HistoTx: Early fusion of H&E images and spatial transcriptomics at varying spatial transcriptomics resolution with self-supervised learning

Alice Driessen [*1] **Emil Malmsten** [1] **Steffen Wolf** [1] **Bernd Illing** [2 3] **Isinsu Katircioglu** [2 3] **Adriano Martinelli** [2 3] **Raphael Gottardo** [2 3 4 5] **Marianna Rapsomaniki** [2 3 4] **Fei Tang** [1]

Abstract

Spatial transcriptomics maps gene expression across tissue sections and, when integrated with whole slide images, jointly captures both morphological architecture and molecular activity to provide a rich, multimodal view of tissue biology. ML approaches combining spatial transcriptomics and WSIs typically learn a mapping from morphological features to molecular profiles via contrastive alignment or supervised fine-tuning of pathology foundation models. In both cases, transcriptomics serves as a training signal rather than a true input modality, leaving inference fundamentally image-driven. We propose an early-fusion vision transformer architecture in which transcript tokens are constructed at arbitrary spatial resolution, merged with image patch tokens, and jointly processed through a shared transformer stack for deep cross-modal interaction. Starting from a pretrained vision-only pathology foundation model, we train `HistoTx` by continuing self-supervised pretraining on paired image and spatial transcriptomics data and demonstrate that: 1) at inference with image only, performance remains on par with the baseline on most tasks while improving on tasks that inherently benefit from molecular knowledge, such as gene expression prediction; 2) when both modalities are available at inference, jointly providing image and transcript inputs outperforms either modality alone.

[1]kaiko.ai, Zürich, Switzerland [2]Biomedical Data Science Center, Lausanne University Hospital, Lausanne, Switzerland [3]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland [4]Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland [5]School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. Correspondence to: Alice Driessen <alice.driessen@kaiko.ai>, Fei Tang <fei@kaiko.ai>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Hematoxylin and eosin (H&E)-stained whole slide images (WSIs) are the gold standard for cancer diagnosis, routinely used for tumor detection, subtyping, and clinical decision making, yet their interpretation remains a labor-intensive task demanding years of specialized pathologist training (Thomas & Button, 2015). Recent advances in artificial intelligence have led to the emergence of large pathology foundation models trained on vast collections of WSIs, achieving strong performance across a broad range of clinically relevant tasks (Wang et al., 2022; Chen et al., 2024; Xu et al., 2024), yet their capabilities remain limited for tasks that could benefit from additional molecular knowledge.

In parallel, advances in spatial transcriptomics (ST) technologies now enable molecular profiling directly on tissue sections, facilitating precise cell type identification, detection of functional cell states, and richer tumor characterization (Jain & Eadon, 2024). ST data comes in different formats depending on the platform. For example, high-resolution Xenium data consists of individual transcript detections, each annotated with its spatial location and gene identity but limited to a fixed gene panel. By contrast, Visium data is whole-transcriptome based and consists of a regular grid of barcoded spots, each with a diameter of 55 µm, where transcripts are aggregated into a single expression profile per spot. Due to the different nature and granularity of the ST data it is nontrivial to train a model on both Xenium and Visium data. In fact most models use only on spot-based (Chen et al., 2025; Chung et al., 2024; Pang et al., 2021; Xie et al.; Hemker et al., 2026; Zhan et al.) or probe-based data (Fu et al., 2025; Martinelli et al., 2026; Gindra et al., 2025) for training. To our best knowledge, only MINT (Lee et al., 2026) is able to encode both modalities with one model, however this model is limited to one learnable ST token.

To leverage molecular information alongside histology, early approaches trained supervised models to predict gene expression directly from H&E images (Schmauch et al., 2025; Chung et al., 2024; Nonchev et al., 2025; Pang et al., 2021; Hu et al., 2026). More recently, methods inspired by CLIP have employed contrastive learning to align image and ST

embeddings in a shared latent space, using the transcriptomic modality merely as a supervision signal to guide visual representation learning (Chen et al., 2025; Zhang et al., 2026; Xie et al.; Hemker et al., 2026). Multimodal fusion is reduced to late-stage feature-level concatenation or averaging, rather than being deeply integrated into the model architecture. We argue that this image-centric strategy is well motivated in practice: given the high cost of ST, molecular profiles are rarely available in routine clinical workflows, whereas H&E images are inexpensive and virtually ubiquitous. Leveraging limited ST data as a supervision signal to endow the image encoder with molecular knowledge is therefore a natural and clinically practical design choice. However, existing contrastive approaches apply this supervision crudely, typically at the coarse patch level, failing to exploit the fine-grained spatial correspondence between transcript distributions and image content. At the same time, in the cases where transcriptomics data are available at inference, deeper and more principled multimodal fusion beyond simple concatenation or averaging remains desirable to fully exploit the complementary information carried by both modalities, as can be achieved by early-fusion (Martinelli et al., 2026).

We present `HistoTx`, an early-fusion multimodal transformer building on (Martinelli et al., 2026) that integrates H&E images and ST. We extend the architecture to accommodate both Visium and Xenium platforms by constructing spatially aligned image and transcript tokens at multiple spatial resolutions, merging them into a joint token sequence, and processing them through a shared transformer stack for deep cross-modal interaction. `HistoTx` is trained by continuing DINOv2 (Oquab et al., 2024) self-supervised pretraining on paired image and ST data, warm-started from a pretrained vision-only pathology foundation model. By constructing tokens at multiple spatial resolutions, the model captures biological structures across scales, from subcellular transcript patterns to tissue-level morphology. The two main contributions of this work are:

- **Multimodal self-supervised pretraining:** a training strategy that continues DINOv2 pretraining on paired image and ST data from a vision-only pathology foundation model, yielding a model that retains strong image-only performance, improves on tasks requiring molecular knowledge, and further benefits when both modalities are available at inference.

- **Flexible transcript tokenization strategy:** This enables a single backbone to handle both Xenium and Visium data, allowing training on the full HEST-1k (Jaume et al., 2024) dataset and broadening exposure to diverse tissues, gene panels, and spatial scales.

2. Methods

2.1. Model architecture and training

`HistoTx` is a multimodal foundation model that jointly encodes tissue morphology from histology images and gene expression from spatial transcriptomics data (see figure 1). It comprises a vision transformer based morphology encoder as the backbone, and a gene expression encoder.

2.1.1. Unimodal tokenization

We train on a mixture of Visium spot-based and probe-based Xenium data. Xenium provides individual transcript coordinates, enabling tokens to be constructed at arbitrary spatial resolution by counting per-gene detections within non-overlapping $T \times T$ pixel regions, each yielding one transcript count vector. By varying the size of T, we can adapt the number of transcript tokens. For an image of size 256×256, this yields 256, 64, or 1 tokens for $T = 16, 32,$ or 256 respectively. The physical area covered by each token depends jointly on T and the image resolution: at mpp $= 0.25$, $T = 16$ corresponds to a $4 \times 4\mu m^2$ region, while at mpp $= 0.5$, $T = 64$ covers 32×32 μm^2. By varying T and mpp, tokens can be constructed at different spatial scales, capturing structures from subcellular to tissue-level.

