MINT: Molecularly Informed Training with Spatial Transcriptomics Supervision for Pathology Foundation Models

Minsoo Lee, Jonghyun Kim, Juseung Yun, Sunwoo Yu, and Jongseong Jang(✉)

LG AI Research
{minsoo.lee, kimjh0107, js.yun, ysw0920, j.jang}@lgresearch.ai

Abstract. Pathology foundation models learn morphological representations through self-supervised pretraining on large-scale whole-slide images, yet they do not explicitly capture the molecular state of the tissue. Spatial transcriptomics technologies bridge this gap by measuring gene expression in situ, offering a natural cross-modal supervisory signal. We propose MINT (Molecularly Informed Training), a fine-tuning framework that uses spatial transcriptomics as molecular supervision for pretrained pathology Vision Transformers. MINT appends a learnable ST token to the ViT input to encode transcriptomic information separately from the morphological CLS token, preventing catastrophic forgetting through DINO self-distillation and explicit feature anchoring to the frozen pretrained encoder. Gene expression regression at both spot-level (Visium) and patch-level (Xenium) resolutions provides complementary supervision across spatial scales. Trained on 577 publicly available HEST samples, MINT achieves the best overall performance on both HEST-Bench for gene expression prediction (mean Pearson r=0.440) and EVA for general pathology tasks (0.803), demonstrating that spatial transcriptomics supervision complements morphology-centric self-supervised pretraining.

Keywords: Pathology Foundation Model · Spatial Transcriptomics

1 Introduction

Foundation models pretrained on large-scale whole-slide image (WSI) collections have become central to computational pathology [1,2,4,5]. Through self-supervised objectives such as DINO [6] and DINOv2 [7], these models learn morphological representations from millions of histopathology tiles without manual annotation, enabling effective transfer to diverse downstream tasks including tissue classification, biomarker prediction, and survival analysis. Recent scaling efforts—H-optimus-0 [1] (ViT-G on 500K slides), UNI2-h [3] (ViT-H on 350K slides), and Virchow2 [20] (ViT-H on 3.1M slides)—have progressively improved performance across standardized benchmarks.

However, these models are trained *exclusively* on visual patterns. Histopathology images implicitly encode the molecular state of the tissue microenvironment: cellular composition, signaling pathway activity, and gene expression programs

all manifest in tissue morphology. Spatial transcriptomics (ST) technologies provide direct measurements of this link. Spot-level platforms [9], from the original Spatial Transcriptomics to its commercial successor Visium (10x Genomics), profile transcriptomes at spot-level resolution ($\sim$55 μm, roughly 10–50 cells per spot), while Xenium [10] provides single-molecule transcript detection at subcellular resolution. Recently, the HEST dataset [8] has aggregated over 1,100 such paired histology–transcriptomics samples spanning 131 organs, making large-scale training data for this modality publicly available.

A natural question arises: *can spatially-resolved gene expression serve as supervisory signal to improve the representations of pretrained pathology foundation models?* Prior work has primarily explored histology-to-gene-expression prediction as a task-specific supervised problem [11,12,13,14], but has not used spatial transcriptomics as molecular supervision to enhance pathology foundation model representations. The key challenge is that fine-tuning on gene expression objectives risks catastrophic forgetting of the morphological representations acquired during large-scale pretraining.

We introduce MINT (Molecularly Informed Training with Spatial Transcriptomics Supervision), a multi-task fine-tuning framework that addresses this challenge through three design principles. First, rather than modifying the CLS token that encodes morphological features, we append a dedicated learnable *ST token* to the ViT input sequence, providing the model with a separate channel for molecular information. Second, we combine DINO self-distillation for continued visual learning with explicit feature distillation from the frozen pretrained model, creating a dual mechanism against catastrophic forgetting. Third, we exploit both spot-level and Xenium patch-level gene expression as complementary supervision at different spatial resolutions.

Our contributions are as follows:

1. We propose MINT that uses spatial transcriptomics as molecular supervision for pretrained pathology ViTs while preventing catastrophic forgetting through a dedicated ST token and dual distillation mechanisms.
2. We show that the ST and CLS tokens capture complementary information—the ST token specializes for molecular signals while the CLS token retains morphological transferability—and that combining both yields consistent improvements in a backbone-agnostic manner.
3. Under the official HEST-Bench and EVA evaluation framework, MINT achieves the best overall performance on both benchmarks (mean Pearson r=0.440 and EVA average 0.803).

