A Novel Pathology Foundation Model by Mayo Clinic, Charité, and Aignostics

Maximilian Alber [*,1,12], Stephan Tietz [*,1], Jonas Dippel [*,1,6,7], Timo Milbich [*,1], Timothée Lesort [*,1], Panos Korfiatis [#,3], Moritz Krügener [#,1], Beatriz Perez Cancer [#,1], Neelay Shah [#,1], Alexander Möllers [1,6,7], Philipp Seegerer [1], Alexandra Carpen-Amarie [1], Kai Standvoss [1], Gabriel Dernbach [1,7,12], Edwin de Jong [1], Simon Schallenberg [12], Andreas Kunft [1], Helmut Hoffer von Ankershoffen [1], Gavin Schaeferle [5], Patrick Duffy [4], Matt Redlon [4], Philipp Jurmeister [10,11], David Horst [10,12], Lukas Ruff [1], Klaus-Robert Müller[†,6,7,8,9], Frederick Klauschen[†,7,10,11,12,13], Andrew Norgan[†,2]

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

Abstract

Recent advances in digital pathology have demonstrated the effectiveness of foundation models across diverse applications. In this report, we present a novel vision foundation model based on the RudolfV approach. Our model was trained on a dataset comprising 1.2 million histopathology whole slide images, collected from two medical institutions: Mayo Clinic and Charité - Universitätsmedizin Berlin. Comprehensive evaluations show that our model achieves state-of-the-art performance across twenty-one public benchmark datasets, even though it is neither the largest model by parameter count nor by training dataset size.

1 Introduction

Anatomical pathology plays a central role in clinical medicine for tissue-based diagnostics and in biomedical research as a basis for the understanding of mechanisms of disease. Although molecular and omics-based data complement histological assessments, the microscopic evaluation of morphological changes remains the cornerstone of pathology practice. Consequently, with the advent of routine slide digitization, computational pathology has focused on making the analysis of histopathology slide images more precise, scalable, and reliable.

However, despite artificial intelligence (AI) having led to promising proof-of-concepts and applications (e.g., [24, 8, 33, 5, 23]), generalization, application variety, and robustness remain challenging

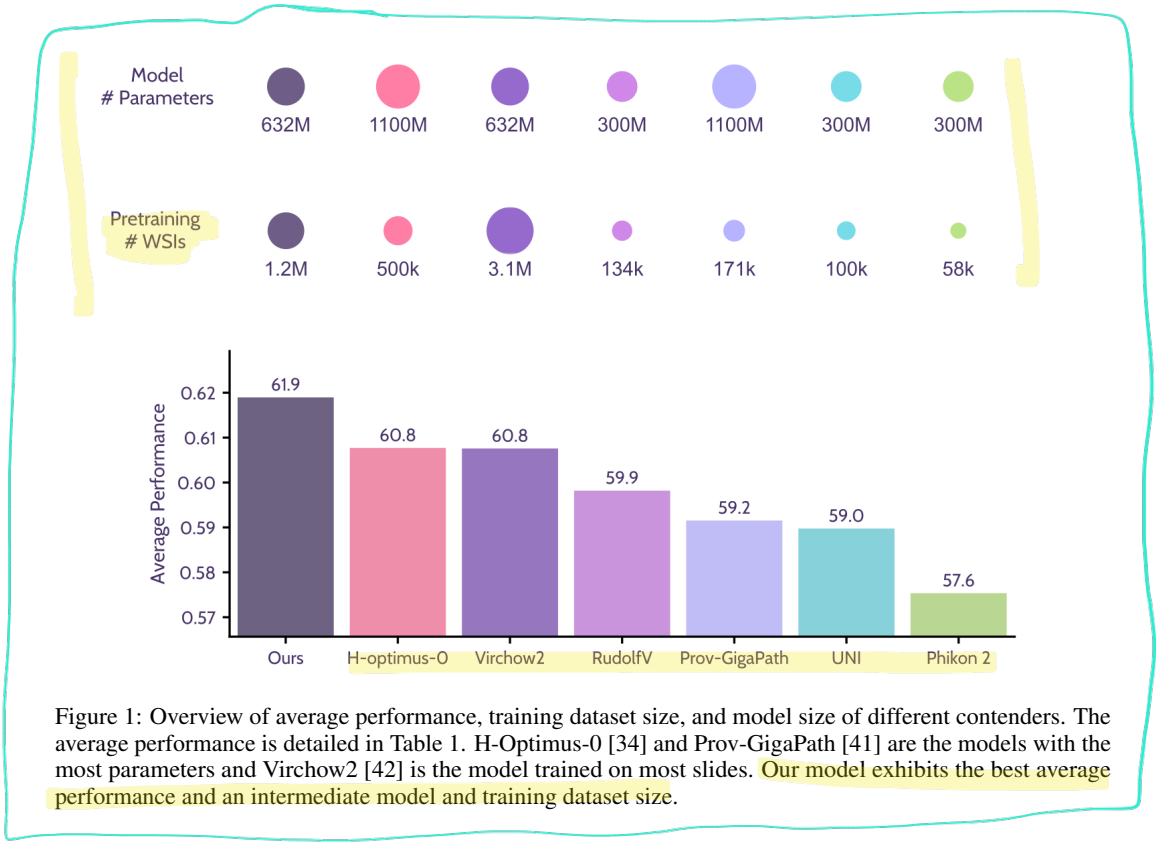

Figure 1: Overview of average performance, training dataset size, and model size of different contenders. The average performance is detailed in Table 1. H-Optimus-0 [34] and Prov-GigaPath [41] are the models with the most parameters and Virchow2 [42] is the model trained on most slides. Our model exhibits the best average performance and an intermediate model and training dataset size.

