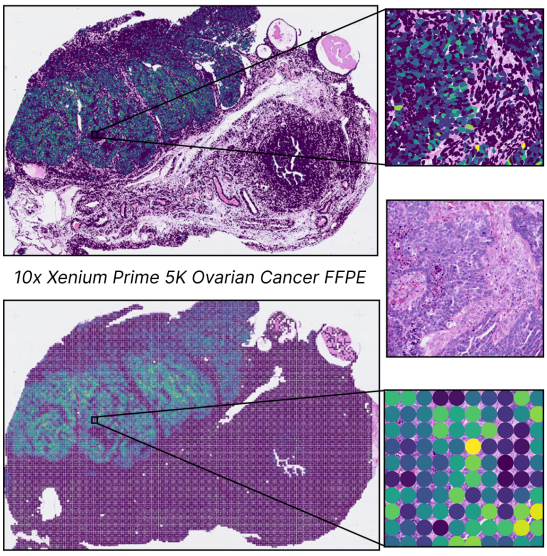

Figure 3. Top: Overlay of TP53 gene counts as observed in individual cells in Xenium Ovarian Cancer tissue. Bottom: Overlay of TP53 gene on simulated pseudo-bulk spot used to form the HESCAPE dataset

7. Datasets

This section summarizes the spatial transcriptomics datasets used in this study. The datasets are organized by gene panel type. A summary table is provided in main Tab. 1. All tissues sections were imaged at 40x magnification and preserved using FFPE method (except where noted).

Since we use 10x Xenium samples with sub-cellular transcript detection, we need to simulate a 10x Visium patch of size $55\mu m$, called pseudo-spot, for patch-based image-gene pairs. We achieve this by sum pooling the transcripts of cells within the simulated spot. The aggregated gene expressions in the simulated spots follow similar distributions to the xenium gene expressions. This is additionally validated visually by plotting expression of several biomarker genes like TP53 on the spatial samples as seen in Fig. 3.

To enable efficient benchmarking, for each dataset above, we have also created an image patch–spatial expression dataset in the Hugging Face Arrow format, comprising a total of 7229962 image-gene expression pairs combined. Each pair is accompanied by patient-specific metadata and is available on the Hugging Face Hub. For the final benchmark, we filter out replicates with different gene panels and single gene panel datasets resulting in $\sim 620k$ image-gene expression pairs.

Patient-based stratification was employed for each dataset to create training, validation, and test splits-ensuring that each patient's samples are confined to a single split. Since a lot of 10x Xenium samples are part of the HEST-benchmark test set, for fair evaluation we design our own testing splits and test all image models from HEST-benchmark on these new splits. All reported results are based on the provided test split to clearly expose the impact of batch effects.

All dataset preprocessing and creation was done using SpatialData and Huggingface Datasets library

7.1. Detailed Description of Dataset Groups

7.1.1. Human 5K Panel

This panel contains 6 datasets spanning multiple organs including Skin, Prostate, Lymphoid, Lung, Breast, Cervix, and Ovary. All samples are from cancer or diseased tissue preserved with FFPE, except one Fresh Frozen ovarian cancer sample. The datasets have uniform pixel size around 0.274 μm.

7.1.2. Human Colon Panel

This panel includes 5 datasets exclusively focused on bowel tissue, with both cancerous and healthy samples. All samples are FFPE-preserved with pixel sizes ranging from 0.137 μm to 0.274 μm. Three datasets come from the same study on immune cell populations in colorectal cancer, all from the same patient.

7.1.3. Human Lung Healthy Panel

The most extensive group with 19 datasets, all focused on lung tissue. Samples include both healthy (6 datasets) and diseased (13 datasets) states from 19 different patients with pulmonary fibrosis. All datasets are from the study Vannan et al. [47]. All images have consistent pixel size of 0.213 μm.

7.1.4. Human Immuno-Oncology Panel

This panel contains 5 datasets from 5 different organs (Ovary, Lung, Pancreas, Bowel, Brain), all studying cancerous tissue. Three datasets include identified patient information. All samples have consistent pixel size around 0.274 μm and include custom add-ons to the base panel.

7.1.5. Human Multi-Tissue Panel

A diverse group of 14 datasets covering 8 different organs (Lymphoid, Bone, Pancreas, Skin, Liver, Heart, Lung, Kidney). Includes all three disease states (Cancer, Healthy, Diseased) with pixel sizes ranging from 0.137 μm to 0.274 μm. Most datasets are organ-specific paired samples (cancer/healthy or diseased/healthy).

