PathoCellBench: A Comprehensive Benchmark for Cell Phenotyping

Jérôme Lüscher[*,4,5], Nora Koreuber[*,1,4,5], Jannik Franzen[*,1,3,4,5], Fabian H. Reith[*,1,2,4,5], Claudia Winklmayr[*,4,5], Elias Baumann[6], Christian M. Schürch[7,8,9], Dagmar Kainmüller[†,3,4,5], Josef Lorenz Rumberger[†,2,4,5,✉]

[1] Charité Universitätsmedizin, Berlin, Germany; [2] Humboldt-Universität zu Berlin, Berlin, Germany; [3] Universität Potsdam, Potsdam, Germany; [4] Helmholtz Imaging; [5] Max-Delbrück Center, Berlin, Germany; [6] Institute of Tissue Medicine and Pathology, University of Bern, Bern, Switzerland; [7] Department of Pathology and Neuropathology, University Hospital and Comprehensive Cancer Center Tübingen, Tübingen, Germany; [8] Cluster of Excellence iFIT (EXC 2180) "Image-Guided and Functionally Instructed Tumor Therapies", University of Tübingen, Germany
[*, †] authors contributed equally; author order was determined at random,
[✉]{firstnames.lastname}@mdc-berlin.de

Abstract. Digital pathology has seen the advent of a wealth of foundational models (FMs), yet to date their performance on cell phenotyping has not been benchmarked in a unified manner. We therefore propose *PathoCellBench*: A comprehensive benchmark for cell phenotyping on Hematoxylin and Eosin (H&E) stained histopathology images. We provide both *PathoCell*, a new H&E dataset featuring 14 cell types identified via multiplexed imaging, and ready-to-use fine-tuning and benchmarking code that allows the systematic evaluation of multiple prominent pathology FMs in terms of dense cell phenotype predictions in a range of generalization scenarios. We perform extensive benchmarking of existing FMs, providing insights into their generalization behavior under technical vs. medical domain shifts. Furthermore, while FMs achieve macro F1 scores > 0.70 on previously established benchmarks such as *Lizard* and *PanNuke*, on *PathoCell*, we observe scores as low as 0.20. This indicates a much more challenging task not captured by previous benchmarks, establishing *PathoCell* as a prime asset for future benchmarking of FMs and supervised models alike. Code and data are available on GitHub.

Keywords: Digital Pathology · Cell Phenotyping · Foundation Models

1 Introduction

Automated cell type classification in Hematoxylin and Eosin (H&E) stained histopathology images is a crucial task in computational pathology, with applications in disease diagnosis, prognosis, and treatment planning [10]. Recent advances in large-scale pre-trained foundation models (FMs) have enabled the development of general-purpose feature extractors demonstrating strong performance across various pathology tasks [2,29,6,25,24]. However, it remains unclear