and have hindered the broad translation of AI applications into clinical routine diagnostics. Here, the limited adoption of digitization in clinical practice results in a scarcity of training data, particularly for infrequent and rare diseases [11]. Furthermore, generating sufficient labeled data representing the full spectrum of human disease, biological, and technical variability inherent to morphology, tissue processing, staining, and slide scanners has proven logistically and financially challenging.

Addressing these problems, foundation models have quickly gained traction in the pathology domain based on their promise to achieve robust and generalizable data representations by incorporating the diversity found in pathology through large-scale self-supervised training. The generalizability and robustness of foundation models are particularly relevant to the performance of downstream tasks in pathology—a domain that has both high variation in input data and tasks, such as disease diagnoses, outcome prediction, detection of morphological structures, and quantification of biomarkers.

In this report, we utilized a corpus of 1.2 million histopathology whole slide images derived from more than 490,000 cases from Mayo Clinic and Charité - Universitätsmedizin Berlin to train a ViT-H/14 pathology foundation model using the training paradigm from RudolfV [10]. Our model incorporates a broad diversity of diseases, staining types, and scanners, and utilizes multiple image magnifications during training. We provide an assessment of its performance compared to other leading models available for testing using twenty-one benchmark datasets assessing a variety of downstream pathology tasks. An overview of model characteristics and results can be found in Figure 1.

2 Related Work

Multiple previous works on histopathology foundation models have developed a variety of models with increasingly large parameter counts and on increasingly expansive pathology datasets curated from public and private sources [10, 42, 13, 41, 34, 7, 9, 26, 17, 28, 32]. Pathology foundation models can be separated into tile- and slide-based models. Tile-based models, which represent the majority of the models, generate embeddings from fixed-sized image tiles derived from gigapixel whole

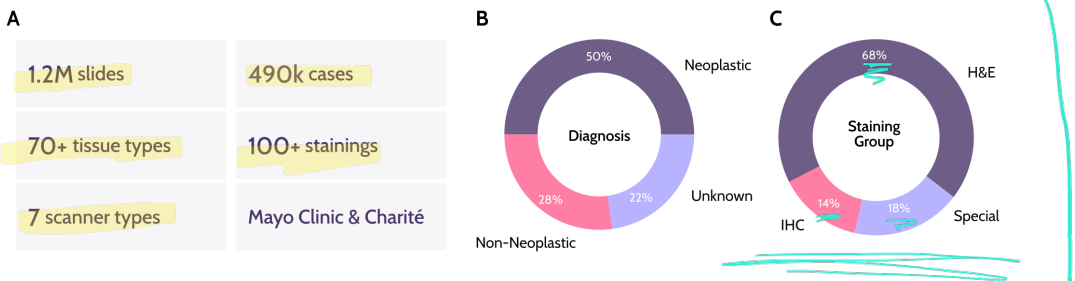

Figure 2: (A) shows the key training dataset statistics. The dataset was derived from 1.2 million pathology slides from 490k cases. The dataset contains data from over 70 tissue/organ types, over 100 different staining types, and 7 scanner types. The data was sourced from Mayo Clinic and Charité - Universitätsmedizin Berlin. (B) shows the distribution of neoplastic vs. non-neoplastic diseases. (C) shows the distribution of the staining groups H&E, IHC, and special stains.

slide images (WSIs). Tile-based models have typically been trained via self-supervised-learning [10, 42, 13, 41, 34, 7, 1, 29, 38, 19, 28] and most studies [10, 42, 13, 41, 34, 7, 1, 29] are based on the DINOv2 framework [30]. We include the leading and available tile-based models in our study, i.e. Virchow2 (632 million parameters; 3.1 million slides; [42]), H-Optimus (1.1 billion parameters; 500k slides; [34]), RudolfV (300 million parameters; 134k slides; [10]), UNI (300 million parameters; 100k slides; [7]), and Phikon 2 (300 million parameters; 58k slides; [13]).

Slide-based foundation models seek to create global representations of whole slide images, which requires encoding features that may span individual tiles or otherwise represent global phenomena not readily evident on single tiles. To address this, most current slide-based models operate in two steps [9, 26, 39, 41, 36]. First, by encoding individual tiles and secondly by aggregating the resulting tile representations into a slide-level representation. For such models, the tile-encoder is a key performance gateway. Accordingly, this study compares the tile-based encoder of Prov-GigaPath (1.1 billion parameters; 171k slides; [41]).