7.1.6. Human Breast Panel

Includes 5 datasets all focused on breast cancer tissue. Two datasets are from a study using the entire sample area from one patient, while three are from a high-resolution mapping study of tumor microenvironment. Pixel sizes range from 0.213 μm to 0.364 μm.

Table 7. Complete Test Recall@1 Results for both Image-to-Gene (I2G) and Gene-to-Image (G2I) tasks across different tissue panels. Experiments with "—" indicate out of memory issues during training

	5K		Multi		ImmOnc		Colon		Breast		Lung	
Model	I2G	G2I	I2G	G2I	I2G	G2I	I2G	G2I	I2G	G2I	I2G	G2I
MLP-CTransPath	0.022(0.002)	0.023(0.004)	0.029(0.004)	0.020(0.003)	0.023(0.002)	0.019(0.003)	0.020(0.004)	0.026(0.008)	0.026(0.005)	0.025(0.002)	0.022(0.004)	0.028(0.004)
MLP-CONCH	0.058(0.007)	0.058(0.006)	0.056(0.005)	0.041(0.005)	0.044(0.004)	0.028(0.002)	0.074(0.009)	0.062(0.005)	0.106(0.019)	0.067(0.011)	0.134(0.007)	0.127(0.013)
MLP-Gigapath	0.069(0.006)	0.067(0.002)	0.074(0.006)	0.051(0.005)	0.042(0.002)	0.028(0.002)	0.077(0.012)	0.073(0.010)	0.109(0.015)	0.079(0.007)	0.173(0.001)	0.165(0.005)
MLP-H0mini	0.059(0.002)	0.060(0.001)	0.048(0.009)	0.033(0.004)	0.041(0.005)	0.027(0.001)	0.071(0.006)	0.072(0.012)	0.086(0.005)	0.060(0.004)	0.099(0.005)	0.093(0.006)
MLP-UNI	0.065(0.011)	0.064(0.009)	0.061(0.006)	0.039(0.006)	0.060(0.009)	0.029(0.006)	0.079(0.001)	0.080(0.007)	0.091(0.015)	0.065(0.014)	0.164(0.016)	0.164(0.019)
scFoundation-CTransPath	—	—	—	—	0.024(0.002)	0.028(0.003)	0.022(0.002)	0.019(0.005)	0.028(0.002)	0.026(0.004)	0.029(0.002)	0.028(0.002)
scFoundation-CONCH	—	—	—	—	0.055(0.006)	0.043(0.003)	0.077(0.008)	0.068(0.011)	0.085(0.012)	0.082(0.016)	0.166(0.005)	0.147(0.014)
scFoundation-Gigapath	—	—	—	—	0.065(0.007)	0.052(0.009)	0.079(0.013)	0.064(0.004)	0.090(0.005)	0.102(0.015)	0.225(0.022)	0.195(0.027)
scFoundation-H0mini	—	—	—	—	0.050(0.003)	0.042(0.005)	0.080(0.007)	0.067(0.005)	0.112(0.006)	0.104(0.011)	0.131(0.011)	0.094(0.005)
scFoundation-UNI	—	—	—	—	0.058(0.005)	0.040(0.004)	0.079(0.008)	0.060(0.021)	0.094(0.008)	0.094(0.015)	0.191(0.011)	0.158(0.011)
Nicheformer-CTransPath	0.023(0.001)	0.024(0.001)	0.027(0.003)	0.024(0.002)	0.027(0.001)	0.027(0.001)	0.018(0.003)	0.021(0.004)	0.025(0.004)	0.027(0.003)	0.027(0.004)	0.031(0.003)
Nicheformer-CONCH	0.061(0.001)	0.076(0.006)	0.067(0.005)	0.077(0.003)	0.056(0.003)	0.063(0.004)	0.069(0.006)	0.068(0.003)	0.101(0.008)	0.123(0.004)	0.125(0.002)	0.164(0.003)
Nicheformer-Gigapath	0.062(0.002)	0.067(0.001)	0.075(0.007)	0.079(0.005)	0.063(0.004)	0.072(0.005)	0.067(0.005)	0.073(0.006)	0.121(0.008)	0.139(0.016)	0.155(0.006)	0.203(0.005)
Nicheformer-H0mini	0.065(0.004)	0.076(0.001)	0.069(0.006)	0.077(0.008)	0.053(0.002)	0.054(0.002)	0.076(0.007)	0.075(0.009)	0.125(0.011)	0.142(0.003)	0.129(0.007)	0.168(0.013)
Nicheformer-UNI	0.071(0.006)	0.082(0.003)	0.073(0.002)	0.079(0.002)	0.060(0.008)	0.067(0.007)	0.063(0.002)	0.066(0.004)	0.122(0.013)	0.136(0.004)	0.144(0.003)	0.188(0.006)
DRVI-CTransPath	0.023(0.000)	0.025(0.001)	0.026(0.001)	0.033(0.002)	0.031(0.003)	0.028(0.003)	0.022(0.004)	0.026(0.004)	0.036(0.004)	0.032(0.006)	0.032(0.004)	0.037(0.003)
DRVI-CONCH	0.073(0.005)	0.093(0.002)	0.084(0.000)	0.110(0.004)	0.083(0.005)	0.081(0.002)	0.098(0.010)	0.108(0.010)	0.117(0.022)	0.124(0.011)	0.195(0.004)	0.240(0.002)
DRVI-Gigapath	0.097(0.011)	**0.115(0.003)**	**0.097(0.005)**	**0.134(0.007)**	**0.103(0.004)**	**0.101(0.005)**	0.111(0.008)	0.121(0.010)	0.146(0.010)	0.142(0.008)	**0.283(0.002)**	**0.329(0.006)**
DRVI-H0mini	0.085(0.005)	0.096(0.003)	0.077(0.008)	0.100(0.009)	0.079(0.006)	0.072(0.004)	**0.123(0.001)**	**0.126(0.004)**	**0.147(0.010)**	**0.152(0.009)**	0.208(0.001)	0.249(0.002)
DRVI-UNI	**0.097(0.002)**	0.106(0.007)	0.092(0.007)	0.126(0.005)	0.094(0.007)	0.090(0.003)	0.122(0.004)	0.124(0.004)	0.142(0.023)	0.145(0.021)	0.242(0.009)	0.301(0.014)

