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, Boston, MA, USA
[2] Mass General Brigham, Boston, MA, USA
gjaume,avaidya,faisalmahmood@bwh.harvard.edu

Abstract. Developing self-supervised learning (SSL) models that can learn universal and transferable representations of H&E gigapixel whole-slide images (WSIs) is becoming increasingly valuable in computational pathology. These models hold the potential to advance critical tasks such as few-shot classification, slide retrieval, and patient stratification. Existing approaches for slide representation learning extend the principles of SSL from small images (e.g., 224×224 patches) to entire slides, usually by aligning two different augmentations (or *views*) of the slide. Yet the resulting representation remains constrained by the limited clinical and biological diversity of the *views*. Instead, we postulate that slides stained with multiple markers, such as immunohistochemistry, can be used as different *views* to form a rich task-agnostic training signal. To this end, we introduce Madeleine, a multimodal pretraining strategy for slide representation learning. Madeleine is trained with a dual global-local cross-stain alignment objective on large cohorts of breast cancer samples (N=4,211 WSIs across five stains) and kidney transplant samples (N=12,070 WSIs across four stains). We demonstrate the quality of slide representations learned by Madeleine on various downstream evaluations, ranging from morphological and molecular classification to prognostic prediction, comprising 21 tasks using 7,299 WSIs from multiple medical centers. Code at https://github.com/mahmoodlab/MADELEINE.

Keywords: Computational pathology; Slide Representation Learning

1 Introduction

Self-supervised learning (SSL) via multimodal pretraining is increasingly adopted in medical AI for constructing universal image representations that can be used for diagnosis, prognosis, and treatment response prediction [2, 13, 38]. The core idea is to *align* an image (*e.g.,* a histology region-of-interest of a tumor) with other corresponding modalities (*e.g.,* the morphological text description of the

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

tumor) into a shared latent space via contrastive learning or other similarity matching losses [54]. Intuitively, the richer the contrasting modality employed, the more detailed and nuanced the image representations become, enabling better generalization and transferability to downstream tasks.

In computational pathology [62], multimodal pretraining has mostly focused on building visual-language models of small images [21, 25], capitalizing on their success in computer vision [54, 69]. However, the scale of whole-slide images (WSIs), often exceeding 100,000 × 100,000 pixels at 20× magnification (0.5 μm/pixel), presents a significant challenge for adapting such techniques to pathology. To address this issue, most intra-modal and multimodal SSL methods focus on embedding small patches (*e.g.,* 224 × 224), which can then be aggregated using multiple instance learning (MIL) for downstream tasks [26, 48, 58]. Alternatively, the aggregation stage can also be pretrained via SSL to create a *slide embedding* from the patch embeddings [12, 39, 49, 63]. The hierarchical construction from patches to patch embeddings to a slide embedding in a two-stage training pipeline enables self-supervised *slide representation learning*, without utilizing labels from pathologists or learning task-specific representations.

However, most existing slide representation learning methods are intra-modal, thus limiting the richness and diversity of the training signal to learning visual invariances within the slide [12, 39]. Instead, we propose to leverage additional modalities that naturally form clinically and biologically relevant *pairs* suitable for pretraining. In this study, we hypothesize that WSIs stained with various markers, such as immunohistochemistry (IHC), can constitute a strong task-agnostic training signal for multimodal pretraining. Each stain can be seen as a different *view* of the H&E slide by highlighting spatially-resolved expression levels of relevant markers. In addition, unlike bulk gene expression data or text captions [29], H&E and other stains offer fine-grained morphological correspondences, which can be leveraged for enhanced representational power.

To this end, we introduce Madeleine, an SSL approach for multistain-guided *slide representation learning*. Madeleine uses a multihead attention-based MIL [26, 48] to encode pre-extracted patch embeddings into a *slide embedding*. Madeleine is pretrained on large collections of multistain tissue using a dual global-local cross-stain objective. The global objective, based on a symmetric contrastive loss [15], learns slide-level correspondences between the H&E slide and the other stains. This alignment guides the H&E embedding to encapsulate the global morphological composition of the tissue. The local objective, based on the Graph Optimal Transport framework [11, 51], learns patch-level correspondences between the H&E and the other stains, thereby enabling cross-stain matching of fine-grained morphological features. The resulting latent space (i) can encode all stains encountered during pretraining, as the same network is employed for encoding each stain, and (ii) can be used for diverse downstream applications, as the training signal and resulting model are task-agnostic.

To summarize, our contributions are (1) Madeleine, a multimodal pretraining strategy for *slide representation learning* in computational pathology; (2) a large-scale demonstration of Madeleine pretraining on two organs, breast

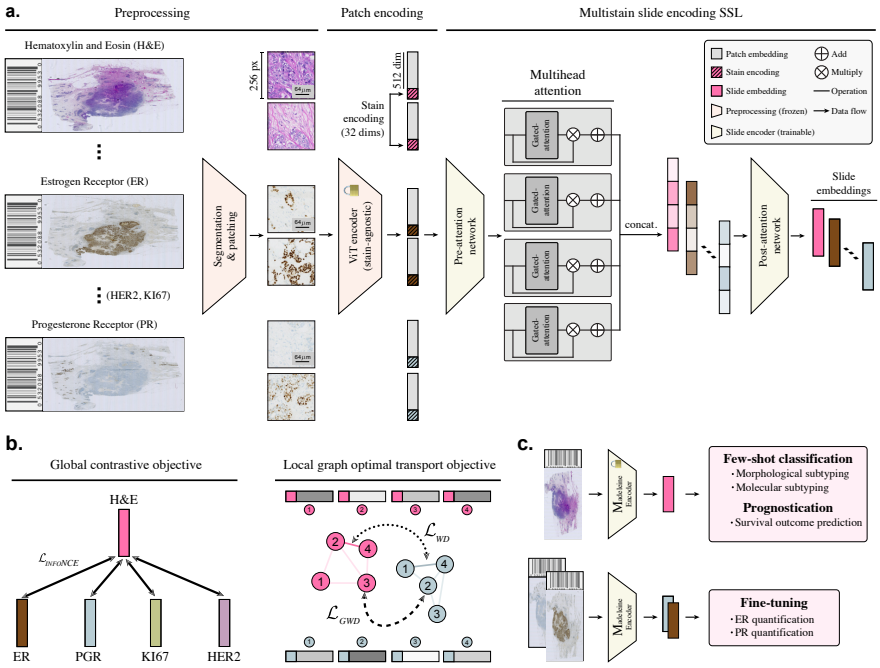

Fig. 1: Overview of Madeleine. a. Preprocessing: WSIs from various stains undergo tissue segmentation and patching into 256×256-pixel tiles. **Patch encoding:** All patches are passed through a stain-agnostic Vision Transformer to extract patch embeddings augmented with a learnable stain-specific encoding. **Slide encoding:** Embeddings from each stain are sequentially passed through a pre-attention, a multi-head attention, and a post-attention module, resulting in stain-specific slide embeddings. **b.** Madeleine is trained with a combination of global and local objectives. **Global objective:** Slide embeddings are aligned using a cross-modal contrastive objective (infoNCE). **Local objective:** Patch embeddings are aligned using a cross-modal local Graph Optimal Transport objective. **c.** The resulting stain-agnostic slide encoder can be used for various downstream tasks in few-shot and full fine-tuning settings.

