Supplementary Material: Multistain Pretraining for Slide Representation Learning in Pathology

Guillaume Jaume[1,2,⋆], Anurag Vaidya[1,2,⋆], Andrew Zhang[1,2,⋆⋆], Andrew H. Song[1,2,⋆⋆], Richard J. Chen[1,2], Sharifa Sahai[1,2], Dandan Mo[2], Emilio Madrigal[2], Long Phi Le[1,2], and Faisal Mahmood[1,2]

[1] Harvard Medical School, Cambridge, MA, USA
[2] Mass General Brigham, Boston, MA, USA
{gjaume, avaidya, faisalmahmood}@bwh.harvard.edu

We provide complementary information on the model architecture and training, additional results, and interpretability examples:

1. **Sec. 1**: Implementation details around pretraining, aggregator architecture, and graph optimal transport loss.
2. **Sec. 2**: Detailed descriptions of datasets used for pretraining and the downstream evaluations.
3. **Sec. 3**: Supervised Multiple Instance Learning (MIL) baselines.
4. **Sec. 4** and **5**: Additional results on downstream breast and kidney tasks.
5. **Sec. 6**: Ablations of loss and aggregator architecture.
6. **Sec. 7**: Additional interpretability examples of breast cancer cases.
7. **Sec. 8**: Limitations of MADELEINE.

1 Implementation details

1.1 Contrastive loss

Formally, we define a batch of B cases, where each case includes K pairs $(\mathbf{h}_i^{\text{HE}}, \mathbf{h}_i^{s_k})_{k=1}^K$, where s_k represents a non-H&E stain. The objective $\mathcal{L}_{\text{INFONCE}}$ is given by:

$$\mathcal{L}_{\text{INFONCE}} = -\frac{1}{K}\sum_{k=1}^{K}\frac{1}{2B}\left(\sum_{b=1}^{B}\log\frac{\exp\left(\frac{1}{\tau}\left(\boldsymbol{h}_b^{\text{HE}}\right)^{\text{T}}\boldsymbol{h}_b^{s_k}\right)}{\sum_{b'=1}^{B}\exp\left(\frac{1}{\tau}\left(\boldsymbol{h}_b^{\text{HE}}\right)^{\text{T}}\boldsymbol{h}_{b'}^{s_k}\right)} + \sum_{b=1}^{B}\log\frac{\exp\left(\frac{1}{\tau}\left(\boldsymbol{h}_b^{s_k}\right)^{\text{T}}\boldsymbol{h}_b^{\text{HE}}\right)}{\sum_{b'=1}^{B}\exp\left(\frac{1}{\tau}\left(\boldsymbol{h}_b^{s_k}\right)^{\text{T}}\boldsymbol{h}_{b'}^{\text{HE}}\right)}\right), \tag{1}$$

where the first and second terms represent the H&E-to-s_k and s_k-to-H&E contrastive loss, respectively. τ represents the Softmax temperature parameter.

⋆ Co-first authorship.
⋆⋆ Co-second authorship.

1.2 Multi-head attention architecture

MADELEINE uses a multi-head attention-based Multiple Instance Learning (MIL) architecture. Before applying each head, the patch embeddings are passed through a common pre-attention network consisting of 3 layers with 512 hidden units, layer normalization, GELU activation, and 0.1 dropout. Each attention head comprises a gated-attention network, consisting of a 2-layer MLP with 512 hidden units with Softmax activation and 0.25 dropout. The attention score $a_{i,j}^m$ for each patch derived from the m^{th} attention head of M total heads are defined as:

$$a_{i,j}^m = \frac{\exp\left(\mathbf{w}_m(\tanh(\mathbf{V}_m\widetilde{\mathbf{H}}_{i,j}^{\mathrm{HE}}) \odot \mathrm{sigm}(\mathbf{U}_m\widetilde{\mathbf{H}}_{i,j}^{\mathrm{HE}})\right)}{\sum_{j'=1}^{N_{\mathrm{HE}}} \exp\left(\mathbf{w}_m(\tanh(\mathbf{V}_m\widetilde{\mathbf{H}}_{i,j'}^{\mathrm{HE}}) \odot \mathrm{sigm}(\mathbf{U}_m\widetilde{\mathbf{H}}_{i,j'}^{\mathrm{HE}})\right)}, \ \forall m. \tag{2}$$

The output of each head is concatenated, and a post-attention network consisting of two linear layers with 2048 and 512 units is applied to get a slide embedding for each stain.

1.3 Additional information on the GOT objective

Additional details of the Graph Optimal Transport objectives are as follows,

1) Graph building: Each stain-specific graph is defined by instantiating 256 randomly sampled patches as nodes from the slide (sampling is done as each slide can have $> 10,000$ patches, making it computationally infeasible to calculate the complete optimal transport-based loss). Then, an edge is built between two nodes (*i.e.,* two patches) if the cosine similarity between their patch embeddings is larger than a threshold. The threshold is dynamically constructed and is set at the lowest similarity value, increasing by 0.1 times the difference between the highest and lowest similarity values.

