How Much Information Is Enough? A Benchmark of Foundation Models for Thyroid Histopathology under Limited Patch Sampling

Katja Löwenstein[1]✉, Lea Maria Stangassinger[2], Christina Kreutzer[3], Sebastien Couillard-Despres[3,4], Gertie Janneke Oostingh[2], Anton Hittmair[5], and Michael Gadermayr[1]

[1] Department of Information Technologies and Digitalisation, Salzburg University of Applied Sciences, Salzburg, Austria
katja.loewenstein@fh-salzburg.ac.at
[2] Department of Biomedical Sciences, Salzburg University of Applied Sciences, Salzburg, Austria
[3] Institute of Experimental Neuroregeneration, Paracelsus Medical Private University, Salzburg, Austria
[4] Austrian Cluster for Tissue Regeneration, Vienna, Austria
[5] Department of Pathology and Microbiology, Kardinal Schwarzenberg Klinikum, Schwarzach, Austria

Abstract. Whole slide images in digital pathology are commonly represented as large sets of image patches, resulting in computationally and memory-intensive inference. While dense patch sampling reduces the likelihood of missing sparsely distributed diagnostic features, it is often computationally infeasible in practice. In particular, limited tissue availability (e.g., small biopsy specimens), variable data quality, and resource-constrained deployment motivate the need to quantify how much patch-level information is required for reliable prediction.
In this work, the robustness of feature extraction models to reduced patch coverage in thyroid cancer subtyping is systematically studied. Using a fixed multiple instance learning framework, a diverse set of histopathology foundation models and domain-agnostic encoders is evaluated under controlled random subsampling. Starting from full patch coverage, the number of patches per slide is progressively reduced to assess performance degradation under increasing sparsity. Experiments are conducted on three independent thyroid histopathology datasets with different subtype distributions, enabling assessment of cross-dataset consistency. The results demonstrate that foundation models exhibit reduced performance degradation under reduced patch coverage compared to domain-agnostic encoders, indicating improved information efficiency and robustness under sparse sampling conditions. Overall, the findings suggest that reliable slide-level prediction can be maintained even with substantially reduced patch availability, supporting more efficient and flexible deployment in practical settings.

Keywords: Thyroid Histopathology · Foundation Models · Sparse Patch Sampling.

1 Introduction

Whole slide image (WSI) classification with multiple instance learning (MIL) has become a standard paradigm in computational pathology, enabling learning from gigapixel images via aggregation of patch-level representations [7,15,20,24]. Model performance strongly depends on the feature extraction backbone, which encodes local morphology into latent embeddings prior to aggregation [6,25]. Recent histopathology-specific foundation models have improved representation quality [22], and prior studies have reported systematic differences between general-purpose and domain-specific encoders across tasks [6,27]. However, their behavior under reduced patch coverage remains insufficiently characterized. At a high level, a central question arises: whether the strongest-performing model under full information conditions remains the strongest when only limited input information is available. MIL-based WSI classification relies on the sampled instances providing sufficient information to distinguish the target classes, although the extent to which random sampling captures diagnostically relevant morphology is task-dependent. Moreover, relevant structures may be sparsely distributed, suggesting that a small subset of patches can already be informative. Recent evidence supports this redundancy in WSIs, indicating that compact subsets may preserve most diagnostic signal [2,18], while data diversity may be more important than patch quantity [5]. At the same time, excessive subsampling risks omitting rare but critical patterns, motivating a systematic analysis of performance under varying patch coverage. From a clinical perspective, reduced patch sampling is relevant in resource-limited settings such as small or punch biopsy specimens and in the presence of domain shift caused by staining variability, acquisition differences, artifacts, or dataset bias [3,25]. Robustness under sparse sampling may therefore improve real-world applicability and enable more efficient inference pipelines, including preselection and augmentation strategies such as Mixup-based MIL [13] where more efficient patch usage could improve diversity within training batches under fixed computational budgets. In this work, the efficiency with which different feature extraction models utilize patch-level information under systematically reduced patch coverage is investigated, with the aim of quantifying the robustness of both general-purpose and pathology-specific encoders across independent histopathology datasets and sampling regimes. A single MIL architecture is used for controlled comparison: CLAM [20] is adopted due to its widespread use and strong empirical performance in weakly supervised histopathology and its class-specific attention mechanism. This design isolates the effects of the feature extractor and patch sampling strategy while minimizing architectural confounding. To this end, the following research questions are addressed:

RQ1 (Patch Efficiency). How does reducing patch coverage at inference affect classification performance in multiple instance learning?