(N=4,211 slides, five stains) and kidney (N=12,070 slides, four stains); and (3) extensive evaluation of Madeleine across 21 tasks including morphological subtyping, molecular subtyping, survival prediction, and IHC quantification, tested in various scenarios for few-shot learning (using linear probing and prototyping) and model fine-tuning.

2 Related work

2.1 Vision representation learning

Training a Vision Transformer (ViTs) [18,65] with self-supervised learning (SSL) [10, 80] is now the preferred approach for learning task-agnostic image repre-

sentations, such as based on visual-language models [4, 31, 43–45, 54, 59, 69, 78]. Visual-language models are usually based on contrastive learning [54], where the objective is to maximize the similarity between an image and its textual description, or as recently proposed, using Optimal Transport (OT) for fine-grained cross-modal alignment [16, 36, 52]. This approach is framed as a distribution matching objective, where the aim is to minimize the cost associated with a transport plan to match a token distribution of one modality to the other. Differently, multimodal training can leverage other *spatial* modalities, such as depth maps or bounding box annotations [8]. Drawing on these methodologies, our model, MADELEINE, integrates various high-resolution "views" of the *same* tissue stained with different markers, such as estrogen or progesterone receptor stainings.

2.2 Representation learning of histology images

SSL for learning representations of histology images is an active field with efforts in (i) developing models that can extract embeddings from small patches, typically 256×256 in size, and (ii) creating models designed to derive representations from entire WSIs, a task we denote as *slide representation learning*, and which constitutes the central contribution of our study.

Patch representation learning Using SSL to encode histology patches has so far been the main focus with increasingly large models trained on larger datasets [7, 13, 20, 33, 37, 66, 71, 77] (e.g., [9] used 3 billion patches from 423,000 slides). Simultaneously, vision-language models for histopathology have been developed using large datasets from sources such as social media and educational textbooks [21, 25, 47]. Similar to MADELEINE, [24] proposed multimodal fine-tuning by aligning H&E and IHC patches. However, their method focuses on encoding patches, whereas MADELEINE focuses on encoding WSIs.

Slide representation learning Developing pretrained encoders that extend beyond simple regions of interest to gigapixel whole slides is the next frontier in representation learning of histology images. Several works [6, 12, 32, 39, 49, 60, 61, 63, 67, 79] proposed hierarchical slide pretraining, first by transforming each patch into a patch embedding and then into a slide embedding (or region embedding). The slide encoder is typically trained using image augmentation techniques to define different *views* of the slide followed by contrastive or reconstruction objectives. Concurrent to this work, multimodal pretraining for slide representation learning was explored using bulk transcriptomics [29] and pathology reports [56, 77].

2.3 Beyond H&E staining

While H&E staining remains the gold standard in standard-of-care, it is often complemented with immunohistochemistry (IHC) and special stains. Several works have been proposed for automatic IHC quantification [22, 35, 53, 64], often leveraging cell segmentation networks. Differently, IHC status can be predicted from H&E slides, such as for HER2 (human epidermal growth factor receptor 2)

status prediction in invasive breast cancer or EGFR (epidermal growth factor receptor) prediction in lung cancer [3, 5, 17, 19, 23, 34, 50, 55, 57, 68].

3 Methods

We introduce MADELEINE for multistain-guided slide representation learning (Fig. 1). MADELEINE is composed of (1) a stain-agnostic patch encoder that transforms histology patches into *patch embeddings* (Sec. 3.1 and 3.2), (2) a multihead attention-based MIL to learn a *slide embedding* (Sec. 3.3), and (3) a cross-stain alignment module based on a dual global-local objective (Sec. 3.4).

3.1 Pre-processing and notation

Given a histology slide $\mathbf{X}_i \in \mathbb{R}^{d_x \times d_y \times 3}$ (H&E or another stain) for the i^{th} patient, we follow the MIL paradigm [26, 40, 41, 48, 58], which consists of tessellating the slide into small patches, using a pretrained vision encoder to extract patch embeddings, and pooling the resulting patch embeddings into a slide embedding. We use s_k to refer to the k^{th} stain with $\{s_k\}_{k=1}^{K}$ collectively referring to all non-H&E stains, *e.g.,* in breast cases, $s_k \in \{\text{ER}, \text{PR}, \text{HER2}, \text{KI67}\}$ with $K = 4$ denoting estrogen receptor, progesterone receptor, human epidermal growth factor receptor 2, and antigen kiel 67, respectively. We start by detecting and segmenting tissue regions to discard any background information. We use the CLAM toolbox [48] to detect H&E tissue and employ a deep learning-based tissue detector trained on mask annotations to detect non-H&E tissue. We then extract non-overlapping 256×256 patches on all stains.

3.2 Patch encoding

As MADELEINE is trained on multiple stains, this renders most SSL models for patch feature extraction trained on H&E suboptimal [66, 70]. Instead, we use CONCH, the image encoder of a visual-language model pretrained on 1M histology image-caption pairs curated from existing publications, which includes various histology stains [47]. We obtain the H&E patch embeddings $\mathbf{H}_i^{\text{HE}} \in \mathbb{R}^{N_{\text{HE}} \times d}$, with N_{HE} and $d = 512$ denoting the number of H&E patches and the embedding dimension, respectively. The j^{th} row entry, $\mathbf{H}_{i,j}^{\text{HE}}$, corresponds to the j^{th} patch embedding. We perform the same procedure for other non-H&E stains $\{s_k\}_{k=1}^{K}$ to obtain patch embeddings, i.e., $\mathbf{H}_i^{s_k} \in \mathbb{R}^{N_{s_k} \times d}$.

3.3 Slide encoding

Pre-attention & stain encoding The patch embeddings $\mathbf{H}_i^{\text{HE}}$ are first passed through a *pre-attention* network, $f^{\text{pre}} : \mathbb{R}^d \to \mathbb{R}^d$, resulting in $\widetilde{\mathbf{H}}_i^{\text{HE}} \in \mathbb{R}^{d \times d}$. As the same pre-attention module is used for encoding *all* stains, having a stain-specific signature in the input can be beneficial. To do so, we define a learnable

stain-specific encoding (denoted as SE, 32 dims) that is concatenated to each patch token before pre-attention, with $d = d + 32$. This is inspired by modality-specific token augmentation schemes in multimodal fusion [27, 46, 61].

Multi-head attention-based MIL We subsequently pass the resulting patch embeddings $\widetilde{\mathbf{H}}_i^{\text{HE}}$ to a multihead (MH) attention network with M heads [26], resulting in an attention score $a_{i,j}^m \in [0, 1]$ for each patch (Appendix Equation 2). Using multiple attention heads allows each head to focus on different yet morphologically important regions, similar to multi-head attention in Transformers [18,65]. Once computed, we form head-specific slide embeddings by taking the weighted average of the transformed patch embeddings, *i.e.,* $\mathbf{h}_{i,m}^{\text{HE}} = \sum_{j=1}^{N_{\text{HE}}} a_{i,j}^m \widetilde{\mathbf{H}}_{i,j}^{\text{HE}}$. The resulting slide embedding $\mathbf{h}_i^{\text{HE}}$ is formed by concatenating the M slide embeddings and passing it through a *post-attention* network for dimension reduction, $f^{\text{post}} : \mathbb{R}^{Md} \to \mathbb{R}^d$,

$$\mathbf{h}_i^{\text{HE}} = f^{\text{post}}([\mathbf{h}_{i,1}^{\text{HE}}, \ldots, \mathbf{h}_{i,M}^{\text{HE}}]). \tag{1}$$

The slide embeddings for other stains $\{\mathbf{h}_i^{s_k}\}_{k=1}^K$ are computed analogously. We emphasize that we apply the *same* model to all stains, instead of stain-specific modules. This way, we reduce memory requirements by a factor of K (the number of stains) and constrain the network to learn stain-agnostic representations.

