Patch-Level Glioblastoma Subregion Classification with Foundation Model and Hierarchical Balanced Sampling

Qifeng Zhong [0009−0004−9249−5027], Ying Weng*[0000−0003−4338−713X], Ke Chen [0000−0002−2046−0034]

University of Nottingham Ningbo China, Ningbo 315100, China
ying.weng@nottingham.edu.cn

Abstract. Recognizing heterogeneous histomorphological regions within glioblastoma (GBM) at the patch level remains difficult because of class imbalance, variation between patients, and similarities in local tissue morphology. We evaluate whether a frozen pathology foundation model could provide useful representations for classifying ten GBM tissue categories. Each 512 × 512-pixel haematoxylin-and-eosin-stained patch is encoded by the ALICE vision stage as a 3,840-dimensional feature vector. These features are used to train a lightweight classifier with a SiLU activation and a linear output layer. Hierarchical class–patient–slide balanced sampling is applied during training. We conduct patient-disjoint five-fold cross-validation using 1,402,607 patches from 244 slides of 121 patients. The trained model is then evaluated on the official online validation set. In the online validation set, our model achieves F1 score 0.757 and MCC 0.896. Performance differs considerably between tissue categories. Some categories are recognized reliably, while others remain challenging. Overall, the results show that frozen ALICE vision-stage features can be combined with hierarchical balanced sampling for patch-level classification of GBM subregions. They also indicate that rare categories and regions with similar local morphology require further investigation. The source code is publicly available at https://github.com/zqf1234588/Brats2026_Pathology.

Keywords: Deep learning · Digital Pathology · BraTS 2026 · Glioblastoma

1 Introduction

Glioblastoma (GBM) contains heterogeneous tumor, necrotic, vascular, infiltrative and immune-associated microenvironments within digitized tissue sections [1]. The dataset used in this study defines ten patch-level classes, including nine expert-annotated histomorphological regions and a none-of-the-above (NOTA) class. The intrinsic morphological similarity among these classes, and therefore their potential confusion at patch level, remains an open empirical question [1].

Pathology foundation models (PFMs) generally learn patch-level representations through self-supervised learning (SSL) pretraining and aggregate them using multiple-instance learning or dedicated slide encoders for slide-level analysis [2], [3]. In brain tumor pathology, PFMs have performed strongly on slide-level tasks such as tumor detection of IDH mutation prediction [2], [4], while strong patch or region of interest

*Corresponding author: Ying Weng

(ROI) level results have mainly been demonstrated in nonbrain tissues [2], [5]. Whether PFM features can precisely separate heterogeneous and potentially morphologically similar GBM subregions therefore remains unclear [1].

Several studies have explored combining knowledge from multiple PFMs. GPFM [6], Meta-Encoder [7], and ALICE [8] adopt different approaches to integrate representations from multiple pathology foundation models. These approaches improved performance on selected downstream tasks. However, they did not directly establish that fusion improves the representation of subtle morphological differences among brain tumor patches. Domain shift remains another limitation. Tissue appearance varies across patients, slides and institutions, and models may learn source shortcuts [9], [10]. Although several methods have been proposed to reduce domain shift, it remains unresolved [11].

We encode GBM patches with the ALICE [8] vision stage and train a linear classifier using hierarchical class-patient-slide balanced sampling. The model achieves an online-validation F1 score 0.757 and MCC 0.896. These results support the feasibility of combining representations from a multi-expert-distilled PFM with hierarchical balanced sampling for patch-level GBM subregion classification.

2 Methods

2.1 Dataset

The labelled glioblastoma patch dataset is derived from haematoxylin-and-eosin-stained, formalin-fixed paraffin-embedded sections in the TCGA-GBM and TCGA-LGG collections. Cases are reclassified using the 2021 World Health Organization criteria to identify IDH-wildtype glioblastoma, CNS WHO grade 4; one section is selected per case. The cohort includes material from 11 institutions [1].

The current patch-level task comprises ten prediction labels:

1. cellular tumor (CT)
2. pseudopalisading necrosis (PN)
3. microvascular proliferation (MP)
4. geographic necrosis (NC)
5. infiltration into cortex (IC)
6. penetration into white matter (WM)
7. leptomeningeal infiltration (LI)
8. dense macrophage regions (DM)
9. presence of lymphocytes (PL)
10. none of the above (NOTA)

The ten subregions are manually delineated under a predefined annotation protocol and reviewed by a certified neuropathologist. The approved regions are divided into 512 × 512-pixel patches, and each patch retains its class, patient and slide identifiers[1].