RQ2 (Encoder Robustness). Do histopathology foundation models retain predictive performance more effectively than general-purpose encoders when patch coverage is reduced?

RQ3 (Cross-Dataset Consistency). To what extent are degradation patterns under reduced patch coverage consistent across different histopathology datasets?

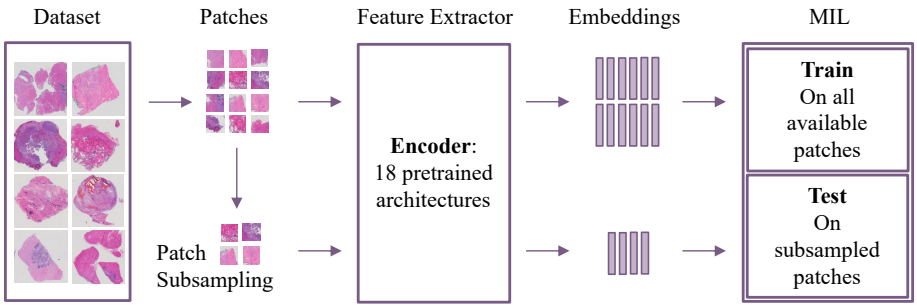

Fig. 1. Graphical Abstract: WSIs are partitioned into patches and encoded into embeddings, which are aggregated by the CLAM MIL model. At inference, patch subsets ranging from 1% to 100% of the available patches are evaluated, with the full 100% setting serving as the reference condition. Performance is evaluated across 18 pretrained encoders, including both domain-specific foundation models and domain-agnostic encoders.

2 Methods

Figure 1 provides an overview of the computational pipeline. Within each dataset, WSIs are assigned to one of two diagnostic classes and partitioned into patches using a fixed extraction scheme, with patch size determined by the respective encoder. Pretrained encoders (see Table 1) generate patch-level embeddings, which are aggregated by the CLAM MIL model for slide-level classification. To evaluate robustness under varying patch coverage, a systematic subsampling strategy is applied during inference. Patch subsets ranging from 1% to 100% of the available instances per slide are randomly sampled in a reproducible manner, with the full 100% setting serving as the standard reference condition.

Three hematoxylin and eosin (H&E)-stained thyroid WSI datasets (Table 2) were used for classification of malignant and benign tissue subtypes with expert-provided slide-level annotations. Representative examples are shown in Fig. 2. While sharing the same tissue origin, the datasets differ in subclass composition, size, and class balance, and exhibit variability in staining, background, and imaging conditions due to laboratory- and acquisition-specific factors, reflecting realistic clinical heterogeneity. The WSIs from all datasets were resampled to a common spatial resolution of 0.25 µm per pixel (mpp) prior to patch extraction.

Classification was performed using the CLAM-SB framework [20], a weakly supervised MIL method for WSIs with slide-level labels. Each slide is represented

Table 1. Overview of the evaluated encoders. Feature Dimensionality, Patch size (Dims); Number of parameters (Pars).