3.4 Loss

MADELEINE is trained using a combination of two cross-modal objectives: (1) a global objective to align slide embeddings of all stains in a shared latent space, and (2) a local objective for matching cross-stain patch embeddings. We optionally complement these two objectives with an intra-modal loss.

Cross-modal global alignment (INFONCE) We align the latent space induced by each stain through a global symmetric cross-modal contrastive learning objective, commonly referred to as INFONCE [15]. This is a widely employed representation learning formulation [54], especially in visual-language pretraining. This objective enforces slide embeddings from the same case to be closer to each other while pushing away slide embeddings from different cases. Each term maximizes the dot-product similarity between embeddings from the same pair normalized (with Softmax) by negative pairs, which can be seen as other "classes".

Cross-modal local alignment (GOT) We also perform *local* alignment by matching the empirical distributions of patch embeddings of all stains. Intuitively, as the local morphological structure is preserved across different stains, we can identify fine-grained cross-stain correspondences. The model can consequently learn to distinguish H&E morphologies corresponding to marker-positive and marker-negative regions.

To this end, we leverage the framework of graph optimal transport (GOT) [11, 51]. Formally, we define the empirical distribution of the H&E patch embeddings as $\hat{p}_{\text{HE}} = \frac{1}{N_{\text{HE}}} \sum_{j=1}^{N_{\text{HE}}} \delta(\mathbf{H}_{i,j}^{\text{HE}})$, with $\delta(\cdot)$ denoting the dirac-delta function. We additionally define an H&E graph $\mathcal{G}_{\text{HE}}(V_{\text{HE}}, E_{\text{HE}})$, where the node $v_j^{\text{HE}} \in V_{\text{HE}}$

represents the j^{th} patch embedding from $\mathbf{H}^{\text{HE}}_{i,j}$ and the edge $e^{\text{HE}}_{j,j'}$ is formed if the cosine similarity between v^{HE}_{j} and $v^{\text{HE}}_{j'}$ is above a certain threshold. The same construction is applied to all other stains.

Based on this setup, we aim to cross-align the stain-specific graphs by minimizing two metrics: (i) The Wasserstein distance (WD) $\mathcal{L}_{\text{Node}}(\hat{p}_{\text{HE}}, \hat{p}_{s_k})$ defined between the empirical distributions of patch embeddings (*i.e.,* the nodes of $\mathcal{G}_{\text{HE}}$ and $\mathcal{G}_{s_k}$). Intuitively, WD can be seen as computing the distance between the node embedding distributions of different stains. (ii) The Gromov-Wasserstein distance (GWD) $\mathcal{L}_{\text{Edge}}(\hat{p}_{\text{HE}}, \hat{p}_{s_k})$ between the edges of $\mathcal{G}_{\text{HE}}$ and $\mathcal{G}_{s_k}$. Intuitively, GWD enforces stain-specific graphs to follow a similar structure (or topology). Additional technical information is provided in Appendix 1.3.

The local alignment objective $\mathcal{L}_{\text{GOT}}$ is given as the combination of two metrics over cross-stain pairs, with γ denoting a weighting term,

$$\mathcal{L}_{\text{GOT}} = \gamma \sum_{k=1}^{K} \mathcal{L}_{\text{Node}}(\hat{p}_{\text{HE}}, \hat{p}_{s_k}) + (1-\gamma) \sum_{k=1}^{K} \mathcal{L}_{\text{Edge}}(\hat{p}_{\text{HE}}, \hat{p}_{s_k}). \tag{2}$$

Intra-modal alignment (INTRA) We additionally define an optional intra-modality objective $\mathcal{L}_{\text{INTRA}}$ to align different augmentations of the H&E slide. This objective is similar to existing pretraining strategy [12,39], and can be seen as a direct extension of SSL from patch- to slide-level. Specifically, we generate two distinct slide embeddings of the H&E slide by separately processing two randomly disjoint sets of patch embeddings using MADELEINE. This process yields a pair of slide embeddings, denoted as $\mathbf{h}^{\text{HE},(1)}_{i}$ and $\mathbf{h}^{\text{HE},(2)}_{i}$, which are then aligned using a contrastive objective.

Overall, we train MADELEINE with the composite loss $\mathcal{L} = \mathcal{L}_{\text{InfoNCE}} + \mathcal{L}_{\text{GOT}}$. The INTRA objective is used as a baseline, which can also be combined with $\mathcal{L}$.

Pretraining details MADELEINE was trained for a maximum of 120 epochs (5 warmup epochs) using AdamW optimizer, cosine learning rate schedule (start: 10^{-4}, end: 10^{-8}) and batch size of 90. All models were trained on 3×24 GB 3090Ti. Additional implementation details are provided in Appendix Table 1.

4 Study design

To assess the representative power of MADELEINE pretraining, we design two distinct scenarios: (1) MADELEINE pretraining on breast cancer cases (Sec. 4.1) and (2) MADELEINE pretraining on kidney transplant cases (Sec. 4.2). We then perform downstream evaluations based on public and private cohorts (Sec. 4.3). The evaluation was designed to encompass the variability of tasks found in pathology. We emphasize that MADELEINE pretraining does not involve datasets used for downstream tasks, precluding any data leakage. A detailed description is provided in Appendix 2.

4.1 Breast

Acrobat (multi-stain, pretraining) We pretrain MADELEINE using data from the Automatic registration of breast cancer tissue MICCAI challenge (Acrobat) [72, 73]. Acrobat is a multi-stain dataset comprising 4,211 WSIs from 1,153 primary breast cancer cases. Every case includes an H&E-stained WSI, along with one to four WSIs of tissue from the same tumor that have been stained with immunohistochemistry, either ER (N=844 WSIs), PR (N=837), HER2 (N=534), or KI67 (N=843), such that K=4. The entirety of Acrobat was used for pretraining, with all slides processed at 10× magnification.

TCGA Breast (H&E, downstream) We use the public TCGA Breast cohort for (1) morphological subtyping (N=1,041) into invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC); (2) binary molecular subtyping for predicting ER status (N=996), PR status (N=993), and HER2 status (N=693), and (3) survival prediction (N=1,049).

BCNB (H&E, downstream) We use the public BCNB cohort [76] for binary molecular subtyping by predicting ER, PR, HER2 and KI67 status (N=1,058).

AIDPATH (H&E, downstream) We use the public AIDPATH cohort [1] for binary HER2 status prediction (N=48) and KI67 status prediction (N=50).