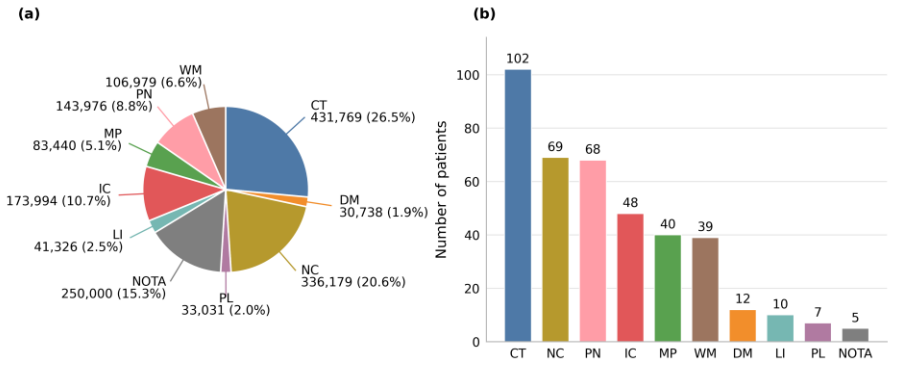

Figure 1. (a) Patch distribution across the ten prediction labels. (b) Number of contributing patients per label.

Figure 1 shows marked imbalance in both patch and patient counts. CT has the most patches, with 431,769 (26.5%), followed by NC with 336,179 (20.6%) and NOTA with 250,000 (15.3%). DM, PL and LI each contribute no more than 2.5% of all patches. CT includes patches from 102 patients, whereas DM, LI, PL and NOTA include patches from 12, 10, 7 and 5 patients, respectively. The imbalance is therefore present at both levels.

2.2 Method Overview

Figure 2 summarizes the workflow. Each H&E patch is encoded by the frozen ALICE vision stage [8]. The resulting embeddings are sampled with the hierarchical class-patient-slide scheme and used to train a ten-class linear classifier.

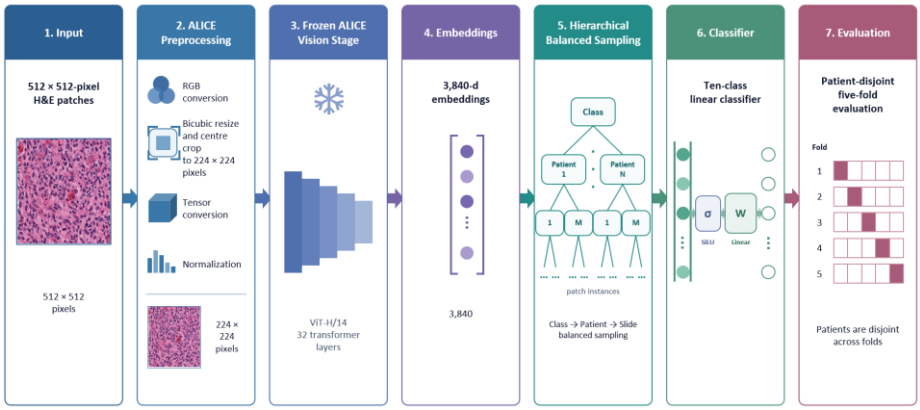

Figure 2. Overview of the ALICE-based framework for patch-level GBM subregion classification.

ALICE maps each preprocessed 512 × 512-pixel H&E patch to a 3,840-dimensional embedding. During training, the hierarchical sampler draws embeddings for a SiLU activation and a ten-output linear layer. Evaluation uses patient-disjoint five-fold cross-validation.

2.3 ALICE-Based Feature Extraction

We use the ALICE vision stage, a ViT-H/14 encoder distilled from the vision-only teachers UNI2[2], Virchow2 [12] and H-optimus-1 [13]. It returns a 3,840-dimensional patch representation from three 1,280-dimensional summary tokens. Only this frozen stage is used; the vision-language and slide-level modules are not included.

Each image is converted to RGB, bicubically resized with antialiasing and centre-cropped to 224 × 224 pixels, converted to a tensor, and normalized using mean (0.48145466, 0.4578275, 0.40821073) and standard deviation (0.26862954, 0.26130258, 0.27577711).