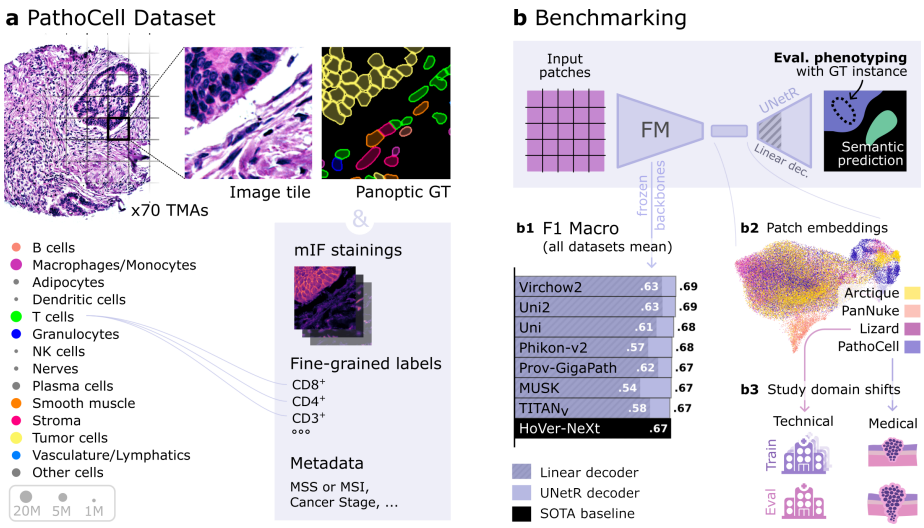

Fig. 1. Overview of the *PathoCell* dataset and FM benchmarking: **(a)** Example image from *PathoCell*, a zoomed-in visualization of a tile and its segmentation mask. Along the pixel-wise panoptic annotations, information such as multiplexed immunofluorescence stainings, pixel-wise labels, or patient data are provided. **(b)** Illustration of the benchmarking pipeline for evaluation of cell phenotyping capabilities of FMs.

to what extent these models outperform supervised baselines for cell type prediction [26]. Prior evaluations have been limited, either focusing on proprietary datasets with restricted access [4] or constrained to a single dataset [2]. Therefore, a systematic benchmark across multiple datasets is needed to determine whether pathology FMs provide an advantage over supervised baselines.

Most existing benchmarking efforts in digital pathology focus on slide- and patch-level prediction tasks [16], whereas datasets for cell phenotyping or segmentation [9,8] are already approaching performance saturation. Furthermore, domain shifts in prior cell-level datasets have been limited to technical factors such as differences in sample preparation, staining protocols, and scanner types [9]. However, tumor and molecular subtypes, as well as different stages of disease, may contribute to a medical domain shift that is currently not captured by existing benchmarks.

To bridge this gap, we (1) publish *PathoCell*, to our knowledge, the largest publicly available dataset for H&E-based cell phenotyping with fine-grained cell type annotations (Fig. 1a), encompassing 14 distinct cell types and 88 million individual cells, much larger than all previous datasets (e.g. *Lizard* [9]: 6 cell types and 495k cells). Our dataset introduces multiple medical domain shifts, capturing variations in tumor subtypes and cancer staging.

(2) We benchmark seven pathology foundation models [2,3,6,24,25,29] and HoVer-NeXt [1], a strong supervised baseline, on the *PathoCell* dataset with three different dataset splits to test generalization performance under domain

shifts (Fig. 1b). We further benchmark on *PanNuke* [8], a pan-cancer dataset, *Lizard* [9], a colon carcinoma dataset, and on the synthetic *Arctique* dataset [7] to evaluate model generalization to a different domain. (3) We test and compare linear probing of ViT patch-level features and a UNetR decoder architecture [12] for dense prediction of cell phenotypes. (4) We publicly release our benchmarking pipeline along with the dataset, enabling more comprehensive benchmarking of FMs in pathology.

2 Datasets

We use four datasets to evaluate cell phenotyping performance: two established H&E datasets, a synthetic dataset and the new *PathoCell* dataset. In selecting the datasets, we made sure that none of them was used in the training procedure of the FMs and that each of them has unique features.

The *Lizard* dataset [9] is one of the most widely used H&E datasets and contains colon tissue histology images with panoptic segmentation annotations generated in a human-in-the-loop fashion. In total, *Lizard* comprises 291 fields of view (FoV) with an average size of 1,016×917 pixels and a total of 495,179 individual nuclei: epithelial cells, connective tissue cells, lymphocytes, plasma cells, neutrophils, and eosinophils. The images in *Lizard* are acquired by six different medical centers, which allows a straightforward domain split (*Center-Split*) where training and validation samples come from five centers, while the test set consists of images from a single center (GlaS) [19].

The *PanNuke* dataset [8] also contains samples from colon tissue which overlap with the *Lizard* dataset but additionally encompasses samples from 18 different tissue types such as breast, liver, or prostate and contains a total of 205,343 nuclei annotations, generated using a human-in-the-loop data annotation pipeline. It consists of 7,901 images, each with a size of 256×256 pixels. The dataset covers epithelial cells, connective tissue cells, inflammatory cells, neoplastic and dead cells. Finally, the *Arctique* dataset [7] is a fully synthetic dataset, constructed in a procedural fashion utilizing 3D rendering. *Arctique* replicates characteristic structures of colon histopathology images and provides exact labels for all cells. We use the standard training split of *Arctique* containing 1450 FoVs of size 512×512 pixel and containing a total of 487,969 cell nuclei modeled after plasma cells, lymphocytes, eosinophils, fibroblasts, and epithelial cells.