BWH Breast (H&E, downstream) We use an in-house breast cohort for two binary tasks (1) morphological subtyping (N=1,265); and (2) molecular subtyping for predicting ER (N=873), PR (N=874), and HER2 (N=816).

MGH Breast (ER and PR, downstream) We use another in-house 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.

4.2 Kidney

BWH Kidney (multi-stain, pretraining) We collected an in-house 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), such that $K = 3$. In total, each case includes 6 to 18 slides. We hold out 20% of this cohort (210 cases, 1,852 slides across all stains, out of which 463 are H&E slides) as an independent test set and used the rest for MADELEINE pretraining. Slides were processed at 20×.

Renal allograft rejection (downstream) We use H&E slides of the held-out cohort (N=463) to screen for Antibody-mediated rejection (AMR, 2 class) and quantify Interstitial Fibrosis and Tubular Atrophy (IFTA, 3 classes) (N=210 cases, N=1,852 WSIs across all stains). 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 1: Few-shot breast cancer subtyping. Evaluation using macro-AUC on TCGA Breast (N=1,041) and BWH Breast cohorts (N=1,265). GigaSSL embeddings for BWH cohort not available. Mean and standard deviation reported over ten runs. Best in **bold**, second best is underlined.

	Model/Data	TCGA Breast (↑)				BWH Breast (↑)			
		k=1	k=5	k=10	k=25	k=1	k=5	k=10	k=25
MIL	ABMIL [26]	79.7 ± 11.8	90.4 ± 3.8	90.4 ± 4.1	93.3 ± 1.2	67.8 ± 13.3	84.7 ± 9.3	93.0 ± 2.9	95.3 ± 2.0
	TransMIL [58]	61.8 ± 5.7	71.5 ± 10.5	72.1 ± 10.6	82.5 ± 6.4	59.5 ± 12.2	72.6 ± 14.0	73.6 ± 9.4	81.2 ± 12.5
	IB-MIL [42]	74.8 ± 10.8	90.1 ± 4.5	91.4 ± 2.2	93.8 ± 1.5	68.7 ± 13.1	82.6 ± 8.0	88.7 ± 6.3	95.2 ± 1.9
	ILRA [74]	68.3 ± 9.3	89.3 ± 2.9	90.9 ± 1.6	92.2 ± 2.1	69.8 ± 8.6	84.7 ± 7.6	91.0 ± 2.4	93.3 ± 3.3
Linear probe	MEAN	70.5 ± 11.0	83.1 ± 3.7	86.6 ± 3.2	91.2 ± 1.2	70.2 ± 9.3	81.8 ± 5.4	87.4 ± 2.8	92.6 ± 1.5
	INTRA	70.8 ± 9.1	82.6 ± 4.2	86.1 ± 2.9	91.2 ± 1.2	69.3 ± 8.7	78.3 ± 7.1	84.8 ± 3.0	92.0 ± 2.4
	HIPT$_{\text{CLS-4k}}$ [12]	62.2 ± 3.9	69.3 ± 5.4	77.5 ± 3.9	83.0 ± 2.3	66.8 ± 12.6	76.6 ± 6.2	80.6 ± 3.3	85.8 ± 2.0
	GigaSSL [39]	68.2 ± 6.6	78.7 ± 4.8	82.8 ± 4.2	88.9 ± 1.9	–	–	–	–
	GigaPath-Mean [77]	59.7 ± 6.6	69.8 ± 5.2	77.0 ± 6.1	85.8 ± 3.4	64.8 ± 9.7	79.7 ± 5.3	84.5 ± 2.9	91.7 ± 1.9
	GigaPath [77]	58.7 ± 6.7	68.6 ± 4.9	75.4 ± 4.8	84.0 ± 2.5	64.0 ± 11.5	78.0 ± 7.5	83.1 ± 2.9	90.3 ± 2.0
	MADELEINE	83.8 ± 8.0	91.2 ± 1.3	91.7 ± 2.0	93.2 ± 0.9	84.0 ± 7.7	91.9 ± 2.4	93.8 ± 1.4	96.0 ± 0.8
	MADELEINE-SE	**87.2** ± 6.6	93.2 ± 1.0	93.1 ± 1.6	94.1 ± 0.8	85.6 ± 7.2	**93.7** ± 1.7	**95.3** ± 0.9	**96.7** ± 0.4
Prototyping	MEAN	69.1 ± 10.3	81.8 ± 6.9	84.2 ± 5.7	91.3 ± 2.8	68.8 ± 9.7	78.5 ± 5.8	83.9 ± 3.2	86.2 ± 3.0
	INTRA	69.3 ± 9.0	81.7 ± 5.9	83.9 ± 5.0	89.6 ± 2.4	68.4 ± 8.4	75.8 ± 6.8	81.9 ± 4.0	85.4 ± 3.0
	HIPT$_{\text{CLS-4k}}$ [12]	62.1 ± 4.1	68.3 ± 6.2	73.6 ± 6.9	78.7 ± 2.4	66.3 ± 12.3	75.7 ± 7.3	79.1 ± 4.2	82.3 ± 0.9
	GigaSSL [39]	67.9 ± 5.9	78.5 ± 5.4	82.6 ± 4.8	88.4 ± 1.6	–	–	–	–
	GigaPath-Mean [77]	58.8 ± 6.2	70.0 ± 5.7	73.5 ± 7.0	81.7 ± 4.1	65.7 ± 9.5	78.8 ± 4.9	81.6 ± 2.7	85.5 ± 2.0
	GigaPath [77]	58.1 ± 6.3	68.4 ± 5.1	71.8 ± 7.0	80.1 ± 3.2	64.4 ± 11.3	76.8 ± 8.3	80.4 ± 2.1	83.3 ± 1.8
	MADELEINE	83.2 ± 7.8	91.6 ± 1.6	92.7 ± 1.3	94.4 ± 0.6	84.6 ± 8.4	91.1 ± 3.0	93.1 ± 1.4	94.9 ± 0.9
	MADELEINE-SE	86.4 ± 6.6	**93.4** ± 1.0	**94.0** ± 1.1	**95.0** ± 0.4	**86.3** ± 7.7	93.2 ± 2.2	94.9 ± 0.8	95.8 ± 0.8

4.3 Evaluation framework

Few-shot classification Following the standard practice in SSL evaluation [10, 13, 80], we benchmark MADELEINE and baselines with k-shot classification ($k = 1, 5, 10, 25$ examples per class) using (1) linear probing and (2) prototyping. All experiments are repeated ten times by randomly sampling k examples per class. Linear probing was conducted without hyper-parameter search using default parameters of the sklearn package.

Survival prediction We assess MADELEINE in survival outcome prediction, where slide embeddings are passed to a Cox proportional hazards loss predicting survival. Following prior work, MIL models are trained using survival negative log-likelihood (NLL) loss [14]. We use site- and survival-stratified five-fold cross-validation evaluated with concordance-index (c-index) [30].

Fine-tuning We assess the performance of MADELEINE encoder for downstream tasks when fine-tuned, compared to when trained from scratch. Evaluations follow a 5-fold label-stratified train-test strategy.