The frozen vision stage runs in 32-bit floating-point precision, and each 3,840-dimensional output is stored with its class, patient and slide identifiers.

2.4 Hierarchical Balanced Sampling

Let $N_{c,h,s}$ denote the number of patches from class c, patient h, and slide s, and N the total number of patches. The empirical distribution under uniform patch sampling is

$$P(c,h,s) = P(c)P(h \mid c)P(s \mid c,h) = \frac{N_{c,h,s}}{N}. \tag{1}$$

To balance the class–patient–slide hierarchy, we define the target distribution as

$$Q(c,h,s) = \frac{1}{C}\frac{1}{H_c}\frac{1}{S_{c,h}}, \tag{2}$$

where C is the number of classes, H_c is the number of patients in class c, and $S_{c,h}$ is the number of slides for patient h in class c. The sampling weight for each patch is therefore

$$w_{c,h,s} = \frac{Q(c,h,s)}{P(c,h,s)} = \frac{N}{CH_cS_{c,h}N_{c,h,s}}. \tag{3}$$

For the i-th patch belonging to class c_i, patient h_i, and slide s_i, the normalized sampling probability is

$$r_i = \frac{w_i}{\sum_{j=1}^{N} w_j} = \frac{1}{CH_{c_i}S_{c_i,h_i}N_{c_i,h_i,s_i}}, \qquad \sum_{i=1}^{N} r_i = 1. \tag{4}$$

This scheme balances classes and, within each class, patients and slides. Before normalization, the weights in Eq. (3) are raised to α, where α = 0 gives uniform sampling and α = 1 full hierarchical balancing. Patches are sampled with replacement, with counts and weights computed from the four training folds only. We tested α from 0 to

1 and selected $\alpha = 0.6$ because it maximized macro-F1 plus MCC. This tempering limits oversampling of rare class-patient-slide groups.

2.5 Experimental Setting

The source code is publicly available at https://github.com/zqf1234588/Brats2026_Pathology. Following consultation with pathologists, we excluded patients P57, P65, P72, P78, and P89 because some annotated classes show low morphological distinguishability in their patches. They were excluded before cross-validation and were not used as a holdout set or for hyperparameter tuning or model selection. The five folds comprise 1,402,607 patches from 244 slides and 121 patients, with 24, 24, 25, 22 and 26 patients per fold. All samples from each patient remain in one-fold, ensuring no training-validation overlap. The official online validation is performed separately through Synapse.

Training uses one epoch, batch size 128, evaluation batch size 512, learning rate 1×10^{-5}, weight decay 0.01, $\alpha = 0.6$ and seed 42. Ablations on the same patient-disjoint folds vary training duration, sampling strategy and patch encoder one at a time, using ALICE with $\alpha = 0.6$ and one epoch as defaults and conventional class-balanced sampling as an additional baseline.

2.6 Evaluation Metrics

We report ACC, AUC, F1, recall, precision, specificity, and MCC. Class specific metrics are computed one versus the rest. Overall ACC and MCC are multiclass, while the remaining overall metrics are macro averaged across classes. Values are reported as mean ± sample standard deviation across five folds.

3 Results

3.1 Local validation

As shown in Table 1, across the five patient-disjoint folds, the classifier achieves an overall ACC (0.803), macro-AUC (0.938), macro-F1 (0.554), macro-recall (0.612), macro-precision (0.569), macro-specificity (0.977) and multiclass MCC (0.753). NC and NOTA show the strongest class-level performance, with high F1 scores (NC, 0.908; NOTA, 0.890), recall (NC, 0.937; NOTA, 0.994) and precision (NC, 0.884; NOTA, 0.833). The next group comprises CT, WM, IC and PN, with F1 scores (CT, 0.749; WM, 0.558; IC, 0.824; PN, 0.651). CT has the lowest specificity among the ten classes (0.916). Performance is lower for MP and DM, with F1 scores (MP, 0.492; DM, 0.344) and recall (MP, 0.521; DM, 0.270). LI and PL are the most difficult classes, with low F1 scores (LI, 0.084; PL, 0.042), recall (LI, 0.171; PL, 0.198) and precision (LI, 0.067; PL, 0.025).