2) $\mathcal{L}_{WD}$: Denoting $\mathbf{T} \in \mathbb{R}_+^{N_{\mathrm{HE}} \times N_{s_k}}$ as the transport plan, we can minimize the Wasserstein Distance (WD) between distributions $\hat{p}_{\mathrm{HE}}$ and $\hat{p}_{s_k}$ by finding the optimal transport plan

$$\mathcal{L}_{\mathrm{WD}}(\hat{p}_{\mathrm{HE}}, \hat{p}_{s_k}) = \min_{\mathbf{T}} \sum_j \sum_m \mathbf{T}_{j,m} \cdot C(v_j^{\mathrm{HE}}, v_m^{s_k}), \tag{3}$$

such that $\sum_{j=1}^{N_{\mathrm{HE}}} \mathbf{T}_{j,m} = 1/N_{s_k}, \forall m$ and $\sum_{m=1}^{N_{s_k}} \mathbf{T}_{j,m} = 1/N_{\mathrm{HE}}, \forall j$. The cost between cross-modal embeddings $C(v_j^{\mathrm{HE}}, v_m^{s_k})$ is computed with the cosine distance metric.

3) $\mathcal{L}_{GWD}$: In addition to the node-matching with $\mathcal{L}_{\mathrm{WD}}$, we also wish to match the graph topology via comparing the edge distance between stain-specific graphs. Denoting $\widetilde{\mathbf{T}} \in \mathbb{R}_+^{N_{\mathrm{HE}} \times N_{s_k}}$ as the transport plan as before,

$$\mathcal{L}_{\mathrm{GWD}}(\hat{p}_{\mathrm{HE}}, \hat{p}_{s_k}) = \min_{\mathbf{T}} \sum_{j,j',m,m'} \widetilde{\mathbf{T}}_{j,m}\widetilde{\mathbf{T}}_{j',m'} \cdot C(v_j^{\mathrm{HE}}, v_m^{s_k}, v_{j'}^{\mathrm{HE}}, v_{m'}^{s_k}), \tag{4}$$

such that $\sum_{j=1}^{N_{\text{HE}}} \widetilde{\mathbf{T}}_{j,m} = 1/N_{s_k}, \forall m$ and $\sum_{m=1}^{N_{s_k}} \widetilde{\mathbf{T}}_{j,m} = 1/N_{\text{HE}}, \forall j$. The cost between the pairs $(v_j^{\text{HE}}, v_{j'}^{\text{HE}})$ and $(v_m^{s_k}, v_{m'}^{s_k})$ is given as $C(v_j^{\text{HE}}, v_m^{s_k}, v_{j'}^{\text{HE}}, v_{m'}^{s_k}) = \|c(v_j^{\text{HE}}, v_{j'}^{\text{HE}}) - c(v_m^{\text{HE}}, v_{m'}^{\text{HE}})\|$, with $c(\cdot, \cdot)$ representing the cosine similarity metric.

1.4 MADELEINE pretraining

We pretrained MADELEINE with AdamW optimizer and a batch size of 90 for 120 epochs. The learning rate is linearly ramped up during a 5-epoch warmup from 1e-9 to 1e-4. Then, we employed a cosine scheduler to reach the final learning rate of 1e-8 after 120 epochs. To increase training diversity and simplify batch processing, we sample a fixed and random subset of patches per slide, specifically 2048 patch embeddings. In slides with fewer patches, we perform random oversampling. All training settings are summarized in Appendix Table 1.

Table 1: MADELEINE pretraining and architectural hyperparameters. 3 × 24GB NVIDIA 3090Ti GPUs were used for training. Batch size refers to the total batch size across all GPUs.

Hyperparameter	Value
Heads	4
Head activation	GELU
Patch embedding dimension	512
Pre-attention hidden dimension	512
Patches sampled (training)	2048
Stain encoding dimension	32
AdamW β	(0.9, 0.999)
Batch size	90
Warmup epochs	5
Max epochs	120
Learning rate schedule	Cosine
Learning rate (start)	0
Learning rate (post warmup)	1e-4
Learning rate (final)	1e-8
Weight decay	0.01
INFONCE Temperature	0.001
Patches sampled for GOT	256
GOT γ	0
Automatic mixed precision	bloaft16
Early stopping criteria	SmoothRank [3]

1.5 Early stopping with rank analysis

Following [3], we use the rank as a measure of the quality of the underlying latent space learned during MADELEINE pretraining and save the model weights from the highest rank iteration (no models are saved during the first 20 epochs of training). We compute the rank as the entropy of the d (assuming $d < n$) L1-normalized singular values of the slide embedding matrix $H \in \mathbb{R}^{n \times d}$. Specifically, we have:

$$\text{RankMe}(H) = \exp(-\sum_{k=1}^{d} p_k \log(p_k)) \ , \tag{5}$$

$$p_k = \frac{\sigma_k(H)}{|\sigma(H)|_1} + \epsilon \tag{6}$$

where σ_k denotes the $k-$th singular of H (sorted from large to low), and ϵ is small constant set to $1e-7$ for numerical stability.

1.6 Additional information on evaluation

Few-shot evaluation is based on linear probing and prototyping classification.

