This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

From Patches to Patients: A study of the tile-to-slide performance transferability in Digital Pathology

Sofiène Boutaj[1,2], Leo Fillioux[1,2], Maria Vakalopoulou[1,2], Stergios Christodoulidis[1,2†], and Pierre Marza[1,2†]

[1] Université Paris-Saclay, CentraleSupélec, Gustave Roussy, INSERM, IHU PRISM, Cancer Data Science Unit, France
[2] Université Paris-Saclay, CentraleSupélec, MICS Laboratory, France
Corresponding author: sofiene.boutaj@centralesupelec.fr

Abstract. Foundation Models (FMs) have recently redefined the state-of-the-art in histopathology by providing robust representations for whole-slide image (WSI) analysis. However, selecting the optimal foundation model (FM) for a specific clinical cohort currently requires multiple pre-processing steps, followed by computationally expensive feature extraction and the training of a Multiple Instance Learning (MIL) aggregator for every model. In this work, we investigate whether efficient tile-level linear probing can serve as a reliable proxy for slide-level performance, reducing the need to run full slide-level pipelines for every candidate encoder. We benchmark 19 state-of-the-art FMs on 42 slide-level and 16 tile-level tasks, comparing tile probing metrics against slide-level outcomes using ABMIL and Mean Pooling aggregations. We observe a high correlation between tile and slide performance across varying task difficulties, indicating that encoder representation quality is the primary determinant of WSI success. Sensitivity analyses show that transferability is stable across models and is more influenced by cohort sizes and numbers of tiles per slide than by average task difficulty.
We also measure the agreement in best performing models between tile and slide-level tasks, showing tile benchmarks reliably shortlist strong candidates. Overall, our study indicates that tile-level benchmarking provides an efficient and practical first step for narrowing down candidate models, while slide-level evaluation remains essential for final validation on clinical tasks.

Keywords: Digital pathology · Foundation models · Benchmark.

1 Introduction

Digital pathology plays a central role in diagnosis and prognosis by automatically extracting relevant information about the cellular environment from tissue imaging. This is particularly true now that many foundation models [7,34,41,29,12,13]

† Denotes equal contribution.

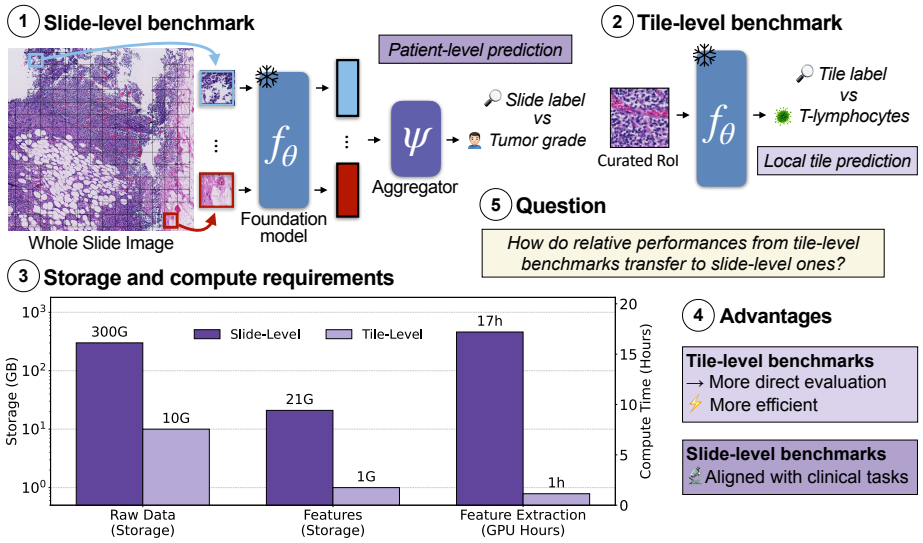

Fig. 1: **Comparison of slide-level and tile-level benchmarks. (1,2)** Overview of the 2 benchmark types. **(3) Storage and compute requirements:** Average storage (log scale) per dataset and compute time per dataset and per model, measured on a single NVIDIA V100 GPU. Slide-level averages were computed across 42 tasks and 19 models, while tile-level averages were computed across 16 tasks and 19 models.

[25,20,22,8,40,17,16,37] were introduced as powerful generalist feature extractors for histopathology images. However, the high-resolution of Whole Slide Images (WSI) collected in histopathology makes it impossible to process them directly. There is a need for a specific processing pipeline: in practice, each WSI is first segmented to remove background and isolate tissue, then divided into tiles; each tile is embedded with a foundation model, and all tile-level features are aggregated to perform a slide-level prediction.