5 Results

We showcase the performance of MADELEINE and MADELEINE-SE (i.e., with stain encoding) on few-shot classification (Sec. 5.1), and full classification (Sec. 5.2), that we complement by a series of ablations (Sec. 5.3). We benchmark MADELEINE

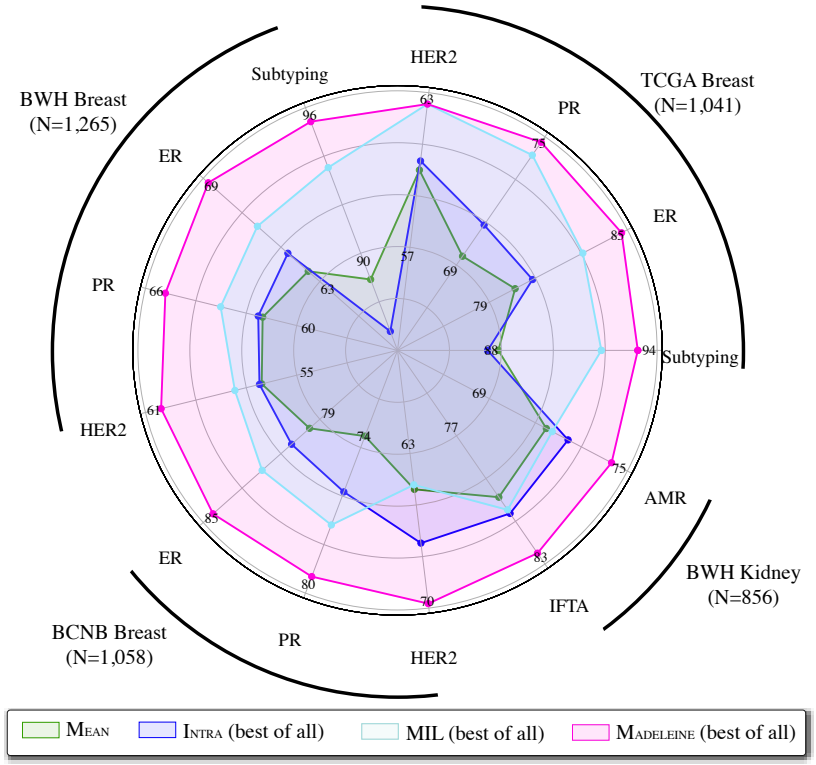

Fig. 2: Few-shot performance of Madeleine against baselines. All tasks are assessed on H&E-stained WSIs. Morphological subtyping is reported for k=10, molecular subtyping for k=25, and kidney transplant rejection for k=50. Each experiment is repeated ten times by sampling k different samples per class. Besides HIPT and GigaSSL, all models use the *same* patch encoder. Each axis represents 10% AUC and each segment a 2% increment. Additional results for all k values are provided in Appendix 4 and 5.

against four MIL methods: single head ABMIL [26], TransMIL [58], IB-MIL [42] and ILRA [75]; four intra-modal SSL methods: Intra, HIPT [12], GigaSSL [39], and GigaPath [77] (work concurrent to Madeleine); and mean pooling (Mean).

5.1 Few-shot results

Fig. 2 highlights the few-shot classification of Madeleine against baselines (for each task: best MIL, best intra-modal, and Mean). Detailed morphological subtyping performance is reported in Table 1, molecular subtyping in Appendix Table 3, and kidney transplant rejection in Appendix Table 7.

Madeleine *vs.* rest Madeleine outperforms all baselines in 13/13 tasks, in some cases by a significant margin, *e.g.,* +10.1% over Intra in TCGA Breast

(k=10, prototyping classification), or +9.0% over ABMIL in BWH Breast (k=5, linear probing). This performance is achieved using simple downstream models based on linear probing or prototyping classification, whereas MIL methods are trained from scratch for each task.

MEAN *vs.* INTRA Despite its simplicity, the MEAN baseline offers high performance, often surpassing INTRA, HIPT and GigaSSL. This highlights (1) the importance of powerful domain-specific patch encoders and (2) the complexity of deriving an information-rich training signal using information from the slide itself, further motivating the exploration of multimodal pretraining for slide representation learning.

MIL comparisons Despite recent advances in MIL, ABMIL remains a strong baseline in a few-shot setting. In some cases, ABMIL is outperformed by IBMIL [42]. TransMIL, which includes patch-to-patch context using self-attention approximation, performs poorly, which we hypothesize is due to overfitting.

MADELEINE *vs.* INTRA MADELEINE outperforms the INTRA baseline for all values of k on all tasks. This highlights the importance of using clinically and biologically meaningful "views" provided by multimodal pretraining.

MADELEINE *vs.* GigaPath *vs.* GigaPath-MEAN MADELEINE outperforms GigaPath on all tasks in both linear probing and prototyping evaluation, in most cases by a significant margin, e.g., +17.7% in TCGA subtyping with linear probing (k=10). Interestingly, GigaPath-(MEAN) (average of all patch embeddings) reaches better performance than GigaPath slide encoder which suggests that intra-SSL can degrade performance, even when scaling to large models and number of samples. While GigaPath is a pan-cancer model, it was trained on a comparable number of breast samples (around 4,500 WSIs), which underscores (i) the quality of multistain pretraining, and (ii) the complexity of building pan-cancer models.

Generalization to other stains As MADELEINE is stain-agnostic, we can use it for encoding non-H&E stains. Specifically, we perform fine-tuning of MADELEINE multi-head encoder (FINETUNE) for quantification of ER and PR slides (framed as a 3-class and 6-class tasks) on the MGH cohort (Fig. 3.a and Appendix Table 2). We compare it against MADELEINE architecture trained from scratch and MEAN. All models trained using k=25 examples per class. Fine-tuning leads to consistently better performance than random weight initialization, with a +7.5% gain on 3-class ER and +5.6% gain on 3-class PR quantification (Fig. 3.a and Appendix Table 4).

5.2 Full classification

Beyond few-shot classification, we assess MADELEINE in a supervised setting using 5-fold cross-validation, where we directly use MADELEINE embeddings for survival prediction and molecular subtyping.

Survival We perform survival outcome prediction on TCGA Breast using a 5-fold site-stratified cross-validation. MADELEINE and other slide-level models (HIPT, GigaSSL and INTRA) are trained using a Cox proportional hazards objective from the slide embedding. All MIL models are trained with a survival

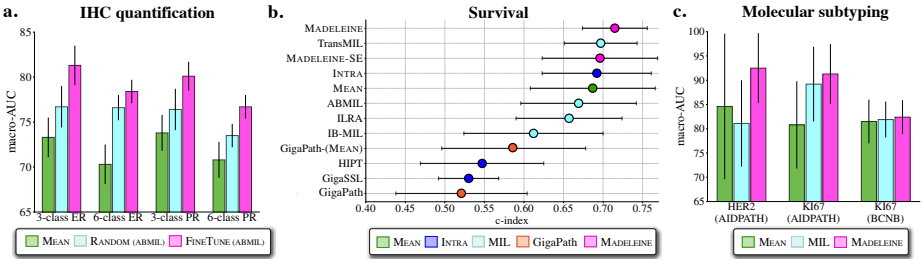

Fig. 3: Evaluation of Madeleine and baselines on IHC quantification and survival prediction. a. We fine-tune Madeleine for IHC quantification on the MGH cohort (N=962 ER and N=1,071 PR slides). 3-class and 6-class variants are derived from IHC scores extracted in pathology reports. Models are trained with k=25 examples per class. Random uses Madeleine architecture trained from scratch; FineTune is initialized with Madeleine pretrained weights. We report the mean and standard deviation (std) on a 5-fold label-stratified train-test study. **b.** Survival prediction on TCGA Breast (N=1,041 slides). We report mean and std using a 5-fold site-stratified cross-validation. "SE" is Madeleine with stain encodings. **c.** Molecular subtyping of Madeleine fine-tuned on AIDPATH (N=48 for HER2 and N=50 for KI67) and BCNB (N=1,058). Evaluation using 5-fold cross-validation. MIL refers to the best of four MIL baselines.