2.1 The *PathoCell* Dataset

The image and metadata comprised in *PathoCell* was originally published without segmentation masks as part of a multiplexed imaging study [21]. The dataset was acquired using multiplexed CO-Detection by indEXing (CODEX) imaging of formalin-fixed paraffin-embedded (FFPE) tissue sections from colon carcinoma patients from the University Hospital of Bern. Tissue samples from 35 patients were used to construct two tissue microarray (TMA) slides, each containing 70

distinct samples. These samples were imaged using 56 antibody markers across multiple cycles of staining, imaging, washing, and re-staining. Cell segmentation was performed using the CODEX toolkit segmenter on the DRAQ5 nuclear channel, followed by the calculation of integrated marker expressions. X-shift clustering [20] was applied to these marker expressions to group cells into phenotypic clusters, which were subsequently refined and merged into 28 cell types through manual expert review.

To create a benchmark for cell type classification in H&E-stained histopathology, we collaborated with the authors of the original study to generate high-quality phenotyping annotations. We constructed dense annotation masks from previously unpublished segmentation data and cleaned the dataset for benchmarking H&E-based cell phenotyping. To this end, through a rigorous quality control process, we visually inspected all segmentation masks and excluded 31 FoVs where significant false merges or false negatives were present in the cell masks. Additionally, in consultation with the original authors, we merged cell phenotypes that are deemed too specific to be distinguished in H&E images (e.g., $CD3^+$, $CD4^+$, and $CD8^+$ T cells were expert-consolidated into a single *T cells* class). The final dataset consists of 109 high-resolution FoVs of size 1440×1920 pixels, with a total of 88 million individual cells and the following 14 distinct cell types: B cells, macrophages, nerves, dendritic cells (DC), plasma cells, granulocytes, tumor cells, T cells, stroma, adipocytes, vasculature, smooth muscle, natural killer (NK) cells, and a residual class named "other cells". The exact merges are documented on GitHub.

PathoCell has some caveats, the segmentation masks were created via an automatic segmentation tool, thus their quality is lower than that of other datasets that were manually or semi-manually segmented. Furthermore, the dataset comes from a single center, thus technical variations are not reflected and it exclusively contains images of colon carcinoma tissues.

To enable detailed benchmarking under domain shifts, we provide three different dataset splits. Besides (1) *Base-Split*, a standard (70/15/15) train/validation/test random split, we also provide: (2) *Tumor-Type-Split* (63/16/21), where the training and validation sets contain data from patients with adenocarcinoma, while the test set consists of patients with mucinous adenocarcinoma, introducing a variation in tumor cell morphology and microenvironment. Finally, in (3) *Tumor-Stage-Split* (64/16/20), training and validation data include patients with pTNM stage 3 tumors, while the test set is composed of pTNM stage 4 tumors. This diversity in dataset splits allows for an in-depth analysis of how well FMs generalize across clinically relevant domain shifts, distinguishing this benchmark from prior datasets that split data on the medical center in which it was acquired.