Table 8. Complete Test Recall@10 Results for both Image-to-Gene (I2G) and Gene-to-Image (G2I) tasks across different tissue panels. Experiments with "—" indicate out of memory issues during training

	5K		Multi		ImmOnc		Colon		Breast		Lung	
Model	I2G	G2I	I2G	G2I	I2G	G2I	I2G	G2I	I2G	G2I	I2G	G2I
MLP-CTransPath	0.194(0.008)	0.203(0.017)	0.265(0.015)	0.191(0.011)	0.215(0.015)	0.183(0.008)	0.197(0.028)	0.235(0.081)	0.220(0.033)	0.220(0.016)	0.196(0.014)	0.234(0.039)
MLP-CONCH	0.388(0.050)	0.387(0.041)	0.414(0.027)	0.314(0.021)	0.333(0.013)	0.241(0.009)	0.501(0.040)	0.440(0.010)	0.586(0.060)	0.421(0.038)	0.642(0.024)	0.605(0.036)
MLP-Gigapath	0.420(0.021)	0.425(0.015)	0.495(0.024)	0.371(0.019)	0.321(0.015)	0.244(0.008)	0.528(0.036)	0.500(0.028)	0.595(0.049)	0.441(0.036)	0.693(0.020)	0.675(0.016)
MLP-H0mini	0.401(0.009)	0.399(0.008)	0.371(0.047)	0.279(0.029)	0.313(0.020)	0.224(0.013)	0.502(0.047)	0.492(0.058)	0.486(0.018)	0.384(0.017)	0.552(0.006)	0.523(0.003)
MLP-UNI	0.409(0.036)	0.409(0.036)	0.435(0.028)	0.306(0.032)	0.412(0.034)	0.243(0.031)	0.532(0.008)	0.524(0.035)	0.545(0.027)	0.403(0.047)	0.684(0.040)	0.677(0.061)
scFoundation-CTransPath	—	—	—	—	0.224(0.013)	0.232(0.017)	0.204(0.036)	0.192(0.037)	0.254(0.011)	0.236(0.020)	0.230(0.008)	0.225(0.017)
scFoundation-CONCH	—	—	—	—	0.373(0.021)	0.314(0.015)	0.473(0.049)	0.439(0.039)	0.517(0.035)	0.497(0.051)	0.691(0.010)	0.656(0.024)
scFoundation-Gigapath	—	—	—	—	0.416(0.028)	0.352(0.029)	0.473(0.075)	0.422(0.014)	0.554(0.004)	0.567(0.030)	0.768(0.024)	0.733(0.043)
scFoundation-H0mini	—	—	—	—	0.358(0.013)	0.304(0.033)	0.513(0.004)	0.455(0.037)	0.581(0.021)	0.578(0.036)	0.616(0.019)	0.536(0.015)
scFoundation-UNI	—	—	—	—	0.398(0.023)	0.312(0.014)	0.482(0.029)	0.413(0.095)	0.557(0.018)	0.552(0.051)	0.726(0.009)	0.670(0.029)
Nicheformer-CTransPath	0.202(0.007)	0.218(0.005)	0.237(0.021)	0.224(0.030)	0.241(0.010)	0.233(0.011)	0.179(0.022)	0.212(0.026)	0.207(0.023)	0.242(0.018)	0.228(0.027)	0.256(0.030)