Model	Dims	Pars	Slides	Data Source	Algorithm	Backbone
DS-FM (Domain-Specific Foundation Model)						
Conch-v1 [19]	512,512	90 M	21 K	Public,Private	iBot,CoCa	ViT-B/16
Conch-v15 [19]	768,512	307 M	21 K	Public,Private	CoCa	ViT-L/16
CTransPath [29]	768,256	27 M	33 K	TCGA,PAIP	SSL	CNN,SwinT
GigaPath [30]	153,256	1.1 B	171 M	Private	DINOv2	ViT-G/14
Hibou-L [21]	102,224	303 M	1 M	Private	DINOv2	ViT-L/16
H-Optimus-v1 [4]	153,224	1.1 B	1 M	Private	SSL	ViT-G/14
Kaiko-L [1]	1024 256	307 M	29 K	TCGA	DINOv2	ViT-L/14
Lunit-S [16]	384,224	22 M	37 K	TCGA,TULIP	DINO	ViT-S/8
Phikon-v1 [11]	768,224	86 M	6 K	TCGA	iBOT	ViT-B/16
Phikon-v2 [12]	1024,224	300 M	60 K	PANCAN-XL	DINOv2	ViT-L/16
Uni-v1 [8]	1024,256	300 M	100 K	Private	DINOv2	ViT-L/16
Uni-v2 [8]	1536,256	681 M	350 K	Private	DINOv2	ViT-H/14
Virchow-v1 [28]	2560,224	632 M	1.5 M	Private	DINOv2	ViT-H/14
Virchow-v2 [31]	2560,224	632 M	3.1 M	Private	DINOv2	ViT-H/14
GP-FM (General-Purpose Foundation Model)						
Dino-v2 [23]	1024,224	300 M	142 M	LVD-142M incl. ImageNet	SSL incl. DINOv2	ViT-L/14
GP-SL (General-Purpose Supervised Learning)						
ResNet-152 [14]	1024,256	60 M	1 K	ImageNet-1k	Supervised	ResNet-152
ResNet-50 [17]	1024,224	25 M	21 K	ImageNet-21k	Supervised	ResNet-50
ViT-H [9]	1280,224	630 M	21 K	ImageNet-21k	Supervised	ViT-H/14

Table 2. Thyroid histopathology datasets: each dataset is described by the number of WSIs, class labels, and corresponding IDs.

Dataset	#WSIs	Classes	ID and #WSIs
SALK [13]	111	Papillary carcinoma	(PC, 55)
		Follicular nodule	(FN, 56)
THCA [26]	111	Papillary carcinoma: Columnar cell	(CC, 37)
		Papillary carcinoma: Follicular variant	(FV, 74)
BUCA [10]	39	Papillary carcinoma	(PC, 22)
		Follicular adenoma	(FA, 17)

as a bag of image patches, which are aggregated via an attention mechanism into a slide-level representation. Compared to standard attention-based MIL, CLAM introduces class-specific attention branches and a clustering-constrained loss to improve separation between representative and non-representative instances, enhancing stability and discriminative power. The resulting representation is passed to a fully connected layer with softmax activation for classifi-

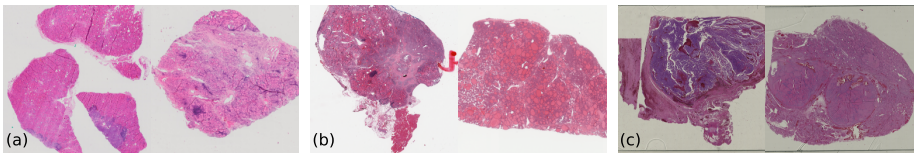

Fig. 2. Images of thyroid datasets: (a) SALK, (b) THCA and (c) BUCA

cation. CLAM was chosen due to its strong performance in weakly supervised histopathology and robustness to varying patch distributions. Its instance-level attention enables identification of relevant regions without explicit modeling of inter-instance dependencies, which is particularly suitable for this study where patch coverage is systematically reduced. Optimization followed the standard CLAM setup using cross-entropy loss and Adam with a learning rate of $2 \cdot 10^{-4}$. Dropout ($p = 0.25$) was applied to the attention module. Training was conducted on NVIDIA A6000 GPUs with CUDA acceleration. Encoder details, including patch size, embedding dimension, and pretraining, are provided in Table 1. Model performance was evaluated using k-fold cross-validation with $k = 10$ for SALK and THCA and $k = 5$ for BUCA. Each dataset was split into patient-level folds, ensuring no patient overlap between splits, with stratification to preserve class distributions. In each fold, one subset served as test data, one as validation, and the remainder for training. For each test fold, patch-level subsampling was applied per slide to generate different retention levels, ensuring evaluation exclusively on unseen data. Each setting was repeated five times with fixed seeds, and results were averaged for robustness. Balanced accuracy was used as the evaluation metric due to class imbalance.