Comparing such foundation models, and better understanding their differences thus becomes important to draw a clear picture of the progress in the field but also from a practical point of view when selecting a feature extractor for a new clinical task. For these reasons, various benchmarks were proposed recently [36,19,26,15,3,14,5,21,23,2,39,6,24]. We can divide them in two categories: *tile-level* and *slide-level* benchmarks. The former evaluates how foundation model embeddings can extract relevant information from specific tiles, or regions of WSIs, while the latter targets slide-level predictions. Figure 1 provides an overview of both types of benchmarks, presents a comparison of their storage and compute efficiency, and summarizes their advantages. As presented in [24], tile-level benchmarks are more efficient, isolate the impact of foundation models from required aggregators to perform slide-level predictions, and leverage denser supervision. On the other hand, slide-level benchmarks require more

pre-processing, more compute, come with sparser supervision, and necessitate training aggregators on top of foundation model features, but are closer to the final clinical tasks of interest.

We are thus presented with the following dilemma: tile-level benchmarks allow a more direct and efficient evaluation of foundation models, which is appealing from a methodological point of view, while slide-level benchmarks better model clinical needs. An important question is thus: *How do relative performances from tile-level benchmarks transfer to slide-level ones?* A good transfer would imply that tile-level benchmarks are effective tools for evaluating new models, providing confidence that their relative ranking will be preserved on slide-level tasks. However, this question is not binary but more nuanced. This paper aims not only to address it, but also to better understand the conditions under which performance transfers.

Our contributions are fourfold: (*i*) to the best of our knowledge, we provide the first large-scale tile-to-slide benchmarking study across 19 open-source pathology foundation models, 16 tile-level tasks, and 42 slide-level tasks from publicly available datasets; (*ii*) we measure tile-to-slide rank correlation using Pearson, Spearman, and Kendall metrics for both mean-pooling and ABMIL slide-level aggregation, and show strong transferability overall; (*iii*) we perform sensitivity analyses showing stable correlations and highlighting the role of cohort size and number of tiles per slide for tile-to-slide transferability; and (*iv*) we show through a top-5 overlap analysis that tile-level benchmarks are useful for shortlisting candidate encoders before expensive slide-level training, and highlight the current limitations of tile-level datasets.

2 Related work

Foundation models in digital pathology — are trained in a self-supervised manner on large datasets spanning multiple organs, scanning protocols and centers. They can be divided into two categories: (*i*) vision-only [7,34,41,29,12], [13,25,20,33,35,38] trained with DINO-style objectives [27], and (*ii*) vision-language encoders [22,8,40,17,16,37,30] optimized with CLIP-style losses [28]. Such models are then used as powerful feature extractors for a variety of downstream tasks such as cancer sub-typing or gene mutation prediction.

Benchmarking foundation models — With the recent surge of foundation models in digital pathology, systematic evaluation has become a bottleneck. This has led the community to search for methods of comparing them to answer two main questions: (*i*) *Which direction should the field take to build better-performing models?*, and (*ii*) *which models should practitioners use to get the best results?* A common approach is to benchmark foundation models on slide-level tasks [36,19,26,15,3,14,5,21,23,2,39,6]. These are clinically relevant, but require to couple a slide-level aggregator to existing tile-level foundation models to perform predictions. This induces additional computational constraints and also makes the assessment of the impact of the features from the foundation model itself less direct as many choices related to the design of the aggregator are to

Table 1: Detailed summary of the 19 slide-level WSI datasets (42 tasks), including average number of slides per task and average number of tiles per slide.

Source	Dataset	Mutation tasks	Molecular tasks	Histological tasks	Immune tasks	Avg Slides / Task	Avg Tiles / Slide
CPTAC [9]	9 datasets	18	1	2	7	212	5089
TCGA [32]	8 datasets	8	2	1	-	443	12919
BRACS [4]	1 dataset	-	-	2	-	545	13610
CAMELYON16 [10]	1 dataset	-	-	1	-	398	14950
Total	**19 datasets**	**26**	**3**	**6**	**7**	**343**	**9922**

be made. Tile-level benchmarks [14,24] are another alternative. They are generally more efficient and provide a more direct comparison of foundation model representation spaces, as they isolate their impact from any aggregation method.