Nicheformer-CONCH	0.395(0.008)	0.443(0.023)	0.422(0.010)	0.459(0.006)	0.375(0.009)	0.410(0.013)	0.444(0.029)	0.409(0.022)	0.569(0.016)	0.612(0.010)	0.609(0.001)	0.692(0.011)
Nicheformer-Gigapath	0.395(0.011)	0.417(0.011)	0.441(0.021)	0.458(0.017)	0.405(0.021)	0.434(0.006)	0.433(0.016)	0.439(0.038)	0.621(0.019)	0.648(0.027)	0.667(0.011)	0.742(0.006)
Nicheformer-H0mini	0.396(0.014)	0.439(0.009)	0.430(0.023)	0.448(0.023)	0.355(0.009)	0.356(0.007)	0.470(0.038)	0.450(0.042)	0.620(0.018)	0.649(0.018)	0.617(0.013)	0.696(0.022)
Nicheformer-UNI	0.422(0.020)	0.466(0.009)	0.436(0.006)	0.462(0.005)	0.406(0.033)	0.427(0.033)	0.423(0.021)	0.402(0.019)	0.623(0.033)	0.639(0.003)	0.644(0.004)	0.726(0.005)
DRVI-CTransPath	0.204(0.006)	0.215(0.008)	0.222(0.013)	0.272(0.012)	0.258(0.015)	0.246(0.014)	0.212(0.025)	0.226(0.033)	0.289(0.015)	0.273(0.030)	0.264(0.017)	0.292(0.018)
DRVI-CONCH	0.431(0.020)	0.488(0.011)	0.471(0.004)	0.549(0.013)	0.471(0.011)	0.463(0.020)	0.561(0.034)	0.556(0.040)	0.591(0.050)	0.591(0.031)	0.714(0.011)	0.764(0.003)
DRVI-Gigapath	0.479(0.041)	**0.534(0.013)**	**0.499(0.018)**	**0.607(0.015)**	**0.519(0.019)**	**0.508(0.014)**	0.594(0.011)	0.597(0.028)	0.660(0.024)	0.626(0.027)	**0.800(0.001)**	**0.849(0.003)**
DRVI-H0mini	0.468(0.012)	0.495(0.010)	0.434(0.022)	0.523(0.018)	0.450(0.014)	0.431(0.015)	**0.617(0.012)**	0.590(0.019)	**0.664(0.023)**	**0.661(0.023)**	0.739(0.001)	0.778(0.004)
DRVI-UNI	**0.496(0.011)**	0.521(0.023)	0.486(0.023)	0.579(0.026)	0.503(0.022)	0.500(0.008)	0.611(0.021)	**0.600(0.013)**	0.648(0.050)	0.629(0.045)	0.772(0.002)	0.826(0.008)

7.2. Additional datasets

We also include 4 datasets studying breast cancer tissue from two patients in the HESCAPE dataset. These datasets were not used for any experiments above, but are can be useful for further downstream tasks. Each patient has two datasets: one with a custom add-on panel and one with a pre-designed panel, allowing for direct comparison. All datasets have consistent pixel size of 0.213 μm.