Since Visium data is spot-based, providing only aggregated expression at fixed spot locations rather than individual transcript coordinates, we follow HEST-1k (Jaume et al., 2024) and extract 224×224 pixel images at mpp $= 0.5$ centered on each spot yielding one transcript token per image. At coarser resolutions, a single image encompasses multiple spots, up to 4 and 16 at mpp $= 1$ and mpp $= 2$ respectively, each contributing its aggregated count vector as a separate transcript token.

Transcript count vectors are encoded by a gene expression encoder: here we use the pretrained gene foundation model Geneformer (Theodoris et al., 2023), producing transcript token embeddings of dimension D_{expr}.

Images are tokenized via the patch embedding layer of the morphology encoder, producing N_I image token embeddings of dimension D_{morph} per image. We deliberately decouple the spatial resolution of transcript tokens from that of image tokens for two reasons: first, high-resolution Xenium data enabling fine-grained transcript tokenization is scarce and decoupling allows Visium data to complement Xenium in training; second, at very high resolution, individual transcript tokens may be dominated by noise rather than meaningful signal.

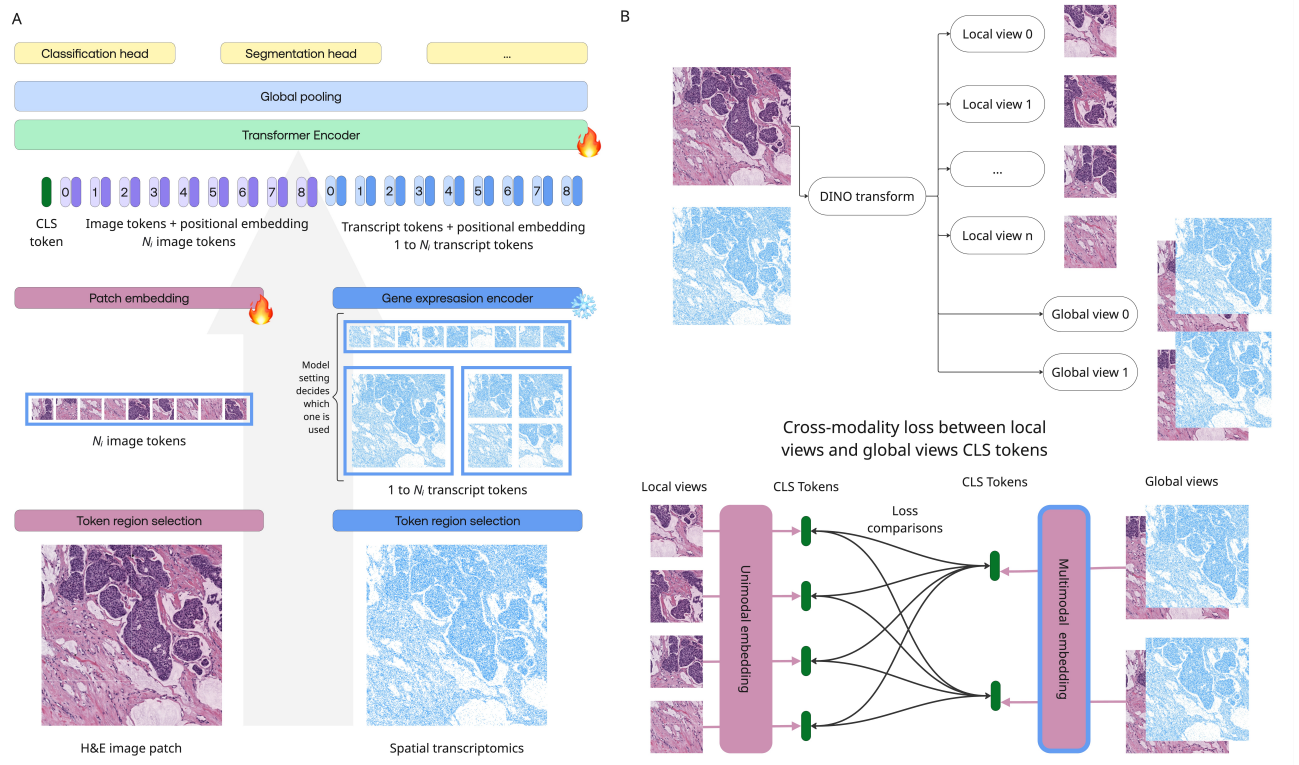

Figure 1. Early-fusion multimodal model architecture and cross-modal loss. A) Input patches contain the H&E image and the corresponding spatial transcriptomics. Both modalities are tokenized into smaller token regions and embedded with modality-specific token encoders. The tokens of both modalities are concatenated and fed through a ViT block. The snow flake emoji indicates the frozen part of the model and the flame the part where model weights are optimized. B) DINO transforms created n local views and 2 global views. By removing transcripts from the local views, the DINO local view loss on the CLS tokens becomes a cross-modality loss.

2.1.2. Multimodal fusion

After tokenization, transcript embeddings are projected to match the image token dimension via a linear layer and rescaled to match the norm of the image tokens. The morphology encoder's positional embeddings are then applied to both modalities; when the number of transcript tokens N_T differs from the number of image tokens N_I, positional embeddings are adapted to the transcript spatial grid via dynamic spatial resampling (Dosovitskiy et al., 2021). The fused sequence $[\text{CLS} \mid \mathbf{z}_I^{1:N_I} \mid \mathbf{z}_G^{1:N_G}]$ of length $1+N_I+N_G$ is passed through the remaining ViT blocks. The output CLS token serves as the global multimodal representation, while patch and transcript tokens occupy the remaining positions. At inference, the model supports CLS, mean, max, or flatten pooling over the output sequence.

2.1.3. Self-supervised learning

Starting from a pathology foundation model pretrained on large WSI collections, we continue pretraining on paired image-transcript data using DINOv2. To accommodate the multimodal setting, we extend the notion of an image view to image-transcript pairs. For Xenium data, the random crop and resize transformation is applied jointly to both modalities, retaining only transcript detections falling within the cropped region; the cropped image then undergoes standard photometric augmentation in DINO transform (colour jitter, Gaussian blur, etc.). To introduce a cross-modality learning signal within the DINOv2 framework, transcripts are included only in global views while local views remain image-only, such that global CLS tokens are multimodal and local CLS tokens are image-only. The local crop loss then enforces consistency between these two, driving cross-modal knowledge distillation from transcripts into the visual encoder. (see figure 1B). For Visium data, where we always obtain one single transcript token per patch, the token is retained for global views and omitted for local views.