NLL objective following prior work [14, 30]. Madeleine leads to the best survival prediction reaching 0.71 c-index outperforming all baselines (Fig. 3.b and Appendix Table 5).

Molecular subtyping In addition, we use logistic regression to predict HER2 status in AIDPATH and KI67 status in AIDPATH and BCNB from H&E. In AIDPATH, Madeleine-SE leads to +11.4% performance boost over the best MIL in HER2, and +1.8% in KI67 (Fig. 3.c and Appendix Table 6.

5.3 Ablation

All ablations were run using Madeleine pretrained on breast cancer slides, evaluated using AUC, and benchmarked using prototyping classification (k=25) on a set of three representative tasks: (1) BWH Breast subtyping, (2) TCGA PR classification, and (3) BCNB ER classification. In addition, we benchmark TCGA Breast survival using a Cox model trained using 5-fold cross-validation. Prototyping is not sensitive to hyper-parameter selection compared to linear probing, making it ideal for ablating components of Madeleine.

Loss ablation We perform a thorough ablation of Madeleine loss function by retraining models with InfoNCE alone, cross-modal Mean-Squared Error (replacing InfoNCE), GOT alone, combining InfoNCE and GOT (Madeleine default), and finally combining InfoNCE, GOT and Intra (Table 2). InfoNCE alone significantly outperforms Mean (+4.4% AUC), Intra (+4.1%) and MSE (+9.3%). GOT alone performs similarly to Mean and Intra. When combining InfoNCE and GOT, we observe an additional gain of +1.4% over

Table 2: Ablation study of MADELEINE loss. 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 ten times with fixed seed across baselines. Standard deviation reported over the 10 runs. "MSE" stands for Mean-Squared Error. Best in **bold**, second best is underlined.

Model/Data	**TCGA Survival** (↑)	**BWH Subtyping** (↑)	**TCGA PR** (↑)	**BCNB ER** (↑)	**Avg**
MEAN	68.8 ± 7.9	86.2 ± 3.0	70.8 ± 2.0	76.9 ± 2.2	75.7
INTRA	69.2 ± 6.9	85.4 ± 3.0	71.6 ± 1.4	77.7 ± 1.8	76.0
MSE	68.0 ± 9.3	80.9 ± 2.8	65.2 ± 2.8	69.0 ± 3.3	70.8
INFONCE	69.9 ± 8.1	93.3 ± 0.9	74.5 ± 1.3	82.8 ± 1.7	80.1
GOT	70.1 ± 3.6	85.9 ± 2.6	70.1 ± 2.0	75.8 ± 2.7	75.5
INFONCE + GOT	**71.5** ± 4.1	**94.9** ± 0.9	**76.4** ± 1.2	83.0 ± 1.6	**81.5**
INFONCE + GOT + INTRA	71.0 ± 6.2	**94.9** ± 1.1	**76.4** ± 1.2	**83.3** ± 1.3	81.4

INFONCE. However, including an INTRA objective on top leads to a similar performance. Overall, the global cross-modal INFONCE objective remains the most critical component, which benefits from the local GOT cross-modal objective.

Feature extractor ablation We further test if the benefits of MADELEINE pretraining generalize when using CTransPath [70], a state-of-the-art patch encoder based on the Swin-Transformer model and that was pretrained on 15 million patches from TCGA and PAIP (Appendix Table 8). When comparing the MEAN baseline, our patch encoder significantly outperforms CTransPath (+5.0% AUC). The same observation holds using MADELEINE embeddings, where using our patch encoder leads to +10.5% AUC gain over CTransPath. Overall, these results further assert that (1) using powerful domain-specific feature encoders trained on large amounts of diverse data is necessary, and (2) MADELEINE pretraining leads to high performance even when using weaker feature encoders.

Architecture ablation MADELEINE explores two architectural features: (1) the use of a multi-head (MH) attention network and (2) the use of a learnable stain encoding (Appendix Table 9). Using multiple ABMIL heads (four in MADELEINE) leads to a consistent performance gain (on average of +2.1%). We hypothesize the gain arises from each head focusing on different morphologies during pretraining. Adding stain-encoding does not have a consistent effect, as it boosts performance in morphological subtyping but decreases performance in molecular subtyping. Replacing the ABMIL architecture with a TransMIL backbone [58] leads to lower performance (on average -2.9% over ABMIL and -5.0% over MH-ABMIL).

5.4 MADELEINE attention visualization

By visualizing head-specific attention weights, we can gain insights into the internal behavior of MADELEINE (Fig. 4). We show that different heads learn to focus on morphologically distinct regions, *e.g.,* Head-3 focuses on tumor while Head-4

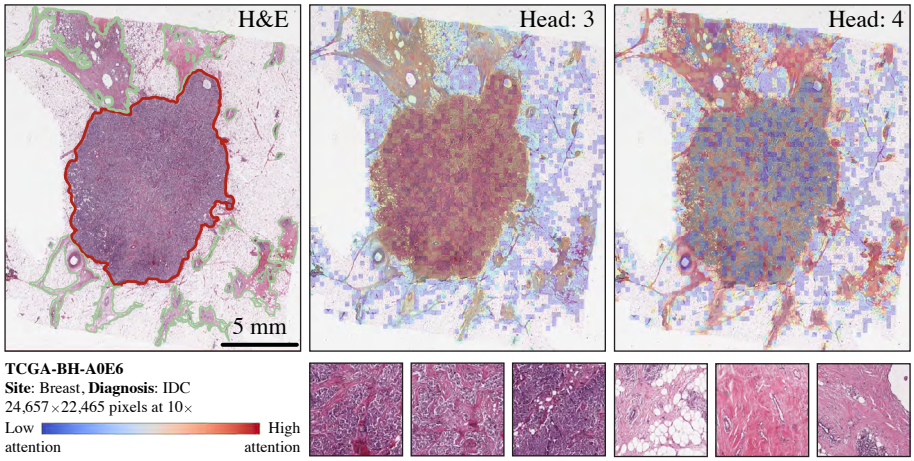

Fig. 4: Madeleine attention weight visualization in a breast cancer case. Attention weights of the third (focusing on tumor, annotated in red) and fourth (focusing on non-tumor regions, annotated in green) heads of Madeleine slide encoder along with high attention patches per head.

focuses on non-tumor stroma. This is a remarkable finding as Madeleine was not given any morphological labels like tumor grade or subtype during training. Additional example heatmaps are provided in Appendix 7.