3 Pathology Foundation Models

We benchmark seven pathology foundation models, each trained with self-supervised (SSL) or contrastive learning on large-scale histopathology datasets. While

all models utilize Vision Transformers (ViTs) as backbones, they differ in architecture, training strategies, and dataset composition. Most models employ DINOv2 [17] for SSL. Of these, Uni (ViT-L/16) [2], and Phikon-v2 (ViT-L/16) [6] are trained on a mix of publicly available and proprietary whole slide images (WSIs). Uni2 (ViT-H/14-reg8) [14] expands upon Uni with a larger dataset and model, while Virchow2 (ViT-H/14-reg4) [29] applies domain-specific augmentations and regularization. In contrast, Prov-GigaPath (ViT-G/14) [25] introduces GigaPath, an SSL framework tailored for whole-slide representations, and trains on Prov-Path, one of the largest proprietary WSI datasets. Beyond DINOv2-based models, TITAN_v (ViT-B/16) [3] is trained using iBOT [28] and contrastive captioning [27], integrating vision-language alignment while supporting vision-only representations. MUSK (ViT-L/16) [24] differs from all others by incorporating a BEiT3-based [23] multi-modal transformer with pathology-specific tokenization.

While models vary in their use of proprietary and public datasets, most models rely on extensive internal datasets. Only Phikon-v2 and MUSK exclusively use publicly available datasets. MUSK is trained on PubMed central, TCGA [15], QUILT-1M [13] and PathAsst [22], while Phikon-v2 uses CPTAC [5], TCGA [15] and GTEx [11]. Uni, Uni2 and TITAN_v are trained on GTEx and different subsets of proprietary data from the Mass General Brigham Hospital. For training Virchow2, proprietary data from the Memorial Sloan Kettering Cancer Center was used, and for Prov-GigaPath, data from the Providence Healthcare System was used. Finally, while most encoders provide embeddings for the 14x14 pixel ViT patches, Prov-GigaPath and TITAN_v focus on whole-slide representations.

4 Experiments and Results

We compare the performance of pathology FMs with the supervised HoVer-NeXt baseline [1], evaluating multiple aspects, including FM decoder architecture, generalization under technical and medical domain shifts, and data efficiency. For all experiments, we report the macro F1 score to reflect class imbalances.

The encoder (FM) weights were kept frozen for all experiments. We fine-tuned for 10k steps with early stopping in all FM experiments. We used a learning rate of $3e^{-5}$, with a linear warm-up of 1k steps, followed by a cosine decay.

To assess the impact of different decoding strategies, we fine-tuned FMs with both a simple linear projection head and the UNetR head [12], which incorporates a U-Net-style decoder architecture. A detailed list of training parameters and additional metrics can be accessed via GitHub.

Choice of decoder: The choice of decoder has a significant effect, where the UNetR outperforms the linear projection head by a considerable margin on most FMs and datasets (Fig. 2a). The greatest performance gain is observed for MUSK, where using UNetR improves the F1 score by up to 0.19 on the *Arctique* dataset. Whereas the smallest effect (no improvement) is seen for Uni2 on *PathoCell*. We also implemented the Dense Prediction Transformer (DPT) [18], which yielded very similar results to UNetR: 0.002 F1 decrease for the best