Our experiments are conducted primarily on a ViT-B/14 backbone initialized from an internal Midnight (Karasikov et al., 2025) checkpoint of the same size. The gene expression encoder uses the *gf-6L-10M-i2048* variant of GeneFormer, kept frozen throughout training. Full model and training hyperparameters are reported in Appendix A.

Table 1. Gene expression tasks: Hbench: HEST-Bench, Hexpr: HEST expression prediction data from (Jaume et al., 2024); HexprIO: Hest expression task with image only input, Classification tasks: BH: BreakHis , GTMA: Gleason TMA , PR: PANDA Radboud , PK: PANDA Karolinska; Segmentation tasks: CNP: CoNSeP , MNS: MoNuSAC, TTS: tan tock seng. See Appendix B for more task details. MCAcc: Multiclass accuracy, BBAcc: Binary Balance accuracy, MDS: Monai Dice Score, PC: Mean Pearson correlation coefficient.

Model	HBench PC	HExprIO PC	BH MCAcc	GTMA MCAcc	PK BBAcc	PR MCAcc	TTS MDS	CNP MDS	MNS MDS	Avg
Image only	0.34	0.26	**0.82**	0.71	0.70	0.61	0.31	0.61	0.61	0.56
HistoTx	**0.40**	**0.32**	0.72	**0.73**	**0.72**	**0.66**	**0.37**	**0.62**	**0.63**	**0.59**
SEAL + CONCH	0.38	0.28	0.68	**0.73**	0.71	**0.66**	**0.37**	**0.62**	**0.63**	0.58

2.2. Training datasets

We use the **HEST1k** (Jaume et al., 2024) and **TCGA** (TCGA, 2025) dataset to train HistoTx. Training details and dataloader weights are listed in Appendix A. TCGA is an image-only dataset used to preserve the model's base image processing capabilities. Since, HEST1k contains H&E images with spot-based (Visium Spatial Transcriptomics) or Xenium data we use this data for multimodal training, as well as image-only training.

3. Results

We trained a family of HistoTx models differing in how expression tokens are constructed from Xenium data; implementation details and full results are provided in Appendix A, Tables 3 and 4. Here we focus on the variant using one expression token per Visium spot and 16 expression tokens per Xenium image of size 224x224 px. We hypothesize that too many small tokens introduce noise as transcript counts at very fine spatial scales are unreliable, while too few tokens sacrifice spatial resolution and molecular detail.

We compare HistoTx against two baselines: the vision-only foundation model used as our starting point, and SEAL (Hemker et al., 2026), which adapts a pretrained image encoder via CLIP-style multimodal fine-tuning. For SEAL we use the CONCH backbone, as CONCH is also a ViT-B model, comparable in scale to our HistoTx backbone. Downstream benchmarks include image-only classification (BreakHis, Gleason TMA, PANDA Karolinska), image-only segmentation (Tan Tock Seng, MoNuSAC), and gene expression prediction in both image-only and multimodal settings (HEST-Bench, HEST-Expr); see Appendix B for task descriptions.

HistoTx outperforms the image-only baseline on most benchmark tasks, with the most notable gains on gene expression prediction. On HEST-Bench, which requires predicting 50 highly variable genes from H&E patches, Pearson correlation improves from 0.34 to 0.40. This improvement is partially attributable to multimodal pretraining, though increased exposure to HEST images during continued pretraining may also contribute. HistoTx also outperforms the contrastive SEAL-CONCH baseline on this task.

On HEST-Expr, where the goal is to predict a stable panel of 380 genes from H&E patches, HistoTx improves the Pearson correlation over the image-only baseline and SEAL-CONCH from 0.26 and 0.28, respectively to 0.32 in the image-only setting. When 1,000 input genes are additionally provided, it increases further to 0.38 (Table 4), demonstrating HistoTx's ability to leverage multimodal information at inference.

On multi-class classification tasks, we observe performance drops on breast cancer benchmarks, shown here for BreakHis (see Table 3 for additional tasks BACH & BRACS) and mixed results on prostate tasks: minor improvement on PANDA-Radboud but a loss on Gleason TMA. These tasks share a common difficulty: distinguishing visually similar cancer subtypes. Continued multimodal pretraining may reorient the encoder toward gene-expression-aligned features, weakening the subtle morphological detail these subtype tasks rely on. Whether this is tissue-specific or task-specific remains open.

In summary we trained HistoTx, an early-fusion multimodal transformer that integrates H&E image and spatial transcriptomics tokens into a shared ViT token sequence, by continuing DINOv2 self-supervised pretraining on paired image and ST data from a vision-only pathology foundation model. HistoTx outperforms the image-only baseline and SEAL-CONCH on gene expression prediction tasks HEST-Bench and HEST-Expr, and shows further improvements when transcript inputs are available at inference. Performance on image-only classification tasks is largely preserved, though drops on breast cancer benchmarks suggest a potential trade-off between multimodal representation learning and fine-grained subtype discrimination.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A., and Campilho, A. Classification of breast cancer histology images using Convolutional Neural Networks. *PLOS ONE*, 12(6):e0177544, June 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0177544. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0177544.

Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S. S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., Fernandez, G., Zeineh, J., Kohl, M., Walz, C., Ludwig, F., Braunewell, S., Baust, M., Vu, Q. D., To, M. N. N., Kim, E., Kwak, J. T., Galal, S., Sanchez-Freire, V., Brancati, N., Frucci, M., Riccio, D., Wang, Y., Sun, L., Ma, K., Fang, J., Kone, I., Boulmane, L., Campilho, A., Eloy, C., Polónia, A., and Aguiar, P. BACH: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56:122–139, August 2019. ISSN 1361-8423. doi: 10.1016/j.media.2019.05.010.

Arvaniti, E., Fricker, K. S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P. J., Rüschoff, J. H., and Claassen, M. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific Reports*, 8(1):12054, August 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-30535-1. URL https://www.nature.com/articles/s41598-018-30535-1.

Brancati, N., Anniciello, A. M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., Gabrani, M., Feroce, F., and Frucci, M. BRACS: A Dataset for BReAst Carcinoma Subtyping in H&E Histology Images. *Database: The Journal of Biological Databases and Curation*, 2022:baac093, October 2022. ISSN 1758-0463. doi: 10.1093/database/baac093.