Overall, the relatively small standard deviations across most metrics indicate that the model achieves generally stable performance. Specificity shows the lowest overall

standard deviation (±0.009), followed by AUC (±0.033) and F1-score (±0.036). In comparison, ACC (±0.080) and MCC (±0.071) exhibits slightly greater variability. At the class level, ACC and specificity are generally stable, whereas larger variations are observed for some classes, particularly in the F1-score, recall, and precision. For example, DM shows relatively high variability in precision (±0.486) and F1-score (±0.362), while PL exhibits substantial variation in recall (±0.438). These results suggest that the model is generally robust, although its performance on certain classes remains less consistent.

Figure 3 shows that CT, IC, NC and NOTA are usually classified correctly. LI and MP are often predicted as CT, PL as CT, DM or NC, PN as NC or CT, and WM as CT or IC. Errors are therefore concentrated among subregions.

Table 1. Mean class-specific metrics and overall performance across five patient disjoint folds with values reported as mean ± sample standard deviation across folds.

Label	ACC	AUC	F1	Recall	Precision	Specificity	MCC
CT	0.897 ±0.059	0.952 ±0.042	0.749 ±0.169	0.842 ±0.065	0.712 ±0.228	0.916 ±0.067	0.704 ±0.157
DM	0.983 ±0.008	0.939 ±0.072	0.344 ±0.362	0.270 ±0.302	0.528 ±0.486	0.994 ±0.006	0.362 ±0.371
IC	0.969 ±0.016	0.984 ±0.010	0.824 ±0.137	0.791 ±0.182	0.871 ±0.102	0.987 ±0.006	0.811 ±0.133
LI	0.968 ±0.031	0.837 ±0.206	0.084 ±0.081	0.171 ±0.149	0.067 ±0.062	0.988 ±0.010	0.085 ±0.086
MP	0.951 ±0.040	0.939 ±0.042	0.492 ±0.169	0.521 ±0.229	0.558 ±0.196	0.983 ±0.010	0.493 ±0.160
NC	0.963 ±0.013	0.990 ±0.006	0.908 ±0.040	0.937 ±0.027	0.884 ±0.083	0.972 ±0.013	0.885 ±0.038
PL	0.997 ±0.003	0.852 ±0.188	0.042 ±0.088	0.198 ±0.438	0.025 ±0.048	0.998 ±0.002	0.067 ±0.147
PN	0.932 ±0.034	0.948 ±0.034	0.651 ±0.144	0.687 ±0.103	0.686 ±0.272	0.969 ±0.018	0.632 ±0.125
WM	0.952 ±0.020	0.940 ±0.019	0.558 ±0.217	0.705 ±0.077	0.524 ±0.297	0.970 ±0.020	0.561 ±0.179
NOTA	0.996 ±0.005	1.000 ±0.000	0.890 ±0.159	0.994 ±0.010	0.833 ±0.239	0.996 ±0.005	0.900 ±0.142
Overall	**0.803 ±0.080**	**0.938 ±0.033**	**0.554 ±0.036**	**0.612 ±0.061**	**0.569 ±0.051**	**0.977 ±0.009**	**0.753 ±0.071**