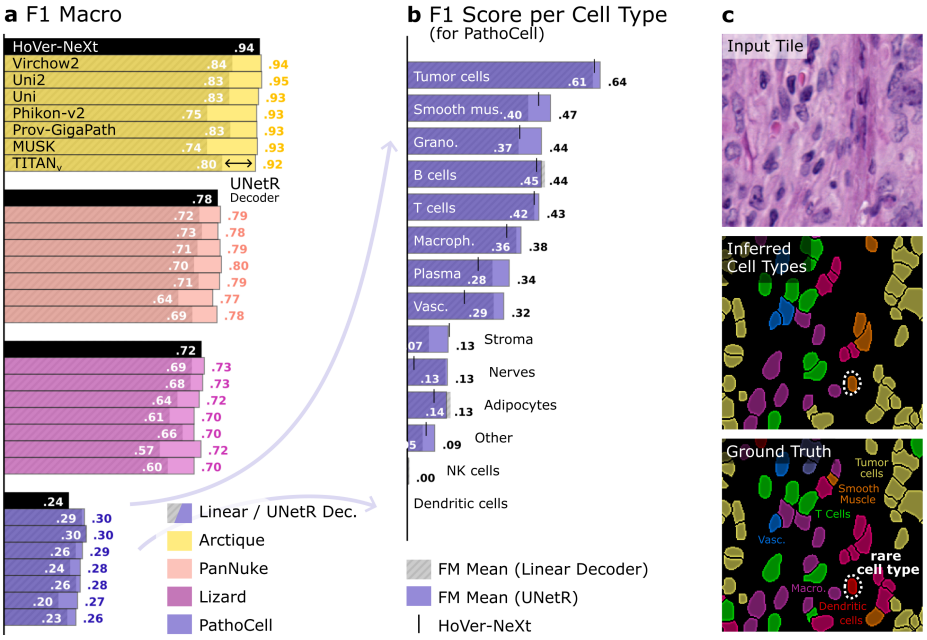

Fig. 2. Overview of benchmarking results on the base-split of each of the four datasets: **(a)** Macro F1 scores of all FMs and HoVer-NeXt, numbers in white indicate results from the linear probe, results from UNetR are in color. **(b)** F1 Scores per cell type for *PathoCell* dataset, highlighting the challenge posed by rare cell types. **(c)** Qualitative example of our predictions and ground truth for a sample with a rare cell type (DC).

model, Uni2, on both *Lizard* and *PathoCell*. This suggests a saturation of the performance of the frozen encoder features. In the following, we report the results for the UNetR head by default.

Performance across datasets: Model performance varies significantly across datasets. As expected, the highest scores are achieved on the synthetic dataset *Arctique*, where models reach an average F1 of 0.93. Performance decreases for *PanNuke*: 0.79 F1, followed by *Lizard*: 0.71 F1, and is lowest for *PathoCell*: 0.28 F1.

To better understand model behavior on *PathoCell*, we analyze the performance across different cell types (Fig. 2b). The dataset is highly imbalanced, with tumor cells being the most abundant, outnumbering the least prevalent cell type, natural killer (NK) cells, by a factor of 153. As a result, classification scores are highly variable, with tumor cells achieving an F1 of 0.65, while NK cells score near zero. These contrasts illustrate the omnipresent challenge of recognizing rare cell types in pathology datasets [19].

Generalization under domain shifts: To assess domain generalization, we evaluate models on dataset splits designed to introduce technical and medical

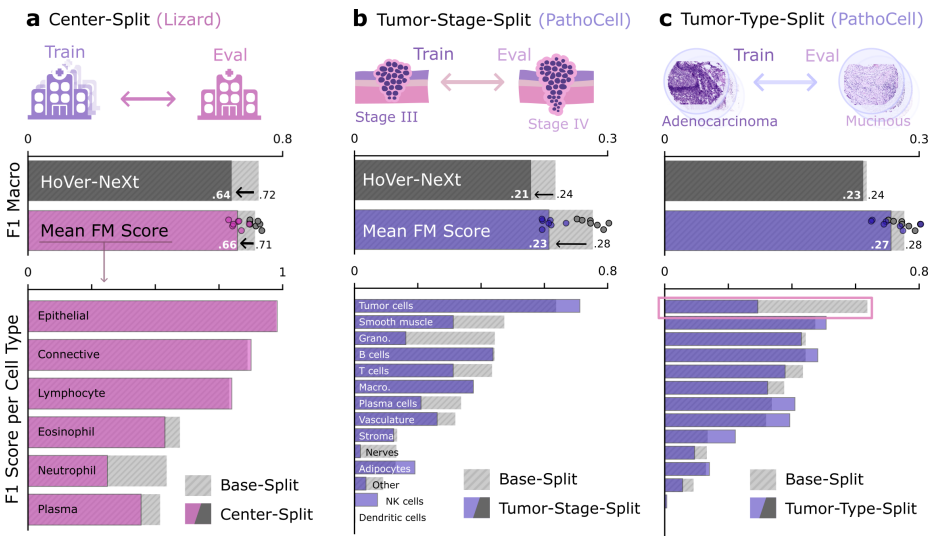

Fig. 3. Generalization of FMs and HoVer-NeXt under technical and medical domain shifts. **Top:** Mean performance on base split vs. domain split. Dots are individual model performance. **Bottom:** Performance by cell type. **(a)** *Lizard Center-Split* a technical domain split by medical centers. **(b)** *PathoCell Tumor-Stage-Split*, a medical split by cancer stages 3 vs. 4. **(c)** *PathoCell Tumor-Type-Split*, a medical split by adenocarcinoma vs. mucinous adenocarcinoma.

domain gaps and compare their results with the random *Base-Split* . Main results are given in Fig. 3.