Bulten, W., Kartasalo, K., Chen, P.-H. C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D. F., van Boven, H., Vink, R., Hulsbergen-van de Kaa, C., van der Laak, J., Amin, M. B., Evans, A. J., van der Kwast, T., Allan, R., Humphrey, P. A., Grönberg, H., Samaratunga, H., Delahunt, B., Tsuzuki, T., Häkkinen, T., Egevad, L., Demkin, M., Dane, S., Tan, F., Valkonen, M., Corrado, G. S., Peng, L., Mermel, C. H., Ruusuvuori, P., Litjens, G., Eklund, M., and PANDA challenge consortium. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine*, 28(1):154–163, January 2022. ISSN 1546-170X. doi: 10.1038/s41591-021-01620-2.

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., and Mahmood, F. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, March 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02857-3. URL https://www.nature.com/articles/s41591-024-02857-3.

Chen, W., Zhang, P., Tran, T. N., Xiao, Y., Li, S., Shah, V. V., Cheng, H., Brannan, K. W., Youker, K., Lai, L., Fang, L., Yang, Y., Le, N.-T., Abe, J.-i., Chen, S.-H., Ma, Q., Chen, K., Song, Q., Cooke, J. P., and Wang, G. A visual–omics foundation model to bridge histopathology with spatial transcriptomics. *Nature Methods*, 22(7):1568–1582, July 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02707-1. URL https://www.nature.com/articles/s41592-025-02707-1.

Chung, Y., Ha, J. H., Im, K. C., and Lee, J. S. Accurate Spatial Gene Expression Prediction by Integrating Multi-Resolution Features. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11591–11600, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.01101. URL https://ieeexplore.ieee.org/document/10657393/.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A. W. M., the CAMELYON16 Consortium, Hermsen, M., Manson, Q. F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M. C., Bult, P., Beca, F., Beck, A. H., Wang, D., Khosla, A., Gargeya, R., Irshad, H., Zhong, A., Dou, Q., Li, Q., Chen, H., Lin, H.-J., Heng, P.-A., Haß, C., Bruni, E., Wong, Q., Halici, U., Öner, M. U., Cetin-Atalay, R., Berseth, M., Khvatkov, V., Vylegzhanin, A., Kraus, O., Shaban, M., Rajpoot, N., Awan, R., Sirinukunwattana, K., Qaiser, T., Tsang, Y.-W., Tellez, D., Annuscheit, J., Hufnagl, P., Valkonen, M., Kartasalo, K., Latonen, L., Ruusuvuori, P., Liimatainen, K., Albarqouni, S., Mungal, B., George, A., Demirci, S., Navab, N., Watanabe, S., Seno, S., Takenaka, Y., Matsuda, H., Ahmady Phoulady, H., Kovalev, V., Kalinovsky, A., Liauchuk, V., Bueno, G., Fernandez-Carrobles, M. M., Serrano, I., Deniz, O., Racoceanu, D., and Venâncio, R. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With

Breast Cancer. *JAMA*, 318(22):2199–2210, December 2017. ISSN 1538-3598. doi: 10.1001/jama.2017.14585.

Fondón, I., Sarmiento, A., García, A. I., Silvestre, M., Eloy, C., Polónia, A., and Aguiar, P. Automatic classification of tissue malignancy for breast carcinoma diagnosis. *Computers in Biology and Medicine*, 96:41–51, May 2018. ISSN 1879-0534. doi: 10.1016/j.compbiomed.2018.03.003.

Fu, X., Cao, Y., Bian, B., Wang, C., Graham, D., Pathmanathan, N., Patrick, E., Kim, J., and Yang, J. Y. H. Spatial gene expression at single-cell resolution from histology using deep learning with GHIST. *Nature Methods*, 22(9):1900–1910, September 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02795-z. URL https://www.nature.com/articles/s41592-025-02795-z.

Gindra, R. H., Palla, G., Nguyen, M., Wagner, S. J., Tran, M., Theis, F. J., Saur, D., Crawford, L., and Peng, T. A Large-Scale Benchmark of Cross-Modal Learning for Histology and Gene Expression in Spatial Transcriptomics, August 2025. URL http://arxiv.org/abs/2508.01490. arXiv:2508.01490 [q-bio].

Graham, S., Vu, Q. D., Raza, S. E. A., Azam, A., Tsang, Y. W., Kwak, J. T., and Rajpoot, N. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, December 2019. ISSN 1361-8423. doi: 10.1016/j.media.2019.101563.

Hemker, K., Song, A. H., Almagro-Pérez, C., Jaume, G., Wagner, S. J., Vaidya, A., Simidjievski, N., Jamnik, M., and Mahmood, F. Towards Spatial Transcriptomics-driven Pathology Foundation Models, 2026. URL https://arxiv.org/abs/2602.14177. Version Number: 1.

Hu, S., Zeng, Q., Bhasker, N., Kather, J. N., and Speidel, S. HistoPrism: Unlocking Functional Pathway Analysis from Pan-Cancer Histology via Gene Expression Prediction, 2026. URL https://arxiv.org/abs/2601.21560. Version Number: 3.

Jain, S. and Eadon, M. T. Spatial transcriptomics in health and disease. *Nature Reviews Nephrology*, 20(10):659–671, October 2024. ISSN 1759-507X. doi: 10.1038/s41581-024-00841-1. URL https://www.nature.com/articles/s41581-024-00841-1.

Jaume, G., Doucet, P., Song, A. H., Lu, M. Y., Almagro-Pérez, C., Wagner, S. J., Vaidya, A. J., Chen, R. J., Williamson, D. F. K., Kim, A., and Mahmood, F. HEST-1k: A Dataset for Spatial Transcriptomics and Histology Image Analysis, November 2024. URL http://arxiv.org/abs/2406.16192. arXiv:2406.16192 [cs].

Karasikov, M., Doorn, J. v., Känzig, N., Cesur, M. E., Horlings, H. M., Berke, R., Tang, F., and Otálora, S. Training state-of-the-art pathology foundation models with orders of magnitude less data. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15967, pp. 573–583. Springer Nature Switzerland, October 2025. doi: 10.1007/978-3-032-04984-1_55.

Kather, J. N., Halama, N., and Marx, A. 100,000 histological images of human colorectal cancer and healthy tissue, April 2018. URL https://zenodo.org/records/1214456.

Lee, M., Kim, J., Yun, J., Yu, S., and Jang, J. MINT: Molecularly Informed Training with Spatial Transcriptomics Supervision for Pathology Foundation Models, March 2026. URL http://arxiv.org/abs/2603.07895. arXiv:2603.07895 [cs].

Martinelli, A., Illing, B., Katircioglu, I., Driessen, A., Tang, F., Berke, R., Gottardo, R., and Rapsomaniki, M. Learning Joint Morpho-Molecular Tissue Representations with a Multimodal Transformer. March 2026. URL https://openreview.net/forum?id=h2GcySraTP.