3 Results

Figure 3 shows encoder-group–level accuracies across sampling levels. The x-axis denotes the percentage of retained patches, and the y-axis the classification accuracy. Curves represent the mean performance per group, with shaded regions indicating ± 1 standard deviation across encoders. The baseline corresponds to 100% patch coverage. Across all datasets, DS-FM consistently achieves the highest baseline performance, followed by GP-FM and GP-SL. In SALK, performance remains stable across subsets with a clear separation of approximately 10 percentage points between groups; DS-FM exceeds 90%, GP-SL 80%, and GP-FM 70%. Minor degradation appears at 20–10% coverage, with a sharper drop at 5–1%. In THCA, all groups exhibit similar performance levels (approximately 85–90%), with only small differences. GP-FM shows earlier degradation starting around 50% coverage, while DS-FM and GP-SL remain stable until 5–10%, followed by a more pronounced decline at 1% relative to SALK. In contrast, BUCA shows lower overall performance. DS-FM and GP-SL reach around 70%, whereas

GP-FM approaches chance level and deteriorates further below 20% patch coverage. Detailed results for the individual encoders, including mean performance and standard deviation across sampling levels, are provided in the supplementary material (Table 3). Across datasets, the standard deviation generally increases at lower sampling levels, indicating greater variability in performance across random sampling seeds as fewer patches are retained. Overall, DS-FM consistently achieve the highest performance across all datasets and sampling levels, while GP-SL follow with lower performance, with the magnitude of the gap depending on the dataset. In contrast, GP-FM consistently yield substantially lower performance and represent the weakest encoder group across all evaluated settings.

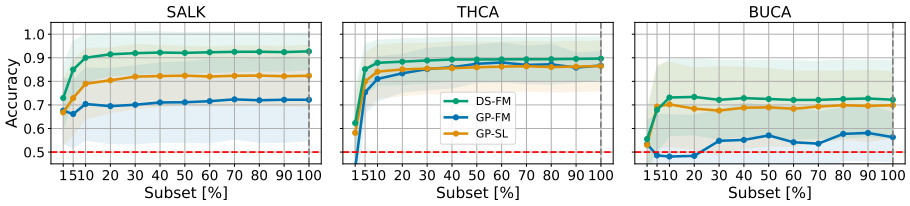

Fig. 3. Degradation curves illustrate the effect of reducing the number of patches on classification accuracy for different encoder groups and datasets. The x-axis shows the percentage of patches used for testing, and the y-axis indicates accuracy. Each curve represents the mean performance of all encoders in a group for a given subset, with shaded regions denoting ±1 standard deviation across encoders. A vertical dashed line at 100% marks full-data baseline performance, while a horizontal dashed line at 0.5 indicates chance-level accuracy.

Figure 4 compares accuracy at reduced patch coverage (10%, 5%, and 1%) against the baseline (100%) for each encoder and dataset. As the x-axis (baseline accuracy) is fixed across rows, deviations are reflected along the y-axis. Higher coverage levels (>10%) are omitted, as performance remains largely stable in this range (see Fig. 3). Spearman rank correlations (ρ), which quantify the strength and direction of a monotonic relationship between two variables based on their ranked values (here reflecting how consistently encoder rankings under a given subset match the baseline ranking), are reported in the upper left of each subplot. At 10% coverage, most points lie close to the diagonal, indicating performance comparable to the baseline. With decreasing coverage (5% and 1%), points increasingly deviate downward, reflecting performance degradation. SALK shows a near-perfect rank correlation at 10% ($\rho = 0.96$), which remains similar at 5% but breaks down at 1%. THCA and BUCA exhibit lower initial correlations, yet follow a similar trend: moderate reduction at 5%, and near-complete loss of rank consistency at 1%. The figure also highlights encoder-specific behavior. For instance, Conch-v15 and CTransPath remain notably stable on THCA, maintaining accuracies of around 80% even at 1% patch coverage. This behavior is specific to THCA and is not observed in the other datasets. In contrast,

although Virchow-v1 achieves a comparatively high baseline performance on BUCA, its accuracy degrades at a similar rate to the other encoders under reduced patch coverage. In addition to within-dataset rank stability across patch coverage levels, cross-dataset consistency at full patch availability was assessed. Spearman rank correlations between datasets at 100% patch coverage indicate weak to moderate agreement in encoder rankings, with $\rho_{\mathrm{SALK,THCA}} = 0.28$, $\rho_{\mathrm{SALK,BUCA}} = 0.27$, and $\rho_{\mathrm{THCA,BUCA}} = -0.04$.