However, it remains unclear whether these two benchmarking paradigms lead to consistent conclusions when comparing foundation models. To the best of our knowledge, we present the first large-scale study of tile-to-slide (patch-to-patient) performance transferability.

3 Patches to patients performance transferability

3.1 Experimental setup

Foundation Models — We benchmark 19 recent pathology Foundation Models (FMs), covering both visual-only encoders and vision-language models. The set includes five vision-language models – CONCH (v1, v1.5) [22], KEEP [40], PLIP [16], QuiltNet [17] – and 14 vision-only encoders – ProvGigaPath [38], H-Optimus (0 and 1) [29], Hibou-Base [25], H0-mini [11], Kaiko (ViT-S and ViT-B variants) [1], Phikon (v1, v2) [12,13], UNI (v1, v2) [7], and Virchow (v1, v2) [34].
Benchmarks and Tasks — For the tile-level evaluation, we include all patch-level *classification* tasks from the `THUNDER` benchmark [24], totaling 16 datasets (16 tasks) covering a broad range of tissue origins and tasks (e.g., tumor detection, histologic pattern recognition). In total, all datasets contain 2,202,752 tiles, with dataset sizes ranging from 408 to 367,229 samples. The details about the considered datasets are presented in [24] and on the `THUNDER` documentation[†].

The slide-level analysis involves 42 tasks drawn from 19 WSI datasets (see Table 1). Collectively, these slide-level datasets cover 10 anatomical sites (breast, lung, colon/rectum, kidney, ovary, uterus, brain, stomach, pancreas, and head & neck). Following the standard `PathoBench` protocol [39], we evaluate the *CPTAC* tasks using a 50-fold train/test split, while employing 5-fold cross-validation for all other WSI datasets. For all splits, we ensure there is no data leakage between the training and test sets.
Downstream tasks evaluation — The tile-level tasks derived from `THUNDER` [24] are implemented as linear probing on frozen tile embeddings. For slide-level tasks,

[†] https://mics-lab.github.io/thunder/#available-datasets

we rely on the open-source `TRIDENT` library [39] to preprocess raw WSIs, including tissue segmentation, patching, and feature extraction. All WSIs are processed at 20× magnification. Tile size is automatically selected according to the architectural requirements of each foundation model, using one of $\{224 \times 224,\ 256 \times 256,\ 512 \times 512\}$ pixels. Slide-level evaluation is performed with the `PathoBench` framework [39], using its default implementations of two aggregation strategies: (*i*) **mean pooling with linear probing**, where we compute a mean feature vector and train a linear classifier on the pooled vector; and (*ii*) **attention-based MIL (ABMIL)** [18], where we learn a gated attention mechanism that assigns a weight to each tile embedding and aggregates them into a slide representation, followed by a linear classification head (single attention head, projection dimension 512, and pre-classification dropout 0.25). We select ABMIL as it remains highly effective in practice, with prior studies demonstrating that its performance is often on par with more complex MIL architectures [31], making it a robust and sufficient baseline for our evaluation. Overall, the pre-processing, feature extraction, and downstream training across all *slide-level* tasks and foundation models required more than 15,000 V100 GPU-hours.

Metrics — For the slide-level evaluation, as it involves multiple datasets with varying numbers of tasks, we first compute the mean macro-F1 within each dataset, and subsequently average these intra-dataset means, preventing any single dataset from skewing the global WSI score, that we denote S_m. For the tile-level benchmark, the aggregated score T_m is computed as the macro-F1 averaged across the 16 individual datasets in `THUNDER`.

Conducted analyses — Our main experiment is about (*i*) quantifying the overall tile-to-slide transferability. For that, we analyze the relationship between the tile summary T_m and the slide summary S_m across all evaluated FMs. We report the correlation of performance values across models using three metrics: Pearson's ρ_P for the tile-to-slide performance linear relationship, and Spearman's ρ_S and Kendall's τ for tile-to-slide rank-order consistency. We assess significance using a two-sided permutation test. Beyond this, we perform the following complementary analyses: (*ii*) **Leave-one-model-out sensitivity analysis:** recompute the correlation after removing each FM to ensure results are not driven by a single model. (*iii*) **Task-ablation sensitivity:** track correlation as slide tasks are removed based on cohort size, number of tiles per slide, and average task performance to identify key dataset factors for tile-to-slide performance transferability. (*iv*) **Top-5 shortlist utility:** compute the overlap between top-5 tile-level and top-5 slide-level models for each tile/slide task pair. (*v*) **Rank-sum consensus:** aggregate ranks across Tile, Mean Pooling, and ABMIL to highlight consistently strong models and where reordering concentrates.