6 Conclusion

In this study, we present Madeleine, a method exploring multimodal pretraining for *slide representation learning* based on multistain alignment. Our method utilizes extensive datasets of multistain slides, where we consider each stain as a unique perspective of the standard H&E-stained slide, each revealing different aspects of the tissue's biological state. We demonstrate that Madeleine slide encoder outperforms multiple instance learning and intra-modal pretraining models in few-shot and full classification scenarios across various tasks, ranging from morphological and molecular subtyping to prognosis prediction to IHC quantification. Our method currently incorporates four to five different stains per sample, yet in clinical practice, more stains can be available to assist pathologists. This opens up promising avenues for expanding Madeleine pretraining to include a broader range of stains. Furthermore, while our focus has been on multimodal pretraining with multiple stains, there exists a potential to explore other spatial modalities, such as those based on immunofluorescence, mass spectrometry, or spatial transcriptomics [28] for slide representation learning.

References

1. AIDPATH DB (Mar 2024), https://mitel.dimi.uniud.it/aidpath-db/app/login.php, [Online; accessed 6. Mar. 2024] 8
2. Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical AI. Nat. Med. **28**, 1773–1784 (Sep 2022) 1
3. Akbarnejad, A., Ray, N., Barnes, P.J., Bigras, G.: Predicting Ki67, ER, PR, and HER2 Statuses from H&E-stained Breast Cancer Images. arXiv (Aug 2023) 5
4. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022) 4
5. Anand, D., Kurian, N.C., Dhage, S., Kumar, N., Rane, S., Gann, P.H., Sethi, A.: Deep Learning to Estimate Human Epidermal Growth Factor Receptor 2 Status from Hematoxylin and Eosin-Stained Breast Tissue Images. Journal of Pathology Informatics **11** (2020) 5
6. Aryal, M., Yahyasoltani, N.: Context-Aware Self-Supervised Learning of Whole Slide Images. arXiv (Jun 2023) 4
7. Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Chen, T., Tomasev, N., Mitrović, J., Strachan, P., et al.: Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nature Biomedical Engineering pp. 1–24 (2023) 4
8. Bachmann, R., Mizrahi, D., Atanov, A., Zamir, A.: MultiMAE: Multi-modal Multi-task Masked Autoencoders. arXiv (Apr 2022) 4
9. Campanella, G., Kwan, R., Fluder, E., Zeng, J., Stock, A., Veremis, B., Polydorides, A.D., Hedvat, C., Schoenfeld, A., Vanderbilt, C., Kovatch, P., Cordon-Cardo, C., Fuchs, T.J.: Computational Pathology at Health System Scale – Self-Supervised Foundation Models from Three Billion Images. arXiv (Oct 2023) 4
10. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294 (2021) 3, 9
11. Chen, L., Gan, Z., Cheng, Y., Li, L., Carin, L., Liu, J.: Graph optimal transport for cross-domain alignment. In: International Conference on Machine Learning. pp. 1542–1553. PMLR (2020) 2, 6
12. 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) 2, 4, 7, 9, 10
13. 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., Mahmood, F.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024) 1, 4, 9
14. Chen, R.J., Lu, M.Y., Williamson, D.F., Chen, T.Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., Mahmood, F.: Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell **40**(8), 865–878 (Aug 2022) 9, 12
15. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (Nov 2020) 2, 6

16. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: UNiversal Image-TExt Representation Learning. In: Computer Vision – ECCV 2020, pp. 104–120. Springer, Cham, Switzerland (Sep 2020) 4
17. Couture, H.D., Williams, L.A., Geradts, J., Nyante, S.J., Butler, E.N., Marron, J.S., Perou, C.M., Troester, M.A., Niethammer, M.: Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. npj Breast Cancer **4**(30), 1–8 (Sep 2018) 5
18. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021) 3, 6
19. Farahmand, S., Fernandez, A.I., Ahmed, F.S., Rimm, D.L., Chuang, J.H., Reisenbichler, E., Zarringhalam, K.: Deep learning trained on hematoxylin and eosin tumor region of Interest predicts HER2 status and trastuzumab treatment response in HER2+ breast cancer. Mod. Pathol. **35**(1), 44–51 (Jan 2022) 5
20. Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Kain, A., Saillard, C., Schiratti, J.B.: Scaling self-supervised learning for histopathology with masked image modeling. medRxiv (07 2023) 4
21. Gamper, J., Rajpoot, N.: Multiple instance captioning: Learning representations from histopathology textbooks and articles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16549–16559 (2021) 2, 4
22. Ghahremani, P., Li, Y., Kaufman, A., Vanguri, R., Greenwald, N., Angelo, M., Hollmann, T.J., Nadeem, S.: Deep learning-inferred multiplex immunofluorescence for immunohistochemical image quantification. Nat. Mach. Intell. **4**, 401–412 (Apr 2022) 4
23. Gil Shamai, M.: Artificial Intelligence Algorithms to Assess Hormonal Status From Tissue Microarrays in Patients With Breast. JAMA Netw. Open **2**(7), e197700 (Jul 2019) 5
24. Hua, S., Yan, F., Shen, T., Zhang, X.: Pathoduet: Foundation models for pathological slide analysis of h&e and ihc stains (2023) 4
25. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature Medicine **29**, 1–10 (08 2023) 2, 4
26. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018) 2, 5, 6, 9, 10
27. Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., Henaff, O.J., Botvinick, M., Zisserman, A., Vinyals, O., Carreira, J.: Perceiver IO: A general architecture for structured inputs & outputs. In: International Conference on Learning Representations (2022) 6
28. Jaume, G., Doucet, P., Song, A.H., Lu, M.Y., Almagro-Perez, C., Wagner, S.J., Vaidya, A.J., Chen, R.J., Williamson, D.F.K., Kim, A., Mahmood, F.: HEST-1k: A Dataset for Spatial Transcriptomics and Histology Image Analysis. arXiv (Jun 2024) 14
29. Jaume, G., Oldenburg, L., Vaidya, A.J., Chen, R.J., Williamson, D.F., Peeters, T., Song, A.H., Mahmood, F.: Transcriptomics-guided slide representation learning in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 2, 4

30. Jaume, G., Vaidya, A., Chen, R., Williamson, D., Liang, P., Mahmood, F.: Modeling dense multimodal interactions between biological pathways and histology for survival prediction. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 9, 12
31. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916. PMLR (2021) 4
32. Jiang, S., Hondelink, L., Suriawinata, A.A., Hassanpour, S.: Masked pre-training of transformers for histology image analysis. arXiv preprint arXiv:2304.07434 (2023) 4
33. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3344–3354 (06 2023) 4
34. Kather, J.N., Heij, L.R., Grabsch, H.I., Loeffler, C., Echle, A., Muti, H.S., Krause, J., Niehues, J.M., Sommer, K.A.J., Bankhead, P., Kooreman, L.F.S., Schulte, J.J., Cipriani, N.A., Buelow, R.D., Boor, P., Ortiz-Brüchle, N., Hanby, A.M., Speirs, V., Kochanny, S., Patnaik, A., Srisuwananukorn, A., Brenner, H., Hoffmeister, M., van den Brandt, P.A., Jäger, D., Trautwein, C., Pearson, A.T., Luedde, T.: Pan-cancer image-based detection of clinically actionable genetic alterations. Nat. Cancer **1**(8), 789–799 (Aug 2020) 5
35. Khameneh, F.D., Razavi, S., Kamasak, M.: Automated segmentation of cell membranes to evaluate HER2 status in whole slide images using a modified deep learning network. Comput. Biol. Med. **110**, 164–174 (Jul 2019) 4
36. Kim, M.: Differentiable expectation-maximization for set representation learning. In: International Conference on Learning Representations (2022) 4
37. Koohbanani, N.A., Unnikrishnan, B., Khurram, S.A., Krishnaswamy, P., Rajpoot, N.: Self-path: Self-supervision for classification of pathology images with limited annotations. IEEE Transactions on Medical Imaging (2021) 4
38. Krishnan, R., Rajpurkar, P., Topol, E.J.: Self-supervised learning in medicine and healthcare. Nature Biomedical Engineering (2022) 1
39. 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) 2, 4, 7, 9, 10
40. Lee, Y., Park, J., Oh, S., et al.: Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. Nat. Biomed. Eng (2022) 5
41. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14318–14328 (2021) 5
42. 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) 9, 10, 11
43. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023) 4
44. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021) 4