To date, many foundation models in pathology have been trained exclusively on hematoxylin and eosin (H&E) stained slides with tiles extracted at a single magnification level. While H&E staining represents the bulk of routine pathology, alternative histochemical and immunohistochemical (IHC) stains are the basis for many biomarker evaluations and accordingly have been increasingly represented in recent studies [42, 10, 41]. Similarly, data from multiple image magnifications have been included in recent work [42, 1]. The presented model incorporates these features, and uses data obtained from H&E, IHC, and special stains, as well as multiple magnifications.

3 Data and Methods

3.1 Dataset and Preprocessing

A curated set of 1.2 million de-identified WSIs, derived from 490k pathology cases, from the digital archives of Mayo Clinic and Charité - Universitätsmedizin Berlin, was utilized to generate 3.4 billion image tiles for training. Tiles were extracted at multiple resolutions, namely 0.25, 0.5, 1.0, and 2.0 microns per pixel, corresponding to objective microscopic magnifications of 40×, 20×, 10×, and 5×, respectively. An overview of relevant dataset statistics are given in Figure 2. Slides and tiles were sampled for training using the sampling algorithm of RudolfV [10].

3.2 Model Framework and Compute Environment

A sampled dataset of ca. 520m tiles was used to train a ViT-H/14 (632 million parameters) model [12] using an adapted RudolfV [10] approach, which is based on the DinoV2 framework [30]. Model training was performed with Nvidia H100 GPUs within the Mayo Clinic Platform environment[1].

[1] https://www.mayoclinicplatform.org/

3.3 Evaluation Protocols

We evaluated model performance using linear probing protocols as established in the literature [10, 34, 42], with all models evaluated on extracted embeddings from frozen backbones. We use both public benchmarks as well as public evaluation frameworks where available, to foster reproducibility and comparability. Results were computed for 5 seeds over data split, shuffling, and "probing" model initialization per foundation model and task, if not stated otherwise in a specific task description (see Appendix A.1). We report the mean performance and standard errors over the seeds. Dataset splits are described in detail in the respective task descriptions in Appendix A.1. No augmentations were applied when extracting embeddings. As all models use Vision Transformer architectures [12], we always evaluate both the CLS token and the CLS+Mean[2] token embeddings for every model and report the better (maximum) performance of the two in Table 1. This accounts for potential systematic differences between information encoded in different tokens between different models, as also done in previous works (see e.g. [42]). Results of using the CLS and CLS+Mean token only, are reported in Table 4 and Table 5, respectively.

Patch-level classification For patch-level classification tasks, which make up the majority of benchmark datasets, linear probing (LP) is the default protocol. Balanced accuracy was utilized as the performance metric for all patch-level classification tasks.

We use *eva* [14], an open-source evaluation framework for pathology foundation models, for LP evaluation where available[3]. This includes the BACH [3], CRC-100k [20], MHIST [40], and PCAM [37] datasets. Linear classification on extracted embeddings in *eva* is done by training a single-layer neural network with a batch size of 256 patches using stochastic gradient descent (SGD) with a cosine learning rate schedule and base learning rate of 0.0003 for a total of 12.5k iterations.

For patch-level classification tasks not implemented in *eva*, we use an internal LP evaluation framework. These include the MSI CRC (patch), MSI STAD (patch), Pan-cancer TIL, TCGA Uniform (10x), and TCGA Uniform (20x) datasets. In the internal framework, we perform LP on extracted embeddings using `scikit-learn`'s Logistic Regression [31] with balanced class weights. The L2 regularization parameter is chosen by performing a cross-validated grid search over 15 different values between 10^{-8} and 10^{4}. Using the best parameter, a final model is fit and applied to the test set.

Slide-level classification For all slide-level classification tasks, we use the *eva* [14] framework, which applies the default Attention-based Multiple Instance Learning (ABMIL) [15] protocol. Here, an ABMIL head with ReLU activations is trained with a batch size of 32 slides using the AdamW optimizer [27] with a cosine learning rate schedule and base learning rate of 0.001 for a total of 12.5k iterations. Balanced accuracy was utilized as a performance metric for all slide-level classification tasks.

Patch-level regression The HEST-Benchmark [16] is composed of tasks for gene expression prediction and designed as multivariate regression. We follow and use the default evaluation protocol as recommended and implemented[4] by the authors. We provide respective details in the description of the HEST-Benchmark in Appendix A.1.

4 Results

The following analysis is based on 21 public benchmark datasets from two public foundation model evaluation frameworks *eva* [14] and HEST [16] as well as additional tumor-micro-environment (TME) and cancer typing benchmark datasets. The tasks range from TME tissue- and cell-typing over identifying morphological patterns, identifying cancer subtypes, to classifying molecular mutations. The task descriptions and evaluation protocols are detailed in Appendix A.1 and Section 3.3, respectively.

