gleason Arvaniti dataset was introduced for automated Gleason grading of prostate cancer tissue microarrays (TMAs) [5]. The task involves multi-class classification of prostate tissue images into four categories: "benign", "Gleason pattern 3", "Gleason pattern 4", and "Gleason pattern 5". The dataset comprises 22,752 H&E-stained images (750×750 pixels at 40× magnification) from 886 patients across two cohorts: a training cohort of 641 patients and an independent test cohort of 245 patients, with annotations provided by two experienced pathologists. Following the *eva* [29] protocol, the dataset is split into 15,303 training images, 2,482 validation images selected from TMA 76 due to its balanced distribution of Gleason scores, and 4,967 test images.

MoNuSAC The MoNuSAC (Multi-Organ Nuclei Segmentation and Classification Challenge) dataset was introduced for the task of nuclear segmentation and classification across multiple organs and tissue types [82]. The task involves segmenting and classifying cell nuclei into five categories: "epithelial cells", "lymphocytes", "macrophages", "neutrophils", and "ambiguous" (cells that cannot be definitively classified into the other categories). The dataset consists of 294 H&E-stained tissue images from four organs: breast, kidney, lung, and prostate, with variable dimensions, from 113×81

to 1,398×1,956 pixels at 40× magnification. The images were collected from 37 hospitals and 71 patients, containing over 46,000 annotated nuclei. Instance-level segmentation masks are provided, indicating both the spatial extent and class label of each nucleus. The dataset is split into 209 training images and 85 test images. Following the *eva* [29] evaluation protocol, we use the test split as the validation split.

MSI CRC (patch) This dataset contains 173,630 H&E images (224×224 pixels at 20× magnification) extracted from $N = 360$ colorectal cancer (CRC) tissue scans from TCGA (TCGA-CRC-DX). The task is binary classification of microsatellite instability (MSI) vs. microsatellite stability (MSS), which is a clinically important prognostic marker [43, 37]. The dataset is split into 56,044 (28,022 MSI + 28,022 MSS) training images, 18,682 (9,341 MSI + 9,341 MSS) validation images, and 98,904 (28,335 MSI + 70,569 MSS) test images.

MSI STAD (patch) This dataset contains 198,464 H&E images (224×224 pixels at 20× magnification) extracted from $N = 315$ stomach adenocarcinoma (STAD) tissue scans from TCGA (TCGA-STAD). The task is binary classification of microsatellite instability (MSI) vs. microsatellite stability (MSS), which is a clinically important prognostic marker [43, 37]. The dataset is split into 60,342 (30,171 MSI + 30,171 MSS) training images, 20,114 (10,057 MSI + 10,057 MSS) validation images, and 118,008 (27,904 MSI + 90,104 MSS) test images.

TCGA Uniform (10×) and (20×) The TCGA Uniform dataset [49] contains 264,110 to 271,700 patches per resolution with 256×256 pixels. The task is to differentiate between 32 cancer classes from different tissue types (e.g., Colon adenocarcinoma). Only patches showing the specific cancer type were extracted from the TCGA WSIs based on annotations from two trained pathologists. As there is no official train and test split, we generate five folds with no overlapping patients and sampled 100 patches per class and fold, resulting in a total dataset size of 16,000 patches. We generate one dataset containing patches with $0.5\,\mu\text{m}$/pixel (20×) and one with $1.0\,\mu\text{m}$/pixel (10×) to test the performance of the foundation models at different zoom levels. The results represent the mean balanced accuracy for the five-fold cross-validation evaluation.

PathoROB The PathoROB Robustness Index was introduced by [47] as a novel metric to quantify the robustness of pathology foundation models to non-biological confounding features, specifically medical center differences arising from variations in staining procedures, scanner hardware, surgical techniques, and laboratory protocols. The Robustness Index measures the degree to which biological features (e.g., tissue type, cancer type) dominate over confounding non-biological features (e.g., medical center signatures) in the neighborhood structure of a foundation model's embedding space. The benchmark comprises three datasets sourced from CAMELYON[6, 11], TCGA[49], and Tolkach ESCA[79], covering 28 biological classes from 34 medical centers.

Plismbench The Plismbench robustness benchmark was introduced by [24] as a framework to quantify the robustness of pathology foundation models to technical variations arising from different scanning devices and staining procedures. The benchmark measures robustness through cosine similarity and top-k retrieval accuracy metrics, evaluating the ability of foundation models to maintain consistent representations of matching tissue tiles across different technical conditions. The evaluation is performed on registered tile pairs from the PLISM dataset [61], which consists of 46 human tissue types processed with 13 different H&E staining conditions and captured using 7 WSI scanners, resulting in 16,278 registered tissue tile pairs. Robustness is assessed across three conditions: cross-scanner robustness (fixed staining), cross-staining robustness (fixed scanner), and combined cross-scanner and cross-staining robustness.