2 Method

2.1 Overview

Given a pretrained ViT encoder f_θ with embedding dimension D, MINT constructs a student–teacher framework and fine-tunes with four complementary objectives (Fig. 1). The student and teacher are initialized from f_θ; the student

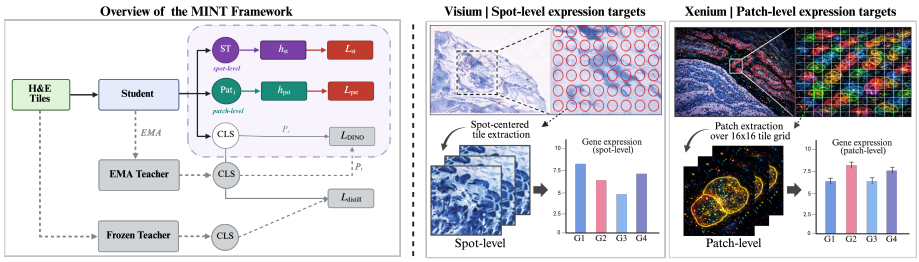

Fig. 1. MINT framework. The student ViT, augmented with a learnable ST token, outputs CLS, ST, and patch token representations. Gene expression regression from the ST token ($\mathcal{L}_{\text{ST}}$) and patch tokens ($\mathcal{L}_{\text{pST}}$) provides transcriptomic supervision, while DINO self-distillation ($\mathcal{L}_{\text{DINO}}$) and feature anchoring ($\mathcal{L}_{\text{distill}}$) preserve morphological representations. Only the student is updated via backpropagation.

is updated via gradient descent while the teacher tracks the student through exponential moving average (EMA). A frozen copy of f_θ serves as a distillation anchor. The key idea is to augment the ViT with a learnable *ST token* that encodes transcriptomic information, while the original CLS and patch tokens retain their morphological roles.

2.2 ST Token Design

We augment the ViT input with a learnable *ST token* $\mathbf{t}^{\text{st}} \in \mathbb{R}^D$, appended alongside the CLS token and N patch tokens (with N=256 for ViT-g/14 at 224×224). After processing through all transformer layers, the model produces three output representations: the CLS token $\mathbf{z}^{\text{cls}} \in \mathbb{R}^D$ retaining morphological features, the ST token $\mathbf{z}^{\text{st}} \in \mathbb{R}^D$ encoding transcriptomic features, and patch tokens $\{\mathbf{z}_i^{\text{pat}}\}_{i=1}^N$ providing spatially-localized representations.

The motivation for a separate token, rather than supervising the CLS token itself for gene expression, is twofold. First, it preserves the CLS token's morphological representations acquired through large-scale WSI pretraining. Directly supervising the CLS token for gene expression risks overwriting these features; the separate ST token instead specializes for gene expression while the CLS token continues to serve its original role in the self-supervised and feature anchoring objectives. Second, the ST token participates in self-attention with CLS and patch tokens across all transformer layers, enabling it to learn molecular representations from the full spatial context. At inference, concatenating both tokens ($[\mathbf{z}^{\text{cls}} \| \mathbf{z}^{\text{st}}] \in \mathbb{R}^{2D}$) combines both types of information, while using CLS alone recovers the original feature space.

2.3 Training Objectives

DINO Self-Distillation. Following DINO [6], we apply multi-crop augmentation: 2 global crops (224×224) and 8 local crops (96×96). The student processes

all 10 views while the teacher sees only global views. Both project their CLS tokens through a shared MLP head to a 65,536-dimensional probability space with centering and sharpening:

$$\mathcal{L}_{\text{DINO}} = -\sum_{x \in \mathcal{V}_g} \sum_{\substack{x' \in \mathcal{V} \\ x' \neq x}} P_t(x) \log P_s(x') \tag{1}$$

where $\mathcal{V}_g$ and $\mathcal{V}$ denote global and all crops, and P_t, P_s are teacher and student distributions. This loss maintains the self-supervised learning dynamics of the original pretraining, enabling continued visual representation learning during fine-tuning.

Feature Distillation. To explicitly anchor the student's CLS representation near the pretrained feature space, we minimize the L_2 distance to the frozen model's output on all crops:

$$\mathcal{L}_{\text{distill}} = \frac{1}{|\mathcal{V}|} \sum_{x \in \mathcal{V}} \left\| \mathbf{z}^{\text{cls}}_{\text{student}}(x) - \mathbf{z}^{\text{cls}}_{\text{frozen}}(x) \right\|_2^2 \tag{2}$$

This provides a direct regularization against catastrophic forgetting, complementing the implicit regularization from DINO's EMA teacher. The combination of both mechanisms—EMA-based self-distillation and explicit feature anchoring—provides redundant safeguards for representation stability.