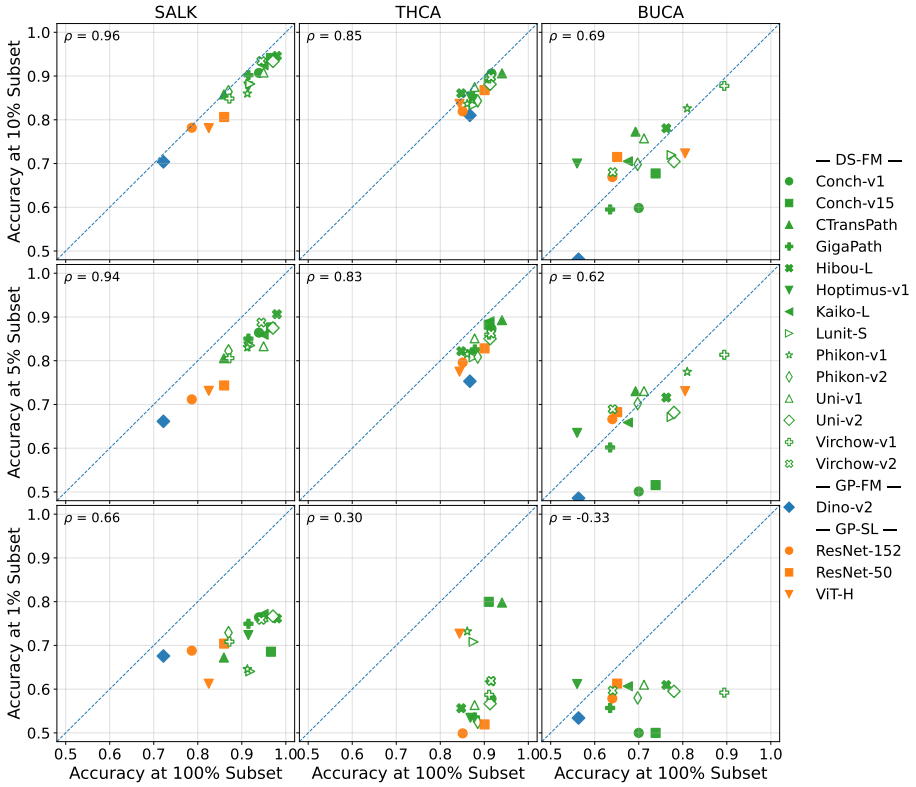

Fig. 4. Accuracy of Baseline (subset=100%) against Accuracy at reduced patch availability (rows: subset=1%, 5% and 10%) for each dataset (columns). Encoder instances are color-coded according to their respective model groups. For each subplot, the Spearman rank correlation ρ is reported in the upper-left corner.

4 Discussion

The study investigates MIL stability under reduced patch coverage across three histopathology datasets for thyroid cancer subtyping with varying subtype distributions, structured around three research questions: patch efficiency (RQ1),