On *Lizard* we use the *Center-Split*, which captures a technical domain shift primarily driven by variations in sample preparation, staining, and imaging across different medical centers. Under these conditions, F1 for HoVer-NeXt drops by 0.09, whereas FM performance decreases on average by 0.06 F1 (Fig. 3a), with the most robust models Uni2 and Virchow2 losing less than 0.03 F1, and the least robust model TITAN_v dropping by 0.10. Detailed results of the individual FMs can be accessed on GitHub.

On *PathoCell*, we evaluate robustness to domain shifts using the medical data splits *Tumor-Stage-Split* (Fig. 3b) and *Tumor-Type-Split* (Fig. 3c). On *Tumor-Stage-Split*, FMs exhibit a higher absolute F1 score than HoVer-NeXt despite experiencing a larger performance drop when compared to the random *Base-Split* of 0.05 F1 (compared to HoVer-NeXt's decrease of 0.03 F1). On *Tumor-Type-Split*, HoVer-NeXt's performance remains stable, decreasing by only .005 F1, whereas FMs drop by 0.02 F1 on average. Despite this drop, FMs still maintain a higher overall F1 of 0.27 compared to HoVer-NeXt with 0.23 F1 in this setting.

Parameter- and Data Scaling: We plot the Base-Split performance of the fine-tuned FMs against the number of parameters and the amount of data used for their original training (Fig. 4a). Both show a general increase with regard to their respective parameters.

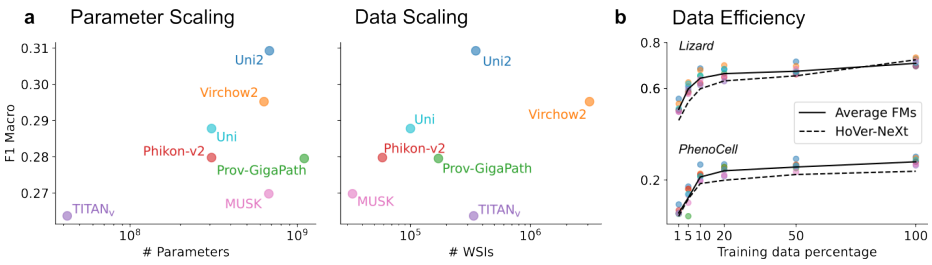

Fig. 4. (a) Performance of FMs with respect to the number of trainable parameters on the left and the number of WSIs in the original training data on the right. **(b)** Performance of FMs with UNetR head and HoVer-NeXt with a reduced finetuning training set size on *Lizard* and *PathoCell*.

To further evaluate the FMs, we conduct a data efficiency analysis (Fig. 4b), by progressively reducing the training set to 50, 20, 10, and 5% of its original size while keeping the validation and test sets unchanged. On *Lizard*, FMs show a clear advantage in low-data regimes, particularly at 5% of the training data, where they outperform HoVer-NeXt by 0.06 F1 on average. This advantage is inversely proportional to the amount of training data used. On *PathoCell*, FMs maintain a consistent advantage across data reductions $\geq 10\%$. Their reduced advantage at 1% and 5% training data sizes may be due to high label imbalance. The most data-efficient model on both *Lizard* and *PathoCell* is Uni2. Detailed results of the individual FMs can be accessed on GitHub.

5 Discussion and Conclusion

Our experiments show that while FMs generally achieve competitive results in settings without domain gap (Fig. 2), their advantage over the HoVer-NeXt baseline remains limited. The choice of decoder plays a crucial role, with the UNetR decoder consistently outperforming linear projection heads, due to its ability to better recover spatial details.