Linear probing Linear probing is implemented using a logistic regression objective based on sklearn. We use the default sklearn L2 regularization (set to 1.0) with an lbfgs solver. We set the maximum number of training iterations to 10,000 for all experiments.

Prototyping We define positive and negative slide "prototypes" p^+, p^- as the average of k (k=1,5,10,25) slide embeddings using downstream task labels. Subsequently, we measure the similarity between a query slide embedding q_i and the two prototypes using the L2 distance. We apply this evaluation for morphological and molecular subtyping, and allograft rejection prediction.

2 Datasets

Overall, our study comprises a total of 23,580 whole slide images (WSIs) from two organs (breast and kidney) and includes eight different immunohistochemistry and special stains. We use 16,281 of these WSIs for pretraining and 7,299 for downstream evaluation. We now detail all the sources of the WSIs used in the study.

Acrobat (multi-stain, pretraining) Acrobat is a multi-stain dataset originally proposed as part of the AutomatiC Registration Of Breast cAncer Tissue MICCAI challenge [12, 13]. It comprises 4,211 whole slide images (WSI) sourced from 1,153 patients diagnosed with primary breast cancer. These WSIs are available at a magnification of 10× (equivalent to 1μm/px) and show tissue resections, which have been processed using either hematoxylin and eosin (H&E) staining or immunohistochemistry (IHC). For every patient included in the dataset, there is one WSI that has been stained with H&E, along with a minimum of one and a maximum of four WSIs of tissue from the same tumor that has been stained with ER (N=844), PR (N=837), HER2 (N=534), or KI67 (N=843). The collection of slides was digitized at Karolinska Institutet in Stockholm, Sweden, during routine clinical workflows. Data can be downloaded at https://acrobat.grand-challenge.org/data/.

TCGA Breast (H&E, downstream) We collected N=1,041 primary cases from the TCGA Breast Invasive Carcinoma (BRCA) cohort, which comprises N=831 Invasive Ductal Carcinoma (IDC) and N=210 Invasive Lobular Carcinoma (ILC). For each case, we downloaded the corresponding disease-specific survival and associated censorship status, subtype (IDC and ILC), and molecular status: ER (N=996; 780 positive, 216 negative), PR (N=993; 678 positive, 315 negative), and HER2 (N=693; 158 positive, 535 negative) from UCSC Xena [4] and cBioPortal [1]. WSIs can be downloaded from Genomics Data Commons (https://portal.gdc.cancer.gov/).

BCNB (H&E, downstream) The Cancer Core-Needle Biopsy WSI (BCNB) dataset comprises N=1,058 patients, with a single side associated with each patient [17]. BCNB includes the molecular status of each patient: ER (N=1,058; 831 positive, 227 negative), PR (N=1,058; 790 positive, 268 negative), HER2 (1,058; 277 positive, 781 negative), and KI67 (1,058; 156 positive, 902 negative). The dataset was originally collected from hospital systems in Beijing, China, and is made publicly available at (https://paperswithcode.com/dataset/bcdalnmp).

AIDPATH (H&E, downstream) AIDPATH dataset contains 50 breast cancer WSIs stained with H&E. The dataset additionally provides HER2 expression (positive or negative, where equivocal cases are analyzed with FISH) (7 positive, 41 negative) and KI67 expression (provided as a percentage). We convert the continuous KI67 expression values to a binary task using 50% as a threshold for

IHC status prediction (19 positive, 31 negative). The dataset is made publicly available at https://mitel.dimi.uniud.it/aidpath-db.

BWH Breast (H&E, downstream) We collected an invasive breast cancer cohort (N=1,265) from the archives of Hopsital-A, which comprises N=982 IDC and N=283 ILC cases. All cases were primary breast cancers and included resections and biopsies. All slides were scanned at 20× or 40× magnification. Using patient reports, we additionally collected molecular status: ER (N=874; 613 positive, 261 negative), PR (N=874; 504 positive, 370 negative), and HER2 (N=816; 151 positive, 665 negative).

MGH Breast (Estrogen and Progesterone receptor stains, downstream) We use another private breast cohort for IHC quantification of ER abundance (N=962) and PR abundance (N=1,071). We frame both ER and PR quantification as 3- and 6-class problems. For a detailed breakdown, see Appendix Table 2.

BWH Kidney (multi-stain, pretraining) We collected a private renal transplant cohort comprising kidney biopsies from 1,069 renal transplant cases. Each case includes one to three tissue blocks, where each block consists of one to two H&E-stained (N=4,638) and one to two periodic acid-Schiff (PAS) (N=4,630) slides, one Jones-stained slide (N=2,326) and one Trichrome-stained slide (N=2,328). In total, each case includes 6 to 18 slides. We held out 20% of the cohort (210 cases, 463 H&E slides) as an independent test set and used the rest for MADELEINE pretraining. All slides were processed at 20×. We use H&E slides of the held-out cases to screen for Antibody-mediated rejection (AMR, 2 class; 107 positive, 356 negative) and quantify Interstitial Fibrosis and Tubular Atrophy (IFTA, 3 classes; mild : 292, moderate : 104, advanced : 67). As each case includes several H&E slides, we define two sub-tasks: “single-slide" prediction, where we use a single slide per case (N=463 H&E slides), and “all-slides" prediction, where we use all available slides per case (N=305 H&E slides).