Our model achieves an average performance score of 61.9%, a 1.1 p.p. improvement over the two closest contenders, Virchow2 [42] and H-Optimus-0 [34] (see Table 1). It displayed the highest

[2] CLS+Mean being the concatenation of the CLS token and the mean over all "mini-patch" tokens

[3] Available at `https://github.com/kaiko-ai/eva`, accessed on Oct 31, 2024 (version: 0.1.3)

[4] Available at `https://github.com/mahmoodlab/HEST`, accessed git commit '5a0cbba'.

Table 1: Overview of results on different benchmark datasets. The benchmark datasets are split into morphology- and molecular-related benchmarks. The metric for the first 10 benchmark datasets is Pearson correlation, and the metric for the other benchmark datasets is balanced accuracy. Higher values are better, the highest value per row is bold, and the second-highest value is underlined. The evaluation protocols and descriptions for each benchmark dataset can be found in the methods 3.3 and appendix A.1 sections, respectively.

Group	Benchmark	Phikon v2 [13]	UNI [7]	Gigapath [41]	RudolfV [10]	Virchow2 [42]	H-optimus-0 [34]	Ours
Molecular-related	HEST-COAD	25.6	26.2	30.7	**31.0**	25.9	30.9	29.4
	HEST-HCC	7.8	8.3	7.1	9.4	9.6	8.4	**10.7**
	HEST-IDC	56.6	58.5	56.8	57.4	59.3	**61.0**	60.4
	HEST-LUAD	54.8	55.2	55.8	57.7	56.9	57.3	**58.0**
	HEST-LYMPH_IDC	24.8	25.8	25.1	25.6	25.9	**26.8**	26.4
	HEST-PAAD	47.9	48.8	49.5	51.1	47.3	50.9	**51.8**
	HEST-PRAD	37.7	32.2	38.4	37.7	35.1	**38.5**	38.4
	HEST-READ	18.5	18.4	19.6	19.9	21.1	**24.1**	22.8
	HEST-SKCM	58.4	63.5	58.8	61.8	63.7	**66.1**	62.5
	HEST-ccRCC	27.3	25.3	24.9	25.3	27.4	29.0	**29.4**
	MSI CRC (patch)	68.8	69.5	70.4	69.9	**74.0**	71.2	73.6
	MSI STAD (patch)	71.2	70.5	71.0	74.1	74.8	73.6	**76.0**
	Molecular-Average	41.6	41.8	42.3	43.4	43.4	44.8	**44.9**
Morphology-related	Pan-cancer TIL	92.9	92.6	92.3	92.6	**93.1**	93.0	93.0
	TCGA Uniform (10x)	64.0	68.6	69.1	70.6	**73.0**	70.4	71.8
	TCGA Uniform (20x)	69.8	67.8	68.0	**78.1**	71.5	72.4	67.8
	BACH	73.8	80.1	80.2	76.9	88.7	75.8	**93.1**
	CRC-100k	95.5	95.4	95.9	96.0	96.7	96.2	**97.1**
	MHIST	78.4	84.4	83.1	80.5	85.9	85.0	**86.4**
	PCAM	90.0	93.6	94.5	94.6	93.9	94.3	**94.9**
	CAMELYON16	79.8	85.0	82.1	77.1	86.5	84.0	**86.8**
	PANDA	65.3	69.6	69.6	69.6	66.4	68.0	**70.5**
	Morphology-Average	78.8	81.9	81.6	81.8	84.0	82.1	**84.6**
	Overall Average	57.6	59.0	59.2	59.9	60.8	60.8	**61.9**

performance on 6 out of 9 morphology-related tasks with Virchow2 being the closest contender performing best on 2 tasks. For the molecular-related biomarker tasks H-Optimus and our model perform both best on 5 out of 12 tasks. Overall, our model displayed the best performance on 11 of 21 assessed tasks across both molecular- and morphology-related tasks. In 7 of the 10 benchmarks in which ours was not the leading model it was the second best model by performance. The model displayed below-average performance for a single benchmark, TCGA Uniform (20×), where ours and UNI were the poorest performers; interestingly, our performance on the 10× version of that benchmark was second only to Virchow2.

5 Discussion

The presented model showed consistently good results across a diverse set of benchmarks covering molecular and morphology related tasks. Related work using a larger data base [42] or more model parameters [42, 34, 41] suggests that scaling data and model size even further might yield additional improvements.

To accurately assess the quality, generalization, and robustness of pathology foundation models, it is essential to utilize a wide range of (public) benchmark datasets, covering the entire spectrum of diseases, morphologies, scanners, etc. Despite first standardized frameworks such as *eva* [14] and HEST [16] being available, foundation model development would benefit from a larger and more diverse pool of benchmark datasets. Access to such diverse datasets would provide deeper insights into model performance and robustness, thereby enhancing our understanding of the capabilities and limitations of pathology foundation models.