Nonchev, K., Dawo, S., Silina, K., Moch, H., Andani, S., Consortium, T. P., Koelzer, V. H., and Rätsch, G. DeepSpot: Leveraging Spatial Context for Enhanced Spatial Transcriptomics Prediction from H&E Images, July 2025. URL https://www.medrxiv.org/content/10.1101/2025.02.09.25321567v3. ISSN: 3067-2007 Pages: 2025.02.09.25321567.

Oner, M. U., Ng, M. Y., Giron, D. M., Chen Xi, C. E., Yuan Xiang, L. A., Singh, M., Yu, W., Sung, W.-K., Wong, C. F., and Lee, H. K. An AI-assisted tool for efficient prostate cancer diagnosis in low-grade and low-volume cases. *Patterns*, 3(12):100642, November 2022a. ISSN 2666-3899. doi: 10.1016/j.patter.2022.100642. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9768677/.

Oner, M. U., Ng, M. Y., Giron, D. M., Xi, C. E. C., Xiang, L. A. Y., Singh, M., Yu, W., Sung, W.-K., Wong, C. F., and Lee, H. K. Digital Pathology Dataset for Prostate Cancer Diagnosis, February 2022b. URL https://zenodo.org/records/7152243.

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., and Bojanowski, P. DINOv2: Learning Robust Visual Features without Supervision, February 2024. URL http://arxiv.org/abs/2304.07193. arXiv:2304.07193 [cs].

Pang, M., Su, K., and Li, M. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors, November 2021. URL https://www.biorxiv.org/content/10.1101/2021.11.28.470212v1. Pages: 2021.11.28.470212 Section: New Results.

Polónia, A., Eloy, C., and Aguiar, P. BACH Dataset : Grand Challenge on Breast Cancer Histology images, May 2019. URL https://zenodo.org/records/3632035.

Schmauch, B., Herpin, L., Olivier, A., Duboudin, T., Dubois, R., Gillet, L., Filiot, A., Schiratti, J.-B., Di Proietto, V., Le Corre, D., Bourgoin, A., Taïeb, J., Emile, J.-F., Fridman, W. H., Pronier, E., Laurent-Puig, P., and Durand, E. Y. A deep learning-based multiscale integration of spatial omics with tumor morphology. *Nature Communications*, 16(1):11674, November 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-66691-y. URL https://www.nature.com/articles/s41467-025-66691-y.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. A Dataset for Breast Cancer Histopathological Image Classification. *IEEE transactions on bio-medical engineering*, 63(7):1455–1462, July 2016. ISSN 1558-2531. doi: 10.1109/TBME.2015.2496264.

TCGA. The Cancer Genome Atlas, May 2025. URL https://registry.opendata.aws/tcga.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., and Ellinor, P. T. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, June 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06139-9. URL https://www.nature.com/articles/s41586-023-06139-9.

Thomas, H. and Button, M. Pathology in cancer. In Macbeth, F., Hanna, L., and Crosby, T. (eds.), *Practical Clinical Oncology*, pp. 42–53. Cambridge University Press, Cambridge, 2 edition, 2015. ISBN 978-1-107-28025-0. doi: 10.1017/CBO9781107280250.006. URL https://www.cambridge.org/core/books/practical-clinical-oncology/pathology-in-cancer/84A6267051133C371CEFCF093449C41F.

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M. Rotation Equivariant CNNs for Digital Pathology. In Frangi, A. F., Schnabel, J. A., Davatzikos, C., Alberola-López, C., and Fichtinger, G. (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pp. 210–218, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00934-2. doi: 10.1007/978-3-030-00934-2_24.

Verma, R., Kumar, N., Patil, A., Kurian, N. C., Rane, S., Graham, S., Vu, Q. D., Zwager, M., Raza, S. E. A., Rajpoot, N., Wu, X., Chen, H., Huang, Y., Wang, L., Jung, H., Brown, G. T., Liu, Y., Liu, S., Jahromi, S. A. F., Khani, A. A., Montahaei, E., Baghshah, M. S., Behroozi, H., Semkin, P., Rassadin, A., Dutande, P., Lodaya, R., Baid, U., Baheti, B., Talbar, S., Mahbod, A., Ecker, R., Ellinger, I., Luo, Z., Dong, B., Xu, Z., Yao, Y., Lv, S., Feng, M., Xu, K., Zunair, H., Hamza, A. B., Smiley, S., Yin, T.-K., Fang, Q.-R., Srivastava, S., Mahapatra, D., Trnavska, L., Zhang, H., Narayanan, P. L., Law, J., Yuan, Y., Tejomay, A., Mitkari, A., Koka, D., Ramachandra, V., Kini, L., and Sethi, A. MoNuSAC2020: A Multi-Organ Nuclei Segmentation and Classification Challenge. *IEEE transactions on medical imaging*, 40 (12):3413–3423, December 2021. ISSN 1558-254X. doi: 10.1109/TMI.2021.3085712.

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., and Han, X. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, October 2022. ISSN 1361-8415. doi: 10.1016/j.media.2022.102559. URL https://www.sciencedirect.com/science/article/pii/S1361841522002043.

Wei, J., Suriawinata, A., Ren, B., Liu, X., Lisovsky, M., Vaickus, L., Brown, C., Baker, M., Tomita, N., Torresani, L., Wei, J., and Hassanpour, S. A Petri Dish for Histopathology Image Analysis, January 2021. URL https://arxiv.org/abs/2101.12355v2.

Xie, R., Pang, K., Chung, S. W., Perciani, C. T., MacParland, S. A., Wang, B., and Bader, G. D. Spatially Resolved Gene Expression Prediction from H&E Histology Images via Bi-modal Contrastive Learning.

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., and Poon, H. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, June 2024. ISSN 1476-4687. doi: 10.1038/

s41586-024-07441-w. URL https://www.nature.com/articles/s41586-024-07441-w.

Zhan, X., Xu, J., Zheng, Y., Good, Z., and Gevaert, O. SAGE-FM: A lightweight and interpretable spatial transcriptomics foundation model.

Zhang, D., Gong, Z., Pang, X., Liu, J., Lu, J., Cui, H., Ge, J., Zeng, Z., Yi, K., Li, Y., Liu, S., Yu, T., Wang, H., Crispin-Ortuzar, M., Yu, W., Li, C., and Gao, Z. CARE: A Molecular-Guided Foundation Model with Adaptive Region Modeling for Whole Slide Image Analysis, March 2026. URL http://arxiv.org/abs/2602.21637. arXiv:2602.21637 [cs] version: 3.

A. Model details

A.1. Training data

- 7924 **TCGA** (TCGA, 2025) diagnostic slides, also used in training the baseline backbone, are used to preserve the model's base image processing capabilities.
- **HEST-1k** (v1.1.0): 1,229 WSIs spanning 26 organs and two organisms (human and mouse), comprising 65 Xenium samples and 1,164 spot-based expression samples (Visium and Spatial Transcriptomics) (Jaume et al., 2024). From this we constructed paired image-transcript samples from human Xenium and Visium data at varying image and transcript token resolutions, as well as image-only samples from the WSIs. We follow the official HEST-Bench train/test split, using only the training split during pretraining.