Table 2: MGH Breast label distribution.

Model/Data	**ER**		**PR**	
	$\mathcal{C}=6$	$\mathcal{C}=3$	$\mathcal{C}=6$	$\mathcal{C}=3$
0	175	335	168	337
$<1\%$	160		169	
$1-10\%$	120	292	219	389
$10-50\%$	172		170	
$50-90\%$	176	335	175	344
$>90\%$	159		169	

3 Baselines

3.1 Supervised multiple instance learning baselines

We provide a detailed description of the four multiple-instance learning (MIL) approaches used in the study.

1. **ABMIL** [6]: Attention-based multiple instance learning (ABMIL) is a popular MIL architecture. ABMIL operates as follows: first, it assigns patch-level importance scores through a gated-attention mechanism. Attention scores are used to weigh patch embeddings, which are subsequently summed to build a slide representation used for classification.
2. **TransMIL** [10]: Transformer-based multiple instance learning (TransMIL) replaces the gated attention from ABMIL with a low-rank Transformer. TransMIL first squares the sequence of low dimensional representations then applies a Pyramidal Positional Encoding module to encode spatial information and finally uses Nystrom attention [16] to approximate self-attention scores between patches. The CLS token is finally taken as the slide-level representation.
3. **Information Bottleneck MIL** [8]: Information bottlenecks (IB) are used to compress the WSI by removing uninformative instances (patch embeddings). IB aims to find patch instances that minimize the mutual information between the distribution of patches and patch representations. By only keeping such instances, [8] postulate that the most informative patches can be retained, which can then be aggregated into a slide embedding.
4. **Low-rank MIL** [15]: While TransMIL tries to learn slide-level representations by encoding patch correlations, it does not leverage the redundancy in WSIs, which [15] used to propose iterative low-rank attention (ILRA). Each ILRA block consists of two layers: one aims to project the sequence of patch representations to a low-rank space by cross-attending it with a latent matrix, and the second reconstructs the input. Performing max-pooling over the output of k such layers yields a low-rank slide-level representation.

3.2 Slide-level baselines

1. **Mean**: The Mean baseline is defined by taking the arithmetic average of the patch embedding constituting the slides.
2. **HIPT** [2]: Hierarchical Image Pyramid Transformer (HIPT) proposes a 3-level slide encoding schema, where each level is independently trained with a Transformer. The first level transforms patches into patch embeddings, which are then aggregated in region embeddings and finally into a slide embedding.
3. **GigaSSL** [7]: GigaSSL, similar to the Intra baseline, is a method for learning slide representations based on different views of the same slide. It creates different views of a slide by sampling patches and applying augmentations such as random cropping. The different views are then pulled using a contrastive loss to learn the slide representation. Author-provided GigaSSL slide embeddings for TCGA Breast were taken from https://data.mendeley.com/datasets/d573xfd9fg/3.

4. **GigaPath** [18]: GigaPath is a concurrent work to ours scaling intra-SSL to large cohorts. It includes its own pan-cancer patch encoder that was pre-trained on 171,000+ WSIs (> 30,000 patients) using DINOv2. The slide encoder was trained using a LongNet model with masked auto encoding. We used the official GigaPath demo[3] using (1) 256×256-pixel patching at 20× and the latest HuggingFace tile and slide encoders. GigaPath-(MEAN) is defined by taking the average of all patch embeddings, and GigaPath is defined by building a slide embedding using global pooling of all LongNet tokens at the 11th Transformer layer.

[3] https://github.com/prov-gigapath/prov-gigapath/blob/main/demo/run_gigapath.ipynb

4 Additional Breast results

Table 3: Few-shot molecular status prediction from H&E in TCGA Breast. Evaluation using Macro-AUC. Standard deviation reported over ten runs. All results for $k = 25$ training samples per class. MADELEINE refers to INFONCE + GOT. Besides HIPT and GigaSSL, all models use the *same* patch encoder. GigaSSL embeddings for BWH and BCNB cohorts were not available. Best in **bold**, second best is underlined.