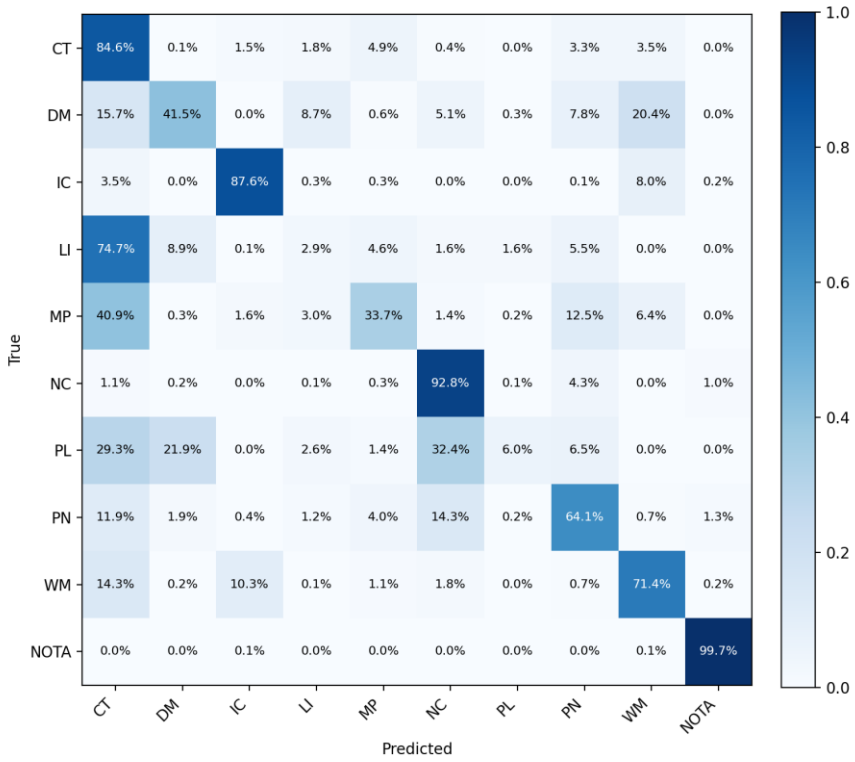

Figure 3. Row normalized out of fold confusion matrix for the ten patch level prediction labels.

3.2 Ablation results

As shown in Table 2, ALICE gives the highest mean macro-F1 and MCC in their respective comparisons for one epoch and $\alpha = 0.6$. The selected sampler outperforms standard sampling without replacement and other sampling strategies with replacement, including uniform, class balanced, and full hierarchical sampling.

Table 2. Ablation of training epochs, sampler strength, and patch encoder with values reported as mean ± sample standard deviation across five folds, where bold indicates the selected configuration.

Component	Setting	Macro-F1	MCC
Epoch	**1**	**0.554 ± 0.036**	**0.753 ± 0.071**
Epoch	2	0.553 ± 0.048	0.749 ± 0.077

Component	Setting	Macro-F1	MCC
Epoch	3	0.548 ± 0.047	0.747 ± 0.078
Epoch	4	0.545 ± 0.049	0.745 ± 0.079
Epoch	5	0.544 ± 0.052	0.744 ± 0.079
Sampler	Standard	0.498 ± 0.032	0.727 ± 0.088
Sampler	α=0	0.500 ± 0.034	0.729 ± 0.088
Sampler	Class-balanced	0.524 ± 0.046	0.726 ± 0.097
Sampler	**α=0.6**	**0.554 ± 0.036**	**0.753 ± 0.071**
Sampler	1	0.553 ± 0.033	0.746 ± 0.068
Encoder	**ALICE**	**0.554 ± 0.036**	**0.753 ± 0.071**
Encoder	Virchow2	0.541 ± 0.040	0.734 ± 0.074
Encoder	UNI2	0.523 ± 0.035	0.715 ± 0.086
Encoder	H-optimus-1	0.526 ± 0.041	0.726 ± 0.082

3.3 Online validation

As shown in Table 3, the model achieves ACC 0.925, AUC 0.956, F1 score 0.757, recall 0.755, specificity 0.989 and MCC 0.896 on the official online validation set. The official test-set results are unavailable.

Table 3. Overall performance on the official online validation set. The official test-set results are unavailable.

Online	ACC	AUC	F1	Recall	Specificity	MCC
valid	0.925	0.956	0.757	0.755	0.989	0.896
Test	N/A	N/A	N/A	N/A	N/A	N/A

4 Discussion

Frozen ALICE embeddings with hierarchical sampling yield macro-F1 0.554 ± 0.036 and MCC 0.753 ± 0.071 in patient-disjoint cross-validation, and F1 0.757 and MCC 0.896 on official online validation. The ablations select ALICE, $\alpha = 0.6$ and one epoch for the final model.

Performance is class dependent. NC performs strongly (F1 0.908), although 32.4% of PL patches are classified as NC; this confusion does not by itself prove morphological similarity. NOTA also achieves F1 0.890 despite only five patients. Because NOTA comprises 250,000 patches that contain none of the nine target subregions, its performance may reflect both sample size and broad morphological separation. LI and PL remain the most difficult classes (F1 0.084 and 0.042). An isolated patch of LI may

contain little recognizable meningeal tissue and resemble CT. For PL, a few small lymphocytes may be obscured by surrounding tumor, necrotic or macrophage-rich morphology. These interpretations require targeted neuropathological review.

The large fold-to-fold variation for DM, LI and PL likely reflects sparse patient coverage and morphological ambiguity. These classes contain only 11, 9 and 6 patients. Under patient-disjoint validation, different held-out folds may therefore contain very few patients from a given class, producing marked variation in recall and precision. Combined with limited context for LI and sparse cellular cues for PL, this may explain both the large standard deviations and low mean fold performance of LI and PL.