7.2.1. Preservation Methods

Only one dataset, Xenium_Prime_Human_Ovary_FF, uses Fresh Frozen preservation for ovarian cancer tissue.

7.2.2. Imaging Parameters

All datasets were imaged at 40x magnification. Pixel sizes range from 0.137 μm to 0.364 μm, with most datasets having pixel sizes around 0.213 μm or 0.274 μm.

8. Pretraining

8.1. Implementation details

For consistency, both encoders output embeddings of dimension $d = 128$. The training is performed with gradient clipping set to 5.0, a batch size of 256 distributed across 4 GPUs for $20,000$ steps, and an initial warmup phase of 780 steps. We use an AdamW optimizer β: $(0.9, 0.95)$ with a learning rate starting at 1×10^{-5} reduced over iterations via a cosine scheduler, and a weight decay of 0.01. The full contrastive pretraining is performed with a finetuning image-tuning and locked gene encoding objective, allowing the image encoders to learn robust, gene-biomarker specific features. All experiments are conducted on a Slurm GPU cluster equipped with A100 GPUs. We use PyTorch and Hydra for all our experiments.

Table 9. Projection head ablation study: Experiments performed with CLIP loss and frozen encoders. Best results are in **bold**, second-best are underlined.

Model	projection	5K		Colon	
		I2G	G2I	I2G	G2I
DRVI-Gigapath	linear	0.170	0.180	0.295	0.274
DRVI-UNI	linear	0.172	0.180	0.243	0.251
DRVI-Gigapath	mlp	**0.201**	**0.210**	**0.317**	**0.328**
DRVI-UNI	mlp	**0.201**	0.201	0.289	0.231
DRVI-Gigapath	transformer	0.114	0.130	0.250	0.277
DRVI-UNI	transformer	0.135	0.166	0.179	0.245

Table 10. Loss function ablation study: Experiments performed with same configuration as the benchmark. Best results are in **bold**, second-best are underlined.

Model	loss	5K		Colon	
		I2G	G2I	I2G	G2I
DRVI-Gigapath	CLIP	0.315	**0.359**	0.388	0.394
DRVI-UNI	CLIP	**0.322**	0.341	**0.404**	**0.401**
DRVI-Gigapath	SIGLIP	**0.322**	0.352	0.377	0.345
DRVI-UNI	SIGLIP	0.292	0.293	0.359	0.359

Table 11. Encoder Finetuning ablation study: Experiments performed with CLIP loss and MLP projection head. Experiments with "—" indicate out of memory issues during training. Best results are in **bold**, second-best are underlined.

Model	finetune		5K		Colon	
	img	gene	I2G	G2I	I2G	G2I
Nicheformer-UNI	✗	✗	0.153	0.174	0.201	0.221
Nicheformer-Gigapath	✗	✗	0.149	0.169	0.195	0.221
DRVI-UNI	✗	✗	0.172	0.180	0.243	0.251
DRVI-Gigapath	✗	✗	0.170	0.180	0.295	0.274
Nicheformer-UNI	✗	✓	0.187	0.205	0.261	0.266
Nicheformer-Gigapath	✗	✓	0.188	0.197	0.282	0.292
DRVI-UNI	✗	✓	0.204	0.225	0.318	0.319
DRVI-Gigapath	✗	✓	0.198	0.208	0.311	0.343
Nicheformer-UNI	✓	✗	0.262	0.282	0.238	0.244
Nicheformer-Gigapath	✓	✗	0.277	0.296	0.249	0.266
DRVI-UNI	✓	✗	0.289	0.293	0.326	0.284
DRVI-Gigapath	✓	✗	0.269	0.335	0.333	0.362
Nicheformer-UNI	✓	✓	0.308	0.317	0.323	0.326
Nicheformer-Gigapath	✓	✓	—	—	—	—
DRVI-UNI	✓	✓	**0.358**	0.342	**0.335**	0.336
DRVI-Gigapath	✓	✓	0.299	**0.370**	0.334	**0.376**

8.2. Ablation study

To measure the contributing factors of different hyperparameters and architectural changes towards contrastive alignment performance, we performed 3 independent ablation studies. The ablations were performed on the pan-organ 5K gene panel and the Colon panel.

8.2.1. Image Projection head

The ablation was to see which of the 3 heads, the basic linear projection, MLP or a transformer based projection helps improve the image encodings during contrastive pretraining. The results in Tab. 9 show, that MLP as an image projection head performs consistently well across both datasets.