Acknowledgments

We would like to thank the teams at Aignostics, Charité - Universitätsmedizin Berlin - Pathology Department, Mayo Clinical Platform, Mayo Clinic Generative Artificial Intelligence Program, Mayo Clinic Department of Laboratory Medicine and Pathology, and Mayo Clinic Digital Pathology for supporting this work. Special thanks goes to Alexander Baxendale, Amelie Froessl, Amin Abbasloo, Barbara Feulner, Carl Andersson, Dharma Indurthy, Erinç Argımak, Fabian Spieß, Gerrit Erdmann, Jay Tolley, Jeannine Korp, Jeff Anderson, Jennifer Flores, Joshua Pankratz, Mark Ibrahim, Olja Smilić, Rob Blundo, Tom Lehmann, Sara Then, Steele Clifton-Berry, and Valentin François for their organizational, technical, and other support.

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 for Education and Research (BMBF) under Grants 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] Shahira Abousamra, Rajarsi Gupta, Le Hou, Rebecca Batiste, Tianhao Zhao, Anand Shankar, Arvind Rao, Chao Chen, Dimitris Samaras, Tahsin Kurc, et al. Deep learning-based mapping of tumor infiltrating lymphocytes in whole slide images of 23 types of cancer. *Frontiers in oncology*, 11:806603, 2022.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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, Drew F. K. Williamson, Bowen Chen, Cristina Almagro-Perez, Paul Doucet, Sharifa Sahai, Chengkuan Chen, Daisuke Komura, Akihiro Kawabe, Shumpei Ishikawa, Georg Gerber, Tingying Peng, Long Phi Le, and Faisal Mahmood. Multimodal whole slide foundation model for pathology, November 2024. arXiv:2411.19666.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[28] 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.

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

[30] 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.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[32] 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. From pixels to gigapixels: Bridging local inductive bias and long-range dependencies with pixel-mamba, 2024. arxiv:2412.16711.

[33] 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.

[34] 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.

[35] Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca Batiste, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, 23(1):181–193, 2018.

[36] 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.

[37] 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.

[38] 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, October 2022.

[39] 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*, pages 1–9, September 2024.

[40] 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.

[41] 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.

[42] 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 [16] 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 used the default Ridge Regression with PCA (256 factors) evaluation protocol to solve the multivariate regression on extracted embeddings [16]. Specifically, the 10 tasks are 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 [16]. Performance is evaluated using the Pearson correlation between the predicted and measured gene expression and reported results are the mean and standard deviation across folds.

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 [22, 18]. 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 [22, 18]. 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.

Pan-cancer TIL The pan-cancer tumor-infiltrating lymphocytes (TIL) detection dataset contains 304,097 H&E images (100×100 pixels at 20× magnification) extracted from tissue sample scans of 23 different cancer types from TCGA [2, 35]. The task is to classify an image into TIL positive vs. negative. An image is labeled positive if it contains at least two TILs. The dataset is split into 209,221 training images and 56,275 test images.

TCGA Uniform (10×) and (20×) The TCGA Uniform dataset [25] 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 generated 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 generated 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 on the five-fold cross-validation evaluation.

BACH The BACH dataset comprises 400 H&E microscopy images (2048×1536 pixels at 20× magnification) of breast cancer biopsies. It originates from the ICIAR 2018 Grand Challenge on BreAst Cancer Histology images [3]. The classification 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 [21]. 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 9 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 take the original (no-norm) images without Macenko color normalization [20].

MHIST The task of MHIST is to classify images of colorectal polyps into hyperplastic polyps (HPs) vs. sessile serrated adenomas (SSAs) [40]. 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 and 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 [37]. 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 80:10:10 into training, validation, and test sets with no overlap of WSIs/cases between the splits and every split having a 50:50 balance of positive and negative examples. For evaluation, we resize each image to 224x224 pixels. PCam has been derived from the CAMELYON16 Challenge [4].

CAMELYON16 The task of the CAMELYON16 challenge [4] is to classify whole slide images (WSIs) of lymph node tissue sections into having metastatic breast cancer tissue 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 [4]. 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* [14], which randomly samples max. 1,000 patches per slide.

PANDA The PANDA challenge [6] considers 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 use 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* [14], which considers a fewer number of total slides (955 train, 477 validation, 477 test) as well as fewer randomly sampled patches per slide (200).

Table 2: Summary of benchmark datasets and evaluation frameworks.

Dataset	Pathological Task	ML Task	Input Type	Size	Implementation
HEST-Benchmark	Gene Expression Prediction	Regression	Patch	72	HEST
MSI CRC (patch)	Microsatellite Instability Prediction	Classification (binary)	Patch	173,630	Internal
MSI STAD (patch)	Microsatellite Instability Prediction	Classification (binary)	Patch	198,464	Internal
Pan-cancer TIL	Tumor-Infiltrating Lymphocytes Detection	Classification (binary)	Patch	304,097	Internal
TCGA Uniform (10×)	Cancer Subtyping	Classification	Patch	264,110	Internal
TCGA Uniform (20×)	Cancer Subtyping	Classification	Patch	271,700	Internal
BACH	Breast Cancer Classification	Classification	Patch	400	*eva*
MHIST	Colorectal Polyp Classification	Classification	Patch	3,152	*eva*
PCAM	Metastasis Detection	Classification	Patch	327,680	*eva*
CRC-100k	Tissue Classification	Classification	Patch	107,180	*eva*
CAMELYON16	Metastasis Detection	Classification	WSI	399	*eva*
PANDA (small)	Tumor Grading	Classification	WSI	1909	*eva*

A.2 Additional Results

Table 3: Model results per task, maximum over CLS token and CLS+MEAN token. Same table as Table 1 but with standard deviation over the 5 data splits.