45. Li, Y., Fan, H., Hu, R., Feichtenhofer, C., He, K.: Scaling language-image pre-training via masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23390–23400 (2023) 4
46. Liang, P.P., Lyu, Y., Fan, X., Tsaw, J., Liu, Y., Mo, S., Yogatama, D., Morency, L.P., Salakhutdinov, R.: High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. Transactions on Machine Learning Research (2023) 6
47. 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) 4, 5
48. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021) 2, 5
49. Mukashyaka, P., Sheridan, T., pour, A., Chuang, J.: Sampler: unsupervised representations for rapid analysis of whole slide tissue images. eBioMedicine **99**, 104908 (01 2024) 2, 4
50. Naik, N., Madani, A., Esteva, A., Keskar, N.S., Press, M.F., Ruderman, D., Agus, D.B., Socher, R.: Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains. Nat. Commun. **11**(5727), 1–8 (Nov 2020) 5
51. Pramanick, S., Jing, L., Nag, S., Zhu, J., Shah, H., LeCun, Y., Chellappa, R.: Volta: Vision-language transformer with weakly-supervised local-feature alignment. arXiv preprint arXiv:2210.04135 (2022) 2, 6
52. Pramanick, S., Jing, L., Nag, S., Zhu, J., Shah, H., LeCun, Y., Chellappa, R.: VoLTA: Vision-Language Transformer with Weakly-Supervised Local-Feature Alignment. TMLR (Oct 2023) 4
53. Qaiser, T., Mukherjee, A., P. B., C.R., Munugoti, S.D., Tallam, V., Pitkäaho, T., Lehtimäki, T., Naughton, T., Berseth, M., Pedraza, A., Mukundan, R., Smith, M., Bhalerao, A., Rodner, E., Simon, M., Denzler, J., Huang, C.H., Bueno, G., Snead, D., Ellis, I.O., Ilyas, M., Rajpoot, N.: HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues. Histopathology **72**(2), 227–238 (Jan 2018) 4
54. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 2, 4, 6
55. Rawat, R.R., Ortega, I., Roy, P., Sha, F., Shibata, D., Ruderman, D., Agus, D.B.: Deep learned tissue "fingerprints" classify breast cancers by ER/PR/Her2 status from H&E images. Sci. Rep. **10**(7275), 1–13 (Apr 2020) 5
56. Shaikovski, G., Casson, A., Severson, K., Zimmermann, E., Wang, Y.K., Kunz, J.D., Retamero, J.A., Oakley, G., Klimstra, D., Kanan, C., Hanna, M., Zelechowski, M., Viret, J., Tenenholtz, N., Hall, J., Fusi, N., Yousfi, R., Hamilton, P., Moye, W.A., Vorontsov, E., Liu, S., Fuchs, T.J.: PRISM: A Multi-Modal Generative Foundation Model for Slide-Level Histopathology. arXiv (May 2024) 4
57. Shamai, G., Livne, A., Polónia, A., Sabo, E., Cretu, A., Bar-Sela, G., Kimmel, R.: Deep learning-based image analysis predicts PD-L1 status from H&E-stained histopathology images in breast cancer. Nat. Commun. **13**(6753), 1–13 (Nov 2022) 5

58. 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) 2, 5, 9, 10, 13
59. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Flava: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15638–15650 (2022) 4
60. Song, A.H., Chen, R.J., Ding, T., Williamson, D.F., Jaume, G., Mahmood, F.: Morphological prototyping for unsupervised slide representation learning in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 4
61. Song, A.H., Chen, R.J., Jaume, G., Vaidya, A.J., Baras, A., Mahmood, F.: Multimodal prototyping for cancer survival prediction. In: Forty-first International Conference on Machine Learning (2024) 4, 6
62. Song, A.H., Jaume, G., Williamson, D.F.K., Lu, M.Y., Vaidya, A., Miller, T.R., Mahmood, F.: Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering (Oct 2023) 2
63. Tavolara, T., Gurcan, M., Niazi, M.: Contrastive multiple instance learning: An unsupervised framework for learning slide-level representations of whole slide histopathology images without labels. Cancers **14**, 5778 (11 2022) 2, 4
64. Vandenberghe, M.E., Scott, M.L.J., Scorer, P.W., Söderberg, M., Balcerzak, D., Barker, C.: Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. Sci. Rep. **7**(45938), 1–11 (Apr 2017) 4
65. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. In: Neural Information Processing Systems (NeurIPS) (2017) 3, 6
66. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Liu, S., Mathieu, P., van Eck, A., Lee, D., Viret, J., Robert, E., Wang, Y.K., Kunz, J.D., Lee, M.C.H., Bernhard, J., Godrich, R.A., Oakley, G., Millar, E., Hanna, M., Retamero, J., Moye, W.A., Yousfi, R., Kanan, C., Klimstra, D., Rothrock, B., Fuchs, T.J.: Virchow: A million-slide digital pathology foundation model (2023) 4, 5
67. Vu, Q.D., Rajpoot, K., Raza, S.E.A., Rajpoot, N.: Handcrafted Histological Transformer (H2T): Unsupervised representation of whole slide images. Med. Image Anal. **85**, 102743 (Apr 2023) 4
68. Wang, J., Zhu, X., Chen, K., Hao, L., Liu, Y.: HAHNet: a convolutional neural network for HER2 status classification of breast cancer. BMC Bioinf. **24**(1), 1–16 (Dec 2023) 5
69. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19175–19186 (2023) 2, 4
70. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Huang, J., Yang, W., Han, X.: Transpath: Transformer-based self-supervised learning for histopathological image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 186–195. Springer (2021) 5, 13
71. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**, 102559 (2022) 4

72. 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) 8
73. 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) 8
74. 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) 9
75. 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 (2023) 10
76. 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) 8
77. 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) 4, 9, 10
78. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. Transactions on Machine Learning Research (2022) 4
79. Yu, Z., Lin, T., Xu, Y.: Slpd: Slide-level prototypical distillation for wsis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 259–269. Springer (2023) 4
80. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: Image BERT pre-training with online tokenizer. In: International Conference on Learning Representations (2022) 3, 9