encoder robustness (RQ2), and cross-dataset consistency (RQ3). A fixed CLAM architecture is used to isolate effects of the feature extractor and sampling strategy, enabling controlled comparison of DS-FM, GP-FM, and GP-SL encoders. Inference-time subsampling is applied from 100% (baseline) down to 1% patch coverage. Patch efficiency (RQ1) is higher than anticipated, as performance degrades only moderately under reduced patch coverage. This suggests that a small subset of patches may already capture sufficient diagnostic information, particularly when relevant tissue is prevalent within a slide. Further interpretation is limited by the lack of pixel-level annotations. While the use of simple random sampling with a fixed random seed ensures reproducibility, it does not account for spatial structure and may influence efficiency. Despite variations in baseline performance across datasets, the relative performance drop remains largely consistent across encoder groups, with the exception of the GP-FM group on BUCA. Analysis of encoder groups (RQ2) shows that foundation models consistently achieve the highest performance in both baseline accuracy and patch-subsetting experiments. This supports the hypothesis that self-supervised foundation models maintain strong predictive performance even under reduced slide-level information, compared to encoders trained on general-purpose datasets. Although DS-FM performance decreases at similar patch coverage levels as the other encoder groups, it remains consistently higher in absolute terms across all settings. While absolute accuracies and inter-group gaps vary between datasets, DS-FM achieve the highest average performance across all evaluated datasets. However, there is no consistent evidence that individual encoders achieving the best baseline performance also remain optimal under reduced patch coverage. Instead, performance rankings under subsampling depend on both the dataset and the specific encoder, with several cases showing reordering of top-performing models as patch coverage decreases. The overall degradation patterns are largely consistent across datasets (RQ3), despite differences in absolute performance levels. All datasets exhibit stable performance at moderate coverage levels, followed by a sharper decline at very low retention. However, dataset-specific variations are evident. In particular, BUCA shows generally lower performance and higher sensitivity to reduced coverage, while SALK and THCA exhibit more stable behavior. Despite these differences, the relative ranking of encoder groups remains consistent, with DS-FM outperforming GP-FM and GP-SL across datasets. The rank correlation between datasets at the baseline is suggesting limited transferability of encoder rankings across datasets despite stable within-dataset behavior. The observed encoder hierarchy is likely driven by differences in pretraining domain alignment, with DS-FM benefiting from pathology-specific representations. Dataset-level performance differences may further stem from variations in size, class imbalance, and class separability, leading to differing levels of inherent task difficulty. The increasing variability at lower sampling levels suggests that sparse patch sampling makes the classification outcome more sensitive to the specific spatial regions retained in the sampled subset. This may indicate that, as fewer patches are retained, the probability of omitting diagnostically relevant regions increases, resulting in greater dependence on the particular sampling

realization. However, the standard deviation alone does not provide direct evidence that all regions of the ROI are adequately covered; rather, it quantifies the sensitivity of model performance to different random sampling realizations. A limitation of this study is that differences in baseline performance may be influenced by variations in class distributions and separability across datasets. The use of heterogeneous datasets with differing class definitions was a deliberate choice, as thyroid-specific H&E datasets are scarce and rarely share consistent label structures. Furthermore, the analysis was restricted to a single MIL architecture, which may limit the generalizability of the findings across alternative model designs. Although all datasets were resampled to a common spatial resolution, the encoder-specific input patch sizes resulted in different physical fields of view, such that the image inputs were not spatially identical across encoders. A further limitation is the absence of region-level annotations, which prevents assessment of whether sampled subsets contain diagnostically relevant regions. Therefore, stable performance at reduced sampling rates cannot be taken as evidence that rare diagnostic regions are reliably captured. Nevertheless, the study primarily enables a comparative assessment of the robustness of different encoders to reduced instance sampling under the evaluated conditions.

To conclude, multiple instance learning exhibits strong robustness to reduced patch coverage, with stable performance even at low sampling levels. Domain-specific foundation models consistently outperform general-purpose encoders and retain this advantage under subsampling. Despite dataset-specific differences in absolute performance, domain-specific foundation models consistently outperform other encoder groups at 5% input information, demonstrating robust effectiveness under constrained sampling conditions. At 1% input information, foundation models remain the strongest-performing group; however, their performance advantage over competing encoder groups becomes less pronounced. Future work may explore whether targeted or content-aware patch selection strategies, for example ROI-based or attention-guided sampling, can further improve robustness by prioritizing diagnostically relevant regions instead of uniform patch coverage.

Acknowledgments. This work was funded by the County of Salzburg under the project number 20102/F2300525-FPR (AIBIA).