	Model/Data	**TCGA** (↑)			**BCNB** (↑)			**BWH** (↑)		
		ER	PR	HER2	ER	PR	HER2	ER	PR	HER2
MIL	ABMIL [6]	82.7 ± 3.1	72.8 ± 2.5	62.4 ± 3.3	81.4 ± 4.5	75.9 ± 2.9	64.5 ± 3.4	65.4 ± 3.5	62.6 ± 2.6	57.0 ± 2.4
	TransMIL [10]	75.1 ± 5.5	63.5 ± 7.1	55.0 ± 2.9	68.7 ± 7.0	63.4 ± 8.2	56.0 ± 4.1	54.8 ± 4.2	54.3 ± 3.0	51.8 ± 2.7
	IB-MIL [8]	81.6 ± 2.8	73.1 ± 2.5	<u>62.6</u> ± 2.6	81.0 ± 2.8	76.7 ± 2.2	64.3 ± 4.7	65.8 ± 3.1	61.7 ± 2.5	55.2 ± 3.7
	ILRA [15]	82.3 ± 2.7	74.0 ± 2.4	62.2 ± 2.6	79.0 ± 7.2	74.2 ± 1.9	63.6 ± 3.9	63.2 ± 3.9	59.9 ± 4.7	55.1 ± 3.5
Linear probe	MEAN	79.4 ± 4.7	68.5 ± 4.2	59.6 ± 3.1	78.5 ± 3.2	72.4 ± 3.8	64.7 ± 4.3	62.8 ± 2.8	60.7 ± 3.6	55.8 ± 4.3
	Intra	80.1 ± 3.3	70.2 ± 3.2	60.0 ± 2.5	79.6 ± 2.7	75.1 ± 2.7	67.1 ± 3.7	64.0 ± 2.2	60.9 ± 3.1	55.9 ± 3.4
	$\text{HIPT}_{\text{CLS-4k}}$ [2]	74.1 ± 3.0	63.9 ± 2.9	61.9 ± 3.7	65.3 3.7	62.2 4.4	54.0 3.0	60.3 ± 3.6	57.0 ± 3.6	52.9 ± 3.7
	GigaSSL [7]	77.7 ± 3.1	69.4 ± 2.8	59.9 ± 3.2	– –	– –	– –	– –	– –	– –
	GigaPath-(MEAN) [18]	77.7 ± 4.8	68.1 ± 2.8	58.3 ± 3.4	76.2 ± 3.9	71.0 ± 3.9	63.7 ± 3.8	64.6 ± 2.2	59.3 ± 3.1	54.4 ± 3.4
	GigaPath [18]	76.0 ± 5.0	66.7 ± 2.7	57.4 ± 3.4	74.1 ± 3.7	68.9 ± 4.0	61.8 ± 3.7	63.0 ± 2.4	58.2 ± 3.2	53.2 ± 3.4
	MADELEINE	<u>84.7</u> ± 2.1	74.7 ± 3.4	61.3 ± 2.4	**84.4** ± 1.3	<u>79.2</u> ± 2.3	<u>68.5</u> ± 3.1	**68.7** ± 1.8	<u>65.1</u> ± 3.2	**60.3** ± 3.8
	MADELEINE-SE	84.6 ± 2.2	74.6 ± 2.8	62.5 ± 2.6	81.8 ± 2.0	76.9 ± 2.8	**69.8** ± 2.3	<u>68.5</u> ± 2.0	64.3 ± 2.2	<u>59.7</u> ± 3.9
Prototyping	MEAN	79.6 ± 4.0	70.8 ± 2.0	61.1 ± 3.4	76.9 ± 2.2	74.8 ± 2.7	65.5 ± 2.9	62.5 ± 2.8	58.7 ± 3.9	54.3 ± 4.2
	Intra	79.9 ± 2.8	71.6 ± 1.4	60.3 ± 2.4	77.7 ± 1.8	74.7 ± 2.1	66.8 ± 3.2	63.3 ± 2.7	59.4 ± 3.8	54.2 ± 4.4
	$\text{HIPT}_{\text{CLS-4k}}$ [2]	69.9 ± 2.9	64.9 ± 2.1	60.7 ± 3.3	56.4 ±2.8	56.2 ±4.1	53.3 ±2.3	60.7 ± 2.0	57.0 ± 2.6	55.0 ± 2.7
	GigaSSL [7]	80.0 ± 2.4	71.5 ± 2.0	62.8 ± 2.4	– –	– –	– –	– –	– –	– –
	GigaPath-(MEAN) [18]	71.7 5.6	66.1 3.8	56.4 2.6	69.7 ± 3.2	68.3 ± 5.1	61.6 ± 3.7	62.0 ± 2.7	58.8 ± 2.8	53.8 ± 3.8
	GigaPath [18]	70.5 5.4	63.9 2.7	56.4 1.9	68.5 ± 3.3	66.5 ± 4.8	59.4 ± 4.0	60.4 ± 2.9	57.4 ± 2.4	52.5 ± 3.4
	MADELEINE	**85.1** ± 1.4	**76.4** ± 1.2	<u>62.6</u> ± 2.9	<u>83.0</u> ± 1.6	**80.7** ± 1.8	<u>68.5</u> ± 2.2	68.2 ± 2.7	**65.7** ± 3.8	57.1 ± 3.2
	MADELEINE-SE	83.3 ± 1.5	<u>74.9</u> ± 1.1	**62.9** ± 3.1	80.5 ± 1.6	77.3 ± 1.5	**69.8** ± 2.0	67.2 ± 2.6	64.7 ± 3.2	56.9 ± 3.8

Table 4: IHC quantification. We quantify the abundance of estrogen (ER) (N=962) and progesterone (PR) (N=1,071) receptor expression in 3-class and 6-class scenarios using IHC. We compare MADELEINE fine-tuned against MADELEINE architecture trained from scratch and MEAN. Results using 5-fold cross-validation with k=25 examples per class and evaluated using macro-AUC. Best in **bold**, second best is underlined.