A.2. Architecture

We trained a family of multimodal foundation models that jointly encode tissue morphology and gene expression. The morphology branch is a ViT-B/14 backbone, initialized from a DINOv2 checkpoint pretrained in-house on pathology data and fine-tuned during training. The gene expression branch is a frozen Geneformer encoder (gf-6L-10M-i2048) (Theodoris et al., 2023). Its output passes through a `BatchNorm1d` layer over the embedding dimension, which re-scales each feature to zero mean and unit variance (with a learned affine transform), and is then linearly projected to the morphology embedding dimension. Before fusion, every expression token is additionally rescaled in L2 norm to match its corresponding morphology token (when expression and morphology tokens are aligned one-to-one) or to the mean norm of the morphology patch tokens (when expression tokens are aggregated). The two streams are then fused by token concatenation before being projected by the DINO head. The morphology features were sized as 16×16 tokens per patch.

Variants of the model differ only in how gene expression is aggregated into tokens (Section A.6); the rest of the architecture, optimization, and data pipeline is shared across all configurations.

A.3. Self-supervised framework

Models were trained with the DINOv2 objective including the iBOT masked-image-modeling auxiliary loss (Oquab et al., 2024). The full criterion is a weighted sum of a DINO loss (weight 1.0), an iBOT loss (weight 1.0), and a KDE regularizer (weight 0.05). Teacher targets were produced with Sinkhorn–Knopp centering and a warmup teacher temperature of 0.07.

The DINO projection head followed the standard three-layer MLP design with input dimension 768, hidden dimension 2048, bottleneck dimension 384, and output dimension 131'072. The last layer of the head was kept frozen for the first three epochs of training. Both the morphology backbone and the projection head were initialized from the same in-house DINOv2 pathology run; the head's prototypes were therefore warm-started rather than learned from scratch.

A.4. Optimization

We optimized with AdamW ($\beta = (0.9, 0.999)$) at a base learning rate of 2×10^{-4} and an initial weight decay of 0.04. The learning rate followed a cosine schedule with a 10-epoch linear warmup, decaying to 10% of the base value over 200 epochs. Weight decay was cosine-annealed from 0.04 to 0.2, and the EMA teacher momentum was cosine-annealed from 0.994 to 1.0. Training used BF16 mixed precision, gradient clipping at 1.0, distributed data parallelism (DDP), and a fixed random seed of 42. All models were trained for at most 200 epochs.

A.5. Data and patch extraction

Training mixed four data streams via weighted sampling: HEST–Xenium with paired spatial transcripts (15%), TCGA image-only (15%), HEST image-only (15%), and HEST spot-level expression supplied as cell sentences (55%). HEST–Xenium patches were filtered to retain at least 100 transcripts and 10 unique genes; HEST cell-sentence inputs were truncated to a maximum of 200 genes per spot.

Image patches were 224×224 pixels and were sampled at multiple resolutions: 0.5, 1.0, and 2.0 µm/px for HEST–Xenium, and 0.25, 0.5, 1.0, and 2.0 µm/px for the TCGA and HEST image-only streams. We applied DINOv2 multi-crop augmentation with two global views (224×224) and six local views (112×112). iBOT block-wise masking was applied with probability 0.5 and a mask ratio drawn uniformly from $[0.1, 0.5]$.

A.6. Configuration-specific details

The four variants differ in how spatial transcripts are aggregated into expression tokens before fusion. A sliding window with kernel size k and stride k (non-overlapping) is applied over the 224×224 patch, yielding $(224/k)^2$ expression tokens per patch (Table 2). The 1-token variant aggregates the entire patch into a single token, the 16- and 64-token variants produce coarse spatial grids (4×4 and 8×8), and the 256-token variant produces a 16×16 grid that aligns one expression token with each morphology patch token.

To keep optimization comparable across variants, gradient accumulation and per-epoch batch counts were adjusted so that all configurations converged on an effective batch size of 1024 samples and approximately 78 optimization steps per epoch ($\approx 640'000$ training samples per epoch), absorbing differences in the number of GPUs available in each compute environment. The per-GPU micro-batch size was 4 in all configurations.

Table 2. Configuration-specific hyperparameters for the four expression token variants.

Parameter	**1 token**	**16 tokens**	**64 tokens**	**256 tokens**
Expression aggregation kernel	224×224	56×56	28×28	14×14
Expression aggregation stride	224	56	28	14
Expression tokens per patch	1	16 (4×4)	64 (8×8)	256 (16×16)
GPUs	8	2	8	4
Gradient accumulation steps	32	128	32	64

B. Benchmark tasks details

B.1. Classification tasks

B.1.1. BACH

This dataset consists of 400 (268 train / 132 val) WSIs of breast cancer, split by patient ID to avoid leakage (Polónia et al., 2019; Aresta et al., 2019; Araújo et al., 2017; Fondón et al., 2018). The task is patch-level multiclass classification task, predicting 4 classes: normal, benign, in situ, invasive, measured by multiclass accuracy. We do linear probing using nn.Linear(out=4), CrossEntropyLoss, AdamW lr=3e-4, batch size 256, max 12,500 steps, early stopping patience 1250 on val/MulticlassAccuracy.

B.1.2. BRACS

Data contains 3,657 train / 312 val / 570 test ROI patches of breast lesions (Brancati et al., 2022). The task is patch-level multiclass classification task, predicting 7 subtypes: Normal, Pathological Benign, Usual Ductal Hyperplasia, Flat Epithelial Atypia, Atypical Ductal Hyperplasia, Ductal Carcinoma In Situ, Invasive Carcinoma, measured by multiclass accuracy We do linear probing with nn.Linear (out=7), CrossEntropyLoss, AdamW lr=3e-4, batch size 256, max 12500 steps, early stopping patience 167 on val/MulticlassAccuracy.

B.1.3. BreakHis

Data has 1,132 train / 339 val WSIs of breast cancer patients, split by patient ID (14 patients held out for val) to prevent data leakage (Spanhol et al., 2016). The task is patch-level multiclass classification task, predicting 4 classes: Tubular Adenoma (TA), Mucinous Carcinoma (MC), Fibroadenoma (F), Ductal Carcinoma (DC), measured by multiclass accuracy We performed linear probing with nn.Linear (out=4), CrossEntropyLoss, AdamW lr=3e-4, batch size 256, max 12500 steps, early stopping patience 500 on val/MulticlassAccuracy.