Spot-Level ST Regression. Each training tile corresponds to a Visium spot. We define a shared vocabulary of G=38,710 genes as the union across all slides; since each slide measures only a subset, the loss is computed only over genes present in the current slide. The ST token is passed through an MLP head $h_{\text{st}} : \mathbb{R}^D \to \mathbb{R}^G$ to predict $\log(1+x)$-transformed expression values. To further focus on informative genes, we apply stochastic highly variable gene (HVG) selection [17]: at each iteration, with probability p_{hvg} we restrict the loss to the top-k HVGs, and otherwise use all measured genes:

$$\mathcal{L}_{\text{ST}} = \frac{1}{|\mathcal{G}_s|} \sum_{g \in \mathcal{G}_s} \left(h_{\text{st}}(\mathbf{z}^{\text{st}})_g - y_g \right)^2 \tag{3}$$

where $\mathcal{G}_s$ is the (stochastically) selected gene subset and y_g the target expression.

Patch-Level Xenium Regression. Spot-level platforms provide one expression profile per spot, averaging over 10–50 cells. Xenium technology [10], in contrast, detects individual transcripts at sub-cellular resolution, enabling construction of patch-level targets. For each 16×16 grid of ViT patches within a tile, we aggregate Xenium transcripts falling within each patch's spatial extent to create per-patch expression vectors.

Each patch token $\mathbf{z}_i^{\text{pat}}$ is passed through a dedicated MLP head $h_{\text{pst}} : \mathbb{R}^D \to \mathbb{R}^{G_{\text{xen}}}$, where $G_{\text{xen}} = 5{,}783$ is the union of genes across available Xenium panels. Since each panel covers a different gene subset, the loss for a given sample is restricted to its panel's genes. We further apply *all-zero patch filtering*: the loss is computed only over patches containing at least one detected transcript across the sample's measured Xenium genes, while zero-valued gene entries within these non-empty patches are retained in the regression target, avoiding penalization for regions without detected transcripts:

$$\mathcal{L}_{\text{pST}} = \frac{1}{|\mathcal{P}^+|} \sum_{p \in \mathcal{P}^+} \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \left(h_{\text{pst}}(\mathbf{z}_p^{\text{pat}})_g - y_{p,g} \right)^2 \tag{4}$$

where $\mathcal{P}^+$ is the set of patches containing at least one detected transcript and $\mathcal{G}$ the set of genes measured by the sample's Xenium panel.

Total Objective. The four losses are combined as:

$$\mathcal{L} = \mathcal{L}_{\text{DINO}} + \lambda_{\text{distill}} \mathcal{L}_{\text{distill}} + \lambda_{\text{ST}} \mathcal{L}_{\text{ST}} + \lambda_{\text{pST}} \mathcal{L}_{\text{pST}} \tag{5}$$

where λ_{distill}, λ_{ST}, and λ_{pST} are scalar weights for each auxiliary objective.

3 Experiments

3.1 Training Setup

We fine-tune H-optimus-0 [1] (ViT-g/14, 1.1B parameters) on the HEST dataset [8], which provides paired histology tiles and spot-level spatial transcriptomics measurements (Visium and earlier ST platforms) across diverse tissue types. From the 649 human samples in HEST, we exclude 72 samples included in HEST-Bench to prevent data leakage, yielding 577 training samples. For patch-level Xenium supervision (§2.3), we use 44 Xenium-profiled samples from HEST, oversampled 5× to balance against the more abundant spot-level data. We train for 90K iterations using AdamW with learning rate $5{\times}10^{-5}$, cosine schedule, batch size 64 per GPU across 8 GPUs, and bf16 mixed precision. The loss weights λ_{distill}, λ_{ST}, and λ_{pST} in Eq. 5 are each set to 100. For HVG selection, we use k=200 and p_{hvg}=0.5. We denote the resulting model as **MINT**. At inference, we concatenate the CLS and ST tokens to form $[\mathbf{z}^{\text{cls}} \| \mathbf{z}^{\text{st}}] \in \mathbb{R}^{3072}$ as the default feature representation.