3.2 Results

Global rank correlation between tile-level and slide-level benchmarks — Fig. 2 reports the rank agreement between tile-level linear probing and slide-level performance across 19 foundation models. Using mean pooling for slide aggregation (*left*), the model ordering is highly preserved, with points tightly con-

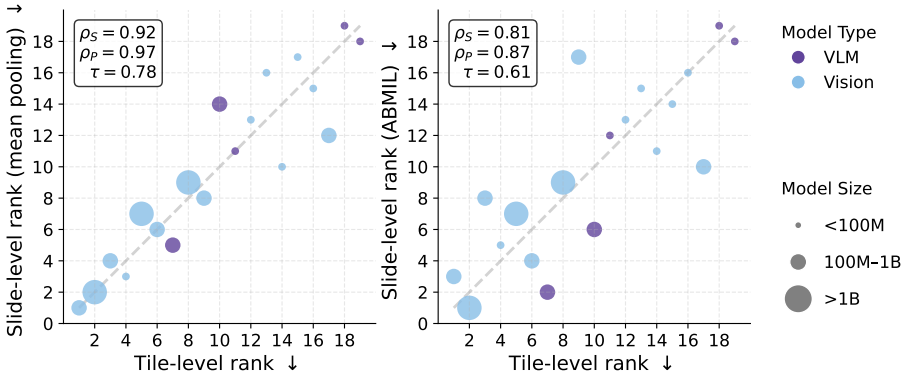

Fig. 2: **Rank correlation between slide-level and tile-level benchmarks.** Comparing the rank across 19 models on a set of tile-level and slide-level tasks. Slide-level aggregation is performed via mean-pooling (*left*) and ABMIL (*right*); measured by Spearman (ρ_S), Pearson (ρ_P), and Kendall's τ.

centrated around the identity line and strong correlations ($\rho_S = 0.925$, $\tau = 0.778$, $\rho_P = 0.967$; permutation $p = 2 \times 10^{-4}$ for all). This indicates that, under a simple aggregation scheme, slide-level success is largely determined by the intrinsic quality of the frozen tile representations. When switching to ABMIL (right), the correspondence remains significantly positive but is weaker ($\rho_S = 0.814$, $\tau = 0.614$, $\rho_P = 0.874$; permutation $p = 4 \times 10^{-4}$ for all), with larger departures from the diagonal. This suggests that learning a more expressive MIL aggregator introduces additional variability that can alter relative performance, particularly among mid-ranked models. The results are robust to the choice of evaluation metric: when replacing macro-F1 with balanced accuracy, the resulting correlations remain within a maximum absolute deviation of $\leq 2\%$ from the macro-F1 correlations (with permutation p-values remaining highly significant). Moreover, the very high Pearson correlation ($\rho_P = 0.874$ for ABMIL and $\rho_P = 0.967$ for mean pooling) indicates a strong linear relationship between tile-level and slide-level performance. This confirms that tile-level benchmarking provides reliable quantitative information about downstream slide-level performance. Overall, these results support tile-level probing as an efficient proxy for slide-level model benchmarking while highlighting that rank transferability depends on the complexity of the aggregation method. Importantly, as slide-level predictions are performed from aggregated tile embeddings, our results validate that high-quality tile representations (evaluated on tile-level benchmarks) naturally lead to strong slide-level performance.

Comprehensive sensitivity analysis of ABMIL transferability — Fig. 3 shows that ABMIL tile-to-slide rank correlation is robust and not driven by any single encoder. In the leave-one-model-out analysis (left), Spearman remains near the baseline ($\rho_s \approx 0.814$) with limited variation. Task-ablation trajectories further show that correlation is more sensitive to **test cohort size** and **number of**