Model/Data	**ER** (↑)		**PR** (↑)	
	$\mathcal{C}=3$	$\mathcal{C}=6$	$\mathcal{C}=3$	$\mathcal{C}=6$
MEAN (linear probe)	74.6 ± 1.9	69.5 ± 1.2	73.1 ± 1.8	69.1 ± 1.0
ABMIL (Random)	82.1 ± 2.0	83.4 ± 1.3	83.8 ± 1.4	83.9 ± 1.4
ABMIL (FineTune)	**89.6** ± 1.3	**86.0** ± 0.8	**89.4** ± 0.9	**85.5** ± 0.9

Table 5: Survival outcome prediction in TCGA Breast. Models are trained using site-stratified 5-fold cross-validation. Evaluation using Concordance index (c-index). Besides HIPT, GigaSSL and GigaPath, all models use the *same* patch encoder. Best in **bold**, second best is underlined.

	Model/Data	**TCGA Breast** (↑)
MIL	ABMIL [6]	0.669 ± 0.073
	TransMIL [10]	0.697 ± 0.046
	IB-MIL [8]	0.612 ± 0.088
	ILRA [15]	0.657 ± 0.067
Slide level	MEAN	0.687 ± 0.079
	INTRA	0.692 ± 0.069
	$\text{HIPT}_{\text{CLS-4k}}$ [2]	0.547 ± 0.078
	GigaSSL [7]	0.530 ± 0.038
	GigaPath-(MEAN) [18]	0.587 ± 0.091
	GigaPath [18]	0.521 ± 0.083
	MADELEINE	**0.715** ± 0.041
	MADELEINE-SE	0.696 ± 0.073

Table 6: Molecular subtyping from H&E. Detection of HER2 and KI67 status (binary) from H&E in AIDPATH and BCNB datasets. Results of MADELEINE and MADELEINE-SE obtained using linear probing. "SL" stands for slide level. Results using 5-fold stratified cross-validation evaluated using macro-AUC. Best in **bold**, second best is underlined.

	Model/Data	**HER2** (↑)	**KI67** (↑)	
		AIDPATH	AIDPATH	BCNB
MIL	ABMIL [6]	81.1 ± 8.9	89.2 ± 7.7	81.9 ± 3.7
	TransMIL [10]	46.4 ± 10.7	65.1 ± 19.9	74.9 ± 10.1
	IB-MIL [8]	73.2 ± 11.1	87.7 ± 6.1	81.6 ± 3.5
	ILRA [15]	76.1 ± 7.8	84.9 ± 4.9	78.8 ± 3.6
SL	MEAN	77.4 ± 20.5	80.2 ± 2.8	79.6 ± 4.0
	INTRA	85.8 ± 17.4	80.2 ± 5.9	80.9 ± 3.6
	MADELEINE	81.5 ± 9.9	**91.3** ± 4.9	81.4 ± 4.2
	MADELEINE-SE	**92.5** ± 7.2	83.0 ± 8.6	**82.0** ± 3.6

5 Additional kidney results

Table 7: Kidney rejection tasks. Linear probe and prototyping for $k = 50$ reported. HIPT and GigaSSL are not available for non-cancer datasets. Best in **bold**, second best is underlined.

	Model/Data	**IFTA** (↑)		**AMR** (↑)	
		Slide	**Patient**	**Slide**	**Patient**
MIL	ABMIL [6]	74.5 ± 1.9	78.2 ± 4.9	69.6 ± 3.7	71.2 ± 5.8
	TransMIL [10]	57.5 ± 3.7	60.4 ± 6.6	55.9 ± 5.4	55.4 ± 10.5
	IB-MIL [8]	73.0 ± 3.7	80.0 ± 5.1	67.4 ± 4.3	69.6 ± 8.5
	ILRA [15]	70.9 ± 5.0	77.5 ± 4.5	62.4 ± 6.8	63.5 ± 10.4
Linear probe	MEAN	73.6 ± 2.6	79.3 ± 2.1	67.8 ± 3.2	70.9 ± 3.2
	INTRA	74.5 ± 2.2	80.2 ± 1.8	67.9 ± 3.1	72.0 ± 3.2
	MADELEINE	<u>75.3</u> ± 2.5	<u>80.9</u> ± 2.3	**71.2** 2.9	<u>73.8</u> ± 3.2
	MADELEINE-SE	**76.1** ± 2.0	**82.4** ± 1.8	<u>70.0</u> ± 2.9	**74.2** ± 3.5
Prototyping	MEAN	70.2 ± 2.4	75.0 ± 2.8	63.8 ± 6.0	67.5 ± 8.1
	INTRA	70.6 ± 2.5	75.7 ± 3.2	63.4 ± 4.3	66.5 ± 5.4
	MADELEINE	72.1 ± 2.4	77.0 ± 3.0	66.7 ± 3.7	71.2 ± 4.3
	MADELEINE-SE	73.1 ± 2.4	79.8 ± 2.0	65.5 ± 4.1	70.4 ± 5.4

6 Additional ablations

Table 8: Ablation study of Madeleine feature extractor. Survival was evaluated using c-index and site-stratified 5-fold cross-validation. Subtyping and molecular status prediction were evaluated using macro-AUC and prototyping evaluation (k=25) repeated five times with fixed seed across baselines. Standard deviation reported over the 5 runs. Madeleine refers to pretraining on breast cancer using InfoNCE + GOT without stain encoding. CONCH is the patch encoder of the Vision+Language model proposed in [9]. Best in **bold**, second best is underlined.