3.2 Benchmarks and Protocols

We evaluate on two complementary benchmarks, strictly following each benchmark's official protocol. First, we report results on HEST-Bench, which assesses how well histology features support gene expression prediction across 9 cancer types. Following the benchmark specification, we use the standard PCA+Ridge

Table 1. HEST-Bench results (PCA+Ridge with 256 PCA components; Pearson correlation; higher is better). Best in **bold**, second-best underlined.

Model	IDC	PRAD	PAAD	SKCM	COAD	READ	ccRCC	LUAD	LYMPH	Avg
CTransPath [15]	0.511	0.343	0.438	0.511	0.229	0.110	0.228	0.499	0.235	0.345
Phikon [19]	0.533	0.342	0.443	0.536	0.259	0.152	0.242	0.547	0.237	0.366
CONCH [5]	0.536	0.355	0.448	0.579	0.253	0.167	0.218	0.531	0.251	0.371
REMEDIS [22]	0.529	0.347	0.464	0.582	0.286	0.115	0.265	0.534	0.247	0.374
CONCH 1.5 [5]	0.544	0.381	0.457	0.552	0.280	0.160	0.218	0.551	0.270	0.379
GigaPath [16]	0.551	0.371	0.477	0.554	0.301	0.186	0.239	0.540	0.249	0.385
UNI [2]	0.570	0.314	0.476	0.625	0.263	0.176	0.243	0.551	0.257	0.386
H0-mini [23]	0.586	0.368	0.492	0.601	0.249	0.186	0.267	0.548	0.263	0.396
Virchow [4]	0.570	0.331	0.488	0.609	0.311	0.202	0.264	0.546	0.259	0.398
Virchow2 [20]	0.592	0.347	0.466	0.617	0.258	0.208	0.279	0.561	0.258	0.398
UNI2-h [3]	0.590	0.357	0.500	0.660	0.301	0.223	0.264	0.558	0.273	0.414
H-optimus-0 [1]	0.598	0.385	0.493	0.643	0.299	0.229	0.265	0.558	0.260	0.415
MINT	**0.655**	**0.389**	**0.509**	**0.695**	**0.314**	**0.244**	**0.283**	**0.597**	**0.275**	**0.440**

Table 2. EVA evaluation framework results (higher is better). Best in **bold**, second-best underlined. Average is the arithmetic mean over 9 benchmarks.

Model	BHis	CRC	Glea	MHIST	PCam	Cam16	PANDA	CoNS	MoNuS	Average
Phikon-v2 [24]	0.713	0.939	0.757	0.777	0.894	0.805	0.625	0.627	0.644	0.753
Phikon [19]	0.717	0.940	0.729	0.803	0.920	0.798	0.644	0.627	0.635	0.757
CONCH [5]	0.681	0.953	0.729	0.832	0.922	0.841	0.628	0.618	0.632	0.760
hibou-L [25]	0.735	0.932	0.764	0.839	**0.955**	0.823	0.634	**0.646**	0.669	0.777
UNI [2]	0.785	0.944	0.750	0.843	0.937	0.833	0.659	0.628	0.659	0.782
Pv-GigaPath [16]	0.827	0.951	0.724	0.829	0.945	0.815	0.653	0.626	0.680	0.783
Midnight-12k [26]	0.819	0.966	**0.800**	0.804	0.929	0.849	0.649	0.623	0.659	0.789
H-optimus-0 [1]	0.801	0.955	0.770	0.843	0.943	0.827	0.671	0.644	**0.685**	0.793
UNI2-h [3]	0.859	0.965	0.775	0.824	0.950	0.849	0.657	0.630	0.642	0.795
Virchow2 [20]	0.821	**0.967**	0.783	**0.861**	0.938	**0.861**	0.646	0.640	0.669	0.798
MINT	**0.864**	0.960	0.769	0.834	0.943	0.851	**0.673**	**0.646**	**0.685**	**0.803**

evaluation pipeline with 256 PCA components and report Pearson correlation (higher is better). Second, we evaluate general-purpose pathology transferability using the EVA evaluation framework, which consists of nine pathology benchmarks spanning classification (BreakHis, CRC, Gleason, MHIST, PCam), weakly-supervised whole-slide tasks (Cam16, PANDA), and nuclei segmentation (CoNSeP, MoNuSAC); we refer to [18] for full protocol details. For classification and whole-slide tasks, we use the [CLS ‖ ST] concat feature. For segmentation tasks, we follow the EVA protocol using spatially arranged patch features concatenated with the input image under online linear evaluation, without using the ST token. For fair comparison, all baseline numbers reported in Tables 1 and 2 (except MINT) are taken from the official benchmark leaderboards. We additionally run the official evaluation pipelines to verify that these reported baseline results are reproducible under the benchmark settings. All MINT results are obtained using the same official evaluation pipelines without any modifications.