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Aben, N., de Jong, E.D., Gatopoulos, I., Känzig, N., Karasikov, M., Lagré, A., Moser, R., van Doorn, J., Tang, F.: Towards large-scale training of pathology foundation models (2024). https://doi.org/10.48550/arxiv.2404.15217
2. Alfasly, S., Alabtah, G., Hemati, S., Kalari, K.R., Garcia, J.J., Tizhoosh, H.R.: Validation of histopathology foundation models through whole slide image retrieval. Scientific Reports (2025). https://doi.org/10.1038/s41598-025-88545-9
3. Asadi-Aghbolaghi, M., Darbandsari, A., Zhang, A., Contreras-Sanz, A., Boschman, J., Ahmadvand, P., Köbel, M., Farnell, D., Huntsman, D.G., Churg, A., Black, P.C., Wang, G., Gilks, C.B., Farahani, H., Bashashati, A.: Learning generalizable ai models for multi-center histopathology image classification. npj Precision Oncology (2024). https://doi.org/10.1038/s41698-024-00652-4
4. Bioptimus: H-optimus-1 (2025), https://huggingface.co/bioptimus/H-optimus-1
5. Bosch, C., Wong, J.K., Paulikat, M., Zapukhlyak, M., Arora, B., Aichmüller-Ratnaparkhe, M., et al.: Diversity over scale: Whole-slide image variety enables h&e foundation model training with fewer patches. Journal of Pathology Informatics (2026). https://doi.org/10.1016/j.jpi.2026.100648
6. Breen, J., Allen, K., Zucker, K., Godson, L., Orsi, N.M., Ravikumar, N.: A comprehensive evaluation of histopathology foundation models for ovarian cancer subtype classification. npj Precision Oncology (2025). https://doi.org/10.1038/s41698-025-00799-8
7. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine (2019). https://doi.org/10.1038/s41591-019-0508-1
8. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
10. Eftimie, L.G., Glogojeanu, R.R., Tejaswee, A., Gheorghita, P., Stanciu, S.G., Chirila, A., Stanciu, G.A., Paul, A., Hristu, R.: Differential diagnosis of thyroid nodule capsules using random forest guided selection of image features. Scientific Reports (2022). https://doi.org/10.1038/s41598-022-25788-w
11. Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Kain, A.M., Saillard, C., Schiratti, J.B.: Scaling self-supervised learning for histopathology with masked image modeling. medRxiv (2023). https://doi.org/10.1101/2023.07.21.23292757
12. Filiot, A., Jacob, P., Kain, A.M., Saillard, C.: Phikon-v2, a large and public feature extractor for biomarker prediction (2024), https://arxiv.org/abs/2409.09173
13. Gadermayr, M., Koller, L., Tschuchnig, M., Stangassinger, L.M., Kreutzer, C., Couillard-Despres, S., Oostingh, G.J., Hittmair, A.: MixUp-MIL: Novel Data Augmentation for Multiple Instance Learning and a Study on Thyroid Cancer Diagnosis (2023). https://doi.org/10.1007/978-3-031-43987-2_46
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (2016). https://doi.org/10.1109/cvpr.2016.90

15. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Proceedings of the 35th International Conference on Machine Learning (2018)
16. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
17. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (bit): General visual representation learning. In: European Conference on Computer Vision (2019)
18. Li, S., Chen, Z., Chen, F.J., Gao, X.: Whole slide image redundancy reduction for efficient pathological diagnosis. Biomedical Signal Processing and Control (2026). https://doi.org/10.1016/j.bspc.2025.109289
19. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine (2024)
20. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering (2021)
21. Nechaev, D., Pchelnikov, A., Ivanova, E.: Hibou: A family of foundational vision transformers for pathology (2024)
22. Neidlinger, P., El Nahhas, O.S.M., Muti, H.S., Lenz, T., Hoffmeister, M., Brenner, H., et al.: Benchmarking foundation models as feature extractors for weakly supervised computational pathology. Nature Biomedical Engineering (2025). https://doi.org/10.1038/s41551-025-01516-3
23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., et al.: Dinov2: Learning robust visual features without supervision (2023)
24. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in Neural Information Processing Systems (2021)
25. Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.M., Ciompi, F., van der Laak, J.: Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Medical Image Analysis (2019). https://doi.org/10.1016/j.media.2019.101544
26. The Cancer Genome Atlas Research Network: THCA Dataset. https://www.cancer.gov/tcga (2014), accessed via GDC Data Portal
27. Vadori, V., Peruffo, A., Graïc, J.M., Finos, L., Grisan, E.: Mind the gap: Evaluating patch embeddings from general-purpose and histopathology foundation models for cell segmentation and classification. In: 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2025). https://doi.org/10.1109/embc58623.2025.11253185
28. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., et al.: A foundation model for clinical-grade computational pathology and rare cancers detection. Nature Medicine (2024)
29. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical Image Analysis (2022). https://doi.org/10.1016/j.media.2022.102559
30. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature (2024)
31. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., et al.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint arXiv:2408.00738 (2024)