Model/Data	**TCGA Survival** (↑)	**BWH Subtyping** (↑)	**TCGA PR** (↑)	**BCNB ER** (↑)	**Avg**
CTransPath+Mean	68.6 ± 4.0	81.1 ± 3.9	65.0 ± 1.5	67.7 ± 2.1	70.6
CONCH+Mean	68.7 ± 7.9	86.2 ± 7.9	70.8 ± 3.0	76.9 ± 2.0	75.6
CTransPath+Madeleine	65.4 ± 6.5	83.1 ± 4.6	66.9 ± 1.6	68.6 ± 3.5	71.0
CONCH+Madeleine	**71.5** ± 4.1	**94.9** ± 0.9	**76.4** ± 1.9	**83.0** ± 1.6	**81.5**

Table 9: Ablation study of MADELEINE architecture. Survival was evaluated using c-index and site-stratified 5-fold cross-validation. Subtyping and molecular status prediction were evaluated using macro-AUC and prototyping evaluation (k=25) repeated five times with fixed seed across baselines. Standard deviation reported over the 10 runs. "MH" refers to multi-head attention, "SH" to single-head attention, and "SE" to stain encoding. MADELEINE refers to pretraining on breast cancer using INFONCE + GOT. Best in **bold**, second best is underlined.

Model/Data	**TCGA Survival** (↑)	**BWH Subtyping** (↑)	**TCGA PR** (↑)	**BCNB ER** (↑)	**Avg**
MADELEINE-SH	70.1 ± 7.1	91.8 ± 1.7	75.0 ± 1.4	80.5 ± 2.5	79.4
MADELEINE w. TransMIL	55.7 ± 7.8	90.8 ± 1.7	75.6 ± 1.3	82.1 ± 1.3	76.5
MADELEINE w. SE	69.6 ± 7.3	**95.8** ± 0.8	74.9 ± 1.1	80.5 ± 1.6	80.2
MADELEINE-MH	**71.5** ± 4.1	94.9 ± 0.9	**76.4** ± 1.2	**83.0** ± 1.6	81.5

7 Additional interpretability examples

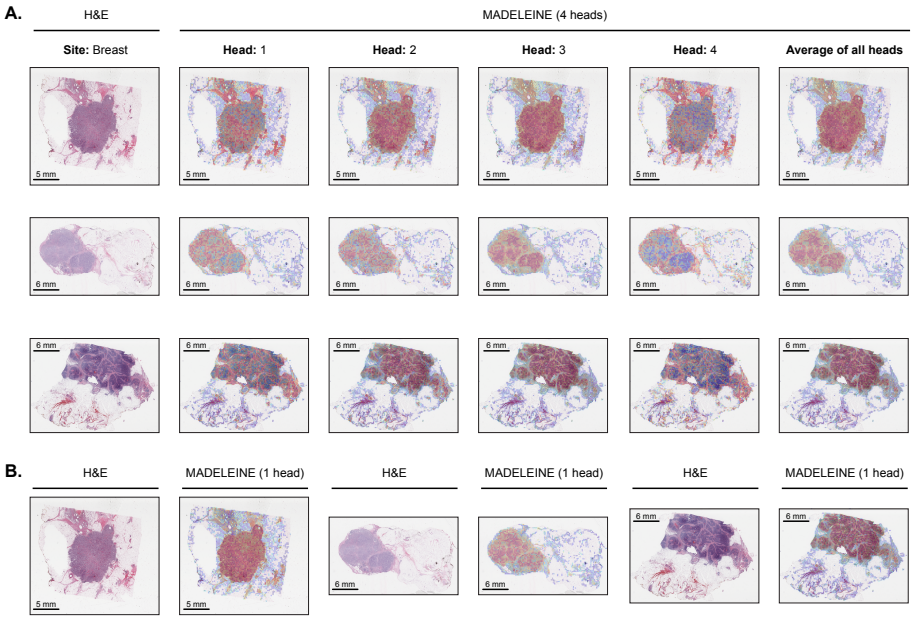

Fig. 1: Additional heatmap examples obtained with Madeleine A. Attention weights of multi-headed (frozen) ABMIL slide encoder pretrained with Madeleine overlaid on three randomly chosen samples for TCGA Breast cohort. We show all heads and the average of heads. **B.** Attention weights of a single head (frozen) ABMIL slide encoder pretrained with Madeleine overlaid on three randomly chosen samples for TCGA Breast cohort. Multi-headed ABMIL trained with Madeleine can focus on different morphologies, whereas single-headed ABMIL focuses only on tumor morphology.

8 Limitations

MADELEINE is a multimodal pre-training strategy for slide representation learning. It operates under the assumption that representation learning of H&E images can be guided by other stains (immunohistochemistry and special stains). This premise is directly inspired by the standard practice in clinical settings, where H&E staining is routinely performed as the *gold standard* procedure, along with complementary stains. Though this approach is principled, we highlight some limitations of our study and this methodology more broadly.