3.3 Comparison with Existing Models

Table 1 presents HEST-Bench results. MINT achieves the best overall mean Pearson correlation (0.440), surpassing all compared models including H-optimus-0 (0.415) and UNI2-h (0.414), and ranking first across all 9 cancer types. On EVA (Table 2), MINT also attains the highest average (0.803), exceeding Virchow2 (0.798) and H-optimus-0 (0.793). Compared to the pretrained H-optimus-0 backbone, MINT maintains comparable or improved performance across all EVA tasks except MHIST, where a marginal decrease is observed.

Notably, MINT ranks first on both benchmarks simultaneously. HEST-Bench measures molecular prediction capability while EVA assesses general-purpose pathology transferability—two axes that are not inherently aligned. The fact that MINT improves on both indicates that spatial transcriptomics supervision provides complementary information to morphology-centric pretraining, rather than introducing a trade-off between molecular and morphological performance.

These gains are achieved not by scaling up image data, but by introducing a new supervisory modality: MINT fine-tunes on only 577 paired histology–transcriptomics samples from the publicly available HEST dataset, yet improves both molecular and morphological benchmarks. This suggests that cross-modal supervision offers a distinct and complementary axis for improving pathology foundation models beyond image-only self-supervised scaling, and that further growth of paired histology–transcriptomics datasets may yield additional gains.

3.4 Representation Analysis

Table 3 examines how the choice of feature representation affects downstream performance across two backbone architectures. Since the EVA segmentation benchmarks (CoNSeP, MoNuSAC) rely on spatially arranged patch embeddings rather than token-level representations, the EVA average here is computed over classification and whole-slide tasks only. The CLS and ST tokens exhibit clear specialization: ST-only outperforms CLS-only on HEST-Bench (0.428 vs. 0.413 for H-optimus-0), while CLS-only outperforms ST-only on EVA (0.828 vs. 0.823), indicating that each token captures distinct information—molecular and morphological, respectively. Combining both tokens improves performance on both benchmarks. Notably, element-wise summation (CLS $\oplus$ ST) uses the same dimensionality as CLS-only (1536-d) yet substantially outperforms it (e.g., 0.431 vs. 0.413 on HEST-Bench), confirming that the gains arise from genuinely complementary information rather than increased feature dimensionality. Concatenation ([CLS $\|$ ST]) achieves the best results across both benchmarks. These patterns are consistently observed across both H-optimus-0 and UNI2-h backbones, confirming that the benefit of the ST token is backbone-agnostic.

To validate the necessity of the dedicated ST token, we ablate it by applying $\mathcal{L}_{\text{ST}}$ directly to the CLS token, with and without feature distillation ($\mathcal{L}_{\text{distill}}$). Both variants achieve higher HEST-Bench scores than the pretrained baseline, indicating that the CLS token successfully learns molecular information regardless of distillation. However, without distillation, EVA-CLS drops substantially

Table 3. Representation variants across backbones, including pretrained baselines and ablations. HEST-Bench reports mean Pearson correlation; EVA-CLS averages over classification and whole-slide tasks (7 benchmarks). Best in **bold**, second-best underlined.

Backbone	Training	Representation	HEST-Bench	EVA-CLS
H-opt-0	No training	CLS-only	0.415	0.830
	L_{ST} on [CLS] w/o L_{distill}	CLS-only	0.424	0.811
	L_{ST} on [CLS] + L_{distill}	CLS-only	0.426	0.827
	MINT	CLS-only	0.413	0.828
		ST-only	0.428	0.823
		CLS $\oplus$ ST (sum)	0.431	0.837
		[CLS $\|\|$ ST] (concat)	**0.440**	**0.842**
UNI2-h	No training	CLS-only	0.414	0.840
	L_{ST} on [CLS] w/o L_{distill}	CLS-only	0.420	0.824
	L_{ST} on [CLS] + L_{distill}	CLS-only	0.426	0.836
	MINT	CLS-only	0.410	0.839
		ST-only	0.426	0.834
		CLS $\oplus$ ST (sum)	0.433	0.844
		[CLS $\|\|$ ST] (concat)	**0.435**	**0.848**