8.2.2. Loss function

Here, we test both CLIP and SigLip losses across the 5K and Colon pretraining experiments to understand their performance in our data and batch size configurations. Our ablations in Tab. 10 suggests, there was no substantial improvement from using SigLip as the loss function.

The SigLip loss for the image-gene expression pair `v2g` is:

$$L_{\texttt{SIGLIP}^{\texttt{v2g}}} = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\log\frac{1}{1+\exp[z_{ij}(-\tau\langle \mathbf{v}_i, \mathbf{g}_j\rangle + b)]} \tag{3}$$

s with an additional learnable bias b. Unlike CLIP loss, SIGLIP avoids computing a global normalization and instead formalizes the objective as a logistic regression task, where the label z_{ij} is 1 for the positive pair and is -1 for all the other pairs.

8.2.3. Encoder Finetuning

In the contrastive pretraining stage, HESCAPE is capable of performing both full fine-tuning for small models and parameter-efficient fine tuning (PEFT) for large transformer-based models. These approaches help to align the pretrained models to the other modality, while potentially helping the encoders to adapt to specific tasks and potentially mitigating problems arising from batch effects.

To evaluate whether frozen pretrained encoders alone are sufficient for multimodal image-gene alignment, we conducted ablation experiments using various combinations of image-gene finetuning. We fine-tune the self-attention query-key-value embeddings and projection layers of the image encoder using LoRA [16]. In our ablation study, we find that both unlocked image and gene models enable better image-gene alignment when possible Tab. 11. However, we can often be restricted by the compute resources for large multimodal Foundation Model finetuning. Additionally, since the gene modality is deeply affected by batch effects, we decided to keep the gene models frozen for the HESCAPE benchmark.

8.3. Batch effects on gene expression modality

During dataset curation and preprocessing, we observed strong batch effects across samples of the same tissue type in all datasets under consideration. To systematically investigate how these batch effects impact contrastive pretraining performance, we employed the silhouette-batch metric from the single-cell integration benchmark `scib` [31].

We computed this metric for all datasets using Leiden clustering results after standard scanpy preprocessing [54] as the *label key*, and the train-validation-test split as the *batch key*. By treating the dataset split as batch information, we quantified how well integrated the gene expression profiles are across different data splits, a measure that reflects the presence of technical artifacts in the gene expression modality.

Fig. 2 shows the relationship between silhouette-batch values and average Recall@5 performance for the Gigapath-DRVI model across all datasets. Notably, we observe a clear linear relationship between retrieval performance and the batch integration metric, directly supporting our hypothesis that batch effects significantly impact contrastive pretraining effectiveness. Specifically, our analysis confirms that cancer tissues exhibit significant heterogeneity in cellular composition and transcriptomic profiles relative to healthy or non-cancerous diseased samples. Our findings reveal pronounced batch effects in cancer samples as seen in the organ specific datasets, especially in breast and colon tissues, whereas lung samples with homogeneous disease conditions, in particular, patients with pulmonary fibrosis, show minimal batch variation. Technical variability can further exacerbating these differences - 10x Visium samples, for instance, are particularly prone to batch effects compared to Xenium.

8.4. Downstream task: gene mutation prediction

For the evaluation of gene mutation prediction, we use a weakly-supervised learning approach for predicting the slide-level mutation targets from frozen patch-level embeddings of the pretrained models.

For feature extraction, we use the pipeline TRIDENT[1] with default parameters, extracting embeddings from patches of size 256×256 pixels at 20× magnification. The pretrained HESCAPE models have a latent dimension of 128.

For slide-level mutation prediction, we follow the HistoBistro pipeline[2]. Concretely, we employ Transformer-based feature aggregation using a two-layer Transformer architecture with eight heads of dimension 64 and latent dimension 512 [49]. We split the TCGA cohorts into five site-preserving folds for five-fold cross validation, using three folds for training, one for validation, and one for testing. We train the models for 10 epochs using the optimizer AdamW with learning rate of 2×10^{-5}, weight decay of 2×10^{-5}, and batch size 1. The best model is chosen based on the validation loss, evaluated every 500 iterations.

[1] https://github.com/mahmoodlab/TRIDENT
[2] https://github.com/peng-lab/HistoBistro