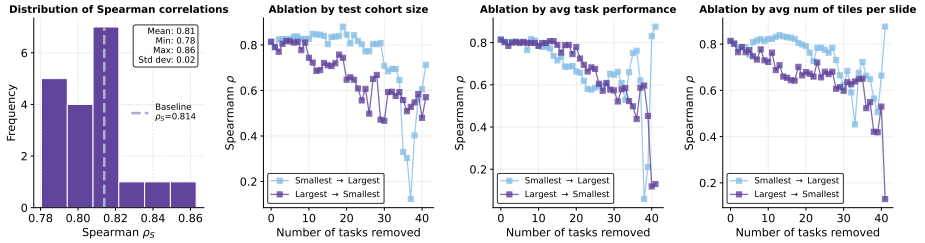

Fig. 3: **Comprehensive sensitivity analysis of tile-to-slide performance correlation on ABMIL.** *From left to right*: (*1*) Leave-One-Model-Out sensitivity distribution demonstrating a stable Spearman correlation (ρ_s) when individual foundation models are removed. The remaining plots show correlation trajectories when slide-level tasks are iteratively removed (smallest-to-largest vs. largest-to-smallest) based on: (*2*) test cohort size, (*3*) average task performance, and (*4*) average number of tiles per slide.

.

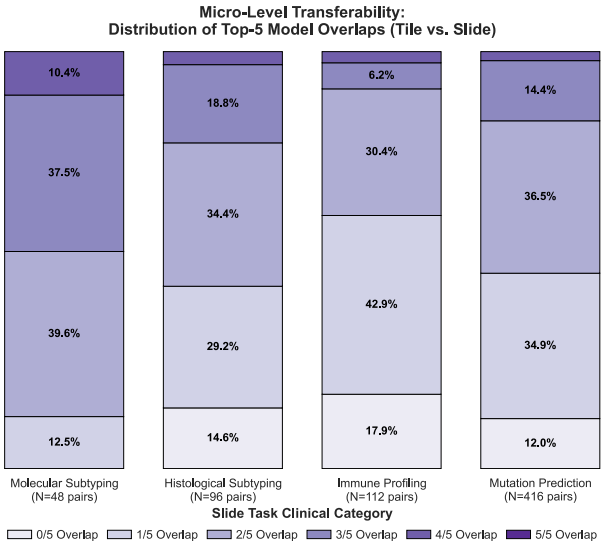

Fig. 4: **Micro-Level Transferability of Top-5 Histopathology Models.** Distribution of model overlap between tile-level and slide-level tasks, stratified by slide task clinical category. For each evaluated pair of tile and slide tasks, the intersection of the top 5 highest-performing models was computed.

Table 2: Rank Sum Analysis of Histopathology Models (Tile vs Mean Pooling vs ABMIL) – T: tile-level rank, S-M: Meanpool slide-level rank, S-A: ABMIL slide-level rank

Model	T	S-M	S-A	Sum
hopt1	2	2	1	**5**
uni2h	1	1	3	**5**
h0mini	4	3	5	**12**
keep	7	5	2	**14**
virch2	3	4	8	**15**
uni	6	6	4	**16**
hopt0	5	7	7	**19**
gigapath	8	9	9	**26**
conch1.5	10	14	6	**30**
conch	11	11	12	**34**
virch	9	8	17	**34**
hiboub	14	10	11	**35**
kaiks16	12	13	13	**38**
phik2	17	12	10	**39**
kaikb16	13	16	15	**44**
phik	15	17	14	**46**
kaiks8	16	15	16	**47**
quilt	19	18	18	**55**
plip	18	19	19	**56**

tiles per slide than to **average slide-task performance**. Indeed, removing large-cohort tasks first, or high-tile-count tasks first, causes an earlier drop in correlation (including an approximate 10% decrease before 10 removed tasks),

whereas performance-based ablation is initially stable. This suggests transferability is supported more by statistical reliability and bag complexity than by task difficulty alone.

Micro-level transferability of top-5 models — Complementing the *global correlation* analysis, Fig. 4 studies the agreement in best-performing models between tile- and slide-level tasks. For each tile/slide task pair, we compute the overlap between the top-5 tile- and slide-level models. Overlap is mostly *partial* (typically 1/5–3/5 shared models), indicating that tile-level benchmarks help *shortlist* strong candidates even without exactly recovering the slide-level top-5.

Overlap is higher for molecular subtyping and mutation prediction and lower for immune profiling. A likely reason is that tile and slide tasks may require different encoder information: tile tasks emphasize local morphology, whereas slide tasks may depend more on tissue architecture, spatial organization, rare patterns, and long-range context. Thus, lower overlap does not contradict transferability; it highlights the *limits* of using tile-level signals for exact top-model identification on heterogeneous slide tasks.