B.1.4. CRC-100K

The dataset contains 100'000 train / 7180 val patches, split by patient ID (Kather et al., 2018). The task is patch-level multiclass classification task, predicting 9 classes: Adipose (ADI), Background (BACK), Debris (DEB), Lymphocytes (LYM), Mucus (MUC), Smooth Muscle (MUS), Normal Mucosa (NORM), Cancer-associated Stroma (STR), Colorectal Adenocarcinoma (TUM), measured by multiclass accuracy We do linear probing with nn.Linear (out=9), CrossEntropyLoss, AdamW lr=3e-4, batch size 256, max 12500 steps, early stopping patience 7 on val/MulticlassAccuracy.

Table 3. Benchmark results of all models on all tasks Bach (Polónia et al., 2019; Aresta et al., 2019; Araújo et al., 2017; Fondón et al., 2018), BRACS (Brancati et al., 2022), BH: BreakHis (Spanhol et al., 2016), CRC: CRC-100K (Kather et al., 2018), GTMA: Gleason TMA (Arvaniti et al., 2018), MHIST (Wei et al., 2021), PC: PatchCamelyon (Veeling et al., 2018), C16: Camelyon16 (Ehteshami Bejnordi et al., 2017), Panda (Bulten et al., 2022), CNP: CoNSeP (Graham et al., 2019), MNS: MoNuSAC (Verma et al., 2021), Hbench: Hest-bench (Jaume et al., 2024), Hexpr: Hest expression prediction data from (Jaume et al., 2024), HexprIO: Hest expression task with image only input, PR: panda radboud (Bulten et al., 2022), PK: panda karolinska (Bulten et al., 2022). TTS: tan tock seng (Oner et al., 2022a;b). Conch+SEAL: (Hemker et al., 2026). Avg: Average over all tasks, excluding Hexpr, but including HexprIO

MODEL	BACH	BRACS	BH	CRC	GTMA	MHIST	PC	C16	PANDA	CNP
IMAGE ONLY	**0.88**	0.64	**0.82**	**0.96**	**0.77**	0.77	**0.91**	0.79	**0.63**	0.61
1 EXPR TOKEN	0.85	**0.65**	0.75	**0.96**	0.74	0.82	0.90	0.83	0.61	0.61
16 EXPR TOKENS	0.86	0.60	0.72	**0.96**	0.73	0.80	0.90	0.83	**0.63**	**0.62**
64 EXPR TOKENS	0.83	0.64	0.76	**0.96**	0.75	0.80	**0.91**	0.82	**0.63**	0.61
256 EXPR TOKENS	0.84	0.63	0.79	**0.96**	0.74	0.78	**0.91**	0.84	**0.63**	0.61
MIXING MPP AND N TOKENS	0.87	0.62	0.75	**0.96**	0.74	0.80	**0.91**	0.82	0.60	0.61
CONCH + SEAL	0.82	0.61	0.68	0.95	0.73	**0.83**	**0.91**	**0.89**	0.62	**0.62**

Table 4

MODEL	MNS	HBENCH	HEXPR	HEXPRIO	PK	PR	TTS	AVG
IMAGE ONLY	0.61	0.34	-	0.26	0.70	0.61	0.31	0.60
1 EXPR TOKEN	0.61	0.40	0.31	0.30	**0.72**	0.66	0.36	**0.62**
16 EXPR TOKENS	0.63	0.40	**0.38**	**0.32**	**0.72**	0.66	**0.37**	**0.62**
64 EXPR TOKENS	0.63	**0.41**	0.30	0.30	**0.72**	0.66	**0.37**	**0.62**
256 EXPR TOKENS	**0.64**	**0.41**	0.34	0.31	**0.72**	**0.67**	**0.37**	**0.62**
MIXING MPP AND N TOKENS	0.61	0.40	0.30	0.29	**0.72**	0.66	0.36	**0.62**
CONCH + SEAL	0.63	0.38	-	0.28	0.71	0.66	**0.37**	0.61

B.1.5. GLEASON MTA

The data consists of 15303 train / 2482 val / 4967 test patches of prostate cancer, split by tissue micro-array ID (Arvaniti et al., 2018). The task is multiclass patch classification of higher level glandular patterns in 4 classes: Benign, Gleason 3, Gleason 4, Gleason 5, measured by multiclass accuracy. We do linear probing with: nn.Linear (out=4), CrossEntropyLoss, AdamW lr=3e-4, batch size 256, max 12500 steps, early stopping patience 42 on val/MulticlassAccuracy.

B.1.6. MHIST

This dataset consists of 3,152 (2,175 train / 977 test) colorectal polyp image patches, labeled by majority vote of 7 pathologists (Wei et al., 2021). The task is a patch-level binary classification task, distinguishing hyperplastic polyps (HP) from sessile serrated adenomas (SSA), measured by binary balanced accuracy. We do linear probing using nn.Linear(out=1), BCEWithLogitsLoss, AdamW lr=3e-4, batch size 256, max 12500 steps, early stopping patience 278 on val/BinaryBalancedAccuracy.

B.1.7. PATCH-CAMELYON

This dataset consists of 327,680 (262,144 train / 32,768 val / 32,768 test) 96x96 px patches extracted from Camelyon16 lymph node WSIs (Veeling et al., 2018). The task is a patch-level binary classification task, detecting tumor metastasis vs. normal tissue, measured by binary balanced accuracy. We do linear probing using nn.Linear(out=1), BCEWithLogitsLoss, AdamW lr=3e-4, batch size 256, max 12,500 steps, early stopping patience 3 on val/BinaryBalancedAccuracy.

B.1.8. CAMELYON16

This dataset consists of 399 (216 train / 54 val / 129 test) lymph node WSIs, with validation slides matching those used to derive PatchCamelyon (Ehteshami Bejnordi et al., 2017). The task is a slide-level binary classification task, detecting tumor metastasis vs. normal tissue, measured by binary balanced accuracy. We use a two-stage offline pipeline: first extract up to 1,000 patch embeddings per slide (224x224 at 0.25 mpp), then train ABMIL (projected dim 128, out=1), BCEWithLogitsLoss, AdamW lr=1e-3, batch size 32, max 100 epochs, early stopping patience 20 on val/BinaryBalancedAccuracy, averaged over 20 runs.

B.1.9. PANDA

This dataset consists of 10,000 prostate cancer WSIs from the Kaggle PANDA challenge, split 70/15/15 stratified by grade, with noisy labels filtered out (Bulten et al., 2022). The task is a slide-level multiclass classification task, predicting 6 ISUP grades (0–5), measured by multiclass accuracy. We use a two-stage offline pipeline: first extract up to 200 patch embeddings per slide (224x224 at 0.5 mpp), then train ABMIL (projected dim 128, out=6), CrossEntropyLoss, AdamW lr=1e-3, batch size 32, max 100 epochs, early stopping patience 20 on val/MulticlassAccuracy, averaged over 20 runs.