Group	Benchmark	Phikon v2 [13]	UNI [7]	Gigapath [41]	RudolfV [10]	Virchow2 [42]	H-optimus-0 [34]	Ours
Molecular-related	HEST-COAD	25.6 ±2.3	26.2 ±3.1	30.7 ±0.0	**31.0** ±2.0	25.9 ±1.6	30.9 ±0.0	29.4 ±1.5
	HEST-HCC	7.8 ±1.2	8.3 ±0.5	7.1 ±1.3	9.4 ±1.7	9.6 ±1.0	8.4 ±1.2	**10.7** ±1.9
	HEST-IDC	56.6 ±7.8	58.5 ±7.7	56.8 ±7.6	57.4 ±8.5	59.3 ±8.5	**61.0** ±8.1	60.4 ±8.3
	HEST-LUAD	54.8 ±2.3	55.2 ±2.2	55.8 ±2.9	57.7 ±1.8	56.9 ±1.7	57.3 ±2.7	**58.0** ±1.5
	HEST-LYMPH_IDC	24.8 ±4.9	25.8 ±4.1	25.1 ±4.2	25.6 ±3.3	25.9 ±3.3	**26.8** ±4.0	26.4 ±4.4
	HEST-PAAD	47.9 ±7.0	48.8 ±5.8	49.5 ±5.8	51.1 ±7.8	47.3 ±6.9	50.9 ±4.3	**51.8** ±7.4
	HEST-PRAD	37.7 ±0.1	32.2 ±8.1	38.4 ±3.1	37.7 ±2.6	35.1 ±3.4	**38.5** ±0.0	38.4 ±0.7
	HEST-READ	18.5 ±5.9	18.4 ±4.9	19.6 ±6.2	19.9 ±6.8	21.1 ±5.0	**24.1** ±2.7	22.8 ±3.1
	HEST-SKCM	58.4 ±6.2	63.5 ±3.6	58.8 ±5.8	61.8 ±4.6	63.7 ±3.1	**66.1** ±5.8	62.5 ±2.4
	HEST-ccRCC	27.3 ±4.0	25.3 ±3.8	24.9 ±4.1	25.3 ±5.2	27.4 ±4.5	29.0 ±3.8	**29.4** ±5.5
	MSI CRC (patch)	68.8 ±0.1	69.5 ±0.0	70.4 ±0.1	69.9 ±0.1	**74.0** ±0.0	71.2 ±0.1	73.6 ±0.0
	MSI STAD (patch)	71.2 ±0.1	70.5 ±0.0	71.0 ±0.1	74.1 ±0.1	74.8 ±0.2	73.6 ±0.0	**76.0** ±0.1
	Molecular-Average	41.6	41.8	42.3	43.4	43.4	44.8	**44.9**
Morphology-related	Pan-cancer TIL	92.9 ±0.0	92.6 ±0.0	92.3 ±0.0	92.6 ±0.1	**93.1** ±0.1	93.0 ±0.1	93.0 ±0.0
	TCGA Uniform (10x)	64.0 ±0.0	68.6 ±0.0	69.1 ±0.0	70.6 ±0.0	**73.0** ±0.0	70.4 ±0.0	71.8 ±0.0
	TCGA Uniform (20x)	69.8 ±0.0	67.8 ±0.0	68.0 ±0.0	**78.1** ±0.0	71.5 ±0.0	72.4 ±0.0	67.8 ±0.0
	BACH	73.8 ±0.3	80.1 ±1.0	80.2 ±0.3	76.9 ±0.3	88.7 ±0.3	75.8 ±1.1	**93.1** ±0.2
	CRC-100k	95.5 ±0.0	95.4 ±0.2	95.9 ±0.0	96.0 ±0.1	96.7 ±0.1	96.2 ±0.1	**97.1** ±0.1
	MHIST	78.4 ±0.4	84.4 ±0.1	83.1 ±0.1	80.5 ±0.0	85.9 ±0.1	85.0 ±0.2	**86.4** ±0.1
	PCAM	90.0 ±0.1	93.6 ±0.1	94.5 ±0.0	94.6 ±0.1	93.9 ±0.1	94.3 ±0.1	**94.9** ±0.0
	CAMELYON16	79.8 ±1.4	85.0 ±0.5	82.1 ±0.7	77.1 ±1.7	86.5 ±0.2	84.0 ±0.8	**86.8** ±0.4
	PANDA	65.3 ±0.4	69.6 ±0.6	69.6 ±0.7	69.6 ±0.4	66.4 ±1.1	68.0 ±0.6	**70.5** ±0.5
	Morphology-Average	78.8	81.9	81.6	81.8	84.0	82.1	**84.6**
	Overall Average	57.6	59.0	59.2	59.9	60.8	60.8	**61.9**

Table 4: Model results per task, CLS token only.