Class imbalance remains a challenge, particularly for rare cell types, where FMs show slightly improved recognition but still struggle with morphologically ambiguous classes such as dendritic and NK cells. Under technical domain shifts, FMs exhibit greater robustness than the supervised baseline (Fig. 3a). However, FMs were more sensitive to medical domain shifts than the baseline (Fig. 3b-c), albeit maintaining higher absolute performance. In low-data regimes, FMs consistently outperform the supervised baseline for the *Lizard* and *PathoCell* dataset (Fig. 4b).

Among the evaluated models, approaches based on DINOv2 demonstrate strong scaling properties, while MUSK and TITAN$_v$ which were trained with alternative strategies, such as iBOT or BEiT-style masked consistency learning, perform less favorably. However, they were also trained with a comparatively small dataset (MUSK) or with less model capacity (TITAN$_v$) and on

different datasets, which limits the interpretability of observed scaling trends. Dataset quality imposes an upper bound on performance, as both *PathoCell* and Lizard [9,1] contain label noise. Finally, the reliance of many FMs on unpublished and proprietary training data limits the ability to analyze their scaling behavior.

This benchmark establishes a foundation for the evaluation of FM for cell phenotyping in digital pathology and underscores the need for more diverse, high-quality data sets with precise annotations from multiple institutions. Future work should focus on expanding *PathoCell* with additional tissue types, improving annotation accuracy, and refining evaluation methodologies to study the representation space more in detail.

Acknowledgments. Funding: DFG Research Unit DeSBi (KI-FOR 5363, project no. 459422098), (DFG) Research Training Group CompCancer (RTG2424), Synergy Unit of the Helmholtz Foundation Model Initiative, Helmholtz Einstein International Berlin Research School In Data Science (HEIBRiDS), Swiss National Science Foundation (CRSII5_193832).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baumann, E., Dislich, B., Rumberger, J.L., Nagtegaal, I.D., Martinez, M.R., Zlobec, I.: Hover-next: A fast nuclei segmentation and classification pipeline for next generation histopathology. In: Medical Imaging with Deep Learning (2024)
2. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L.L., Wang, J.J., Vaidya, A., Le, L.P., Gerber, G., Sahai, S., Williams, W., Mahmood, F.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
3. Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A., Jaume, G., Shaban, M., Kim, A., Williamson, D.F.K., Chen, B., Almagro-Perez, C., Doucet, P., Sahai, S., Chen, C., Komura, D., Kawabe, A., Ishikawa, S., Gerber, G.K., Peng, T., Le, L.P., Mahmood, F.: Multimodal whole slide foundation model for pathology. arXiv preprint arXiv:2003.10778 (2024)
4. Dippel, J., Feulner, B., Winterhoff, T., Milbich, T., Tietz, S., Schallenberg, S., Dernbach, G., Kunft, A., Heinke, S., Eich, M.L., et al.: Rudolfv: a foundation model by pathologists for pathologists. arXiv preprint arXiv:2401.04079 (2024)
5. Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Jacob, S., Madhavan, S., Ketchum, K.A.: The CPTAC Data Portal: A Resource for Cancer Proteomics Research. Journal of Proteome Research **14**(6), 2707–2713 (2015)
6. Filiot, A., Jacob, P., Kain, A.M., Saillard, C.: Phikon-v2, a large and public feature extractor for biomarker prediction. arXiv preprint arXiv:2409.09173 (2024)
7. Franzen, J., Winklmayr, C., Guarino, V.E., Karg, C., Yu, X., Koreuber, N., Albrecht, J., Bischoff, P., Kainmueller, D.: Arctique: An artificial histopathological dataset unifying realism and controllability for uncertainty quantification. Advances in Neural Information Processing Systems **37**, 71855–71867 (2025)