Rank-sum consensus across Tile, Mean Pooling, and ABMIL — Table 2 complements the correlation analysis by summarizing model behavior via a rank-sum criterion. It reveals a stable top tier, with `hoptimus1` and `uni2h` tied for best, followed by a compact group (`h0mini`, `keep`, `virchow2`, `uni`, `hoptimus0`) that remains strong across all three settings.

Most reordering occurs in the *middle* of the ranking, particularly under ABMIL. For example, `keep` and `conch1.5` improve under ABMIL, whereas `virchow` drops substantially. This confirms that tile-level benchmarking provides a strong *consensus prior* for model selection, while ABMIL mainly scrambles mid-ranked models. At the lower end, `quilt` and `plip` consistently rank last, confirming that tile-to-slide transferability preserves both the top and bottom of the leaderboard.

4 Conclusion

In this paper, we show that tile-level benchmarking is a strong proxy for slide-level model selection in digital pathology. Across 19 foundation models, tile-level linear probing correlates strongly with slide-level performance, with higher correlation for mean pooling and slightly lower (but still clear) correlation for ABMIL, where learned aggregation introduces additional variability. Our sensitivity analyses indicate that the ABMIL correlation is stable and not driven by a single model, and that transferability is more sensitive to dataset properties (e.g., cohort size and number of tiles per slide) than to average task difficulty. The top-5 overlap analysis shows that tile-level benchmarks can help identify promising models for a new clinical slide-level task, even when the exact best models differ because tile and slide-level tasks may rely on different types of information. Overall, tile-level benchmarks provide an efficient and practical first step for pathology FM selection, while slide-level validation remains important for context-heavy tasks and for distinguishing between closely ranked models.

Limitations — *(i)* Our tile-level tasks are predominantly oriented toward local morphology, which may explain the lower top-5 overlap observed for immune profiling tasks that rely on broader spatial tissue architecture. *(ii)* Beyond dataset scope, more expressive aggregators such as ABMIL introduce learned task-specific parameters that reshape the frozen FM's representation space, introducing aggregation-specific variability that may affect transferability for certain tasks. *(iii)* Tile-level benchmarking requires laborious manual annotations. However, this remains a fixed, one-time cost (already absorbed by public benchmarks like `THUNDER`). In contrast, the recurring hardware cost of evaluating new FMs or conducting hyperparameter searches is a bottleneck that tile-level benchmarking can substantially reduce across development cycles.

We view these remaining gaps as opportunities, and would like to emphasize that data quality might be a major factor to consider: richer, context-aware tile-level datasets will strengthen tile-to-slide transferability, further consolidating tile-level benchmarking as the first step for FM selection in digital pathology.

Acknowledgments — This work has been supported by the *Agence Nationale de la Recherche* through *ANR-23-IAHU-0002*, *ANR-21-CE45-0007*, *ANR-23-CE45-0029*, *ANR-23-IACL-0003* (*DATAIA CLUSTER*) and the *Health Data Hub* as part of the second edition of the *France-Québec* call for projects *Intelligence Artificielle en santé*. This work was performed using HPC resources from *GENCI-IDRIS* (Grant *2025-AD011015593R1*).

Disclosure of Interests — The authors declare no competing interests.

References

1. Aben, N., de Jong, E.D., Gatopoulos, I., et al.: Towards large-scale training of pathology foundation models. arXiv (2024)
2. Alfasly, S., Alabtah, G., Hemati, S., et al.: Validation of histopathology foundation models through whole slide image retrieval. Scientific Reports (2025)
3. Alfasly, S., Nejat, P., Hemati, S., et al.: Foundation models for histopathology—fanfare or flair. Mayo Clinic Proceedings: Digital Health (2024)
4. Brancati, N., Anniciello, A.M., Pati, P., et al.: Bracs: A dataset for breast carcinoma subtyping in h&e histology images. Database (2022)
5. Breen, J., Allen, K., Zucker, K., et al.: A comprehensive evaluation of histopathology foundation models for ovarian cancer subtype classification. NPJ Precision Oncology (2025)
6. Campanella, G., Chen, S., Singh, M., et al.: A clinical benchmark of public self-supervised pathology foundation models. Nature Communications (2025)
7. Chen, R.J., Ding, T., Lu, M.Y., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024)
8. Ding, T., Wagner, S.J., Song, A.H., et al.: Multimodal whole slide foundation model for pathology. arXiv (2024)
9. Edwards, N.J., Oberti, M., Thangudu, R.R., et al.: The cptac data portal: a resource for cancer proteomics research. Journal of proteome research (2015)
10. Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., 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)