Hierarchical sampling limits dominance by classes, patients and slides with many patches, while patient-disjoint splitting prevents leakage. In the ablation, $\alpha = 0.6$ gives the highest combined macro-F1 and MCC; $\alpha = 1$ gives similar macro-F1 but lower MCC, and no sampler, uniform sampling and conventional class balancing perform worse on both metrics. These results provide preliminary support for the selected ALICE representation and tempered hierarchical sampling, but do not show that domain shift or source shortcuts are resolved. The higher online scores are not directly comparable with fold means because the cohort and evaluation differ; cohort difficulty and class composition may contribute to the gap.

Two limitations remain. First, although ALICE outperforms the three individual vision encoders in our ablation, it was not compared with other multi-model fusion methods under the same experimental setting. Second, DM, LI and PL are represented by few patients, making their class-specific estimates sensitive to the composition of the held-out folds. Future work should benchmark alternative fusion methods and expand rare-class patient coverage, together with the targeted neuropathological validation of the observed confusion patterns.

Acknowledgements

This work was supported by Ningbo Major Science & Technology Project under Grant 2022Z126. (Corresponding: Ying Weng.)

References

[1] S. Bakas *et al.*, 'BraTS-path challenge: assessing heterogeneous histopathologic brain tumor sub-regions', May 17, 2024, *arXiv*: arXiv:2405.10871. doi: 10.48550/arXiv.2405.10871.

[2] R. J. Chen *et al.*, 'Towards a general-purpose foundation model for computational pathology', *Nat. Med.*, vol. 30, no. 3, pp. 851–862, Mar. 2024, doi: 10.1038/s41591-024-02857-3.

[3] H. Xu *et al.*, 'A whole-slide foundation model for digital pathology from real-world data', *Nature*, vol. 630, no. 8015, pp. 181–188, Jun. 2024, doi: 10.1038/s41586-024-07441-w.

[4] X. Wang *et al.*, ‘A pathology foundation model for cancer diagnosis and prognosis prediction’, *Nature*, vol. 634, no. 8035, pp. 971–978, Oct. 2024, doi: 10.1038/s41586-024-07894-z.

[5] E. Vorontsov *et al.*, ‘A foundation model for clinical-grade computational pathology and rare cancers detection’, *Nat. Med.*, vol. 30, no. 10, pp. 2924–2935, Oct. 2024, doi: 10.1038/s41591-024-03141-0.

[6] J. Ma *et al.*, ‘A generalizable pathology foundation model using a unified knowledge distillation pretraining framework’, *Nat. Biomed. Eng.*, vol. 10, no. 3, pp. 545–564, Mar. 2026, doi: 10.1038/s41551-025-01488-4.

[7] R. Gao *et al.*, ‘Meta-encoder: a unified integration framework for multiple pathological foundation models in cancer detection’, *Nat. Commun.*, Apr. 2026, doi: 10.1038/s41467-026-71558-x.

[8] J. Li *et al.*, ‘ALICE: learning a general-purpose pathology foundation model from vision, vision-language, and slide-level experts’, Jul. 10, 2026, *arXiv*: arXiv:2607.09526. doi: 10.48550/arXiv.2607.09526.

[9] J. Kömen *et al.*, ‘Towards robust foundation models for digital pathology’, *Nat. Commun.*, vol. 17, no. 1, p. 5218, Jun. 2026, doi: 10.1038/s41467-026-73923-2.

[10] Y. Huang *et al.*, ‘Knowledge-guided adaptation of pathology foundation models effectively improves cross-domain generalization and demographic fairness’, *Nat. Commun.*, vol. 16, no. 1, Dec. 2025, doi: 10.1038/s41467-025-66300-y.

[11] S. Kapse, M. Aygün, E. Cole, E. Lundberg, L. Song, and E. P. Xing, ‘GenBio-PathFM: a state-of-the-art foundation model for histopathology’, Mar. 20, 2026, *bioRxiv*. doi: 10.64898/2026.03.17.712534.

[12] E. Zimmermann *et al.*, ‘Virchow2: scaling self-supervised mixed magnification models in pathology’, Nov. 06, 2024, *arXiv*: arXiv:2408.00738. doi: 10.48550/arXiv.2408.00738.

[13] ‘Bioptimus | news: bioptimus launches H-optimus-1: a state-of-the-art foundation model for pathology’. Accessed: Apr. 26, 2026. [Online]. Available: https://www.bioptimus.com/news/bioptimus-launches-h-optimus-1