8. Gamper, J., Koohbanani, N.A., Benes, K., Graham, S., Jahanifar, M., Khurram, S.A., Azam, A., Hewitt, K., Rajpoot, N.: Pannuke dataset extension, insights and baselines. arXiv preprint arXiv:2003.10778 (2020)
9. Graham, S., Jahanifar, M., Azam, A., Nimir, M., Tsang, Y.W., Dodd, K., Hero, E., Sahota, H., Tank, A., Benes, K., et al.: Lizard: A large-scale dataset for colonic nuclear instance segmentation and classification. In: IEEE/CVF international conference on computer vision. pp. 684–693 (2021)
10. Graham, S., Vu, Q.D., Jahanifar, M., Weigert, M., Schmidt, U., Zhang, W., Zhang, J., Yang, S., Xiang, J., Wang, X., et al.: Conic challenge: Pushing the frontiers of nuclear detection, segmentation, classification and counting. Medical image analysis **92**, 103047 (2024)
11. GTEx Consortium: Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (New York, N.Y.) **348**(6235), 648–660 (2015)
12. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
13. Ikezogwo, W.O., Seyfioglu, M.S., Ghezloo, F., Geva, D.S.C., Mohammed, F.S., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1M: One Million Image-Text Pairs for Histopathology. NeurIps (2023)
14. MahmoodLab:: Mahmoodlab/uni2-h (revision d517a8d) (2025). `https://huggingface.co/MahmoodLab/UNI2-h`, accessed: 2025-03-01
15. NCI: The Cancer Genome Atlas Program (TCGA) - NCI (05/13/2022 - 08:00)
16. Neidlinger, P., El Nahhas, O.S., Muti, H.S., Lenz, T., Hoffmeister, M., Brenner, H., van Treeck, M., Langer, R., Dislich, B., Behrens, H.M., et al.: Benchmarking foundation models as feature extractors for weakly-supervised computational pathology. arXiv preprint arXiv:2408.15823 (2024)
17. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision (2023)
18. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. CoRR **abs/2103.13413** (2021), `https://arxiv.org/abs/2103.13413`
19. Rumberger, J.L., Baumann, E., Hirsch, P., Janowczyk, A., Zlobec, I., Kainmueller, D.: Panoptic segmentation with highly imbalanced semantic labels. In: 2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC). pp. 1–4. IEEE (2022)
20. Samusik, N., Good, Z., Spitzer, M.H., Davis, K.L., Nolan, G.P.: Automated mapping of phenotype space with single-cell data. Nature methods **13**(6), 493–496 (2016)
21. Schürch, C.M., Bhate, S.S., Barlow, G.L., Phillips, D.J., Noti, L., Zlobec, I., Chu, P., Black, S., Demeter, J., McIlwain, D.R., et al.: Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. Cell **182**(5), 1341–1359 (2020)
22. Sun, Y., Zhu, C., Zheng, S., Zhang, K., Sun, L., Shui, Z., Zhang, Y., Li, H., Yang, L.: PathAsst: A Generative Foundation AI Assistant towards Artificial General Intelligence of Pathology. In: AAAI Conference on Artificial Intelligence. vol. 38, pp. 5034–5042 (2024)

23. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., Wei, F.: Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks (2022)
24. Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., Yu, K.H., Willens, S., Olguin, F.M., Nirschl, J.J., Neal, J., Diehn, M., Yang, S., Li, R.: A vision–language foundation model for precision oncology. Nature pp. 1–10 (2025)
25. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., Bower, T., Lee, S., Weerasinghe, R., Wright, B.J., Robicsek, A., Piening, B., Bifulco, C., Wang, S., Poon, H.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**(8015), 181–188 (2024)
26. Xu, Z., Gupta, R., Cheng, W., Shen, A., Shen, J., Talwalkar, A., Khodak, M.: Specialized foundation models struggle to beat supervised baselines. arXiv preprint arXiv:2411.02796 (2024)
27. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. Trans. Mach. Learn. Res. **2022** (2022)
28. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: Image BERT Pre-training with Online Tokenizer. In: International Conference on Learning Representations (2022)
29. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., Fuchs, T., Fusi, N., Liu, S., Severson, K.: Virchow2: Scaling Self-Supervised Mixed Magnification Models in Pathology (2024)