Data scaling Clinical practice is complex and ever evolving. Every year, new IHC and special stains become available, some of which are integrated in the workflow and can be used on a case-by-case basis. In breast cancer, our study focuses on four IHC stains (the most common ones), whereas many more can be employed, such as Epidermal Growth Factor Receptor (EGFR), P53, and E-Cadherin. As each stain offers a different view of a biomarker, increasing the number of stains would make the training signal richer and the resulting representation potentially better.

Lack of large public datasets Acrobat is the only *large-scale* public dataset with H&E and IHC stains. Therefore, without relying on proprietary data, such method cannot be scaled to more stains and other types of cancer. While the NADT-Prostate [14] cohort includes H&E and IHC, it remains limited by its size; for example, 14/18 stains provided have less than 100 examples, preventing efficient pre-training in prostate adenocarcinoma. In addition, TCGA includes known limitations such as site-specific biases [5] and demographic biases [11]. Despite these limitations, TCGA remains the largest public resource for cancer prognostication and survival analyses.

Model scaling MADELEINE is trained using a combination of a global objective using contrastive learning and a local objective using graph optimal transport (GOT). Using our current hardware (3× 3090 GPUs), we are limited by the maximum batch size for contrastive learning, even using efficient parallelization and bfloat16 quantization. In addition, computing GOT is computationally expensive, with significant memory requirements. Because of this constraint, we must use 256 patch embeddings (or tokens) per stain for computing GOT. Scaling to more tokens would allow finer-grained matching between stains. Local alignment through GOT also requires morphological overlap between tissue sections used for different stains.

References

1. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al.: The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery **2**(5), 401–404 (2012) 5
2. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16144–16155 (2022) 7, 9, 10
3. Garrido, Q., Balestriero, R., Najman, L., Lecun, Y.: Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In: International conference on machine learning (10 2022) 3, 4
4. Goldman, M.J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A.N., Zhu, J., Haussler, D.: Visualizing and interpreting cancer genomics data via the xena platform. Nature biotechnology **38**(6), 675–678 (2020) 5
5. Howard, F.M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O.I., Kather, J.N., et al.: The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nature communications **12**(1), 4423 (2021) 15
6. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018) 7, 9, 10, 11
7. Lazard, T., Lerousseau, M., Decencière, E., Walter, T.: Giga-ssl: Self-supervised learning for gigapixel images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4304–4313 (2023) 7, 9, 10
8. Li, H., Zhu, C., Zhang, Y., Sun, Y., Shui, Z., Kuang, W., Zheng, S., Yang, L.: Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (03 2023) 7, 9, 10, 11
9. Lu, M., Chen, B., Williamson, D., Chen, R., Liang, I., Ding, T., Jaume, G., Odintsov, I., Zhang, A., Le, L., Gerber, G., Parwani, A., Mahmood, F.: Towards a visual-language foundation model for computational pathology. Nature Medicine (2024) 12
10. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in Neural Information Processing Systems **34**, 2136–2147 (2021) 7, 9, 10, 11
11. Vaidya, A., Chen, R.J., Williamson, D.F., Song, A.H., Jaume, G., Yang, Y., Hartvigsen, T., Dyer, E.C., Lu, M.Y., Lipkova, J., et al.: Demographic bias in misdiagnosis by computational pathology models. Nature Medicine **30**(4), 1174–1190 (2024) 15
12. Weitz, P., Valkonen, M., Solorzano, L., et al.: A multi-stain breast cancer histological whole-slide-image data set from routine diagnostics. Sci Data **10**, 562 (2023) 5
13. Weitz, P., Valkonen, M., Solorzano, L., Carr, C., Kartasalo, K., Boissin, C., Koivukoski, S., Kuusela, A., Rasic, D., Feng, Y., et al.: A multi-stain breast cancer histological whole-slide-image data set from routine diagnostics. Scientific Data **10**(1), 562 (2023) 5

14. Wilkinson, S., Ye, H., Karzai, F., Harmon, S.A., Terrigino, N.T., VanderWeele, D.J., Bright, J.R., Atway, R., Trostel, S.Y., Carrabba, N.V., Whitlock, N.C., Walker, S.M., Lis, R.T., Sater, H.A., Capaldo, B.J., Madan, R.A., Gulley, J.L., Chun, G., Merino, M.J., Pinto, P.A., Salles, D.C., Kaur, H.B., Lotan, T.L., Venzon, D.J., Choyke, P.L., Turkbey, B., Dahut, W.L., Sowalsky, A.G.: Nascent prostate cancer heterogeneity drives evolution and resistance to intense hormonal therapy (sep 2020) 15
15. Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: The Eleventh International Conference on Learning Representations (2022) 7, 9, 10, 11
16. Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., Singh, V.: Nyströmformer: A nyström-based algorithm for approximating self-attention. In: Proceedings of the AAAI Conference on Artificial Intelligence (2021) 7
17. Xu, F., Zhu, C., Tang, W., Wang, Y., Zhang, Y., Li, J., Jiang, H., Shi, Z., Liu, J., Jin, M.: Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. Frontiers in oncology **11**, 759007 (2021) 5
18. 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., Poon, H.: A whole-slide foundation model for digital pathology from real-world data. Nature (2024) 8, 9, 10