B.1.10. PANDA RADBOUD

This dataset is derived from the Radboud subset of the PANDA prostate cancer challenge (Bulten et al., 2022). Patches of 256x256 at 0.25 mpp are extracted from WSIs and filtered to contain exactly one Gleason grade covering at least 30% of the patch area. The task is a patch-level multiclass classification task, predicting 5 Gleason grades (stroma, benign epithelium, Gleason 3, 4, 5), with up to 25,000 balanced patches sampled (capped at 200 patches/WSI, 20 patches/WSI/class). Labels come from pixel-level segmentation masks where prostate glands are individually annotated. We use an offline two-stage pipeline with 5-fold cross-validation: first extract and cache patch embeddings with a frozen encoder, then do linear probing using nn.Linear(out=5), CrossEntropyLoss, AdamW lr=3e-4, batch size 256, max 12,500 steps, early stopping patience 278 on val/MulticlassAccuracy.

B.1.11. PANDA KAROLINSKA

This dataset is derived from the Karolinska subset of the PANDA prostate cancer challenge (Bulten et al., 2022). Patches of 256x256 at 0.25 mpp are extracted from WSIs and filtered to contain exactly one tissue class covering at least 30% of the patch area. The task is a patch-level binary classification task, predicting 2 classes (benign tissue vs. cancerous tissue), with up to 1,000,000 balanced patches sampled (capped at 350 patches/WSI, 200 patches/WSI/class). Karolinska provides coarser region-level annotations compared to Radboud's gland-level labels. We use an offline two-stage pipeline with 5-fold cross-validation: first extract and cache patch embeddings with a frozen encoder, then do linear probing using nn.Linear(out=1), BCEWithLogitsLoss, AdamW lr=3e-4, batch size 256, max 12,500 steps, early stopping patience 278 on val/BinaryBalancedAccuracy.

B.2. Segmentation tasks

B.2.1. CONSEP (COLORECTAL NUCLEAR SEGMENTATION AND PHENOTYPES)

This task includes 41 H&E-stained tiles (27 train / 14 val images, 1000x1000px) from 16 colorectal adenocarcinoma patients, containing 24,319 annotated nuclei (Graham et al., 2019). The goal is to semantically segment into five classes: background, other, inflammatory, epithelial, spindle-shaped. We used patch_size 224x224 and batch_size=64. Light-weight decoder probe is trained with early stopping with the MonaiDiceScore, best decoder probe is used for the test set.

B.2.2. MONUSAC (MULTI-ORGAN NUCLEI SEGMENTATION AND CLASSIFICATION)

Dataset consists of 46,000 annotated nuclei from TCGA WSIs across multiple organs at 40x magnification and 209 train/ 85 test images (Verma et al., 2021). Semantic segmentation task into 5 classes: background, epithelial, lymphocyte, neutrophil, macrophage. Light-weight decoder probe is trained with early stopping with the MonaiDiceScore, best decoder probe is used for the test set.

B.2.3. TAN TOCK SENG

This dataset is from Tan Tock Seng Hospital, Singapore (Oner et al., 2022a;b). H&E-stained WSIs from 99 prostate cancer patients (40 prostatectomy + 59 core needle biopsy specimens) were scanned at 40x magnification (0.25 µm/px) using an Aperio AT2 scanner. Pre-extracted patches of 512x512 at the 50 µm resolution level are used from the gland classification subset (46 WSIs), with patient-level train/val/test splits. The task is a patch-level semantic segmentation task predicting 9 classes (background, Gleason pattern 3, Gleason pattern 4, benign, tangential benign, tangential malignant, unknown, PIN, artefact). We use an offline two-stage pipeline with 5 runs: first extract and cache patch embeddings with a frozen encoder (resized to 224x224, ImageNet normalization), then train a ConvDecoderMS segmentation decoder using DiceLoss (softmax, batch mode), AdamW lr=2e-3, batch size 64, max 2,000 steps, early stopping patience 200 on val/MonaiDiceScore. The primary metric is classwise MonaiDiceScore (including background).

Table 5. HEST-bench benchmark results for all models, shown in the average Pearson Correlation coefficient

MODEL	CCRCC	COAD	IDC	LUNG	LYMPH_IDC	PAAD	PRAD	READ	SKCM
IMAGE ONLY	0.096	0.270	0.406	0.513	0.298	0.350	0.309	0.171	0.610
1 EXPR TOKEN	0.154	0.330	0.445	0.621	0.358	0.417	0.383	0.239	0.692
16 EXPR TOKENS	0.173	0.284	0.461	0.576	0.338	0.433	0.390	0.244	0.690
64 EXPR TOKENS	0.172	0.339	0.446	0.616	0.356	0.432	0.371	0.241	0.685
256 EXPR TOKENS	0.170	0.322	0.439	0.573	0.324	0.416	0.382	0.232	0.669
MIXING MPP AND N TOKENS	0.185	0.311	0.453	0.582	0.345	0.407	0.353	0.235	0.689
CONCH + SEAL	0.182	0.288	0.425	0.553	0.340	0.385	0.361	0.228	0.652

B.3. Regression tasks

B.3.1. HEST-BENCH

The hest benchmark task is an image to gene expression prediction tasks, made available by (Jaume et al., 2024) with 10 cancer types and Visium and Xenium samples. The image patches, cross-validation folds, and 50 target highly variable genes per cancer type are given. The task is evaluated by Pearson correlation . Below we report the task-specific performance per model.

B.3.2. HEST-EXPR

This benchmark evaluates multimodal reasoning by predicting a target gene expression panel from H&E images and an input gene expression panel. It is inspired by the practical challenge of limited Xenium panels, where measured genes may not overlap across samples. Specifically, we predict the v1-human-immuno-oncology panel (380 genes) from 10X Genomics, chosen for its relevance to cancer-related tasks, from H&E patches and 1,000 stable input genes.

To select the 1,000 input genes, we used all Visium samples in HEST-Bench, which are held out during training. We first applied spot- and sample-level quality control (QC): spots were excluded if out-of-tissue, more than 5 MAD below the per-sample median on log1p(total_counts) or log1p(n_genes_by_counts), or if pct_counts_mito exceeded both 5 MAD above the median and 20%. Samples were excluded if fewer than 1,000 spots passed QC, median counts fell below 1,000, or the fraction of spots passing QC was below 0.75 (one sample was filtered). We then ranked genes within each sample by detection rate (fraction of spots expressing gene G) and mean log1p normalized expression, averaged ranks across samples, and computed a composite rank weighting both criteria equally. The final 1,000 input genes were selected as those with the highest composite rank, a minimum detection rate of 0.05 in every sample, and no overlap with the target v1-human-immuno-oncology panel. Note that while QC is applied during gene selection, all spots and samples are included in the benchmark evaluation itself.