Group	Benchmark	Phikon v2 [13]	UNI [7]	Gigapath [41]	RudolfV [10]	Virchow2 [42]	H-optimus-0 [34]	Ours
Molecular-related	HEST-COAD	25.0 ±1.7	26.2 ±3.1	29.9 ±2.1	23.7 ±6.0	25.9 ±1.6	**30.9** ±0.0	25.9 ±3.1
	HEST-HCC	6.7 ±1.3	7.8 ±0.2	7.1 ±1.3	6.5 ±0.1	8.2 ±1.0	7.9 ±0.6	**9.4** ±0.8
	HEST-IDC	54.1 ±7.7	57.4 ±7.9	55.1 ±7.3	54.5 ±8.9	59.2 ±8.0	**59.8** ±8.5	59.6 ±8.1
	HEST-LUAD	54.2 ±1.1	54.6 ±2.2	54.1 ±3.6	55.5 ±1.2	55.3 ±1.7	55.9 ±3.3	**57.0** ±1.7
	HEST-LYMPH_IDC	24.4 ±4.6	25.6 ±4.4	25.0 ±5.0	24.3 ±3.5	25.5 ±2.6	**25.9** ±4.0	25.7 ±4.7
	HEST-PAAD	44.5 ±6.6	48.1 ±7.0	47.5 ±4.8	45.7 ±5.8	47.2 ±6.5	49.1 ±4.0	**50.7** ±7.2
	HEST-PRAD	35.4 ±1.5	29.4 ±8.5	37.0 ±2.1	37.0 ±2.8	34.8 ±3.1	**38.5** ±0.0	35.3 ±3.2
	HEST-READ	17.5 ±5.9	18.4 ±4.9	19.6 ±6.2	17.6 ±8.1	20.9 ±5.0	**22.2** ±4.8	21.3 ±2.9
	HEST-SKCM	55.5 ±3.6	63.5 ±3.6	56.1 ±6.2	58.0 ±4.1	61.9 ±2.8	**64.5** ±6.2	56.2 ±0.5
	HEST-ccRCC	26.6 ±3.8	24.0 ±4.0	24.3 ±3.3	24.9 ±5.4	27.4 ±4.5	26.8 ±3.2	**27.8** ±3.6
	MSI CRC (patch)	67.5 ±0.0	69.1 ±0.0	69.0 ±0.1	68.0 ±0.0	71.6 ±0.1	69.7 ±0.0	**71.6** ±0.0
	MSI STAD (patch)	68.6 ±0.0	68.6 ±0.0	67.4 ±0.1	72.4 ±0.0	72.8 ±0.0	72.7 ±0.0	**73.5** ±0.0
Molecular-Average		40.0	41.1	41.0	40.7	42.6	**43.6**	42.8
Morphology-related	Pan-cancer TIL	92.6 ±0.0	92.4 ±0.0	91.8 ±0.0	91.9 ±0.0	92.7 ±0.0	92.6 ±0.0	**92.8** ±0.0
	TCGA Uniform (10x)	63.9 ±0.0	68.3 ±0.0	68.7 ±0.0	70.2 ±0.0	**72.9** ±0.0	69.9 ±0.0	71.8 ±0.0
	TCGA Uniform (20x)	69.8 ±0.0	67.4 ±0.0	67.4 ±0.0	**77.7** ±0.0	71.5 ±0.0	72.1 ±0.0	67.2 ±0.0
	BACH	72.7 ±0.3	79.7 ±0.4	76.1 ±0.4	74.9 ±0.5	88.0 ±0.4	75.8 ±1.1	**93.1** ±0.2
	CRC-100k	93.9 ±0.0	94.8 ±0.1	95.2 ±0.1	94.8 ±0.1	96.6 ±0.1	95.8 ±0.1	**97.0** ±0.0
	MHIST	77.5 ±0.1	84.4 ±0.1	82.9 ±0.1	79.8 ±0.1	**85.8** ±0.2	83.9 ±0.1	85.2 ±0.1
	PCAM	89.3 ±0.0	93.6 ±0.1	94.5 ±0.0	94.2 ±0.1	93.6 ±0.1	94.2 ±0.1	**94.9** ±0.0
	CAMELYON16	79.8 ±1.4	84.9 ±0.8	82.1 ±0.7	77.1 ±1.7	86.5 ±0.2	83.2 ±1.4	**86.8** ±0.4
	PANDA	64.3 ±0.4	69.6 ±0.6	66.1 ±0.3	68.2 ±0.7	65.1 ±0.9	67.2 ±0.4	**70.0** ±0.5
Morphology-Average		78.2	81.7	80.5	81.0	83.6	81.6	**84.3**
Overall Average		56.4	58.5	57.9	57.9	60.2	59.9	**60.6**

Table 5: Model results per task, CLS+MEAN token only.