11. Filiot, A., Dop, N., Tchita, O., et al.: Distilling foundation models for robust and efficient models in digital pathology. In: MICCAI (2025)
12. Filiot, A., Ghermi, R., Olivier, A., et al.: Scaling self-supervised learning for histopathology with masked image modeling. medRxiv (2023)
13. Filiot, A., Jacob, P., Mac Kain, A., et al.: Phikon-v2, a large and public feature extractor for biomarker prediction. arXiv (2024)
14. Gatopoulos, I., Känzig, N., Moser, R., et al.: eva: Evaluation framework for pathology foundation models. In: MIDL (2024)
15. Gustafsson, F.K., Rantalainen, M.: Evaluating computational pathology foundation models for prostate cancer grading under distribution shifts. arXiv (2024)
16. Huang, Z., Bianchi, F., Yuksekgonul, M., et al.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine (2023)
17. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., et al.: Quilt-1m: One million image-text pairs for histopathology. NeurIPS (2023)
18. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: ICML (2018)
19. Kang, M., Song, H., Park, S., et al.: Benchmarking self-supervised learning on diverse pathology datasets. In: CVPR (2023)
20. Karasikov, M., van Doorn, J., Känzig, N., et al.: Training state-of-the-art pathology foundation models with orders of magnitude less data. arXiv (2025)
21. Lee, J., Lim, J., Byeon, K., et al.: Benchmarking pathology foundation models: Adaptation strategies and scenarios. Computers in Biology and Medicine (2025)
22. Lu, M.Y., Chen, B., Williamson, D.F., et al.: A visual-language foundation model for computational pathology. Nature Medicine (2024)
23. Majzoub, R.A., Malik, H., Naseer, M., et al.: How good is my histopathology vision-language foundation model? a holistic benchmark. arXiv (2025)
24. Marza, P., Fillioux, L., Boutaj, S., et al.: THUNDER: Tile-level histopathology image understanding benchmark. Neural Information Processing Systems (NeurIPS) D&B Track (2025)
25. Nechaev, D., Pchelnikov, A., Ivanova, E.: Hibou: A family of foundational vision transformers for pathology. arXiv (2024)
26. Neidlinger, P., El Nahhas, O.S., Muti, H.S., et al.: Benchmarking foundation models as feature extractors for weakly-supervised computational pathology. arXiv (2024)
27. Oquab, M., Darcet, T., Moutakanni, T., et al.: Dinov2: Learning robust visual features without supervision. arXiv (2023)
28. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
29. Saillard, C., Jenatton, R., Llinares-López, F.o.: H-optimus-0 (2024), `https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0`
30. Shaikovski, G., Casson, A., Severson, K., et al.: Prism: A multi-modal generative foundation model for slide-level histopathology. arXiv (2024)
31. Shao, D., Chen, R.J., Song, A.H., et al.: Do multiple instance learning models transfer? arXiv preprint arXiv:2506.09022 (2025)
32. Tomczak, K., Czerwińska, P., Wiznerowicz, M.: Review the cancer genome atlas (tcga): an immeasurable source of knowledge. Contemporary Oncology (2015)
33. Vaidya, A., Zhang, A., Jaume, G., et al.: Molecular-driven foundation model for oncologic pathology. arXiv (2025)
34. Vorontsov, E., Bozkurt, A., Casson, A., et al.: A foundation model for clinical-grade computational pathology and rare cancers detection. Nature medicine (2024)

35. Wang, X., Zhao, J., Marostica, E., et al.: A pathology foundation model for cancer diagnosis and prognosis prediction. Nature (2024)
36. Wölflein, G., Ferber, D., Meneghetti, A.R., et al.: Benchmarking pathology feature extractors for whole slide image classification. arXiv (2023)
37. Xiang, J., Wang, X., Zhang, X., et al.: A vision–language foundation model for precision oncology. Nature (2025)
38. Xu, H., Usuyama, N., Bagga, J., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature (2024)
39. Zhang, A., Jaume, G., Vaidya, A., et al.: Accelerating data processing and benchmarking of ai models for pathology. arXiv (2025)
40. Zhou, X., Sun, L., He, D., et al.: A knowledge-enhanced pathology vision-language foundation model for cancer diagnosis. arXiv (2024)
41. Zimmermann, E., Vorontsov, E., Viret, J., et al.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv (2024)