(0.830→0.811 for H-optimus-0; 0.840→0.824 for UNI2-h), revealing catastrophic forgetting of morphological representations. Adding distillation largely recovers EVA performance (to 0.827 and 0.836, respectively), confirming its role in preventing forgetting. Nevertheless, both $\mathcal{L}_{\text{ST}}$ on [CLS] variants fall short of MINT's token-separated design on both benchmarks for both backbones: even CLS $\oplus$ ST summation outperforms the best $\mathcal{L}_{\text{ST}}$ on [CLS] variant (0.431 / 0.837 vs. 0.426 / 0.827 for H-optimus-0; 0.433 / 0.844 vs. 0.426 / 0.836 for UNI2-h), demonstrating that decoupling molecular and morphological learning into separate tokens is more effective than encoding both in a shared representation. MINT's dedicated ST token avoids this trade-off entirely: the CLS-only variant of MINT closely matches the pretrained baseline on EVA (0.828 vs. 0.830 for H-optimus-0; 0.839 vs. 0.840 for UNI2-h), confirming that the ST token successfully decouples molecular learning from the pretrained feature space.

4 Conclusion

We presented MINT, which incorporates spatial transcriptomics supervision into pretrained pathology ViTs through a dedicated ST token and dual distillation. By decoupling molecular and morphological learning, MINT achieves the highest scores on both HEST-Bench and EVA in a backbone-agnostic manner, suggesting that cross-modal supervision offers a complementary axis beyond image-only self-supervised scaling.

Disclosure of Interests. LG AI Research has filed patent application(s) related to the method described in this paper. The authors are employees of LG AI Research. The authors declare no other competing interests.

References

1. Bioptimus: H-optimus-0. GitHub, `https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0`, last accessed 2026/06/24
2. Chen, R.J., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
3. MahmoodLab: UNI2-h model. Hugging Face, `https://huggingface.co/MahmoodLab/UNI2-h`, last accessed 2026/06/24
4. Vorontsov, E., et al.: A foundation model for clinical-grade computational pathology and rare cancers detection. Nature Medicine **30**(10), 2924–2935 (2024)
5. Lu, M.Y., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024)
6. Caron, M., et al.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660 (2021)
7. Oquab, M., et al.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
8. Jaume, G., et al.: HEST-1k: A dataset for spatial transcriptomics and histology image analysis. In: Advances in Neural Information Processing Systems, Datasets and Benchmarks Track (2024)
9. Ståhl, P.L., et al.: Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science **353**(6294), 78–82 (2016)
10. Janesick, A., et al.: High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. Nature Communications **14**(1), 8353 (2023)
11. He, B., et al.: Integrating spatial gene expression and breast tumour morphology via deep learning. Nature Biomedical Engineering **4**(8), 827–834 (2020)
12. Pang, M., et al.: Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. bioRxiv preprint (2021)
13. Xie, R., et al.: Spatially resolved gene expression prediction from H&E histology images via bi-modal contrastive learning. In: Advances in Neural Information Processing Systems (2023)
14. Jia, Y., et al.: THItoGene: A deep learning method for predicting spatial transcriptomics from histological images. Briefings in Bioinformatics **25**(1) (2024)
15. Wang, X., et al.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical Image Analysis **81**, 102559 (2022)
16. Xu, H., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**, 181–188 (2024)
17. Luecken, M.D., Theis, F.J.: Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular Systems Biology **15**(6), e8746 (2019)
18. Gatopoulos, I., et al.: EVA: Evaluation framework for pathology foundation models. In: Medical Imaging with Deep Learning (2024)
19. Filiot, A., et al.: Scaling self-supervised learning for histopathology with masked image modeling. medRxiv preprint 2023.07.21.23292757 (2023)

20. Zimmermann, E., et al.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint arXiv:2408.00738 (2024)
21. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
22. Azizi, S., et al.: Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nature Biomedical Engineering **7**(6), 756–779 (2023)
23. Filiot, A., et al.: Distilling foundation models for robust and efficient models in digital pathology. arXiv preprint arXiv:2501.16239 (2025)
24. Filiot, A., et al.: Phikon-v2, a large and public feature extractor for biomarker prediction. arXiv preprint arXiv:2409.09173 (2024)
25. Nechaev, D., et al.: Hibou: A family of foundational vision transformers for pathology. arXiv preprint arXiv:2406.05074 (2024)
26. Karasikov, M., et al.: Training state-of-the-art pathology foundation models with orders of magnitude less data. arXiv preprint arXiv:2504.05186 (2025)