Group	Benchmark	Phikon v2 [13]	UNI [7]	Gigapath [41]	RudolfV [10]	Virchow2 [42]	H-optimus-0 [34]	Ours
Molecular-related	HEST-COAD	25.6 ±2.3	25.6 ±4.7	30.7 ±0.0	**31.0** ±2.0	25.6 ±3.2	30.6 ±0.3	29.4 ±1.5
	HEST-HCC	7.8 ±1.2	8.3 ±0.5	6.8 ±0.4	9.4 ±1.7	9.6 ±1.0	8.4 ±1.2	**10.7** ±1.9
	HEST-IDC	56.6 ±7.8	58.5 ±7.7	56.8 ±7.6	57.4 ±8.5	59.3 ±8.5	**61.0** ±8.1	60.4 ±8.3
	HEST-LUAD	54.8 ±2.3	55.2 ±2.2	55.8 ±2.9	57.7 ±1.8	56.9 ±1.7	57.3 ±2.7	**58.0** ±1.5
	HEST-LYMPH_IDC	24.8 ±4.9	25.8 ±4.1	25.1 ±4.2	25.6 ±3.3	25.9 ±3.3	**26.8** ±4.0	26.4 ±4.4
	HEST-PAAD	47.9 ±7.0	48.8 ±5.8	49.5 ±5.8	51.1 ±7.8	47.3 ±6.9	50.9 ±4.3	**51.8** ±7.4
	HEST-PRAD	37.7 ±0.1	32.2 ±8.1	**38.4** ±3.1	37.7 ±2.6	35.1 ±3.4	36.1 ±2.0	38.4 ±0.7
	HEST-READ	18.5 ±5.9	17.4 ±6.3	18.8 ±6.3	19.9 ±6.8	21.1 ±5.0	**24.1** ±2.7	22.8 ±3.1
	HEST-SKCM	58.4 ±6.2	63.0 ±2.4	58.8 ±5.8	61.8 ±4.6	63.7 ±3.1	**66.1** ±5.8	62.5 ±2.4
	HEST-ccRCC	27.3 ±4.0	25.3 ±3.8	24.9 ±4.1	25.3 ±5.2	27.4 ±5.4	29.0 ±3.8	**29.4** ±5.5
	MSI CRC (patch)	68.8 ±0.1	69.5 ±0.0	70.4 ±0.1	69.9 ±0.1	**74.0** ±0.0	71.2 ±0.1	73.6 ±0.0
	MSI STAD (patch)	71.2 ±0.1	70.5 ±0.0	71.0 ±0.1	74.1 ±0.1	74.8 ±0.2	73.6 ±0.0	**76.0** ±0.1
Molecular-Average		41.6	41.7	42.3	43.4	43.4	44.6	**44.9**
Morphology-related	Pan-cancer TIL	92.9 ±0.0	92.6 ±0.0	92.3 ±0.0	92.6 ±0.1	**93.1** ±0.1	93.0 ±0.1	93.0 ±0.0
	TCGA Uniform (10x)	64.0 ±0.0	68.6 ±0.0	69.1 ±0.0	70.6 ±0.0	**73.0** ±0.0	70.4 ±0.0	71.7 ±0.0
	TCGA Uniform (20x)	69.4 ±0.0	67.8 ±0.0	68.0 ±0.0	**78.1** ±0.0	71.5 ±0.0	72.4 ±0.0	67.8 ±0.0
	BACH	73.8 ±0.3	80.1 ±1.0	80.2 ±0.3	76.9 ±0.3	88.7 ±0.3	74.2 ±0.5	**92.5** ±0.5
	CRC-100k	95.5 ±0.0	95.4 ±0.2	95.9 ±0.0	96.0 ±0.1	96.7 ±0.1	96.2 ±0.1	**97.1** ±0.1
	MHIST	78.4 ±0.4	84.0 ±0.1	83.1 ±0.1	80.5 ±0.0	85.9 ±0.1	85.0 ±0.2	**86.4** ±0.1
	PCAM	90.0 ±0.1	93.6 ±0.1	94.3 ±0.1	94.6 ±0.1	93.9 ±0.1	94.3 ±0.1	**94.8** ±0.0
	CAMELYON16	79.1 ±1.2	85.0 ±0.5	80.6 ±0.7	76.3 ±0.5	86.0 ±0.3	84.0 ±0.8	**86.7** ±0.6
	PANDA	65.3 ±0.4	69.0 ±0.5	69.6 ±0.7	69.6 ±0.4	66.4 ±1.1	68.0 ±0.6	**70.5** ±0.5
Morphology-Average		78.7	81.8	81.5	81.7	83.9	81.9	**84.5**
Overall Average		57.5	58.9	59.1	59.8	60.8	60.6	**61.9**